

CYBERSECURITY AND PRIVACY IN AI – MEDICAL IMAGING DIAGNOSIS

JUNE 2023

ABOUT ENISA

The European Union Agency for Cybersecurity, ENISA, is the Union's agency dedicated to achieving a high common level of cybersecurity across Europe. Established in 2004 and strengthened by the EU Cybersecurity Act, the European Union Agency for Cybersecurity contributes to EU cyber policy, enhances the trustworthiness of ICT products, services and processes with cybersecurity certification schemes, cooperates with Member States and EU bodies, and helps Europe prepare for the cyber challenges of tomorrow. Through knowledge sharing, capacity building and awareness raising, the Agency works together with its key stakeholders to strengthen trust in the connected economy, to boost resilience of the Union's infrastructure, and, ultimately, to keep Europe's society and citizens digitally secure. More information about ENISA and its work can be found here: www.enisa.europa.eu.

CONTACT

For contacting the authors, please use info@enisa.europa.eu

For media enquiries about this paper, please use press@enisa.europa.eu.

EDITORS

Monika Adamczyk, Apostolos Malatras, Ioannis Agrafiotis, ENISA

LEGAL NOTICE

This publication represents the views and interpretations of ENISA, unless stated otherwise. It does not endorse a regulatory obligation of ENISA or of ENISA bodies pursuant to the Regulation (EU) No 2019/881.

ENISA has the right to alter, update or remove the publication or any of its contents. It is intended for information purposes only and it must be accessible free of charge. All references to it or its use as a whole or partially must contain ENISA as its source.

Third-party sources are quoted as appropriate. ENISA is not responsible or liable for the content of the external sources including external websites referenced in this publication.

Neither ENISA nor any person acting on its behalf is responsible for the use that might be made of the information contained in this publication.

ENISA maintains its intellectual property rights in relation to this publication.

COPYRIGHT NOTICE

© European Union Agency for Cybersecurity (ENISA), 2023

Reproduction is authorised provided the source is acknowledged.

This publication is licenced under CC-BY 4.0. Unless otherwise noted, the reuse of this document is authorised under the Creative Commons Attribution 4.0 International (CC-BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>). This means that reuse is allowed, provided that appropriate credit is given and any changes are indicated.

ISBN: 978-92-9204-641-5 – DOI: 10.2824/25285 – Catalogue Nr: TP-03-23-362-EN-N



TABLE OF CONTENTS

1. INTRODUCTION	6
1.1 STUDY OBJECTIVES	6
1.2 METHODOLOGY	6
1.2.1 Description of the scenario	7
1.2.2 Identification of cybersecurity and privacy threats and vulnerabilities	7
1.2.3 Identification of cybersecurity and privacy controls	7
1.3 TARGET AUDIENCE	7
1.4 USING THIS DOCUMENT	8
2. SCENARIO DESCRIPTION	9
2.1 PURPOSE AND CONTEXT	10
2.2 HIGH-LEVEL DESCRIPTION	10
2.3 ACTORS AND ROLES	12
2.4 PROCESSED DATA	13
2.5 MACHINE LEARNING ALGORITHMS	13
2.6 ASSETS	14
2.7 OVERALL PROCESS	14
2.8 PRIVACY AND CYBERSECURITY REQUIREMENTS	18
3. SECURITY AND PRIVACY THREATS AND VULNERABILITIES	22
3.1 THREAT CONTEXTUALISATION	22
3.1.1 Compromise of diagnostic system components	23
3.1.2 Evasion	24
3.1.3 Human error	24
3.1.4 Data disclosure	24
3.1.5 Poisoning (by label modification)	24
3.1.6 Unlawful Processing	25
3.1.7 Unfair processing	25
3.1.8 Lack of transparency	25
3.1.9 Diversion of purpose	25
3.1.10 No respect of data minimisation	25
3.1.11 No respect of accuracy	25
3.1.12 No respect of storage limitation	26
3.1.13 No respect of compliance of the training model	26



3.1.14	Synthesis of possible impacts and associated threats	26
3.2	VULNERABILITIES ASSOCIATED TO THREATS AND AFFECTED ASSETS	27
4.	CYBERSECURITY AND PRIVACY CONTROLS	32
4.1	IMPLEMENT A SECURITY BY DESIGN PROCESS	33
4.2	DOCUMENT THE DIAGNOSTIC SYSTEM	33
4.3	CHECK THE VULNERABILITIES OF THE COMPONENTS USED AND IMPLEMENT PROCESSES TO MAINTAIN SECURITY LEVELS OF ML COMPONENTS OVER TIME	34
4.4	ADD SOME ADVERSARIAL EXAMPLES TO THE DATASET	34
4.5	CHOOSE AND DEFINE A MORE RESILIENT MODEL DESIGN	35
4.6	INTEGRATE POISONING CONTROL IN THE TRAINING DATASET	35
4.7	ENLARGE THE TRAINING DATASET	35
4.8	SECURE THE TRANSIT OF THE COLLECTED DATA	35
4.9	CONTROL ALL DATA USED BY THE ML MODEL	36
4.10	IMPLEMENT ACCESS RIGHT MANAGEMENT PROCESS	36
4.11	ENSURE ALL SYSTEMS AND DEVICES COMPLY WITH AUTHENTICATION, AND ACCESS CONTROL POLICIES	37
4.12	MONITOR THE PERFORMANCE OF THE MODEL	37
4.13	REDUCE THE AVAILABLE INFORMATION ABOUT THE MODEL	38
4.14	IDENTIFY A DATA CONTROLLER FOR THE MEDICAL DATA PROCESSING	38
4.15	PSEUDONYMISE DATA COMING FROM THE HISTORICAL PATIENT	39
4.16	GENERATE LOGS AND PERFORM INTERNAL AUDIT	39
4.17	IDENTIFY ALL THE DATA PROCESSORS FOR THE MEDICAL DATA PROCESSING AND PERFORM THE CONTROL ACTIONS NECESSARY TO GIVE REASONABLE ASSURANCE THAT THEY ARE COMPLIANT	40
4.18	PERFORM A PRIVACY IMPACT ASSESSMENT	40
4.19	DEFINE AND IMPLEMENT A DATA RETENTION POLICY	41
4.20	STUDY ON DATA FIELDS NECESSITY AND JUSTIFICATION IN THE PRIVACY POLICY	41
4.21	FORMALIZE A LIA (LEGITIMATE INTEREST ASSESSMENT)	41
4.22	MINIMISE DATA AT EACH STEP OF THE PROCESSING; COLLECT ONLY WHAT IS NEEDED	

WHEN NEEDED	42
4.23 IMPLEMENT A PRIVACY BY DESIGN PROCESS	42
4.24 CALL ON ETHICAL COMMITTEE AND EXTERNAL AUDITS	43
4.25 DEFINE ACCURACY CRITERIA	43
4.26 ENSURE THAT THE MODEL IS SUFFICIENTLY RESILIENT TO THE ENVIRONMENT IN WHICH IT WILL OPERATE	43
4.27 RAISE AWARENESS OF SECURITY AND PRIVACY ISSUES AMONG ALL STAKEHOLDERS	44
4.28 USE RELIABLE SOURCES TO LABEL DATA	44
4.29 ENSURE THAT MODELS ARE UNBIASED	44
4.30 SUMMARY	45
5. CONCLUSION	51
A ANNEX: SECURITY AND PRIVACY SCALES AND REQUIREMENTS	52
A.1 CYBERSECURITY AND PRIVACY SEVERITY SCALES	52
A.2 CYBERSECURITY SCALE OF IMPACT	53
A.3 PRIVACY SCALE OF IMPACT	53
A.4 PRIVACY REQUIREMENTS CRITERIA	54

EXECUTIVE SUMMARY

Given the great influence of artificial intelligence (AI) in people's daily lives due to the key role it plays in digital transformation through its automated decision-making capabilities, ENISA aims to raise awareness of cybersecurity and privacy threats related to various scenarios using artificial intelligence. To this end, ENISA, with the support of the Ad-Hoc Working Group on Artificial Intelligence Cybersecurity, has published two reports in the last two years: Cybersecurity Challenges of Artificial Intelligence¹ and Securing Machine Learning Algorithms².

ENISA continues its momentum with a new report on cybersecurity and privacy in medical imaging diagnosis, which is supported by AI. An in-depth study of the scenario has been conducted by identifying first the assets, the actors and their roles, relevant processes, the AI algorithms used, as well as the requirements in terms of cybersecurity and privacy needed for it. Building upon previous ENISA work such as the "Securing Machine Learning Algorithms" report cited above, in addition to legislation such as GDPR and literature searches, this report has identified cybersecurity and privacy threats and vulnerabilities that can be exploited in the examined scenario. While focus is on ML-related threats and vulnerabilities, broader AI considerations were also taken into account. Lastly, corresponding cybersecurity and privacy controls that consider the context of the scenario and the impact of the associated threats/vulnerabilities were defined. The specificities within the implementation of these controls are described, including possible trade-offs between cybersecurity, privacy, and performance. Each control is classified as a cybersecurity control, a privacy control, or a mixture of both, depending on the threats it mitigates and their associated impacts (cybersecurity impacts, privacy impacts, or both).

This report allows better assessment of the reality that artificial intelligence brings its own set of threats, which consequently insists on the search for new security measures to counter them. Finally, it should be noted that this guide strongly emphasises privacy issues in the same way as cybersecurity issues, privacy being one of the most important challenges facing society today. Security and privacy are intimately related, but both equally important, and a balance must be made specific to each use. As a result, as seen in this report, efforts to optimise security and privacy can often come at the expense of system performance.

¹ See <https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges>

² See <https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms>

1. INTRODUCTION

Abuse of Artificial Intelligence (AI), which has developed significantly in recent years, has been identified by ENISA as one of the top emerging threats³. By providing new opportunities to solve decision-making problems intelligently and automatically, AI is being applied to more and more business cases in a growing number of sectors. The associated benefits are significant. However, the development of AI is also accompanied by new threats which project teams will have to face.

In many projects, it is apparent that these new threats are related to aspects of cybersecurity in addition to privacy, particularly when AI is used for innovative projects whereby processing personal data takes place. In this spirit, on 21 April 2021, the Commission published the AI Act proposal⁴, which sets out requirements (including cybersecurity and privacy) for AI systems deemed to be high-risk (e.g., used in biometric identification or critical infrastructure management and operation) in order to mitigate threats to health, safety, and fundamental rights.

To go further into these high-risk AI systems, ENISA proposes this new report: "Cybersecurity and Privacy in AI – Medical Imaging Diagnosis", which builds directly on the work already initiated by ENISA since 2020 on the identification of risks associated with AI. **This new report analyses cybersecurity and privacy requirements and measures in use of AI in medical imaging diagnosis of osteoporosis. The report describes the scenario fundamental principles (assets, actors processes etc.), identifies the security and privacy risks it poses, and finally cybersecurity and privacy controls, which counteract the identified risks.**

1.1 STUDY OBJECTIVES

Findings from the ENISA's report on securing machine learning algorithms⁵ indicate that there is no uniform strategy in applying a specific set of security controls to protect machine learning algorithms and in some cases, deployed security controls may result in trade-offs in security and performance. ENISA therefore recommends that organisations which use AI systems should perform detailed analysis of their own AI systems, and conduct targeted risk assessments to find the appropriate balance between cybersecurity, privacy and performance.

The objectives of this publication are then as follows:

- Provide a detailed description of a **medical imaging diagnosis** scenario
- Identify AI cybersecurity and privacy measures taking into account requirements, threats and vulnerabilities defined for this scenario and practical guidance on how to implement them
- Provide recommendations on how to balance the trade-offs between cybersecurity, privacy, and performance in this scenario

1.2 METHODOLOGY

Production of this publication was undertaken in three stages:

- Identify and describe in detail the medical imaging diagnosis scenario.

³ <https://www.enisa.europa.eu/news/cybersecurity-threats-fast-forward-2030>

⁴ <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>

⁵ <https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms>

- Identify cybersecurity and privacy threats associated with the scenario
- Identify relevant cybersecurity and privacy controls

1.2.1 Description of the scenario

The scenario description is aligned with ENISA's previous work on AI^{6,7} and provides the following information:

- The purpose and the context
- A high-level description of the scenario, highlighting the data encountered
- The involved actors and their associated roles
- A detailed description of the data
- Machine learning algorithms
- Other assets (besides data) associated with the described scenario
- The overall process of the scenario
- Security and privacy requirements

1.2.2 Identification of cybersecurity and privacy threats and vulnerabilities

This section of the report focusses on the threats and vulnerabilities related to use of AI in medical imaging diagnosis scenario. Based on previously mentioned ENISA work, legislation such as GDPR, and desk research, this report identified the cybersecurity and privacy threats and vulnerabilities that can be exploited. Given the prevalence of machine learning (ML) and ENISA's past work on the topic, there is more emphasis placed on threats and vulnerabilities related to ML. Nonetheless, wider considerations of AI have been taking into account when identifying security threats and vulnerabilities as well as dedicated to privacy, for which GDPR data protection principles⁸ were used as a starting point of our analysis.

1.2.3 Identification of cybersecurity and privacy controls

Following the analysis of medical imaging diagnosis scenario, the identification of threats (and their impact), and associated vulnerabilities, this section of the report presents the corresponding cybersecurity and privacy controls that:

- Take into account the context of the described scenario
- Take into account the security and privacy impact of the threat/vulnerabilities (as described in this report)

The specificities of implementation of such controls are described including the possible trade-offs between cybersecurity, privacy and performance. Each control can be either a cybersecurity control, a privacy control, or a mixture of both, depending on the threats it mitigates and their associated impacts (cybersecurity impacts, privacy impacts, or both). Many of the controls are of technical nature, but when appropriate (e.g., per GDPR requirements), organizational controls have also been identified.

1.3 TARGET AUDIENCE

The target audience of this report can be divided into the following categories:

- **All actors (private or public):** to help them in their risk analysis, in the identification of cybersecurity and privacy threats and in the identification of the appropriate security and privacy controls to mitigate the threats related to this scenario.

⁶ See <https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms>

⁷ See <https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges>

⁸ See Article 5 of GDPR



- **AI technical community, AI cybersecurity and privacy experts and AI experts** (designers, developers, ML experts, data scientists, etc.) with an interest in developing secure solutions and in integrating security and privacy by design in their solutions.
- **Cybersecurity and privacy community:** to help in identifying cybersecurity and privacy threats related to medical imaging diagnosis and in identifying the appropriate security and privacy controls to mitigate the threats.

