INCISIVE

Improving cancer diagnosis and prediction with AI and big data

A Multimodal AI-based Toolbox and an Interoperable Health Imaging Repository for the Empowerment of Imaging Analysis related to the Diagnosis, Prediction and Follow-up of Cancer

Deliverable 6.5

Evaluation of INCISIVE Blind studies

WP 6 –INCISIVE System Integration and Pilot Studies

<22-12-2023>

Revision <1.0>

< Final >

Grant Agreement n 952179

European Commission

DOCUMENT CONTROL	
Project reference	Grant Agreement number: 952179
Document name	Evaluation of INCISIVE Blind studies
Work Package	WP 6
Work Package Title	Evaluation of INCISIVE Blind studies
Dissemination level	CO
Revision	<1.0>
Status	< Final>
Reviewers	Gianna Tsakou, MAG, Andreas Charalambous, CUT
Beneficiary(ies)	UNS, ICCS, AUTH, CUT, KU, CERTH

Dissemination level:
PU = Public, for wide dissemination (public deliverables shall be of a professional standard in a form suitable for print or electronic publication) or CO = Confidential, limited to project participants and European Commission.

AUTHORS		
	Name	**Organisation**
Document leader	Tatjana Loncar-Turukalo	UNS
Participants	Niksa Jakovljevic	UNS
	Ivan Lazic	UNS
	Milan Rapaic	UNS
	Jasmina Boban	UNS
	Igor Nosek	UNS
	Gorana Mijatovic	UNS
	Tijana Nosek	UNS
	Goran Martic	UNS
	Nebojsa Bozanic	UNS
	Olga Tsave	AUTH
	Dimitris Filos	AUTH
	Alexandra Kosvyra	AUTH
	Dimitrios Fotopoulos	AUTH
	Ioanna Chouvarda	AUTH
	Ioannis Rallis	ICCS
	Stavros Sykiotis	ICCS
	Ioannis Tzortzis	ICCS

	Nikolaos Temenos	ICCS
	Nikolaos Bakalos	ICCS
	Anastasios Doulamis	ICCS
	Nikolaos Doulamis	ICCS
	Aikaterini Angeli	ICCS
	Matthaios Bimpas	ICCS
	Athanasios Voulodimos	ICCS
	Dimitrios Kalogeras	ICCS
	Ioannis Vergados	ICCS
	Anastasiois Tzelepakis	CERTH
	Paschalis Bizopoulos	CERTH
	Zisis Sakellariou	CERTH
	Shereen El Nabhani	KU
	Lithin Zacharias	KU
	Andreas Charalambous	CUT

REVISION HISTORY				
Revision	**Date**	**Author**	**Organisation**	**Description**
0.1.1.	19.09.2023	Tatjana Loncar-Turukalo	UNS	Definition of ToC
0.1.2.	25.10.2023	Tatjana Loncar-Turukalo	UNS	Input for sections 1 and 2, modification of ToC
0.1.3.	10.11.2023.	Tatjana Loncar-Turukalo	UNS	Integration of inputs for 3.2 and 3.5
0.1.4.	5.12.2023.	Tatjana Loncar-Turukalo	UNS	Integration of inputs for 3.3 and 3.4
0.1.5.	14.12.2023	Tatjana Loncar-Turukalo	UNS	Integration of inputs for 3.2.7 and 3.2.8., inputs in chapters 4 and 5, and formatting revision
0.1.6.	16.12.2023	Andreas Charalambous, Gianna Tsakou	CUT, MAG	Peer review
0.1.7.	21.12.2023	Tatjana Loncar-Turukalo	UNS	Addressing review comments, Final version before QC
1.0.	22.12.2023	Sofia Theodoridou	MAG	QC, Final version

Disclaimer and statement of originality

The content of this deliverable represents the views of the authors only and is their sole responsibility; it cannot be considered to reflect the views of the European Commission and/or the Consumers, Health, Agriculture and Food ExecutiveAgency or any other body of the European Union. The European Commission and the Agency do not accept any responsibility for use of its contents.

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Contents

Table of figures

Table of Tables

Terms and Abbreviations

Term	Description

Abbreviation	Description
EC	European Commission
TCC	Technical and Clinical Committee
WP	Work Package
T	Task
MMG	Mammography
MRI	Magnetic Resonance Imaging
CT	Computerized Tomography
PET	Positron Emission Tomography
CXray	Chest X ray
H&E	Haematoxylin and Eosin
HEIP	H&E Image Processing
KRAS	Kirsten rat sarcoma virus
EFGR	Epidermal Growth Factor Receptor
ISUP	International Society of Urological Pathology
BIRADS	Breast Imaging-Reporting And Data System
DWI	Diffusion-Weighted Images
PSA	Prostate Specific Antigen
DRE	Digital Rectal Exam
T2W	T2-weighted
SVM	Support Vector Machines
BRF	Balanced Random Forest
BA	Balanced Accuracy
PFS	Progression Free Survival

1 Introduction

1.1 Purpose and scope

This deliverable reports the results from the observational pilot study which is focused on estimation of quantitative performance metrics related to AI services offered through the final INCISIVE prototype. It aims to deliver information on how reliable the services are, as measured using an adequate performance metric depending on the service task.