1.4 USING THIS DOCUMENT

Although we based the scenario close to reality, some assumptions have been made for the purposes of its analysis. These assumptions should be reassessed by the reader when doing their own analysis. Please note this list is non-exhaustive:

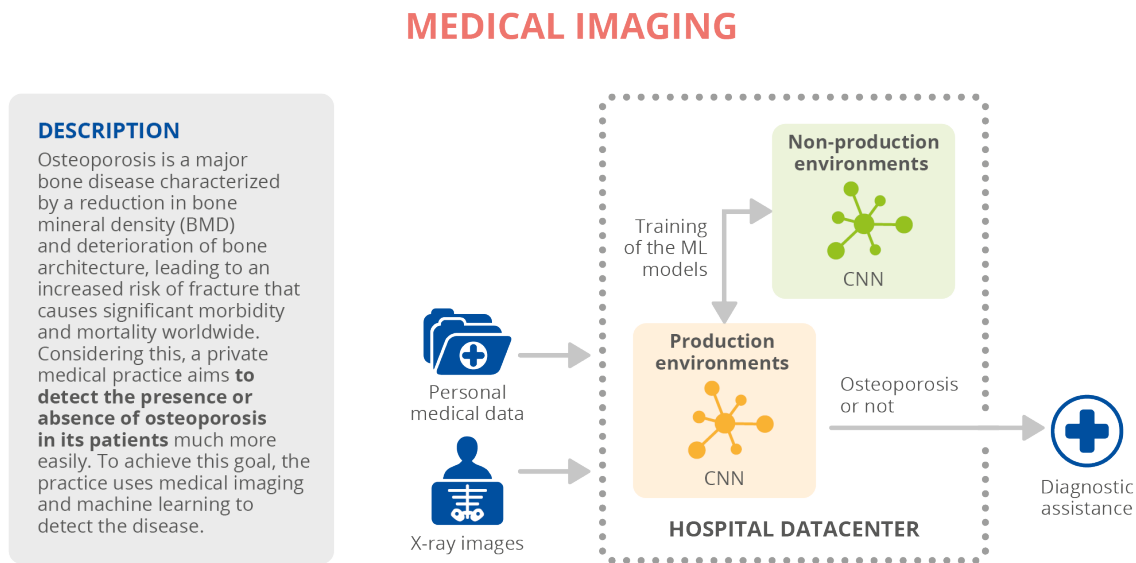
- The algorithms chosen for the scenario are based on desk research, and there may be other algorithms that are better suited for it.
- Regarding the privacy aspects, we put ourselves in the shoes of the data controller, but only based on the hypothesis that we have made. It is left to the attention of the audience that the privacy requirements as well as the control measures must always be adapted to a context and a situation.
- After identifying potential threats, this report identifies security and privacy controls that could be applied to the scenario. However, as is the case for any application of ML, one must also consider traditional security standards (e.g., ISO 27001/2, NIST) because ML applications are also subject to more global threats.

It should also be kept in mind that the elements of this report are valid as of the date of publication, and could evolve over time.

2. SCENARIO DESCRIPTION

The following figure presents an overview of all topics that will be addressed in this chapter.

Figure 1: Scenario overview



DESCRIPTION

Osteoporosis is a major bone disease characterized by a reduction in bone mineral density (BMD) and deterioration of bone architecture, leading to an increased risk of fracture that causes significant morbidity and mortality worldwide. Considering this, a private medical practice aims to **detect the presence or absence of osteoporosis in its patients** much more easily. To achieve this goal, the practice uses medical imaging and machine learning to detect the disease.

DATA

Data used to build the model

- Historical X-rays of patients with or without osteoporosis
- Data related to age, gender, and body mass index of historical patients of the medical cabinet

Data used once the model is in production

- X-rays of patients who come for consultation
- Data related to age, gender, and body mass/ Fairness index of who come for consultation

CYBERSECURITY AND PRIVACY REQUIREMENTS

Cyber requirements

- Availability ● Integrity ● Confidentiality ● Traceability

Privacy Requirements

- Availability ● Integrity ● Confidentiality ● Traceability
- Lawfulness
- Transparency
- Purpose limitation
- Data minimization
- Accuracy
- Storage limitation
- Security of personal data
- Database creation
- Compliance of the training model

- Critical ● High ● Low

ACTORS

- Radiologists/medical practice
- Large tech companies
- Historical Patients
- New Patients
- Cloud provider
- Data scientists
- Developers and Data Engineers
- System and communication network's administrator

ASSETS

- CNN-algorithm used
- Data lake - in the cloud
- Model server - in the cloud
- Scanner
- X-ray computer-aided diagnostic system. on-premises
- Integrated Development Environment
- Libraries
- Communication protocols and network

2.1 PURPOSE AND CONTEXT

Osteoporosis is a major bone disease characterised by a reduction in bone mineral density (BMD) and deterioration of bone architecture, leading to an increased risk of fracture, that causes significant morbidity and mortality worldwide. The diagnosis of the disease is mainly through X-ray scans.

However, analysis of such scans can be very time consuming. For instance, according to a 2017 report from Royal College of Radiologists⁹, 230,000 patients waited more than a month for X-rays results in the UK NHS in 2016. This is largely due to a large increase in the number of X-rays combined with a plateauing number of radiologists. This resulted in an increase in X-ray analysis time.

For this reason, a medical practice¹⁰ specialised in radiology has defined the following purposes:

- Reduce the time it takes for a radiologist to analyse an X-ray scan for osteoporosis suspicion
- Improve the diagnosis of osteoporosis

To reach its objectives, the medical practice seeks to assist radiologists by providing a probability score of presence of osteoporosis in an X-ray scanner. In practical terms, for each patient, this is done with the following three steps:

1. The patient has an X-ray scan with a scanner
2. This X-ray scan is sent back to the radiologist for analysis. Meanwhile, the image is sent to a tool that calculates the probability that the image presented has osteoporosis
3. The radiologist analyses the image together with the probability that the patient has osteoporosis. This probability does not replace the analysis, but it supports the practitioner in their decision, or highlights areas of concern

2.2 HIGH-LEVEL DESCRIPTION

As described above, the final goal is **to detect the potential presence of osteoporosis by giving the radiologist a probability that the bone contains the disease**. To do so, a model of supervised learning is particularly adapted.

Technically, the machine learning model can assist the radiologist in the following way:

- The scanner sends the image to a computer with a dedicated software to which it is connected by wire. This computer is called X-ray computer-aided diagnostic system.
- The radiologist enters the patient's personal information including first name, last name, age, sex, and body mass index (BMI) into the X-ray computer-aided diagnostic system.
- The system sends, via API calls, the image, age, sex, and BMI to a cloud server provided by a cloud provider which hosts the machine learning model.
- The model returns its diagnosis (probability that the bone contains the disease) to the X-ray computer-aided diagnostic system.
- The radiologist makes his diagnosis using the X-ray image, their expertise, and the probability returned by the model.

The patient's file (including first name, image, model probability, age, BMI) is then sent to a database in a cloud data lake to be stored there for a pre-defined period. In case of a trial, this period allows, by the relevant prescribing law, audit of the diagnosis made by the radiologist. The images, age, sex, BMI, and model probability (after they have been pseudonymised) are

⁹ See https://www.rcr.ac.uk/sites/default/files/backlog_survey_feb_2016.pdf

¹⁰ Medical practice means a business entity in which physicians practice medicine together as partners, shareholders, owners, members or employees, or in which a single physician practices medicine. Furthermore, in our scenario, this includes at least public and private hospitals that offer radiology services.

sent to another database, which is accessible to the data scientists and radiologist in charge of the project to improve the model by providing new data.

To build this machine learning model, a database containing radiographic images of former patients from the medical practice has been constructed. This database contains both radiographs of people suffering from the disease and other radiographs of healthy people who do not suffer from any other disease. As noted, all images are from a historical library of images from the medical practice, and the models that are trained to ensure the disease detection are housed in the cloud. It should be added that information about the sex, age, and body mass index of the patient from whom the radiograph was taken is associated with this radiograph. This means that for each radiograph, the information on age, sex, and BMI is visible.

The major actors involved in this scenario – outside the team of data scientists and data engineers, are the medical staff (in which the key stakeholder is the radiologist) and the patients. The radiologists' expertise is used initially to annotate as many images as possible and to confirm the presence or absence of osteoporosis in each annotated image (it is these annotated images that form the database described above).

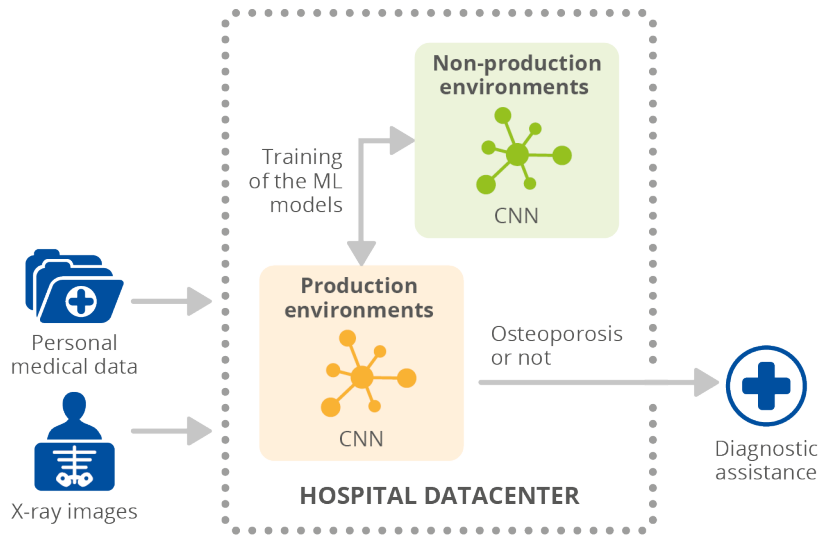
It is essential to note that this scenario refers to two types of patients. The first group is used to build the image libraries (i.e., historical osteoporosis patients or otherwise healthy). The second group are the future patients who will come to the medical office for a consultation. Although the data manipulated in our case are the same (**X-rays of patients**), they come from different sources – one source of **historical patients** whose data is part of the training database, and another source of **new patients** who require diagnosis:

- **X-ray images** and medical data of **historical patients** (**before** the occurrence of the diagnosis) that form the database (these ones are annotated).
- **X-ray images** and medical data of the **new patients** (**during** the occurrence of the diagnosis) who come for consultation.

It should also be noted that a pre-trained machine learning model is used. Training an accurate neural network is costly - large tech companies train heavy model on very large databases, sometimes with billions of datapoints – as such, most medical practices cannot afford to train their own model. Hence, pre-trained models are made publicly available. These only require a small amount of re-training with the osteoporosis database to transfer the knowledge of the neural network to this scenario (this practice is referred to as Transfer Learning). An additional explanation on why transfer learning is used is provided later in this report, in the model deployment section for this report. Security risks associated with transfer learning are described in detail in subsequent sections of this report.

The following sections aim to detail such scenario by describing actors and their roles, processed data, the assets that allow the project to exist, the explanation of the lifecycle, and the privacy and cybersecurity requirements applied to it.

Figure 2: High-level description



2.3 ACTORS AND ROLES

The following actors are involved in the medical imaging diagnosis scenario.

Figure 3: Actors, roles, and their description

Actor	Role	Description
Radiologists/medical practice	<i>End Users and Data Owner (Data Controller)</i>	Radiologists have the historical role of analysing the results of the radiographs. They have an important role to play as they can act as verifiers. The medical staff also has the role of providing the essential data for the training and testing of the algorithm, as it manages the X-rays database. Radiologists also have their role in producing quality x-rays and images to ensure the right diagnosis can be made.
Large tech companies	<i>Model Provider</i>	Large tech companies develop libraries and pre-trained models that are published to be reused by the data science community. In this case, they provide a pre-trained model.
Historical Patients (before the occurrence of the diagnosis)	<i>Data Provider</i>	The historical patients are patients who were consulted in the medical practice before the use of AI for the diagnosis of osteoporosis. The X-ray images from their consultation have been anonymized and now form the database used to train the model.
New Patients (during the occurrence of the diagnosis)	<i>Data Provider</i>	New patients are those patients who have come to the radiologist's office for a consultation since the AI detection of the disease became available. Their data are also used in the future to better train the model.
Cloud Provider	<i>Cloud Provider</i>	The Cloud Provider is a third party that offers computational platforms, data lakes and some data analysis capabilities or "Machine learning as service". ¹¹
Data Scientists	<i>Data Scientists</i>	Data Scientists are the spearhead of the algorithms. They oversee cleaning the data, preparing it, building the models, and ensuring the relevance of the results. They work with the medical staff to build a relevant model.
Developers and Data Engineers	<i>Developers and Data Engineers</i>	Developers and Data Engineers are responsible for routing data, transformation, and other technical operations as required.

¹¹ Inspired from the definition of ENISA, AI Cybersecurity Challenges, December 2020



System and communication Network Administrator	<i>Network Administrators</i>	System and communication Network Administrators are responsible for the network (flows etc.) and configuring the cloud servers that make the application work or the data stored.
---	-------------------------------	---

2.4 PROCESSED DATA

As previously mentioned, data in this scenario comes from two sources and its use is detailed in the following figure.

Figure 4: Data needed for detecting osteoporosis from panoramic X-ray

Data	Data type	Source / data provider	Data Procurement
Data used for building the model			
X-rays of patients with or without osteoporosis (pseudonymised)	Image	Database where historical patients' radiographies are stored. The data provider is former patients .	The data are already available. The images need to be labelled (i.e., analysed by experts) to determine if there are signs of osteoporosis in the image, and measure the performance of the model ^{12 13} .
Data related to age, sex, and body mass index of historical patients of the medical cabinet (pseudonymised)	Structured data	Database where historical patient information is stored. For each patient, their radiograph is associated with this information. The data provider is former patients .	The data are already available. For each X-ray, this information is stored to facilitate the diagnosis (age, sex, BMI) ¹⁴ . Note that such information is pseudonymised ¹⁵ .
Data used once the model is in production			
X-rays of patients who come for consultation	Image	These data are used when a patient is consulted in the medical office. The data provider is patients .	These data are obtained during the consultation when the patient is subject to the X-Ray system
Data related to age, sex, and body mass index of who come for consultation	Structured data	These data are used when a patient is consulted in the medical office. The data provider is the patient , but it is the radiologist who seizes this data.	These data are obtained during the consultation when the patient is subject to the radiography system

2.5 MACHINE LEARNING ALGORITHMS

The objective of using machine learning in this scenario is to detect osteoporosis from panoramic X-rays. The ML model used for this purpose is a **Convolutional Neural Network (CNN)**. Besides the fact that this algorithm is one of the best models used for computer vision, its choice is also justified by the fact that it is widely used for scenarios very close to the one at hand, e.g., chest, breast, brain, musculoskeletal system, and abdomen and pelvis X-rays. CNN systems are widely used in the medical classification task. CNN is an excellent feature extractor, therefore utilising it to classify medical images can avoid complicated and expensive feature engineering.

¹² Soffer, Shelly, et al. Convolutional Neural Networks for Radiologic Images: A Radiologist's Guide. s.l. : RSNA, 2019. <https://pubmed.ncbi.nlm.nih.gov/30694159/>

¹³ Lee, Jae-Seo, et al. Osteoporosis detection in panoramic radiographs using a deep convolutional neural network-based computer-assisted diagnosis system: a preliminary study. 2018. <https://pubmed.ncbi.nlm.nih.gov/30004241/>

¹⁴ *Augmenting Osteoporosis Imaging with Machine Learning*. Valentina Pedoia, Francesco Caliva, Galateia Kazakia, Andrew J. Burghardt, Sharmila Majumdar. 2021 <https://pubmed.ncbi.nlm.nih.gov/34741729/>

¹⁵ See <https://datasaveslives.eu/pseudonymisation#:~:text=In%20many%20cases%20this%20may,always%20suitable%20for%20healthcare%20research.>

Figure 5: Machine learning algorithms used

Learning paradigm	Subtype	Algorithm	Type of data ingested	Description
Supervised Learning	Classification	CNN	Image	A convolutional neural network (CNN) is a class of artificial neural networks mainly focused on applications like object detection, image classification, recommendation systems, and are also sometimes used for natural language processing ¹³ .

2.6 ASSETS

In addition to the previously described data, the scenario is supported by the following additional assets, which are described in the figure below.

Figure 6: asset's description

Type of asset	Asset	Description
Models	CNN, for classification of images between "osteoporosis" and "no osteoporosis"	In this scenario a pre-trained model is used. The weights of the neural networks are calculated on a large public database for the model to be able to classify objects between thousands of classes. These pre-trained models are available open-source and downloadable via the internet. The last layers of this pre-trained neural network are then re-trained. The model takes as inputs: images of bones with osteoporosis, and images of healthy bones with the associated labels to specialise the model in detecting osteoporosis The model is a CNN where personal data such as age, sex or BMI are added just before the affine layer to improve performances.
Environment tools	Data lake – <i>in the cloud</i>	The data ingestion platforms that enable data engineers to store data and exploit them.
	Model server – <i>in the cloud</i>	The model server is used to store the model and process the training and retraining phases.
	Scanner	A diagnostic X-ray machine consisting of a tomography system and a computer that provides the results in the form of images.
	X-ray computer-aided diagnostic system – <i>on-premises</i>	After each X-ray, the image and the results of the model are shown to an expert thanks to this computer and its software.
	Integrated Development Environment	Software application that provides comprehensive development facilities to computer programmers.
	Libraries (with algorithms for transformation)	Collection of precompiled code that can be used in a project to realise well-defined operations.
	Communication protocol and communication Networks	A system of rules allowing a group of entities to communicate with each other to share or ask information.

2.7 OVERALL PROCESS

The purpose of the scenario is predicting the presence or absence of osteoporosis in a patient based on medical imaging data. To do this, Data Engineers, Data Scientists and Radiologists work closely together to enable the system to learn how to detect this bone disease.

In the following paragraphs, our scenario is described by incorporating it into the different stages of the machine learning lifecycle:

Data collection

The first step is to determine how to obtain the data to feed the model. For this, the medical practice already has the necessary medical imaging data (a database of historical patients with or without the disease, with their age, sex, and BMI) **which must be annotated**. As a pre-trained algorithm is used, a relatively small amount of data (reasonably for a medical practice) is needed to re-train the last layers of the neural networks. This is where the expertise of the radiologists is used. Their first role is to annotate many images and for each of them to confirm (or reject) the presence of osteoporosis. Once annotated, these data are routed and stored in a cloud data lake provided by a Cloud Provider and maintained by a Data Engineer and System/Network Administrator. The data thus obtained are images, each with a class describing the state of the bones. The amount of data obtained being small to train a model with satisfactory performance, it is more convenient to use a neural network pre-trained on medical or non-medical imaging to enforce the performance of the model.

Data cleaning and data pre-processing

The collected data are then **cleaned**, ensuring that the data is of high quality and structure, by Data Scientists for **pre-processing**. The pre-processing phase consists of standardising the size of all the panoramic radiographs. For efficient processing in terms of computation and memory requirements, the images are subsampled to a uniform size using bilinear interpolation.

Model design and implementation

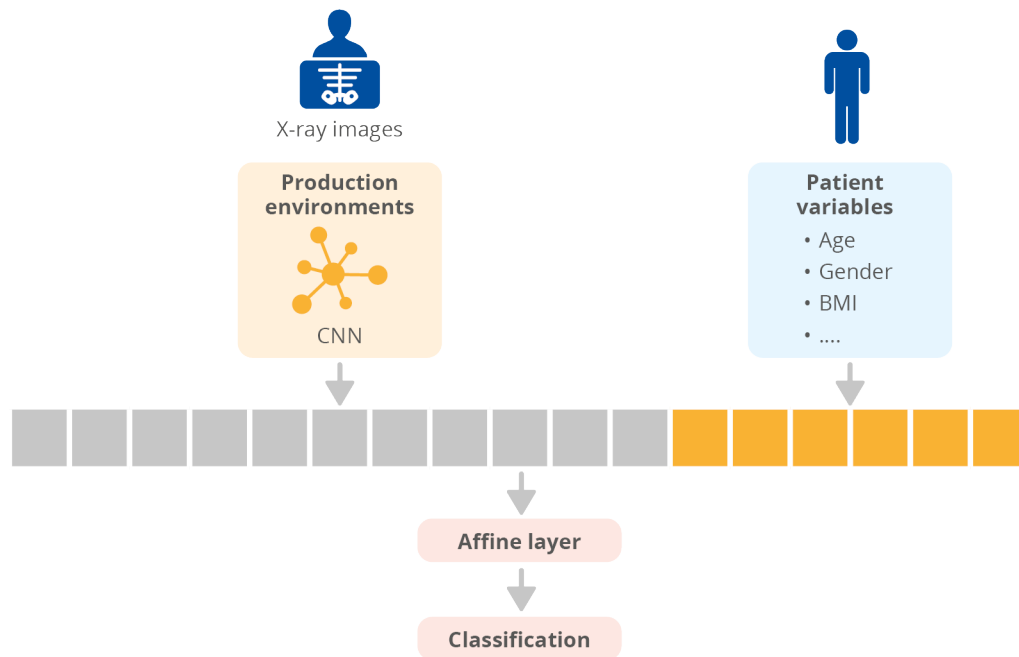
As mentioned earlier, the algorithm used in our scenario is CNN, where personal data such as age, sex or BMI are added just before the affine (fully connected) layer to improve performance. Details on the choice of the algorithm can be found in an earlier section of this report.

One aspect that is important regarding the performance of the model is to be sure that both “normal” and “osteoporosis” images are equally represented. It could be necessary to that extend to proceed to data augmentation, transforming real osteoporosis X-rays to create derivative images and increase the number of examples for this label.

To build the neural network, a pre-trained CNN is used. Pre-trained CNN are publicly made available by large tech companies, and are trained on billions of images, requiring a huge amount of computing power. For radiologists to benefit from Deep Learning without the huge number of resources needed for training and fine-tuning the model, pre-trained models are essential. According to the amount of labelled data available, one or more layers of the network can be re-trained to adapt better to the use it is destined to. Most of the pre-trained CNN architectures available are proven to produce human-level accuracy of medical imaging diagnosis¹⁶. At the end of the CNN, an affine layer is added to use the personal information of the patient to evaluate their bone's health.

¹⁶ Rajpurkar, Pranav, et al. *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*. 2017. <https://ui.adsabs.harvard.edu/abs/2017arXiv171105225R/abstract>

Figure 7: Synthesis of the CNN structure



The input of the network is the grey level of the ROI (Region of Interest) in a defined pixel area and the personal data of the patient (age, sex, BMI, ...), and **the output of the network** is the conditional probability distribution over the two categories, "osteoporosis" and "normal". The network consists of a stack of several convolutional layers, in addition to maximum pooling layers and activation layers.

The topmost layer (or output layer) is regularly a SoftMax classifier, that converts a K-dimensional vector of arbitrary real values into a K-dimensional vector of real values in the range (0, 1) that sum to 1.

Model training, model testing and optimisation

Regarding **the training method** phase (if the available data are sufficient and if it is considered as necessary regarding the performance metrics), the training is done in the cloud servers always provided by the same cloud provider. The dataset is divided into training, validation, and test sets. 80% of the X-rays are randomly selected for training, 10% for validation, and the remaining 10% are reserved for testing. Randomness of the split is made with caution: it is important to respect the osteoporosis images ratio among all the sets.

Model Evaluation

To **evaluate the model**, regular classification metrics are used: accuracy, recall, precision f-score and AUC score¹⁷. In fact, radiologists may prefer to detect osteoporosis even if there is no osteoporosis rather than missing this diagnosis – so emphasis is placed on reducing the number of false negatives¹⁸ (i.e., augmenting recall). Even if the false positive rate must be low, a lot of misdiagnoses could be made, incurring high cost.

Model Deployment

The model is deployed in the cloud by the Developers and the System Network Administrator.

¹⁷ Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. Taha, Allan and Aziz, Abdel. 2015. <https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-015-0068-x>

¹⁸ See <https://towardsdatascience.com/should-i-look-at-precision-recall-or-specificity-sensitivity-3946158aace1>

As described in the high-level description, the model is called by APIs by the computer aided diagnostic system when required. This is usually as the radiologist uses the scanner with a patient.

Monitoring and inference

The neural network by itself is not sufficient to make a diagnosis. Verification by an expert remains necessary, but the expert's work is considerably facilitated by the presence of machine learning. The verification of the medical image by a radiologist allows for an expert evaluation. However, automation bias (when the radiologist overly trusts the results and gradually loses his own skill) is a problem that should be taken into account in the medium term.

For every patient, the images, age, BMI, model probability and the radiologist's diagnostic are sent to a specific database in the cloud data lake (different from the one with the historical data). This is accessible to the Data Scientists and Radiologist in charge of the project to improve the model by providing new data when necessary for retraining the model. Performance of the model is continually monitored¹⁹ to keep accuracy above the criteria defined by experts. Once performance drops below that criterion, the model is retrained.

Eventually, every patient's file (for instance name, first name, image, model probability, age, BMI) is sent to a database (different from the two above) in the cloud data lake to allow audit of the diagnosis made by the radiologist in case of trial.

Figure 8: Synthesis of the ML lifecycle and involved actors

Steps	Description	Actors	Assets
Data Collection	Historical data of images are annotated by radiologist and put in a table in a cloud data lake with information like gender. All those Information (included images) coming from other historical tables in the data lake and are anonymised.	Data Engineer/Developers Data Scientist System and Communication Network Administrators Radiologist, Cloud provider	Data lake Programming Libraries Integrated development environment
Data Cleaning	Data, notably images, are cleaned to keep only the high-quality data.	Data Engineer/Developers Data Scientist Radiologist Cloud provider	Data lake Integrated development environment Programming Libraries
Data pre-processing	The pre-processing phase consists of standardising the size of all the panoramic radiographies. Indeed, the images are subsampled to a uniform size of pixels using bilinear interpolation.	Data Engineer/Developers Data Scientist Cloud provider	Programming Libraries Data lake, Model servers Integrated development environment
Model design and implementation	A pre-trained CNN is used for this scenario. At the end of the CNN, an affine layer is added to use the personal information of the patient to evaluate his bone's health.	Data Scientist Large Company Cloud Provider Large tech company	Programming Libraries Data lake, Model servers Integrated development environment
Model training	The model is then trained on the model servers.	Data Scientist Cloud provider	Programming Libraries Data lake, Model servers Integrated development environment

¹⁹ One way to monitor is to supervise the new database with radiologist and see the diagnostic errors of the model.

Model testing	10% of the most recent data are used for testing.	Data Scientist Cloud provider	Programming Libraries Data lake, Model servers Integrated development environment
Optimization	The hyperparameters of the model are optimised through different techniques such as Early stopping.	Data Scientist Cloud provider	Programming Libraries Data lake, Model servers Integrated development environment
Model evaluation	Model is evaluated by some classification metrics like recall or precision f-test. Radiologists prefer to detect osteoporosis even if there is no osteoporosis – so emphasis will be placed on reducing the number of false negatives.	Data Scientist Cloud provider	Programming Libraries Data lake, Model servers Integrated development environment
Model deployment	The model is put into production on dedicated cloud servers and incorporated in the computer aided diagnostic system.	Data Engineer/Developers Radiologist System and communication network's administrator Cloud Provider	Model server Computer aided diagnostic system Integrated development environment
Monitoring and inference	Radiologist make their diagnostic using the help of probability score return by the model. Data like image are stored in a database in case of re-training. Patients' files are store into a secure database in case of trial.	Data Scientist Radiologist Patients System and communication network's administrator Cloud Provider	Data lake Model server Computer aided diagnostic system Scanner Integrated development environment

2.8 PRIVACY AND CYBERSECURITY REQUIREMENTS

Cybersecurity requirements

The context information given in the previous section enables evaluation of the application's cybersecurity and privacy requirements. The following table summarises the cybersecurity requirements.

Figure 9: Cybersecurity requirements

	Level	Explanation
Availability	Low	The model set up is an aid for the Radiologist. The system can be unavailable for a few days if a radiologist is present to manually analyse the radiographs.
Integrity	Critical	Data, whether used for training or made by the model, must be accurate at all steps of the lifecycle. Alteration could lead to an erroneous prediction and thus errors in the detection and treatment of the disease.
Confidentiality	Critical	Part of the processed data are medical data and are considered as personal data. It is therefore confidential. It must be processed in such a way that it complies with the GDPR.

Traceability	High	Actions linked to the process must at least be logged (traced and dated) to be able to follow up who did what on the algorithm. As personal data are used by the process, actions linked must be imputable. I.e., we must be sure to trace all actions (even consultation actions) and to be able to attribute them without possibility of repudiation.
---------------------	-------------	---

Privacy requirements

On one hand, for the model training, X-rays of medical patients are used. In addition, data indicating the age, sex and BMI of the patients is associated with each radiograph. However, these data are pseudonymised in the medical practice database for the purpose of training machine learning models as outlined as the purpose of this scenario.