The final AI models running behind the AI services are trained, validated in tested in three iterations, as described in D4.1, D4.2 and D4.3, and the performance of the final model versions has been reported in D4.3. In the observational study, INCISIVE has moved on to collect new data samples across multiple data providers in order to evaluate both generalization capacities of the models supporting the AI services and evaluate ground truth annotations provided by health care professionals during the study. The model performance will be compared with those achieved in the evaluation done during model development, presented in D4.3

Data providers involved in the observational study were: AUTH, UoA, HCS, GOC, UNS, IDIBAPS, UNITOV and DISBA.

The observational study precedes the final INCISIVE platform evaluation study: the feasibility study, where AI services developed within the project and offered to health care providers, will be used and user satisfaction and trust measured over time.

1.2 Document structure

The deliverable D6.5 is comprised of the following sections:

Section 2 “Observational study” briefly reviews the observational study protocol, and describes the efforts within INCISIVE to curate and improve the INCISIVE database to facilitate smooth and trustworthy usage of AI services

Section 3 “Performance evaluation of AI toolkit” contains evaluation of all models on the new dataset not used during the model development. This section reviews AI services and pipelines for each cancer type. The evaluation metrics are provided on the level of AI services and compared to those reported in D4.3. Section 3.5 “Analysis of discrepancies in model prediction and annotations” analyses the discrepancy in results between the selected classification and segmentation/detection models and original labels/annotations provided by INCISIVE data providers at the time of data submission. Additional human readers were engaged to evaluate cases with disagreement.

Section 4 provides discussions on flow and results of the observational study, as well as on lessons learned and potential improvements.

Section 5 concludes the deliverable and summarizes the results.

1.3 Relation with other deliverables

This deliverable presents the generalization potential of the AI models behind AI services developed in WP4 and explained through a series of deliverables: D4.1, D4.2 and D4.3, where D4.3 “INCISIVE AI-toolbox, data analytics and user services “contains the description of the final, selected models to be integrated in the INCISIVE AI platform and offered as a service. Moreover, it is related to D2.6 “Study Protocols Definition” and an appendix to D2.6. made in D4.2. “INCISIVE AI-toolbox, data analytics and user services”, which define the observational study protocol. Indirectly, this deliverable is related to other deliverables in WP2 where, based on the user needs, requirements for the AI functionalities and scenarios of INCISIVE platform usage have been defined. These are: D2.1/D2.2 “INCISIVE User requirements”, D2.3 “INCISIVE Scenarios Definition” and D2.5 “User requirements definition and system design”. Finally, this deliverable is related to D6.3 INCISIVE Integrated Prototypes – Final Version, where deployment and integration of the D4.3 final models into INCISIVE platform AI services have been explained.

2 Observational study

2.1 Brief overview of the observational study protocol

The methodological design of the study incorporates a prospective observational study with agreement approach. The study aims to validate specificity and sensitivity of the services provided by the INCISIVE system, which can be grouped in 2 major categories namely a) Initial Diagnosis and b) Disease characterization/Staging.

Based on the aim of the Study Protocols Definition the following research questions have been identified:

- What is the precision (positive predictive value) of the AI output?
- What is the recall (sensitivity) of the AI output?
- What is the F1-Score (harmonic mean of precision and recall) of the AI output?
- What are the False Positive and False Negative Rates of the AI output?

It is worth noting that the research question related to performance metrics will be aligned with the type of the AI model and metrics already reported in D4.3. when models were developed and evaluated. This should facilitate fast performance comparison and evaluation of the generalization potentials of the AI models developed. Sensitivity and specificity provide an indication as to how much confidence we can put into the INCISIVE output, applicable in certain prioritization models. Diseased individuals predicted as positive are called true positives and non-diseased individuals predicted as negative are called true negatives. Diseased individuals predicted as negative are called false negatives and non-diseased predicted as positive are called false positives. The sensitivity of a model is calculated as the number of diseased that are correctly classified, divided by all diseased individuals. The specificity is calculated as the number of correctly classified non-diagnosed patients divided by all non-diagnosed individuals. For specific services, where the metrics defined by the observational study protocol are well suited, such as diagnostic prioritization services where patients with non-oncological findings can be expected, full sets of metrics will be reported (precision, recall, F1).

The data required (=cases .e. a clinical diagnostic episode within a patient journey which is accompanied with an imaging examination) for a fair evaluation were retrieved through a purposive sampling technique. Cases at the participating centers in 5 countries from the Southern, Western and Central Europe (Italy, Spain, Cyprus, Greece, Serbia) during the data collection period of the project were considered eligible and recruited accordingly, given they met the predefined inclusion and exclusion criteria. For the purpose of the clinical studies, a case has been defined as a clinical diagnostic episode within a patient journey which is accompanied with an imaging examination (such as initial diagnosis, disease specification, treatment response evaluation, etc.).