On the other hand, **personal data are processed when patients come to the practice and are diagnosed for osteoporosis, adding information like last name, name, and consultation date in the patient’s file.**

Because the scenario manipulates personal data, **the following privacy requirements and recommendations should be satisfied.**

Figure 10: GDPR Data protection principles

Requirements	Explanation
Lawfulness, fairness, and transparency	<p>Health data are special categories of personal data as highlighted by Article 9 of the GDPR. Personal data collected (both in training and in production) in the purpose of this scenario must be processed lawfully. That is, the processing must be based on one of the legal bases provided by Article 6 of GDPR. In addition, individuals' personal data must be processed only in the way they reasonably expect, and any unexpected but justified processing must be clearly explained. Data processing must also comply with the transparency obligations of the right to be informed.</p> <p>Lawfulness: One of the main actors is the medical practice which is a private medical actor. The medical practice must justify that the processing of these data is necessary to safeguard its legitimate interest²⁰. As noted in the introduction, the processing of this data accelerates and improves the ability of the medical practice to detect osteoporosis. Indeed, radiologists make their diagnosis more quickly and no longer give patients very long appointment times. The data of the patients who come for consultation are stored in a different database than the one built for training the models. These data are kept for consultation in case of legal disputes. The processing of this data is also of vital interest to patients, as the time needed to find an appointment for a consultation is considerably reduced and the consultation is done much more quickly. The patient knows at an early stage if he suffers from the disease and can take a treatment accordingly. In addition, user consent may be required not by the GDPR but by the public health code of the country in which the scenario applies²¹.</p> <p>Fairness: even if the medical practice demonstrates that the data processing has a legal basis, it must only process personal data in the way that patients expect, i.e., only for the detection of osteoporosis, and not use it in a way that has an unjustified negative effect on them.</p> <p>Transparency: General information concerning the activities related to the scenario must be provided to the persons concerned (posting in the premises, mention in the reception booklet, etc.). In addition to this general information, individual information must be provided to the patient. This should be done by the radiologist for each patient during the first interviews with them. The practice must be transparent with patients and clearly inform them how and why their personal data are used throughout the machine learning lifecycle.</p>
Purpose limitation	<p>Personal data collected (both in training and in production) need to be collected for specific, explicit, and legitimate purposes and is not further processed in a manner that is incompatible with those purposes. This means that the medical data collected and treated should only be used to assist the radiologist in his osteoporosis diagnosis.</p>

²⁰ Health data are handled and are special categories of personal data whose processing is prohibited unless there are exceptions, but in the context of our scenario, we fall into one of these two exceptions: article 9.2.h of the GDPR: See <https://gdpr-info.eu/art-9-gdpr/>

²¹ See <https://www.cnil.fr/fr/recherches-dans-le-domaine-de-la-sante-ce-qui-change-avec-les-nouvelles-methodologies-de-reference>



Data minimisation	Personal data collected (both in training and in production) need to be adequate, relevant, and limited to what is necessary in relation to the purpose of detecting osteoporosis. The medical practice should justify and should indicate the procedures in place to ensure that only essential data such as X-rays are manipulated.
Accuracy	Personal data collected (both in training and in production) must be accurate and up to date. The medical practice must ensure that the data manipulated remains accurate. Indeed, the radiographs manipulated by the radiologists must be adequately annotated and verified by the other practitioners in the practice. No mistakes should be made and if there are any, they should quickly be detected and corrected.
Storage limitation	Personal data collected (both in training and in production) must be kept in a form which permits authentication of data subjects for no longer than is necessary for the purposes for which the personal data are processed. Specifically, data retention periods will be applied in accordance with the requirements of the local medical laws of countries and supervisory authorities' recommendations where such scenario occurs.
Security of personal data (Integrity and Confidentiality)	Personal data collected (both in training and in production) must be processed in a manner that ensures appropriate security, including protection against unauthorised or unlawful processing etc. In addition, the risks from data corruption and leakage are critical for patients. The practice must implement adequate security measures to ensure data integrity and confidentiality (as there is a critical need for confidentiality and integrity due to the nature of the data).

As a complement to the GDPR data protection principles listed above, other key privacy topics related to AI systems need to be addressed (some of which are specifically mentioned by national data protection supervisory authorities)²²:

Figure 11: Data protection supervisory authorities' recommendations for AI systems

Recommendations	Explanation
Database creation	After personal data of the consulted patients is collected, they are stored in the medical practice database. This data is kept for consultation in case of legal disputes. Therefore, the medical practice must ensure that only authorised persons have access to the database and prevent data loss. The medical practice must also ensure that only the right data (real radiographs) are injected into the database.
Compliance of the training model (i.e., before production)	The medical practice must justify that the training model it has chosen to achieve its objectives is relevant and unbiased. The practice must ensure that before the system is put into production, it is not discriminatory and that radiologists can verify the effectiveness of the algorithm's decisions to reduce the potential risk of diagnostic errors.

Finally, the project requires a data protection impact assessment²³ as it completes at least one criterion in the following table²⁴:

Figure 12: Is a Data Protection Impact Assessment (DPIA) necessary?

Criteria	Does it match the criteria?	Justification
Evaluation or scoring		
Automated decision making with legal or similar significant effect		
Systematic monitoring		

²² See <https://www.cnil.fr/fr/intelligence-artificielle/la-cnil-publie-ressources-grand-public-professionnels>

²³ See <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/#:~:text=DPIAs%20are%20a%20legal%20requirement,trust%20and%20engagement%20with%20individuals>

²⁴ The condition for a DPIA to occur is that at least two conditions are satisfied.



Sensitive data or data of highly personal nature	X	Medical data is manipulated in addition to the names, age, and sex of the patients. This data is also stored in databases without being anonymised.
Data processed on a large scale	X	If the medical practice has thousands of patients, personal data is used on a large scale.
Matching or combining datasets		
Data concerning vulnerable data subjects	X	The data is from the patients. Patients can be considered vulnerable because they can potentially have an illness and live in a state of stress, willing to do anything to get better.
Innovative use or applying new technological or organizational solutions	X	Prediction of osteoporosis with artificial intelligence is an innovative solution.
Preventing data subjects from exercising a right or using a service or contract		

To assess the security risks in terms of privacy, personal data processing must consider the following requirements:

Figure 13: Privacy requirements regarding data

	Level	Explanation
Availability	Low	The data, for the user, does not necessarily have to be available. An unavailability of several days is acceptable.
Integrity	Critical	The consequences of poor analysis caused by lack of data integrity can be very serious. Alterations could lead to erroneous prediction, therefore errors in disease detection and treatment.
Confidentiality	Critical	Part of the processed data are medical data and are considered as special categories of personal data. It is therefore confidential. It must be processed in such a way that it complies with the GDPR. This data must be kept and protected in a way that it does not leak or get stolen. Only authorised persons should have access to this data. Since these manipulated data are personal data of highly sensitive nature, unauthorised persons could identify patients and make their illnesses public, which may have serious negative impact on the patients.
Traceability	High	At a minimum, the actions related to the process must be recorded (traced and dated) to be able to follow the actors involved during the stages of use of personal data. Particularly in the case of court cases, it is necessary to be able to attribute necessary actions to an individual and to be able to impute them criminally.

3. SECURITY AND PRIVACY THREATS AND VULNERABILITIES

3.1 THREAT CONTEXTUALISATION

This section highlights the threats applicable in medical imaging scenario. Examples in the context of this scenario are also provided. Also outlined are the impacts that each threat can have in terms of security and privacy as well as their severity in the table.

There are two major feared events to consider for this scenario which are the loss of integrity of the data, and the loss of confidentiality of the personal data of the patients.

The loss of integrity of the data could lead to **reputation degradation** and **lawsuit** to the **company and physical and permanent injury for the patient**. Indeed, the impact would be serious in both case:

- On one hand, a patient suffering from osteoporosis could be wrongly diagnosed and treated incorrectly,
- On the other hand, a patient not suffering from osteoporosis could be wrongly diagnosed and treated for the wrong illness.

It should be noted, however, that the impact is lessened because the model only assists and does not make decisions directly. When making a diagnosis, decision-makers (the radiologist in this case) take into account not only the ML output but also their own medical knowledge and experience.

For the medical practice, the loss of confidentiality of personal data could lead to **reputation degradation and a lawsuit**, that can result in patients losing trust. For the patients, the loss of confidentiality of personal data could have several impacts. First, the data stolen (X-rays of patients with their name and age for instance) could be used to perform **Phishing attempts, targeted advertising**. Then, **Unique targeted opportunities** might not be given like real estate loan, studies, or jobs due to the medical condition. For example, companies could be informed of a patient's state of health and therefore refuse certain requests (e.g., insurance). This could lead to a patient **feeling a significant invasion of privacy or discrimination against them**. A lack of regard for privacy requirements by the medical practice (such as in the case of unfair processing) could lead to discrimination in treatment, resulting in better diagnosis of osteoporosis for men than for women, for example. The impact on the user could lead to **infringement of fundamental rights**.

Considering these feared events and associated impacts, the following threats are associated with this scenario.

Figure 14: Summary of threats and vulnerabilities

COMPROMISE OF DIAGNOSTIC SYSTEM COMPONENTS

LOSS REP INV PHISH

- Weak access control
- Use of vulnerable components
- Poor access rights management process

DATA DISCLOSURE

LOSS REP INV PHISH

- Disclosure of sensitive data for ML algorithm training
- Lack of control of Data processor (including external stakeholder)
- Poor data management

LACK OF TRANSPARENCY

INV

- Lack of controls to ensure the adequacy of the purpose and its current use
- Lack of detail on the purposes and justification for their legitimacy
- Lack of privacy by design

NO RESPECT OF STORAGE LIMITATION

PHYS

- Lack of accuracy criteria
- Poor data management
- Lack of privacy by design

EVASION

REP PHYS

- Lack of detection of abnormal inputs
- Lack of training based on adversarial attacks
- Use of a widely known model allowing the attacker to study it

POISONING

REP PHYS

- Lack of control for poisoning
- No detection of poisoned samples in the training dataset
- Use of uncontrolled data

DIVERSION OF PURPOSE

LOSS INV PHISH

- Existing biases in the ML model or in the data
- Lack of controls to ensure that data is used only for the purposes defined
- Lack of privacy by design

NO RESPECT OF STORAGE LIMITATION

LOSS PHISH

- Lack of data deletion mechanisms
- Lack of data retention policy
- Lack of privacy by design

HUMAN ERROR

LOSS REP INV PHISH

- Lack of security by design
- Weak access control
- Poor data management

UNLAWFUL AND UNFAIR PROCESSING

INV FEEL

- Absence of an identified data controller
- Lack of practical means and justification for the legal basis
- Lack of privacy by design

NO RESPECT OF DATA MINIMIZATION

INV

- Lack of measures to prevent further Lack of controls to ensure that the data collected are minimal for the purposes intended
- Lack of necessary data selection
- Lack of pseudonymization

NO RESPECT OF COMPLIANCE OF THE TRAINING MODEL

LOSS REP INV

- Lack of review of treatment by a dedicated committee to check fairness
- Lack of privacy by design

IMPACTS

- LOSS: loss of unique targeted opportunities
- REP: Reputation degradation
- PHISH: Phishing attempts, targeted advertising
- PHYS: Physical and permanent injury
- FEEL: feeling of infringement of fundamental rights
- INV: Significant sense of invasion of privacy

3.1.1 Compromise of diagnostic system components

An attacker could attack the exposed APIs and/or human interfaces during the entire scenario lifecycle. They could, for example, gain access to an account to infiltrate the data lake or the model server. This could be done via brute force password cracking in the case of simple authentication, via a vulnerability exploitation, poor security controls implemented by the cloud provider, credential stuffing, or other password-based attacks. Once access is gained, actions made by the attacker could lead to loss of data integrity (including model parameters) or to loss of data confidentiality (by data leaks). For the medical practice, it could lead to **reputation degradation** and **lawsuit**. For the patient, it could lead to a **significant feeling of invasion of privacy, phishing attempts, or targeted advertising** or the loss of **unique targeted opportunities**.



3.1.2 Evasion

Direct evasion attacks are possible, but only from an attacker compromising one of the components. An attacker could modify the image capturing devices (i.e., the scanner or the X-ray software) to add additional noise to an image, which could trick the model into making incorrect diagnosis. The noise added to the image can be imperceptible for humans and could only influence the output score of the ML program. This could increase the potential diagnostic error of the radiologist if he only relies on the program output. In the long run, the impact could be a misdiagnosis and lead to **reputation degradation, lawsuit, and physical and permanent injury** for the patients.

3.1.3 Human error

A network administrator could incorrectly expose (such as by making public) certain instances of the databases, which would raise the risk profile of the 3 databases. This could leave personal data susceptible to attack, should the database storing personal data and patient files be targeted by attackers. This, in turn, could lead to **reputation degradation and lawsuit** for the medical practice, **significant feeling of invasion of privacy, phishing attempts, or targeted advertising** or the **loss of unique targeted opportunities** for patients.

There is a risk that Radiologists may not be careful when handling patient data by bypassing the tool they have been provided with and filling in other media (text files), the creation of which would have very low visibility (shadow IT). This could open a wide, untraceable area of attack and malicious persons could try and get access to the data stored in unprotected/unrestricted applications. This could make personal data leakage easier, leading to **reputation degradation** for the medical practice, **significant feeling of invasion of privacy, phishing attempts, or targeted advertising** or even the loss of **unique targeted opportunities** for patients.

3.1.4 Data disclosure

Not all data have the same level of sensitivity. On one hand, there is data that is used by the model (when it is trained, and once in production) which is pseudonymised. It contains only the X-Ray scans, BMI, gender, and age. On the other hand, the data used for the traceability of actions, i.e., the patient's file, is very sensitive because it contains all personal information. This sensitive data is manipulated by the computer-aided system and stored in a database in the data lake.

Several scenarios could be conceived:

- An attacker could take advantage of the poor exposure of the patient's file database to extract data from it (by usurping an account for example, or by exploiting a security flaw).
- An attacker manages, by various means such as phishing, to install spyware on the computer aided system and thus harvest patient data.
- The cloud provider does not offer sufficient security guarantees and so an attacker gains access to the data centre and copies the data from the databases because they are not protected (by encryption mechanisms for example)

Those 3 examples would cause harm and lead to **reputation degradation and lawsuit** for the medical practice, **serious feeling of invasion of privacy, phishing attempts, targeted advertising** or even the **loss of unique targeted opportunities** for the patients.

3.1.5 Poisoning (by label modification)

During the data collection phase, a malicious radiologist could annotate the data of the scanner and force the model to make certain targeted predictions, or to underperform. His motivations could be to discriminate against a certain ethnic, racial, or sexual group, fear of being replaced by AI or financial motivation because he was approached by a competing medical practice.

Such actions cause loss of data integrity and lead to **reputation degradation** and **lawsuit** for the medical practice and **physical and permanent injury** for the patient.

3.1.6 Unlawful Processing

The data processing associated with AI accelerates and improves the medical practice's ability to detect osteoporosis and accommodate more patients. The legal basis for the data processing is therefore the legitimate interest of the medical practice. However, the practice may not be able to justify that the processing of the data (or even the use of artificial intelligence) is necessary to safeguard these legitimate interests. In this scenario, patients may question the purpose of the use of their personal data and feel a **significant feeling of invasion of privacy** and a lack of trust in the medical practice.

3.1.7 Unfair processing

Even if the medical practice demonstrates that the data processing has a legal basis, it could process the personal data in a way that is unexpected by patients, i.e., instead of detecting osteoporosis equitably for all patients, it could use the data in a way that has an unjustified negative effect on some patients **with discriminations created by the treatment such as better diagnosis of osteoporosis for men than for women, for example**. Another possible unfair usage of the processing is the case where the radiologist would trust the algorithm too much and spend too little time to examine the radio by himself. This would effectively break the deal made with the patient that the diagnosis is performed by an expert, and assisted by AI, where in reality the diagnosis would mostly be decided by the AI. The impact on the user could lead to a **feeling of infringement of fundamental rights**.

3.1.8 Lack of transparency

The practice may also not be transparent with patients about how their personal data is used and the purpose for which their data is used. The terms governing these aspects may not be included in the information provided to the persons concerned. In addition to general information, individual information may not be provided to patients by radiographers, for example before the start of consultations. The type of information that might not be given to patients is that their data is stored in a database and could be used in case of a lawsuit between a patient and the medical practice. The impact this could have on patients could result in **significant feeling of invasion of privacy**.

3.1.9 Diversion of purpose

The patient data extracted in this scenario is intended only to help diagnostic osteoporosis. However, the medical practice might not respect this principle and use the patient data for ambiguous and non-explicit purposes, even unrelated to the diagnostic of a disease. Indeed, it may be tempting to use these data to offer other services or to resell them to insurers, or for instance, to offer **targeted advertisements** for medicines. The other impact this could have on patients could result in a **significant feeling of invasion of privacy** and a loss of **unique targeted opportunities**.

3.1.10 No respect of data minimisation

The medical practice could collect data from the patients' body other than those solely related to the treatment of osteoporosis. This data could be collected to build other databases for the treatment of other diseases not related to osteoporosis. The impact this could have on patients could lead to a **significant feeling of invasion of privacy**.

3.1.11 No respect of accuracy

Collection of inaccurate data, both in training and in production, can lead to unreliable model behaviour. Inaccurate data could be the result of human error during the data collection phase or by malicious persons who corrupt the data stored in the databases leading a loss of integrity of data. This could lead to **temporary or permanent physical injury of the patient**.

3.1.12 No respect of storage limitation

This threat takes the form of storing personal data in the medical practice database without relevant data suppression rules. As a reminder, there are three databases our case; one for training data to build the model, another to re-train the model, and a last one to store the patients' files in case of lawsuit. The Medical Practice may not comply with data retention periods governed by the requirements of local country laws and the recommendations of regulatory authorities. In this case, private data of the patients involved could be exposed. Such a data breach could lead to **targeted advertising** or the loss of **unique targeted opportunities**.

3.1.13 No respect of compliance of the training model

Medical practices and Data Scientists could develop models not adapted to the problem. The company could develop models that are biased (depending on gender for example). This can be due to insufficiently representative data, malicious acts, or a lack of a testing procedure. Nevertheless, this would lead to **reputation degradation** and **lawsuit**. For the patient, it could lead to a **significant feeling of invasion of privacy, phishing attempts, or targeted advertising** or the loss of **unique targeted opportunities**.

3.1.14 Synthesis of possible impacts and associated threats

The following table sums up the severity of each impact and the associated threats.

Figure 1: Synthesis of possible impacts and associated threats²⁵

Impact	Severity	Type	Associated Threats
Physical and permanent injury and harm	High	Cybersecurity and Privacy	Compromise of diagnostic system components Evasion Poisoning No respect of accuracy No respect of compliance of the training model
Lawsuit	High	Cybersecurity	Compromise of diagnostic system components Evasion Human error Data disclosure Poisoning No respect of compliance of the training model
Reputation degradation	High	Cybersecurity	Compromise of diagnostic system components Evasion Human error Data disclosure Poisoning No respect of compliance of the training model
Phishing attempts, targeted advertising	High	Privacy	Diversion of purpose No respect of storage limitation No respect of compliance of the training model
Loss of unique targeted opportunities	High	Privacy	Diversion of purpose No respect of storage limitation

²⁵ See the severity scales in the Annex

			No respect of compliance of the training model
Significant feeling of invasion of privacy	Moderate	Privacy	Unlawful processing Diversion of purpose No respect of data minimisation No respect of compliance of the training model
Feeling of infringement of fundamental rights	Moderate	Privacy	Unfair processing No respect of compliance of the training model

3.2 VULNERABILITIES ASSOCIATED TO THREATS AND AFFECTED ASSETS

The table below cross evaluates each threat to a set of associated vulnerabilities. The actors involved and the assets possibly affected by the vulnerabilities are also highlighted:

Figure 2: Mapping vulnerabilities to threats and assets/actors on which they rely

Vulnerabilities	Threats	Actors	Assets Involved
Absence of an identified data controller	Unlawful processing Unfair processing Lack of transparency Diversion of purpose No respect of data minimisation No respect of storage limitation	Medical practice	Data
Contract with a low security third party	Compromise of diagnostic system components Data disclosure	Medical practice	N/A
Disclosure of sensitive data for ML algorithm training	Data disclosure	Data scientists	Model
Existing biases in the ML model or in the data	Diversion of purpose	Large tech companies Data scientists	Model Data
Lack of auditability of processing	Unlawful processing Unfair processing Lack of transparency Diversion of purpose No respect of data minimisation No respect of storage limitation	Medical practice Data scientists Developers and data engineers System and communication network administrators	N/A
Lack of accuracy criteria	No respect of accuracy	Data scientists Developers and Data Engineers	Data Model
Lack of documentation	Human error	Medical Practice	All assets

	<ul style="list-style-type: none"> Unlawful processing Unfair processing Lack of transparency Diversion of purpose No respect of data minimisation No respect of storage limitation 	<ul style="list-style-type: none"> Data scientists Developers and Data Engineers System and communication network administrators 	
Lack of pseudonymisation	<ul style="list-style-type: none"> Diversion of purpose No respect of data minimisation 	<ul style="list-style-type: none"> Medical practice Data Engineers 	Data
Lack of consideration of attacks to which diagnostic systems could be exposed	<ul style="list-style-type: none"> Denial of service due to inconsistent data Poisoning Data disclosure Unlawful processing Unfair processing Lack of transparency Diversion of purpose No respect of data minimisation Label modification 	<ul style="list-style-type: none"> Medical practice Large tech companies Data Scientists Developers and Data Engineers 	<ul style="list-style-type: none"> Model Data lake Model server Scanner X-ray computer-aided diagnostic system Integrated Development Environment Libraries Communication protocols and communication networks
Lack of consideration of real-life conditions in training the model	<ul style="list-style-type: none"> Diversion of purpose 	<ul style="list-style-type: none"> Data scientists Developers and Data Engineers System and communication network administrators 	Model
Lack of control for poisoning	<ul style="list-style-type: none"> Poisoning 	<ul style="list-style-type: none"> Data scientists 	Model
Lack of control of Data processor (including external stakeholder)	<ul style="list-style-type: none"> Diversion of purpose Data disclosure 	<ul style="list-style-type: none"> Medical practice Cloud provider 	Data
Lack of control over model performance	<ul style="list-style-type: none"> Unfair processing No respect of accuracy 	<ul style="list-style-type: none"> Medical practice Cloud provider 	Data
Lack of controls to ensure that data is used only for the purposes defined	<ul style="list-style-type: none"> Diversion of purpose 	<ul style="list-style-type: none"> Medical practice 	<ul style="list-style-type: none"> Scanner X-ray computer-aided diagnostic system Data lake Model
Lack of controls to ensure that the data collected are minimal for the purposes intended	<ul style="list-style-type: none"> No respect of data minimisation 	<ul style="list-style-type: none"> Data scientists Developers and Data Engineers 	<ul style="list-style-type: none"> Scanner X-ray computer-aided diagnostic system Data lake Model
Lack of controls to ensure the adequacy of the purpose and its current use	<ul style="list-style-type: none"> Unlawful processing Unfair processing Lack of transparency 	<ul style="list-style-type: none"> Medical practice 	N/A

Lack of data deletion mechanisms	No respect of storage limitation	N/A	Data lake
Lack of data for increasing robustness to poisoning	Poisoning	N/A	Data
Lack of data retention policy	No respect of storage limitation	Medical practice	Data
Lack of detail on the purposes and justification for their legitimacy	Unlawful processing Unfair processing Lack of transparency	Medical practice	N/A
Lack of detection of abnormal inputs	Evasion	Data Scientists Developers and Data Engineers	Data Data lake Model
Lack of justification and traceability of decisions taken	Diversion of purpose	Medical practice	Model
Lack of justification for the collection of individual personal data collected	No respect of data minimisation	Medical practice	Data
Lack of measures to prevent further data collection	No respect of data minimisation	Medical practice	Data
Lack of necessary data selection	No respect of data minimisation	Medical practice Data scientists Data Engineers	Data
Lack of practical means and justification for the legal basis (legitimate interest)	Unlawful processing Unfair processing Lack of transparency	Medical practice Historical Patients New Patients	All the assets
Lack of security by design	Compromise diagnostic system components Evasion Human error Data disclosure Poisoning	Medical practice	All assets
Lack of privacy by design	Unlawful processing Unfair processing Lack of transparency Diversion of purpose No respect of data minimisation No respect of accuracy No respect of storage limitation	Medical practice	All assets
Lack of review of treatment by a dedicated committee to check fairness	Unlawful processing Unfair processing Lack of transparency	Medical practice	N/A

Lack of security process to maintain a good security level of the components of the diagnostic system	Compromise of diagnostic system components Data disclosure	Medical practice	Model Data lake Model server Scanner X-ray computer-aided diagnostic system Integrated Development Environment Libraries Communication protocols and communication networks
Lack of traceability of actions and/or modifications made to the assets on which rely personal data	No respect of accuracy No respect of storage limitation	N/A	Model Data lake Model server Scanner X-ray computer-aided diagnostic system Integrated Development Environment Libraries Communication protocols and communication networks
Lack of training based on adversarial attacks	Evasion	Data scientists	Data
Lack of transparency on the purpose, the exact data that are collected, and how they are processed.	Diversion of purpose	Medical practice	N/A
Lack of verification that the data is adequate, relevant, and not excessive for the purpose of making a diagnostic	No respect of data minimisation	Medical practice	N/A
Model easy to poison	Poisoning	Medical practice Data scientists Developers and data engineers	Model
No detection of poisoned samples in the training dataset	Poisoning	Data scientists Data Engineers	Data
Poor consideration of evasion attacks in the model design implementation	Evasion	Data scientists	Model
Poor access rights management process	Compromise of diagnostic system components	Medical Practice Cloud provider	Data lake Model server Scanner X-ray computer-aided diagnostic system Integrated Development Environment

Poor data management	<p>Poisoning</p> <p>Data disclosure</p> <p>Diversion of purpose</p> <p>No respect of data minimisation</p> <p>No respect of storage limitation</p>	<p>Data scientist</p> <p>Developers and data engineers</p>	Data
Too much information available on the model	<p>Compromise of diagnostic system components</p>	<p>Medical practice</p> <p>Large tech companies</p> <p>Data scientists</p>	Model
Unprotected sensitive data on test environments	<p>Data disclosure</p>	<p>Medical practice</p> <p>Data scientists</p> <p>Developers and Data Engineers</p> <p>System and communication network administrators</p>	Data
Use of uncontrolled data	<p>Poisoning</p>	<p>Medical practice</p> <p>Data scientists</p>	Data
Use of unreliable sources to label data	<p>Label modification</p>	<p>Medical practice</p>	Data
Use of unsafe data or models (e.g., with transfer learning)	<p>Poisoning</p>	<p>Medical practice</p> <p>Large tech companies</p> <p>Historical Patients</p>	Model
Use of vulnerable components (Among the whole supply chain)	<p>Compromise of diagnostic system components</p>	N/A	<p>Model</p> <p>Data lake</p> <p>Model server</p> <p>Scanner</p> <p>X-ray computer-aided diagnostic system</p> <p>Integrated Development Environment</p> <p>Libraries</p> <p>Communication protocols and communication networks</p>
Using of a widely known model allowing the attacker to study it	<p>Evasion</p>	<p>Large tech companies</p> <p>Data Scientists</p> <p>Developers and Data Engineers</p>	Model
Weak access protection mechanisms for machine learning model components and for personal data (encryption, access control mechanism...)	<p>Compromise of diagnostic system components</p> <p>Data disclosure</p> <p>Poisoning</p>	<p>Medical practice</p>	<p>Model</p> <p>Data lake</p> <p>Model server</p> <p>Integrated Development Environment</p> <p>Libraries</p> <p>Communication protocols and communication networks</p>

4. CYBERSECURITY AND PRIVACY CONTROLS

Before expanding on the details of the cybersecurity and privacy controls applied to the scenario, the following figure summarizes all the controls that will be described.

Figure 17: Summary of cybersecurity and privacy controls

SPECIFIC CONTROLS

Pseudonymize data coming from the Historical patient

- Replace names of patients by an ID
- No impact on performance

Add some adversarial examples to the dataset

- include adversarial examples to the algorithm's training
- No impact on performance

Choose and define a more resilient model design

- Perform defensive distillation to avoid evasion attacks
- No impact on performance

Integrate poisoning control

- Employ the STRIP technique
- No impact on performance

Enlarge the training dataset

- Train the model with medical data collected during several years
- Privacy impacts (more personal data collected)

Secure the transit of the collected data

- End-to-end encryption using TLS 1.3 to avoid loss of integrity and confidentiality
- No impact on performance

Ensure all systems and devices comply with authentication, and access control policies

- Active Directory, MFA, Use of OAuth 2.0
- Privacy impacts

Identify all the data processors and perform the control actions necessary to give reasonable assurance that they are compliant


- Contractual clauses, internal and external audits
- Impacts resulting in loss of time and energy

Formalize a LIA (Legitimate Interest Assessment)

- Justify the legal basis
- Impacts resulting in loss of time and energy

Ensure that the model is sufficiently resilient to the environment in which it will operate

- Use real data to train the model, test the model in real life conditions, ...
- Privacy impacts (more personal data collected)



Use reliable sources to label data

- Reliable radiologist to label the data
- Positive impact on accuracy

Check the vulnerabilities of the components

- Regular security audits, vulnerabilities scans, automatic patch management
- Impact on the availability of the system

Monitor the performance of the model


- Ensure the reliability of the model, be sure of the intelligence of the model by selecting quality data, always train and evaluate the model
- Improve the performance of the model

Minimize data at each step of the processing

- Study the necessity of collecting data such as age and body weight, proof the necessity of collecting such data
- Impact on security in case of forensic analysis

GENERIC CONTROLS

- Implement a security by design process
- Implement a privacy by design process
- Generate logs and perform internal audit
- Document the diagnostic system
- Control all data used by the ML Model
- Reduce all the available information about the model
- Define and implement a data retention policy
- Identify a data controller



- Define accuracy criteria
- Raise awareness of security and privacy issues among all stakeholders
- Study on data fields necessity and justification in the privacy anticipation
- Perform a privacy Impact assessment
- Call on ethical committee and external audits
- Implement access right management process
- Ensure that models are unbiased

4.1 IMPLEMENT A SECURITY BY DESIGN PROCESS

Type	Associated Vulnerabilities	Threats it mitigates
Cybersecurity	Lack of Security by Design	<ul style="list-style-type: none"> • Compromise of diagnostic system components • Evasion • Human error • Data disclosure • Poisoning

By default, Security by Design is a methodology to strengthen the cybersecurity of the organisation by automating its data security controls and developing a robust IT infrastructure. This approach focuses on implementing the security protocols from the basic building blocks of the entire IT infrastructure design. The goal of such a methodology is to ensure the risks are mitigated before systems go live, and appropriate security controls are implemented.

The lack of Security by Design increases the likelihood of all threats related to the scenario. Therefore, the medical practice must ensure, from the development phase, to put in place adequate controls to limit the cybersecurity risk. This starts with a global risk analysis in which all risks associated with all assets are identified. The medical practice will try to reduce the attack surface (for example by auditing its cloud provider on a regular basis), to apply the principle of least privilege (by implementing access rights management), to take care of the confidentiality and integrity of the collected data (by encrypting the patients' data for example).