The minimum sample size required (Table 1) has been calculated based on the different values of the prevalence of the four tumor diagnoses and both sensitivity and specificity of the INCISIVE AI output test (while in the meantime, the power is set to be at least 80% and the p-value, is set to be less than 0.05). For consistency purposes all the prevalence rates for the 5 recruitment sites across the 4 tumor diagnoses have been calculated based on the country-specific data provided by the International Agency for Research in Cancer (IARC). The minimum sample size required for sensitivity and specificity test was calculated by using PASS software.

country	Breast Cancer	Prostate Cancer	Colorectal Cancer	Lung Cancer
Cyprus	155	155	50	100
Greece	200	50	231	231
Italy	450	-	350	350
Serbia	50	-	-	-
Spain	-	107	-	-
Total	855	312	631	681

Table 1. Cases per country for the Prospective Observational Study

The total number of samples reported in Table 1 has to be reached in evaluation of all AI services in total for the certain cancer type.

The duration of the study has been defined as 11 months extending from M25 to M35. This period of time will allow for the gradual recruitment of patients, collection of diagnostics and follow up data, data preparation and evaluation of the AI services against labels and annotations made by the human readers in the data collection process.

Due to the multidisciplinary nature of cancer management, the following disciplines have been identified as the most relevant and therefore can be included as human evaluators in the process of data preparation and evaluation of the INCISIVE AI toolkit results:

- **Initial Diagnosis –** Radiologist / Tumor-Specific Specialist (e.g. Urologist) / Medical Oncologist / Radiation Oncologist
- **Disease characterization/Staging –** Radiologist / Histopathologist / Nuclear Medicine Physician / Oncologist (of various specialties)

At minimum 2 human readers can be used per data provider (i.e. recruitment site). The decision to use a minimum number of human readers was based on the principles of consistency and trustworthiness.

The following time points on the patient journey have been set and included in the guidelines for data providers to achieve uniform data structure prior to upload onto the INCISIVE platform:

0. Baseline- before any treatment is given
1. After 1st Treatment,
2. 1st Follow Up (or second line of treatment),
3. 2nd Follow Up (or third line of treatment),
4. For training the algorithm reasons. Any other measurement performed in-between time points 0-3 will be collected (e.g. in between time points)

The modalities that can be included per time point are described in Annex 4

In terms of the INCISIVE services, these correlate as follows to the above time points:

- Initial Diagnosis (Time points 0 + 1)
- Disease characterization/staging (following surgical intervention (where applicable), after radiotherapy (where applicable), and chemotherapy (Time point 2)
- Disease characterization/staging (Secondary Staging) (Time point 3)

Data retrieved at Time point 4 can be correlated at any of the 3 INCISIVE services, depending on the stage at the process where these will be retrieved.

The INCISIVE diagnostic performance within observational study was evaluated against the human performance standards. In cases where there were deviations between AI service predictions and the annotations provided by the medical professionals, another medical professional should have been asked to validate whether the prediction was wrong or whether there was an issue in the original annotation. These external moderators should differ according to tumor type and diagnostic modality. It is worth noting that, as reported in D4.2., during the observational study period, a series of evaluations of annotations made by human evaluators has been performed by external moderators as a data curation process to achieve trustworthy labels, both for the model development and their further evaluations and usage. For these reasons in this deliverable, additional human evaluation was done:

- for two detection/segmentation models where the manual annotations by medical professionals were needed, as an additional security measure to ensure annotation consistency; and
- for one classification model where both accuracy of image annotation and clinical data entries is relevant.

Efforts already invested in annotations correction/revisions and limited time of clinicians/experts, hamper thorough manual analysis of errors in all models.

2.2 Data collection in the prospective observational study

In the period of M25 to M35 the data providers had an extensive task to recruit the patients for INCISIVE observational pilot study, collect the data related to the Baseline, diagnostic, time point, and follow these patients in time in order to collect as much as possible additional time points to facilitate evaluation of disease characterisation and staging services, where applicable.

The data collection over all sites has been supported by the new data collection guidelines, new version of the assistive tools and have included the following steps:

1. extraction of clinical data in the INCISIVE templates
2. selection and anonymization of DICOM images
3. running data quality check and correcting the reported errors
4. annotation of images
5. data upload to the central infrastructure
6. Data correction for folder structure/annotations upon request based on an automated data quality evaluation

The total number of cases (i.e. a clinical diagnostic episode within a patient journey which is accompanied with an imaging examination) per partner for each cancer type has been reported in Table 2 for breast cancer, in Table 3 for lung cancer, in Table 4 for prostate cancer, and in Table 5 for colorectal cancer.

The data collection process has been determined by **the pace of new cases for each cancer type in collaborating hospitals**. Being aware of this risk, all DPs have been advised to start patient recruitment slightly in advance, prior to M25 in order to achieve the overall numbers needed to evaluate the generalization potentials of AI models with the required significance level.

Another **obstacle in the data collection process**, besides the number of the relevant hospital admissions, has been **related to patient consent**. As despite the study uses only the selected anonymized patients' data and does not influence in any way the patient diagnoses and treatment path, some patients were not willing to consent, which was understandable having in mind their difficult state at both psychological and physiological levels.

As seen in Table 2,Table 3,Table 4, and Table 5 some DPs have not reached the target number of the observational data per GA and D2.6 by M37 (October 2023) and data collection has overall been delayed for some DPs during all stages, including observational stage. As a mitigation strategy, the Consortium has decided that the technical partners developing the AI models should separate INCISIVE data used for the final model training, selection and evaluation (as reported in D4.3 in most cases data collected retrospectively, uploaded by M32), and **use only new uploaded data** in the observational study evaluations. This new data was obtained from:
(1) delayed data from the retrospective data collection stage, (2) delayed data from the prospective stage intended for AI training, (3) the collected observational data, (4) unused data from the open data sets, and (5) new open datasets **in order** to perform blind and objective evaluation of the developed models in the observational study **using unseen data samples (images).**

In total for the observational study the following number of cases was used :(1) for lung cancer 435 Xray and 255 CT scans all INCISIVE cases were used (681 needed), for breast cancer 1426 INCISIVE MMG cases and 100 MRI cases from open datasets (6 services were evaluated in total using 4641 samples), for colorectal cancer histopathology model was tested in 7555 image patches from two open data sets, while survival rate prediction model was tested on 70 patients from an open data set , and prostate cancer services were tested on 184 MRI INCISIVE cases and 100 cases from open data set for prostate gland segmentation service.

COUNTRY	To be delivered GA & D2.6	DATA PROVIDER	COLLECTED, UPLOADED and CURATED DATA		Partner/country target reached	
	cases		patients	cases	Target reached	Percentage of collected data (under/over 100%)
GREECE	200	HCS*	847	907	Yes	< 462%
		AUTH	5	18		
SERBIA	50	UNS	35	107	Yes	< 214%
ITALY	225	DISBA	23	57	No	> 25.3%
	225	UNITOV	163	177	No	> 78,6%

CYPRUS	155	GOC	52	159	Yes	< 102,5%
TOTAL	855			1425		

*HCS has 2720 patients/3093 cases reported in the retrospective stage, per M30 management report, the collection is ongoing, patients/cases reported here are new ones and have been used in evaluations during the observational study

Table 2. Breast cancer prospective observational data collection status

COUNTRY	To be delivered GA & D2.6	DATA PROVIDER	COLLECTED, UPLOADED and CURATED DATA		Partner/country target reached	
	CASES		patients	cases	Target reached	Percentage of collected data (under/over 100%)
GREECE	231	AUTH	0	0	No	>0%
		UoA	18	50	Yes	100% (partner target)
ITALY	175	DISBA	19	32	No	>18.29%
	175	UNITOV	95	101	No	>57.7%
Cyprus	100	GOC	19	45	No	>45%
TOTAL	681			228		

Table 3. Lung cancer prospective observational data collection status

COUNTRY	To be delivered GA & D2.6	DATA PROVIDER	COLLECTED, UPLOADED and CURATED DATA		Partner/country target reached	
	CASES		patients	cases	Target reached	Percentage of collected data (under/over 100%)
GREECE	50	UoA	48	48		>96%
CYPRUS	155	GOC	50	134	No	>86,45%
SPAIN	107	IDIBAPS	117	117	Yes	< 106.4%

TOTAL	312			299		

Table 4. Prostate cancer prospective observational data collection status

COUNTRY	To be delivered GA & D2.6	DATA PROVIDER	COLLECTED, UPLOADED and CURATED DATA		Partner/country target reached	
	CASES		patients	cases	Target reached	Percentage of collected data (under/over 100%)
GREECE	231	AUTH	24	76	No	>33%
ITALY	175	DISBA	12	14	No	>8%
	175	UNITOV	140	157	No	>89.7%
CYPRUS	50	GOC	20	42	No	>84%
TOTAL	631			289		

Table 5. Colorectal cancer prospective observational data collection status

2.3 Data curation if the INCISIVE database within the prospective observational study

As data collection in retrospective data cycle took place under WP3:T3.1 and has been the longest data collection phase, it has resulted with the large volume of data collected and reported in the INCISIVE M30 management report. This data has been used for the training of the AI models developed and described in the series of deliverables D4.x, whereas the final models are described in D4.3

In the model development phase, the INCISIVE data have been extensively used and analysed and different quality and structural issues have been noticed by AI developers, which is not uncommon in so large databases with multiple data sources. Despite the data collection guidelines were issued and support in the data upload process ensured, in such a large database errors were expected. Some of the noticed errors are: folder structure and folder naming problems, different annotation styles and inconsistent annotations over different data sites, MRI sequences were not decoupled, thus could not be automatically loaded, in some cases annotations were not correctly placed in the same folder with the corresponding image.