Conceiving a project following a Security by Design methodology requires less effort than adding security on top of an existing project. However, there could be an impact on the functionality of the scenario if the outcome of the Security by Design methodology would lead to some functionality not being implemented due to the security risk it generates (the installation of unsecure scanners for example).

4.2 DOCUMENT THE DIAGNOSTIC SYSTEM

Type	Associated Vulnerabilities	Threats it mitigates
Cybersecurity & Privacy	Lack of documentation on the diagnostic system	<ul style="list-style-type: none"> • Human error • Unlawful processing • Unfair processing • Lack of transparency • Diversion of purpose • No respect of data minimization • No respect of storage limitation

Project and system documentation must be produced to preserve knowledge on the choices made during the project phase, the application architecture, its configuration, its maintenance, how to maintain its effectiveness over time and the assumptions made about the model use.

The system is complex with multiple assets (Scanner, X-ray computer-aided diagnostic system, etc...) which enhance the fact that an exhaustive documentation is mandatory, for each of those assets. This documentation should also include the changes that will be applied, including to the documentation throughout the algorithm's life cycle. Therefore, it is necessary that the Data Scientists, Developers and Data Engineers, and the System and Communication Network Administrators work together to create and sustain the documentation. **This control does not impact system performance, cybersecurity, or privacy.**

4.3 CHECK THE VULNERABILITIES OF THE COMPONENTS USED AND IMPLEMENT PROCESSES TO MAINTAIN SECURITY LEVELS OF ML COMPONENTS OVER TIME

Type	Associated Vulnerabilities	Threats it mitigates
Cybersecurity & Privacy	<ul style="list-style-type: none"> Lack of security process to maintain a good security level of the components of the diagnostic system Use of vulnerable components Disclosure of sensitive data for ML algorithm training 	<ul style="list-style-type: none"> Compromise of diagnostic system components Data disclosure

For the data lake, the model server, and the X-ray computer-aided diagnostic system, it is necessary for the company to ensure that regular security audits are carried out to check that there are no vulnerabilities. Regular vulnerability scans and an automatic patch management process should also be implemented to maintain a good security level. The medical practice should also have a remediation plan that can be implemented quickly. Moreover, this plan needs to be reviewed over time.

For the data lake and the model server (hosted in a cloud), the company will also have to complete a security questionnaire for the cloud service provider. This questionnaire must cover all aspects of security and each answer must be justified. These answers will be attached to the contract and will have to be reviewed annually by both parties and whenever the contract is modified.

To perform audits of scanners, which are IOT devices, specific expertise is required. These devices are connected to the network as regular devices, but due to their scanning functionalities, they are composed of specific hardware and firmware, which can be different from classical IT systems. For the hardware part, the idea is to identify and assess the components that enhance the risk of an attacker gaining administration access to the scanner. To do so, retro engineering techniques such as direct access to electrical components can be used.

This control would impact the availability of the system, thus its performance, as it may be audited or even updated. This could lead to unavailability of the system for several hours on a regular basis. However, as the availability need of the system is low, the impact of this unavailability is only moderate. Of greater concern is the risk that updates could cause the system to malfunction, leading to errors in the diagnostic systems which would have a major impact.

4.4 ADD SOME ADVERSARIAL EXAMPLES TO THE DATASET

Type	Associated Vulnerabilities	Threats it mitigates
Cybersecurity	Lack of training based on adversarial attacks	Evasion

The medical practice must ensure that the model is resilient to evasion attacks to prevent misdiagnosis. To do so, inclusion of adversarial examples to the algorithm's helps enable it to be more resilient to such attacks. Depending on the application domain and ambient conditions, such training could be done continuously. Some examples would be: Adversarial Training, Ensemble Adversarial Training, Cascade Adversarial Training or Principled Adversarial Training. **This control does not impact performance or privacy.**

4.5 CHOOSE AND DEFINE A MORE RESILIENT MODEL DESIGN

Type	Associated Vulnerabilities	Threats it mitigates
Cybersecurity	Poor consideration of evasion attacks in the model design implementation	Evasion

To avoid evasion attack, the chosen model should be robust against such attacks. For instance, defensive distillation is an adversarial training technique that adds flexibility to an algorithm’s classification process, so the model is less susceptible to exploitation. In distillation training, one model is trained to predict the output probabilities of another model that was trained on an earlier, baseline standard, to emphasise accuracy. **This control does not impact performance or privacy.**

4.6 INTEGRATE POISONING CONTROL IN THE TRAINING DATASET

Type	Associated Vulnerabilities	Threats it mitigates
Cybersecurity	<ul style="list-style-type: none"> No detection of poisoned samples in the training dataset Lack of control for poisoning 	Poisoning

The DNN algorithm used in this scenario can be checked for poisoning using the STRIP technique. The principle of this technique is to disturb the inputs and observe the randomness of the predictions. For example, the company could intentionally perturb the incoming input, for instance by superimposing various image patterns, and observe the randomness of predicted classes for perturbed inputs from a given deployed model. A low entropy in predicted classes violates the input-dependence property of a benign model and implies the presence of a malicious input. These tests should be performed before the production phase, during the training phase, ensuring that the models used in production are healthy. **This control does not impact system performance, cybersecurity, or privacy.**

4.7 ENLARGE THE TRAINING DATASET

Type	Associated Vulnerabilities	Threats it mitigates
Cybersecurity	Lack of data for increasing robustness to poisoning	Poisoning

A large time frame (multiple years) of medical data can be collected and used to train the algorithm to prevent it from being vulnerable to poisoning attacks, as modifying only small amounts of medical data would have a low overall impact of the prediction. Using a large time frame (multiple years) of scanner image data could also reduce the probability that poisoning attacks on this data have an impact on the prediction.

In this case, privacy is negatively impacted because enlarging the data set means taking even more personal data which could be stolen by an attacker. Therefore, it would enlarge the attack surface, and it is a trade-off between retaining personal data and deleting them to comply with the GDPR.

4.8 SECURE THE TRANSIT OF THE COLLECTED DATA

Type	Associated Vulnerabilities	Threats it mitigates
------	----------------------------	----------------------

Cybersecurity	Poor data management	Poisoning
---------------	----------------------	-----------

Considering the possible impacts of Poisoning, the transit of the collected data should be fully protected against a loss of integrity. The medical practice should ensure that the transit of such data is secure and protected against loss of integrity. To do so the company must ensure that the protocols used for the transfer are encrypted, using a secure protocol such as TLS 1.3 for example. As such, an attacker won't be able to modify the data during its transit phase and won't be able to poison the model through that vector. Such encryption must be end-to-end, i.e., from the scanner to the model and the data lake. **This control won't have any impact on privacy or performance of the system.**

4.9 CONTROL ALL DATA USED BY THE ML MODEL

Type	Associated Vulnerabilities	Threats it mitigates
Cybersecurity	<ul style="list-style-type: none"> • Use of uncontrolled data • Lack of detection of abnormal input • Poor data management 	Poisoning

Considering the high impacts of Poisoning attacks, several means to control the data must be applied, some specific to data science and others more to business and common sense.

During the data collection phase and especially at the data lake level, the data must be controlled to ensure its relevance. A second check should be a more traditional data science check to eliminate data that deviates too much from the "normal". This would greatly complicate a poisoning attack on the model because attempts using data that deviates from the "normal" would be eliminated by this check. These checks can be done several times including just before training the model. This is to protect against possible data modification in the data lake.

This control could add latency to the system in case of many reported anomalies on the data. But this scenario does not have a strong need for availability, and the control is important to reduce the probability of model poisoning. In our case, the checks that are applied could have an impact on privacy, because they would manipulate personal data from patients. To remediate, the systems used (software, scripts, excel sheets, etc.) to perform these checks should be regularly audited, documented, secured by design, and follow access control rules (authentication to use the system, proper access management, etc.).

4.10 IMPLEMENT ACCESS RIGHT MANAGEMENT PROCESS

Type	Associated Vulnerabilities	Threats it mitigates
Cybersecurity & Privacy	Poor access rights management	<ul style="list-style-type: none"> • Compromise of ML application components • Poisoning • Data disclosure • Human error

Threats related to poor access rights management, such as data disclosure, can induce high impacts for the company (lawsuits, Reputation degradation) or for the patients (temporary or permanent physical injury and harm). This security measure defines and assigns roles to users to respect the principle of least privilege and thus limits the access scenarios for processes and assets only to those for which the users have a justified reason.

For the scanners, the X-ray computer-aided diagnostic system, data lake, model server, integrated development environment, the company should create roles for each user group,

define rules, and instigate an access management process. This enables provision of proper accesses only to actors as required, and prevent unauthorised persons gaining access to these systems. In particular, the different user profiles (for instance Radiologist or Data Scientist) need to clearly define a profile that will be able to access specifically the patient's file database, as this is the most sensitive data set in this scenario.

However, in the case of an incident or failure, the complexity of the management of these accesses could prevent teams from gaining access to the system to correct a problem which could impact availability of the system, thus its performance.

4.11 ENSURE ALL SYSTEMS AND DEVICES COMPLY WITH AUTHENTICATION, AND ACCESS CONTROL POLICIES

Type	Associated Vulnerabilities	Threats it mitigates
Cybersecurity & Privacy	<ul style="list-style-type: none"> Weak access protection mechanisms for machine learning model components Weak access protection mechanisms of the personal data (encryption, access control mechanism...) 	<ul style="list-style-type: none"> Compromise of diagnostic system components Poisoning Data disclosure Human error

For user access in this scenario, the medical practice must ensure that proper user authentication and access control for the Scanner, X-ray computer-aided diagnostic system, data lake, Model server, and Integrated Development Environment is present. This should be managed centrally, with an authentication solution (such as Active Directory for example) linked to all assets. For sensitive assets that manipulate personal data, the medical practice should enforce multi-factor authentication for the Data Scientists and Administrators that should have access to it. The multi-factor authentication solution needs to be linked to the central authentication system.

For devices access, the medical practice should ensure that all device-to-device requests (particularly between the scanner and the X-ray computer-aided diagnostic system) are properly authenticated. This can be done using the OAuth 2.0 protocol for all exposed APIs with control via an API gateway. In addition, the certificates for the authorisation server keys should be managed through the company's own Public Key Infrastructure to avoid any identity theft scenario. It is also necessary to respect an authentication policy, based on the current recommendations of bodies such as ENISA²⁶, NIST²⁷. To protect the sensitive data stored in the scanner, it is necessary to secure the scanners by ensuring that the OS version is up-to-date, and that non-essential ports are closed.

Also, personally authenticating the employees on the system can have an impact on privacy, because personal data (identifier, mail address, etc) would be logged in these authentication systems. These authentication systems must respect the principles of Article 5 of the GDPR (this includes data minimisation etc).

4.12 MONITOR THE PERFORMANCE OF THE MODEL

Type	Associated Vulnerabilities	Threats it mitigates
Cybersecurity	Lack of control over model performance	<ul style="list-style-type: none"> Unfair processing No respect of accuracy

To ensure the reliability of the model, the medical practice must be sure of the intelligence of the model it has developed. To do so, data scientists and data engineers must ensure the quality of

²⁶ <https://www.enisa.europa.eu/news/enisa-news/tips-for-secure-user-authentication>

²⁷ See <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-63-3.pdf>



the model training by selecting quality data. In addition, the model must be trained continuously and routinely evaluated. **This will have an impact on the performance of the model in the sense that by applying this control, the model is always efficient and reliable.**

4.13 REDUCE THE AVAILABLE INFORMATION ABOUT THE MODEL

Type	Associated Vulnerabilities	Threats it mitigates
Cybersecurity & Privacy	<ul style="list-style-type: none"> • Too much information available on the model • Using a widely known model allowing the attacker to study it 	Compromise of diagnostic system components

This control lowers the likelihood of compromise of the ML application components by limiting the knowledge of a malicious user about his target and consequently making it more difficult to launch a cyberattack. It is important to understand that this security control alone is not enough, as security through obscurity does not provide sufficient protection.

Therefore, all documentation concerning the model itself should be protected. It should be encrypted and protected by a Data Right Management mechanism, allowing only the Data Scientists to have access through their computers. Related documentation can be stored in a dedicated file server directory whose rights are regularly reviewed and linked to the access rights management mechanisms defined in 3.2.10 and 3.2.11. This server must be encrypted by default, and located in a network zone dedicated to the company's sensitive resources. In addition, user workstations accessing this file server must be sufficiently protected through encryption of the workstations, a confidentiality filter to be placed on the screen for workstations in a nomadic situation, OS version up-to-date, closure of non-essential ports, amongst other protection mechanisms. To avoid the propagation of documents relating to the model, it is recommended to deploy a data loss protection solution.

This control could have an impact on privacy and it may be thought that this contradicts the principle of transparency required by the GDPR. Therefore, it is important to define documentation that can be communicated to external users (i.e., customers), explaining that their data is not directly used by the model. Moreover, this could impact the performance of the system, because in case of an incident or a failure, the complexity of the access management could prevent Data Scientists gaining access to the documentation to correct the problem and have an impact of availability of the system, thus its performance.

4.14 IDENTIFY A DATA CONTROLLER FOR THE MEDICAL DATA PROCESSING

Type	Associated Vulnerabilities	Threats it mitigates
Privacy	Absence of an identified Data Controller	<ul style="list-style-type: none"> • Unlawful processing • Unfair processing • Lack of transparency

Identifying a Data Controller is essential to prevent impact on the company and its employees through unlawful processing, unfair processing, or lack of transparency. Self-identifying the data controller means performing accountability actions (documentation, assessments, etc.) and making sure that the energy consumption anticipation data processing is compliant with personal data protection GDPR obligations and does not infringe on privacy rights of data subjects. Self-identifying as Data Controller means also taking responsibility in case of non-compliance or adverse privacy effects on data subjects. Such a measure does not have a negative impact on this scenario. On the contrary, **the control improves the privacy of users without impacting the performance of the model.**

4.15 PSEUDONYMISE DATA COMING FROM THE HISTORICAL PATIENT

Type	Associated Vulnerabilities	Threats it mitigates
Privacy	<ul style="list-style-type: none"> Lack of pseudonymisation Unprotected sensitive data on test environments 	No respect of data minimisation

The reader is reminded that data that could identify patients such as name, surname, age, gender, and body mass index of historical patients are pseudonymised. In this case pseudonymisation requires removing the attributes (“name” and “surname”) from the data, replacing it with random IDs, and keeping the correspondence between these removed attributes and their IDs in a dedicated data base. For instance, instead of storing the patient data with the attributes [first name, surname, gender, age, BMI, X-ray image], the medical practice would store the patient data with the attributes [ID, gender, age, BMI, X-ray image]. A different database would be used to store the correspondence composed of [ID, surname, first name]. Usually, market vendor solutions allow ease and security in this process. **This control does not impact on security or performance.**

4.16 GENERATE LOGS AND PERFORM INTERNAL AUDIT

Type	Associated Vulnerabilities	Threats it mitigates
Privacy	<ul style="list-style-type: none"> Lack of auditability of processing Lack of traceability of actions and/or modifications made to the assets on which rely personal data Lack of explanation and traceability of decisions taken 	<ul style="list-style-type: none"> Unlawful processing Unfair processing Lack of transparency Diversion of purpose No respect of data minimisation No respect of storage limitation

Generating logs and performing internal audit allows for improvement of supervision of all model assets, to better understand the decisions made in real time by the algorithm, to understand each incident, but also to audit the internal processes to question the processing carried out in the frame of the project. This reduces the likelihood of many threats in this scenario.