In order to resolve these issued, facilitate model development/usage and improve further data collection in the observational study phase we have improved the existing and developed some new data curation tools to automatically correct and curate the database. The following data curation steps have been made:

- **Improved DICOM image anonymizer (T5.2)–** the correction applies only to preservation of MRI sequence type in DICOM header in order to allow for an automated search and use of a specific MRI sequences, as required by AI models.
- **Improved version of Data Integration quality check tool (T3.1)** For the last part of this study an updated version of the tool was delivered included some new components: (i) Error flags in Content Validity: each error reported is annotated with a flag explaining the type of error: (a) data type (e.g. number instead of string) and (b) value range (e.g. the value inserted is not compliant with the allowable value range), (ii) template Consistency: checks for consistency inside the clinical metadata template (e.g. provision of laboratory examination data in the timepoints stated that there are laboratory examinations), (iii) Burned-in Information: This component checks the dicom images for burnt-in information. If the image detected is a DICOM report, the image is flagged as 'Report' and needs to be removed. If the image detected is an actual DICOM image, the message 'Replace image' appears. The image should be replaced with a cleaned image placed in a folder along with the rest of the data and the full path of the cleaned image is reported in tool, (iv) Annotation Analysis: This component examines the annotations provided for the images. A table lists the annotation files found in the specified path. For quality assurance of the annotation files, the component verifies whether each annotation file is located in the expected folder, specifically within the series folder. Additionally, if more than one annotation file is present in the folder, the corresponding rows are highlighted in orange. Furthermore, each annotation file must include specific labels corresponding to the cancer type and imaging modality. The component checks for the presence of correct annotation labels and flags any invalid labels in the files by highlighting them in red. Moreover, the component ensures consistency between the Image Region of Interest (ROI) and the ROI provided by the annotation, as well as consistency in the number of images and their respective annotation slices. Finally, the component verifies the agreement between the provided files and the information outlined in the template file.

- **Improved annotation guidelines for each cancer type and each imaging modalities (T6.3)** – the initial guidelines are improved through a series of workshops organized with the data providers for each cancer type. As envisaged by the observational study protocol, external validators have checked all annotated images, and all discrepancies, sporadic or systematic errors and all sources of ambiguities have been noted. Based on these observations new set of more precise annotation rules have been discussed in the workshops and jointly approved between radiologists. All images were re-annotated according to the new rules, this correction process started from M26 and has been completed by M29. The process has been described in D4.2, and the refined guidelines have been published internally, and as part of the D 6.4 “INCISIVE training material” to serve as a guideline for new data donors.

The process of annotation correction has significantly improved the quality of annotations. More importantly, the *overall quality of INCISIVE database has been improved* and *usage of potentially erroneously annotated images in the development of the final models has been avoided*.

- **Data curation script for harmonization of data structure and detection of corrupted data sources (T 5.2)**

  The curation script's objective is to ensure the integrity and quality of the INCISIVE data. One of its primary tasks is to address issues related to the folder structure. It identifies instances where the folder structure does not adhere to the prescribed guidelines, particularly in cases where multiple sequence Series folders exist. The script splits these into distinct, single sequence Series folders, by using the Sequence Name and Series Time tags stored in the DICOM files, ensuring consistency. It also deals with NIFTI files that are found outside of their corresponding Series folders. When possible, it automatically resolves this issue by matching NIFTI files with the appropriate Series folders, but it may involve healthcare professionals in cases where clarity is lacking. Furthermore, the script identifies and removes empty Series folders, ensuring the dataset is free from redundant or unused data. Any inconsistencies between the total number of DICOM files and NIFTI slices are reported in .txt files for each cancer type and data provider, indicating potential issues during the annotation process. This information is flagged for further investigation by the healthcare professionals. Files that cannot be effectively examined due to missing header information or invalid dimensions are reported for future re-upload. Finally, the curation script has the functionality to identify cases where the data were not passed through the Quality Check.

Overall, the described, improved and developed tools, new guidelines and new quality control standards have been applied during the data collection phase in the observational study. As data was uploaded gradually, upon patient recruitment, the curation process has been designed as follows (Figure 1):

1. Data collection guidelines were used to instruct the data providers in each step.
2. Data Integration quality check tool has been used to evaluate adherence to clinical data collection, time alignment between clinical data and images and an adequate renaming of the image folders to reflect the time point of on the patient journey.
3. Data upload to the predefined temporary storage where the data are checked by the data curation script and error report is generated.

4. Error correction process is initiated and data providers re-upload the corrected data.
5. The corrected data is transferred to the INCISIVE database.

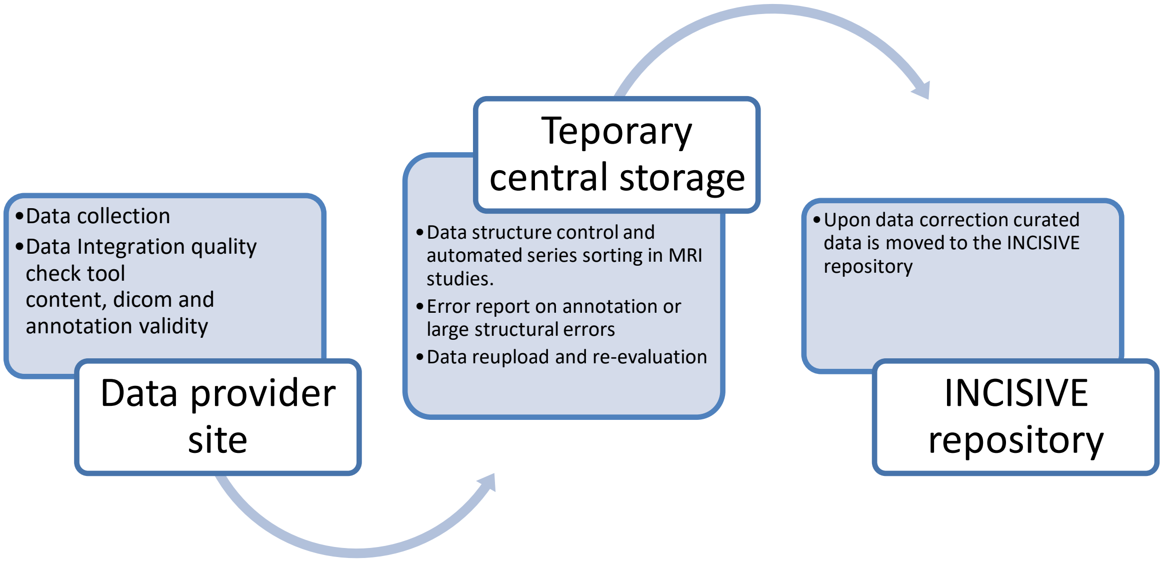

•Data collection
•Data Integration quality check tool content, dicom and annotation validity
Data provider site
Teporary central storage
•Data structure control and automated series sorting in MRI studies.
•Error report on annotation or large structural errors
•Data reupload and re-evaluation
•Upon data correction curated data is moved to the INCISIVE repository
INCISIVE repository

Figure 1. Data quality control and curation process improved and applied during the INCISIVE observational study

3 Performance evaluation of AI toolkit

AI toolkit has foreseen the services for four cancer types: lung, breast, prostate and colorectal cancer. All services have been selected by the HCPs, experts from the participating DPs who took engagement in the AI workshops organized within WP4 in order to identify and prioritize the needs for each cancer type. Pace of data upload, availability of data types and availability of open data have shaped the final service selection, as described in D4.3.

Based on the Data collection reports (M30 Management report), observational data collection summarized in Table 2 to Table 5 and previous deliverables on INCISIVE AI toolbox, the data collection and availability of open data sets has been most advanced for the breast and lung cancer models. Specifically, for **breast cancer** the majority of uploaded images were baseline diagnostic mammography (MMG) images. Interestingly, almost no ultrasound images have been uploaded and small number of MRI images. For these reasons, for breast cancer services, models trained on MMG images have been the most rigorously tested. The performance is realistic, and pitfalls of evaluation on selected data sets of better quality has been elaborated. Since many different DPs combined with multiple open data sets were used, MMG lesion detection and segmentation service provide state-of-the-art results on the open data sets. MMG detection and segmentation services can be considered mature, though further improvements using learning from both projections and exploration of novel segmentation architectures remains to be explored. MMG BIRADS and Breast density classification services experienced decrease in performance on the observational evaluation set when compared to development evaluation results, thus for these models' additional level of robustness is needed in terms of harmonization of image appearance.

The **lung cancer** services focus on chest Xray and CT scans. *Chest x-ray classification* model was comprehensively evaluated on a large test dataset of images. It works well when applied on images processed using a specific pre-processing (same as that found in the INCISIVE dataset from provider VISARIS). On the other INCISIVE DPs datasets, the performance varies, when the data is balanced, the performance is better, while if the model is tested only with cancer positive cases (unbalanced) the performance is worsening. Chest x-ray classification model is not mature enough for wider and more general applications where input images are not of controlled appearance. The evaluation indicated that the model needs improvements related to image harmonization at the model input, and use of data from different sources in the training phase. *CT scan prioritization service* (assigns low/high priority to patients) show high maturity and stable performance both the original test set used in the development stage, and at the prospective data of different INCISIVE DPs. However, lesion detection and segmentation service show decreased performance on the observational data set, which additionally varies among DPs (i.e.

different data sources) implicating that improvements may be achieved if model retraining is done with more heterogeneous image sources. *Cancer staging for lung CT scans* has achieved improved results compared with those reported in D4.3, mainly in terms of specificity, that now predicts correctly all low stage cases. Sensitivity is also improved, up to 70%, but some stage 3 and 4 cases are missed as the model is less focused on the nodules and metastasis dimensions of staging. *Metastasis risk prediction model* shows performance higher than the one observed using the dataset from D4.3, and this can be attributed to the fact that further improvements were made on the model parameters and due to the fact that more data were used during the testing process. However, the fact that more data from class 1 (metastasis group) were added in the dataset, (only 4 cases were considered as having metastasis during the 2 years period) lead to a more specific model, which means that the model is able to predict the metastasis more accurately (from 70% to 87.5% specificity). Cancer staging and Metastasis risk models have been evaluated on low, available, number of cases, and their maturity cannot be easily assessed as they require more complete patient information which is not easily found.