The process must be auditable in the sense that certain questions must always be answered clearly: Who accessed the data? (Especially before the pseudonymisation of the data) Why did the algorithm make a particular decision? How exactly was the data processed? The technical and organisational parts of the process must generate traces (computer logs, activity reports, etc.) so that they can be audited. These logs should be stored in a dedicated tool, such as a log management solution. The logs provided by the equipment must be signed (using their own private key) and their signature stored by the server that centralises the logs to ensure a principle of non-repudiation. In addition, at each (usually daily) backup of the log server, these backups must also be signed by the log sink to ensure once again the principle of non-repudiation. Moreover, these logs should be analysed by a SOC and rules should be defined to detect the slightest anomaly such as a high number of calls from a scanner to the X-ray computer-aided diagnostic system, or an abnormal response. In addition, all processes must be regularly audited by the company's internal control teams to ensure compliance with the requirements. As far as privacy is concerned, these audits should at least ensure the lawfulness of the processing, its transparency, its application, the data minimisation, the respect of the storage limitation and the fairness aspect. These checks should be based in part on the logs collected.

This control could have an impact on privacy because the generated logs can contain personal data. This log management system should be included in the privacy impact

assessment scope of the system. Moreover, the logging system should be carefully implemented to avoid unnecessary overhead due to the extra operation it implies.

4.17 IDENTIFY ALL THE DATA PROCESSORS FOR THE MEDICAL DATA PROCESSING AND PERFORM THE CONTROL ACTIONS NECESSARY TO GIVE REASONABLE ASSURANCE THAT THEY ARE COMPLIANT

Type	Associated Vulnerabilities	Threats it mitigates
Privacy	<ul style="list-style-type: none"> Lack of control of Data processor (including external stakeholder) Contract with a low security third party 	<ul style="list-style-type: none"> Diversion of purpose Data disclosure

As the data are manipulated and stored in resources provided by a cloud provider, the medical practice needs to control the cloud provider and its actions.

Contractual clauses, internal and external audits, and assessments of the processors in the diagnostic prediction processing all contribute to ensure the compliance of the implied processors.

The Medical Practice needs to ensure that the Cloud Provider where the model is trained, and the data stored is also compliant with the security and privacy policy. To do this, the medical practice should have privacy and cybersecurity documents completed to understand the level of security and privacy provided by the cloud provider. For example, the cloud provider will need to explain the processing of data in such documents and the means used to ensure its security. Such documents should be attached to the contract with the cloud provider and should be reviewed annually or whenever the contract is amended by either party. **The impact this control could have for the medical practice would be a loss of time and energy spent to formalise documents, and complete the assessments and audits.**

4.18 PERFORM A PRIVACY IMPACT ASSESSMENT

Type	Associated Vulnerabilities	Threats it mitigates
Privacy	<ul style="list-style-type: none"> Lack of controls to ensure that data is used only for the purposes defined Lack of controls to ensure the adequacy of the purpose and its current use Lack of detail on the purposes and justification for their legitimacy 	Diversion of purpose

This privacy measure allows for an in-depth analysis of the impact of the processing on the privacy of users, compliance with the GDPR, and to identify whether the associated risks are well addressed by the proposed privacy and security measures.

The privacy impact assessments and general accountability actions described above (in internal audit process control) help ensure that the purpose of the processing is well defined (the diagnostic of osteoporosis) and that the actual use of the data remains within the scope of this documentation. To do so, the medical practice could use the requirements described below, and assess the vulnerabilities of the implemented system to illustrate privacy threats and counter measures. **Such analysis may possibly impact the performance of the scenario** if it results in one of the functionalities not permitting minimisation of the data protection risks to the rights and freedoms of natural persons.

4.19 DEFINE AND IMPLEMENT A DATA RETENTION POLICY

Type	Associated Vulnerabilities	Threats it mitigates
Privacy	<ul style="list-style-type: none"> Lack of data deletion mechanisms Lack of data retention policy 	No respect of storage limitation

This privacy control, in which a storage duration value must be defined for each personal data involved in the processing (scanners images, first name, surname, age, etc) and implementation, minimises the risk of keeping the data longer than is strictly necessary.

There are two different situations to address. First, the data used for training can be considered as research data, for which the retention period can be two years. Secondly, the data used for the patient file (which is important in case of legal disputes) for which the retention period can be twenty years. However, in terms of performance, once all that data is deleted it can no longer be used to train models, which is why **this control can have an impact on the performance of this scenario.**

4.20 STUDY ON DATA FIELDS NECESSITY AND JUSTIFICATION IN THE PRIVACY POLICY

Type	Associated Vulnerabilities	Threats it mitigates
Privacy	<ul style="list-style-type: none"> Lack of justification for the collection of individual personal data collected Lack of transparency on the purpose, the exact personal and medical data that are extracted, and how they are processed. 	<ul style="list-style-type: none"> Unlawful processing Lack of transparency No respect of data minimisation

A lack of justification on how data are collected correlates with a lack of transparency on the purpose of the processing and on the accuracy of the data collected., This could result in unlawful processing which would have a moderate impact for the patient (significant sense of invasion of privacy). Therefore, the medical practice must provide rationale for the necessity of data fields and justify this in the privacy policy.

The personal data used are the age, gender, and body mass index, X-ray images of the patient. The medical practice must be able to explain why it is necessary to use these data specifically for the purpose of this scenario and it must also formalise the explanation and make it available to the patient.

Beyond the justification on the use of the data, clear explanation is also very important in the context of this scenario. The medical practice must therefore ensure they are able to explain very clearly why the model it created made the decisions it did (in this case, the detection or not of osteoporosis).

Nevertheless, the only data that can be collected is the one with proper justification, purpose and with clear consent of the patient. This practice may have some impact on the relevance of the data collected. Taking the time to justify the need for the data may lead to the abandonment of the collection of certain data for which justification is difficult, which may in turn reduce the quality of the data at hand.

4.21 FORMALIZE A LIA (LEGITIMATE INTEREST ASSESSMENT)

Type	Associated Vulnerabilities	Threats it mitigates
Privacy	<ul style="list-style-type: none"> Lack of legal basis related to legitimate interest Lack of practical means and justification for the legal basis 	Unlawful processing

	<ul style="list-style-type: none"> Lack of transparency on the purpose, the exact data that are collected, and how they are processed 	
--	--	--

The Legitimate Interest Impact Assessment (LIA)²⁸ is used to determine if an organisation can process data using the legitimate interest lawful basis. In this scenario, the legal basis for the use of personal data and images is legitimate interest. If the medical practice does not justify this legal basis, it faces unlawful processing. To justify the legal basis, the medical practice must put in place a LIA. In this LIA, the medical practice must formalise the reflection on how the data processing is necessary, and how it balances with the rights and freedoms of patients.

The impact this could have on the medical practice would be a loss of time and energy spent on formalising the LIA.

4.22 MINIMISE DATA AT EACH STEP OF THE PROCESSING; COLLECT ONLY WHAT IS NEEDED WHEN NEEDED

Type	Associated Vulnerabilities	Threats it mitigates
Privacy	<ul style="list-style-type: none"> Lack of necessary data selection Lack of verification that the data is adequate, relevant, and not excessive for the purpose of delivering a diagnostic Lack of measures to prevent further data collection 	No respect of data minimisation

Minimisation requires an analysis to correctly identify the subset of fields required, and should be performed as early as possible in the processing. Non-minimisation of data is the lack of selection of necessary, adequate, and relevant data. In this scenario the medical practice gathers data such as age, name, body weight and X-ray images of patients. A study on the necessity of collecting data such as age and body weight must be done, and a proof of its necessity made public. However, as mentioned above, data minimisation also involves data pseudonymisation. In this case, the data is pseudonymised and in the event of a security breach, it may be more difficult to perform a cybersecurity forensic analysis. This has the impact of significantly slowing down the work of those employed to determine the cause of a security issue.

4.23 IMPLEMENT A PRIVACY BY DESIGN PROCESS

Type	Associated Vulnerabilities	Threats it mitigates
Privacy	Lack of Privacy by Design	<ul style="list-style-type: none"> Unlawful processing Unfair processing Lack of transparency Diversion of purpose No respect of data minimisation

The lack of Privacy by Design can cause threats related to the privacy aspect of the scenario. Therefore, the medical practice must ensure that a compliance study and privacy risk assessment is formalised in a Privacy Impact Analysis document, and that the identified action plan is implemented before the scenario is put into operation.

Conceiving a project following a Privacy by Design methodology requires less effort than adding privacy on top of an existing project. However, there could be an impact on the functionality if

²⁸ See [https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/legitimate-interests/how-do-we-apply-legitimate-interests-in-practice/#:~:text=There's%20no%20defined%20process%2C%20but,\(consider%20the%20individual's%20interests\)](https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/legitimate-interests/how-do-we-apply-legitimate-interests-in-practice/#:~:text=There's%20no%20defined%20process%2C%20but,(consider%20the%20individual's%20interests))

the outcome of the Privacy by Design methodology would lead to some functionality not being implemented due to the privacy risk it generates (but in our case, personal data are pseudonymised). The other impact that can be noted is related to security, as it may not be possible to collect some data useful for forensic analysis following a cyber-attack, due to privacy limitations.

4.24 CALL ON ETHICAL COMMITTEE AND EXTERNAL AUDITS

Type	Associated Vulnerabilities	Threats it mitigates
Privacy	Lack of review of treatment by a dedicated committee to check fairness	Unfair processing

The lack of review of treatment by a specialised equity committee can lead to unfair treatment. The Medical Practice can find a solution by employing an ethics committee and external auditors. Therefore, in this scenario, the Medical Practice must involve independent entities to review the inputs and outputs of the processing of data to determine whether the processing is unfair. A review from other Medical Practice and Radiologist using the system with their own images could be a proper way to challenge the model and check his fairness.

The impact it could have on the Medical Practice in implementing this measure is the cost of external auditors or the ethics committee.

4.25 DEFINE ACCURACY CRITERIA

Type	Associated Vulnerabilities	Threats it mitigates
Privacy	Lack of accuracy criteria	No respect of accuracy

In addition to tracking the correct predictions made by the model (i.e., does the model correctly diagnose a patient in most situations) it is important to assess the accuracy of these predictions. Models provide, in addition to a predictive output, the probability of this prediction which can be interpreted as an “accuracy level”. For each prediction, accuracy criteria (probability score) should be stored to help a Data Scientist to better understand the models and analyse errors.

In production, it is recommended to follow these probability scores as they are obtained and to set up indicators of good functioning and alerts. For example, an alert should be raised if several low probability scores are noted for the same patient. With this alert, the data scientists should investigate to find the cause of the poor performance of the model. **Such a measure does not affect this scenario.** It may require a little time from the Data Scientist teams, but it allows further improvement in the performance of the model.

4.26 ENSURE THAT THE MODEL IS SUFFICIENTLY RESILIENT TO THE ENVIRONMENT IN WHICH IT WILL OPERATE

Type	Associated Vulnerabilities	Threats it mitigates
Cybersecurity & Privacy	Lack of consideration of real-life condition in training the model	Failure or malfunction of ML application

The Medical Practice and Data Scientist teams need to work together to ensure that they build the most resilient model possible. This involves several steps:

- Using data from real-life conditions to build the model, i.e., direct patient data.
- Testing the model against real-world behaviour to evaluate it
- Testing the model under real-life conditions before deploying it with a relevant set of scenarios to test (e.g., a wide variety of patients...)

- Investigate and test the impossibility of using the model for purposes other than those intended

This control is a compromise between performance and privacy/feasibility. Having as much data as possible from real-life conditions, with different types of scanners, requires a lot of personal data and is not always feasible depending on the equipment available at the time in the medical practice.

4.27 RAISE AWARENESS OF SECURITY AND PRIVACY ISSUES AMONG ALL STAKEHOLDERS

Type	Associated Vulnerabilities	Threats it mitigates
Cybersecurity & privacy	Lack of consideration of attacks to which diagnostic systems could be exposed	<ul style="list-style-type: none"> • Denial of service due to inconsistent data • Poisoning • Data disclosure • Unlawful processing • Unfair processing • Lack of transparency • Diversion of purpose • No respect of data minimisation • Label modification

The Medical Practice must ensure that such threats are comprehended by the teams. For instance, the company must ensure that all its teams (including Data Scientists) are trained in the specificities of machine learning and the new associated cybersecurity and privacy risks.

This can take the form of regular training sessions or local cyber and privacy risks reporting. This allows the teams to be aware of the issues at stake, and sometimes to find solutions adapted to the specificities of the scenario. For Radiologists, who are not normally trained in computer privacy and security issues, it is necessary to train them in basic computer hygiene rules (handling sensitive data, etc.) and how to respect privacy of personal data they handle. **This control does not directly impact performance of the system.**

4.28 USE RELIABLE SOURCES TO LABEL DATA

Type	Associated Vulnerabilities	Threats it mitigates
Cybersecurity	Use of unreliable sources to label data	Label modification

In this scenario radiologists label the data, and a label modification could lead to a misdiagnosis, which would have a high impact on the Medical Practice and on the patient. To prevent label modification and misdiagnosis, the medical practice must ensure that this operation is done correctly. To make sure that the model is correctly trained and delivers proper diagnostics with a good accuracy level, images provided for training and testing must be correctly labelled. If not, the model could output incorrect diagnoses with a high accuracy score.

The Medical Practice should be responsible for ensuring the labelling of the images used for training is conducted by a reliable Radiologist. Radiologists must be trustworthy, have a proven track record of excellent diagnosis and have the time to do it well. It is also necessary to train and assist them in this task to limit human error in the entry of the diagnosis. **This control will have a positive impact on accuracy since the dataset will be correctly labelled.**

4.29 ENSURE THAT MODELS ARE UNBIASED

Type	Associated Vulnerabilities	Threats it mitigates
Privacy	Existing biases in the ML model or in the data	Unfair processing

		No respect of compliance of the training model
--	--	--

It is necessary for the Medical Practice to take various measures to prevent bias in the models. A biased model can cause discrimination, such as gender discrimination, where the model is more effective in producing diagnoses for one sex over another.

This can be done by subjecting the models to different techniques (classification parity, calibration, anti-classification, or sample bias, for instance) as well as by monitoring different indicators (e.g., percentage error of the model for people with disabilities compared to the average error) to create non-discriminatory models by preventing potential biases. In addition, the Medical Practice must provide the patients with a means of reporting an incident to allow fair use of the AI by avoiding such a difficulty (re-training the model, for example). **This privacy control could improve performance of the system over longer periods**, as unbiased models are more accurate.