Prostate cancer services rely on MRI images and provide: prostate gland segmentation, prostate lesion segmentation and ISUP score classification. *Prostate gland segmentation model* has been developed and tested using external open data sets, as this type of annotation has not been done within INCISIVE, yet the service proved useful as a prerequisite for both lesion segmentation and ISUP score classification. This model exhibits high maturity and stable performance. *Prostate lesion segmentation model* provides for patient prioritization, lesion localization and lesion segmentation. In overall it exhibits lower performance on the new, observational data set, despite the sensitivity in the lesion localization mode is still very high. Finally, the ISUP score classification model, due to the absence of the prostate gland segmentation file in within INCISIVE data, was trained and tested using a large external dataset, demonstrating good performance (79% accuracy). The model is vendor-agnostic, as it was evaluated on images acquired from two different MRI vendors. Additional efforts were made to enhance its performance by analysing various areas of the prostate gland. However, this approach did not yield the expected improvements. The evaluation of the model in the feasibility study and beyond will be conducted on INCISIVE data, while the expected mask will be generated by the INCISIVE Prostate Gland Segmentation model mentioned above.

Colorectal cancer services are focused on MRI images and histopathological images. Since *MRI service for lesion segmentation* in colorectal cancer has been developed on T2W sagittal view MRI sequences, it could not be evaluated since no MRI sequences of this type have been available. *Histopathology image segmentation service*, H&E Image Processing (HEIP), proves similar in performance compared with other instance segmentation methods. This indicates that HEIP is a reliable tool for the segmentation and annotation of different cancer cell types, including colorectal cancer. It is particularly well-suited for identifying and distinguishing lymphocyte and epithelial cells in colorectal cancer. This model shows high maturity. *Survival Rate prediction model* is based on the HEIP, which extracts features from histopathological images that are used in survival prediction. However, this model exhibited poor performance as the limited data and high processing requirements prevented more thorough training and improvement.

3.1 Lung cancer

3.1.1 Lung cancer final AI service overview

Figure 2 presents the overview of the integrated AI services for lung cancer tailored for X-rays and CT scans.

Lung cancer services:

1. **Chest X-rays classification** – offers 2 level classification
   a) if any suspicious finding is present in the image
   b) if the finding is oncological
   c) explainable AI (XAI) service provides insight into models' decision
2. **CT scan lesion segmentation**
   a) prioritization service – classifies CT images based on the presence of lesions
   b) localization service – detects the slices with the presence of lesions
   c) segmentation service – segments the lesion in each located slice
3. **Cancer Staging** – classification model enabling binary classification of CT scans into category C0 (comprising I and II subtypes) and C1 (comprising III and IV subtypes). The service uses patient age and lesion segmentation service to extract radiomics features of both the tumor and the lung lobes. The additional explainable AI (XAI) service based on LIME and Anchors provides insight into models' decision.
4. **Metastatic risk prediction (2 years)** - model predicts the likelihood of metastasis within a two-year period. It utilizes CT images and clinical characteristics, including the patient's age, EGFR and KRAS mutations, the time elapsed between the initial diagnosis and the date of surgery. This service relies on radiomics analysis of both the tumor and its surrounding area. Explainable AI (XAI) service based on LIME and Anchors provides insight into models' decision

Lung Cancer pipelines:

1. **Chest X ray pipeline** – uses full chest X ray service classification service, for which in case some findings are present provides the likelihood of suspicious findings, likelihood of oncological findings and an XAI based explanation of these decision.
2. **CT scan pipeline** - This pipeline is a serial connection of CT lesion segmentation service, binary Cancer staging and Metastasis risk prediction, both with XAI based explanations of decisions based on LIME and Anchors

Medical report - This service combines the predictions of the AI models with additional spatial analysis and outputs a diagnostic report at the end of each pipeline

For the detailed description and methodological foundations of AI models the reader is referred to the deliverable D4.3 “INCISIVE AI-toolbox, data analytics and user services – Final Version”