4.30 SUMMARY

The following table summarises every control described in the previous section. Each control is associated with vulnerabilities, mitigated threats, and addressed privacy and security requirements.

Figure 18: Summary of controls and mitigated threats

Control name and type	Associated Vulnerabilities	Threat mitigated	Privacy and security requirements addressed
Implement a security by design process	Lack of security by design	Compromise of diagnostic system components Evasion Human error Data disclosure Poisoning	Integrity of the data Availability of the data Confidentiality of the data Traceability of the data
Document the diagnostic system	Lack of documentation on the diagnostic system	Human error Unlawful processing Unfair processing Lack of transparency Diversion of purpose No respect of data minimisation No respect of storage limitation	Integrity of the data Availability of the data Confidentiality of the data Traceability of the data Lawfulness of the process Fairness of the process Transparency of the process Purpose limitation of the process Data minimisation of the process Accuracy of the data Storage limitation of the data
Check the vulnerabilities of the components used and implement processes to	Lack of security process to maintain a good security level of the components of	Compromise of diagnostic system components Data disclosure	Integrity of the data Availability of the data

maintain security levels of ML components over time	<p>the diagnostic system</p> <p>Use of vulnerable components</p> <p>Disclosure of sensitive data for ML algorithm training</p>		
Add some adversarial examples to the dataset	Lack of training based on adversarial attacks	Evasion	Accuracy of the data
Choose and define a more resilient model design	Poor consideration of evasion attacks in the model design implementation	Evasion	Accuracy of the data
Integrate poisoning control in the training dataset	<p>No detection of poisoned samples in the training dataset</p> <p>Lack of control for poisoning</p>	Poisoning	Integrity of the data
Enlarge the training dataset	Lack of data for increasing robustness to poisoning	Poisoning	Integrity of the data
Secure the transit of the collected data	Poor data management	Poisoning	Integrity of the data
Control all data used by the ML Model	<p>Use of uncontrolled data</p> <p>Lack of detection of abnormal inputs</p> <p>Poor data management</p>	Poisoning	Integrity of the data
Implement access right management process	Poor access rights management process	<p>Compromise of diagnostic system components</p> <p>Poisoning</p> <p>Data disclosure</p> <p>Human error</p>	<p>Integrity of the data</p> <p>Availability of the data</p> <p>Confidentiality of the data</p> <p>Traceability of the data</p>
Ensure all systems and devices comply with authentication, and access control policies	<p>Weak access protection mechanisms for machine learning model components</p> <p>Weak access protection</p>	<p>Compromise of diagnostic system components</p> <p>Poisoning</p> <p>Data disclosure</p> <p>Human error</p>	<p>Integrity of the data</p> <p>Availability of the data</p> <p>Confidentiality of the data</p> <p>Traceability of the data</p>

	mechanisms of the personal data (encryption, access control mechanism...)		
Monitor the performance of the model	Lack of control over model performance	Unfair processing No respect of accuracy	Integrity of the data Fairness of the process Purpose limitation of the process
Reduce the available information about the model	Too much information available on the model Using a widely known model allowing the attacker to study it	Compromise of diagnostic system components	Confidentiality of data
Identify a data controller for the medical data processing	Absence of an identified data controller	Unlawful processing Unfair processing Lack of transparency	Lawfulness of the process Fairness of the process Transparency of the process
Pseudonymise data coming from the historical patient	Lack of pseudonymisation Unprotected sensitive data on test environments	No respect of data minimization	Confidentiality of data Data minimisation of the process
Generate Logs and perform Internal audit	Lack of auditability of processing Lack of traceability of actions and/or modifications made to the assets on which rely personal data Lack of clear explanations and traceability of decisions taken	Unlawful processing Unfair processing Lack of transparency Diversion of purpose No respect of data minimisation No respect of storage limitation	Lawfulness of the process Fairness of the process Transparency of the process Purpose limitation of the process Data minimisation of the process Accuracy of the data Storage limitation of the data
Identify all the data processors for the medical data processing and perform the control actions necessary to give reasonable assurance that they are compliant	Lack of control of Data processor (including external stakeholder) Contract with a low security third party	Diversion of purpose Data disclosure	Integrity of the data Availability of the data Confidentiality of the data Traceability of the data Lawfulness of the process Fairness of the process Transparency of the process Purpose limitation of the process

			Data minimisation of the process Accuracy of the data
Perform a privacy Impact Assessment	Lack of controls to ensure that data is used only for the purposes defined Lack of controls to ensure the adequacy of the purpose and its current use Lack of detail on the purposes and justification for their legitimacy	Diversion of purpose	Purpose limitation of the process Lawfulness of the process Data minimisation of the process Transparency of the process
Define and implement a data retention policy	Lack of data deletion mechanisms Lack of data retention policy	No respect of storage limitation	Storage limitation of the data
Study on data fields necessity and justification in the privacy policy	Lack of justification for the collection of individual personal data collected Lack of transparency on the purpose, the exact personal and medical data that are extracted, and how they are processed.	Unlawful processing Lack of transparency No respect of data minimisation	Lawfulness of the process Data minimisation of the process Transparency of the process
Formalise a LIA (Legitimate Interest Assessment)	Lack of legal basis related to legitimate interest Lack of practical means and justification for the legal basis Lack of transparency on the purpose, the exact data that are collected, and how they are processed	Unlawful processing	Lawfulness of the process
Minimise data at each step of the processing; collect only what is needed when needed	Lack of necessary data selection Lack of verification that the data is adequate, relevant,	No respect of data minimisation	Data minimisation of the process

	<p>and not excessive for the purpose of delivering a diagnostic</p> <p>Lack of measures to prevent further data collection</p>		
Implement a privacy by design process	Lack of privacy by design	<ul style="list-style-type: none"> Unlawful processing Unfair processing Lack of transparency Diversion of purpose No respect of data minimisation 	<ul style="list-style-type: none"> Lawfulness of the process Fairness of the process Transparency of the process Purpose limitation of the process Data minimisation of the process
Call on ethical committee and external audits	Lack of review of treatment by a dedicated committee to check fairness	<ul style="list-style-type: none"> Unfair processing 	<ul style="list-style-type: none"> Fairness of the process
Define accuracy criteria	Lack of accuracy criteria	<ul style="list-style-type: none"> No respect of accuracy 	<ul style="list-style-type: none"> Accuracy of the data
Ensure that the model is sufficiently resilient to the environment in which it will operate	Lack of consideration of real-life conditions in training the model	<ul style="list-style-type: none"> Failure or malfunction of ML application 	<ul style="list-style-type: none"> Integrity of the data Availability of the data Confidentiality of the data Lawfulness of the process Fairness of the process Transparency of the process Purpose limitation of the process Data minimisation of the process Accuracy of the data
Raise awareness of security and privacy issues among all stakeholders	Lack of consideration of attacks to which diagnostic systems could be exposed	<ul style="list-style-type: none"> Denial of service due to inconsistent data Poisoning Data disclosure Label modification Unlawful processing Unfair processing Lack of transparency Diversion of purpose No respect of data minimisation 	<ul style="list-style-type: none"> Integrity of the data Availability of the data Confidentiality of the data Traceability of the data Lawfulness of the process Fairness of the process Transparency of the process Purpose limitation of the process Data minimisation of the process Accuracy of the data
Use reliable sources to label data	Use of unreliable sources to label data	<ul style="list-style-type: none"> Label modification 	<ul style="list-style-type: none"> Accuracy of the data

Ensure that models are unbiased	Existing biases in the ML model or in the data	Unfair processing No respect of compliance of the training model	Fairness of the process
---------------------------------	--	---	-------------------------

5. CONCLUSION

In this report, analysis of **medical imaging in osteoporosis diagnosis supported by Artificial Intelligence (AI) is presented**. AI is often used as an umbrella term that encompasses the technology behind many smart solutions and devices. It is a constantly evolving field where new innovations appear regularly in different areas of activity.

Even though **AI can be very beneficial for the industries areas it applies to, it can also have quite a significant impact for security and privacy**, especially when these systems have sensitive functionalities, as they are highlighted in this report. Indeed, AI comes with a wide range of privacy and security vulnerabilities causing threats with heavy impacts for organisations.

Regardless how AI is being used in support of business functionality, it should not be a surprise that the many identified threats are similar. However, we must be aware of the fact that **depending on the context of the scenario, the same threats apply differently and have different levels of impact**. Regarding the impact of each threat, it must be also noted that even in the same scenario every instance is unique, and a proper study must be carried out by each company using AI to maintain a proper security and privacy level.

This guide helps in the identification and evaluation of threats in a specific business scenario which uses AI, but it is important to remember that while the analyzed scenario is based on real life examples, it includes assumptions which may not match the business context in which other organizations would like to implement in. Therefore, **the entire cybersecurity and privacy context (requirements, threats, vulnerabilities, and controls) must be adapted to the context and reality of the individual organization**. In addition, the controls proposed in this document are not sufficient on their own, and must be complemented by the standard security measures that already exist.

While security and privacy are not necessarily the same, they are intimately related, and equally important. In their management, a balance must be found between the two in the sense that one must always make sure that the regulations and recommendations concerning the two aspects are always respected. Unfortunately, these two parameters are regularly at the expense of performance. It is therefore an equation with three variables, two of which respond to the need for regulation and risk, which need to be correctly balanced to achieve the desired effect.

A ANNEX: SECURITY AND PRIVACY SCALES AND REQUIREMENTS

A.1 CYBERSECURITY AND PRIVACY SEVERITY SCALES

Availability	
Level	Definition
Low	Service provided must be restored within few days or less .
Moderate	Service provided must be restored within a day or less .
High	Service provided must be restored within half a day or less .
Critical	Service provided must be restored within few hours or less .

Integrity	
Level	Definition
Low	A loss of integrity in the process does not need to be identified or corrected .
Moderate	Any degradation in the process must be identified but not necessarily corrected .
High	Any degradation in the process must be identified and corrected .
Critical	No degradation in the process is tolerated at any time.

Confidentiality	
Level	Definition
Low	Process-related data can be accessed by everyone .
Moderate	Access to process-related data is restricted to internal staff and trusted partners .
High	Access to process-related data is restricted to employees having an organisation or functional link with the process .
Critical	Access to process-related data is restricted to a very limited number of individuals .

Traceability	
Level	Definition
Low	The absence of traces of actions on the service provided is acceptable .
Moderate	Actions related to the service provided must be identified . They must be traced and detected.
High	The actions related to the process and their actors must be identified and dated . They must be imputable .
Critical	Service provided actions must be legally enforceable and time stamped . They must have a probative value .

A.2 CYBERSECURITY SCALE OF IMPACT

Severity ²⁹	
Level	Definition
1 - Low	No operational impact on business performance or on the safety of people and property. The company/entity will overcome the situation without too much difficulty.
2 - Moderate	Degradation of business performance without impact on safety of people and property. The company/entity will overcome the situation despite some difficulties (operation in degraded mode).
3 - High	Severe deterioration in the performance of the business, with possible significant impacts on the safety of people and property. The company/entity will overcome the situation with serious difficulties (operation in very degraded mode).
4 - Critical	Inability of the company/entity to carry out all or part of its business, with possible serious impacts on the safety of people and property. The company/entity is unlikely to overcome the situation (its survival is threatened).

A.3 PRIVACY SCALE OF IMPACT

Severity ³⁰	
Level	Definition
1 - Low	The persons concerned affected will not be affected or may experience some inconvenience, which they overcome without difficulty.
2 - Moderate	The persons concerned may experience significant inconvenience, which they be able to overcome despite some difficulties.
3 - High	The persons concerned may experience significant consequences, which they should be able to overcome, but with real and significant difficulties.
4 - Critical	The people concerned may experience significant consequences, if not irreparable irremediable consequences, that they may not overcome.

²⁹ Based on « Agence Nationale de la Sécurité des Systèmes d'Information » (ANSSI). See: https://www.ssi.gouv.fr/uploads/2019/11/anssi-guide-ebios_risk_manager-en-v1.0.pdf

³⁰ Metrics based on the National Commission on Informatics and Liberty (CNIL) -an independent French Administrative regulatory body. See: <https://www.cnil.fr/sites/default/files/atoms/files/cnil-pia-3-en-knowledgebases.pdf>

A.4 PRIVACY REQUIREMENTS CRITERIA

Regarding the privacy, the applied requirements are based on the GDPR data protection principles, which are summarized in the table below.

Requirements	Article
Lawfulness, fairness, and transparency	Art. 5.a
Purpose limitation	Art. 5.b
Data minimisation	Art. 5.c
Accuracy	Art. 5.d
Storage limitation	Art. 5.e
Security of personal data (integrity and confidentiality)	Art. 5.f

Moreover, as a complement to the GDPR requirements listed above, some key privacy topics related to AI systems need to be addressed (some of them are specifically mentioned by national data protection supervisory authorities³¹). These additional privacy requirements are listed in the next table.

Recommendations	Details
Database creation	Projects must ensure the compliance of the data collected and injected into the database for training and in production. Moreover, they must ensure that only legitimate persons have access to the data in the database, prevent data loss and hide personal data.
Compliance of the training model (i.e., before production)	The data collected and the training model must be processed in accordance with the state of the art of Machine Learning development to guarantee control of all processes carried out. This consists, for instance, of good justifications for the chosen learning method, the reliability of the third-party tools used, paying attention to open-source data, conducting a meticulous training protocol based on the state of the art, and finally verifying the quality of the system once in the experimentation phase. ³²

³¹ Such as the CNIL. It proposes an analysis grid to enable organizations to assess the maturity of their artificial intelligence systems regarding the GDPR available on <https://www.cnil.fr/fr/intelligence-artificielle/guide>.

³² see the various CNIL method sheets on this subject, such as this one: <https://www.cnil.fr/fr/intelligence-artificielle/guide/developper-et-entrainer-un-algorithme>



ABOUT ENISA

The European Union Agency for Cybersecurity, ENISA, is the Union's agency dedicated to achieving a high common level of cybersecurity across Europe. Established in 2004 and strengthened by the EU Cybersecurity Act, the European Union Agency for Cybersecurity contributes to EU cyber policy, enhances the trustworthiness of ICT products, services and processes with cybersecurity certification schemes, cooperates with Member States and EU bodies, and helps Europe prepare for the cyber challenges of tomorrow. Through knowledge sharing, capacity building and awareness raising, the Agency works together with its key stakeholders to strengthen trust in the connected economy, to boost resilience of the Union's infrastructure, and, ultimately, to keep Europe's society and citizens digitally secure. More information about ENISA and its work can be found here: www.enisa.europa.eu.

ENISA

European Union Agency for Cybersecurity

Athens Office

Agamemnonos 14, Chalandri 15231, Attiki, Greece

Heraklion Office

95 Nikolaou Plastira

700 13 Vassilika Vouton, Heraklion, Greece

enisa.europa.eu



ISBN 978-92-9204-641-5
doi: 10.2824/25285