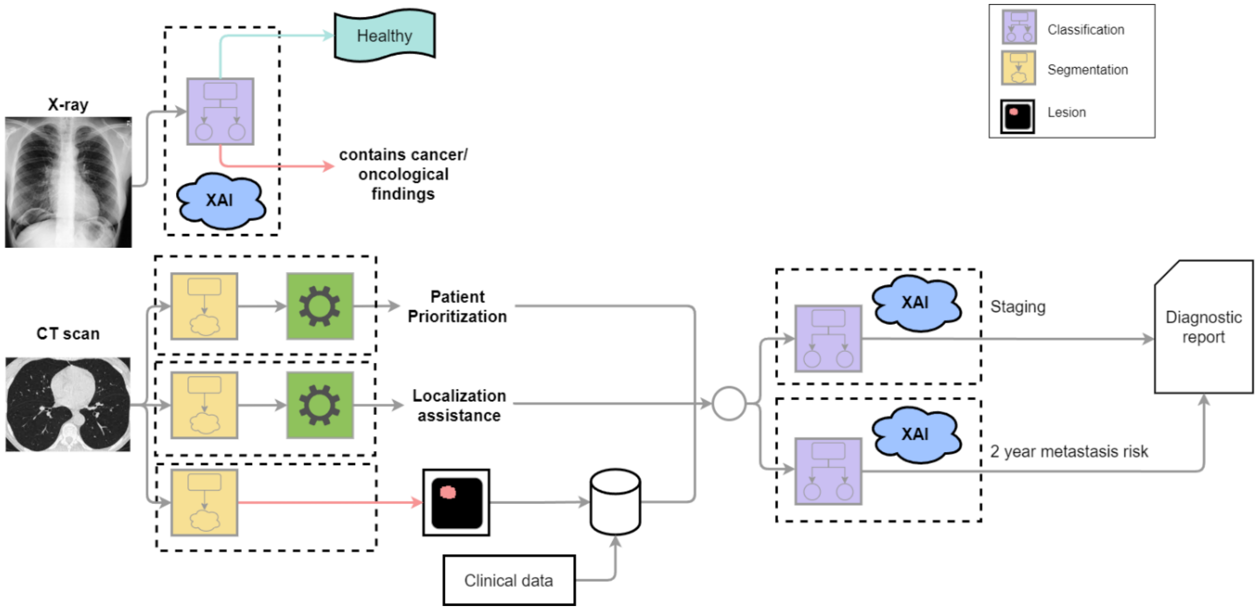


Figure 2: Overview of INCISIVE Lung Cancer AI Services

3.1.2 Chest X-rays classification

Chest classification service performance was evaluated on newly collected chest x-ray data. This new data is described in the following section.

3.1.2.1 Evaluation Chest X-ray Data

Evaluation of the trained models is performed on the collected observational data, on retrospective data not used in the development of the final models reported in D4.3 and on an external source of chest x-ray data not used in the training process. The external source data and a subset of retrospectively collected data uploaded to the INCISIVE platform after M32 has also been used in this evaluation process to improve the accuracy of the evaluation of generalization (prediction) errors made by the final models. Table 6 presents the count of the collected number of patients who have available chest x-ray examinations (studies) across all data partners providing this data under the INCISIVE project.

The data includes both diagnostic (i.e., baseline) and, where available follow-up DX examinations, during treatment or after surgery. Even though these studies are not relevant to the classification service (the patients are known to have a positive diagnosis), due to the nature of data collection it is not possible to exclude these studies from the test sets used in the evaluation described in this chapter.

Partner/Source	Number of annotated patients	Number of annotated studies
AUTH	35	35
DISBA	7	7
UOA	48	48
VIS	345	345
TOTAL	**435**	**435**

Table 6.: Number of curated INCISIVE CX studies and images available for evaluation.

As with the initial image data set used for model creation, the Chest Xray images were annotated with classifications of normal/abnormal finding and the nature of the finding if it is oncological. These annotations can be used to form a binary classification problem for a prioritization service.

AUTH data consists of 35 images of which all are with positive oncological (cancer) findings. DISBA data likewise contains only positive cancer findings, on 7 images and UOA data contains 48 images of which 34 have oncological findings and 14 are normal (healthy) findings. Visaris evaluation dataset consists of 345 chest x-rays. 222 images/studies have abnormal findings of which 68 are oncological and 154 are other pathologies. Remaining 123 images/studies show healthy patients. Several samples of images showing the nature of post-processing included in the images are illustrated in Figure 3.

Additionally, the generalisation ability of the model is tested on the external NIH chest x-ray dataset. This dataset consists of 112120 chest x-rays (Table 7). Of all images, 28,595 contain oncological findings while 23,164 contain examples of other pathologies. The remaining 60,361 images represent normal, healthy findings. A more detailed description of the dataset can be found at https://www.kaggle.com/datasets/nih-chest-xrays/data. Several sample images from this dataset are illustrated in Figure 4 and show a significant difference compared to the VIS collected dataset illustrated in Figure 3 which means they represent a significant challenge to the generalisation ability of the trained models.

Source	Number of annotated studies
INCISIVE	435
External	112,120
TOTAL	**112,555**

Table 7.: Number of curated INCISIVE and External CX studies and images available for evaluation.

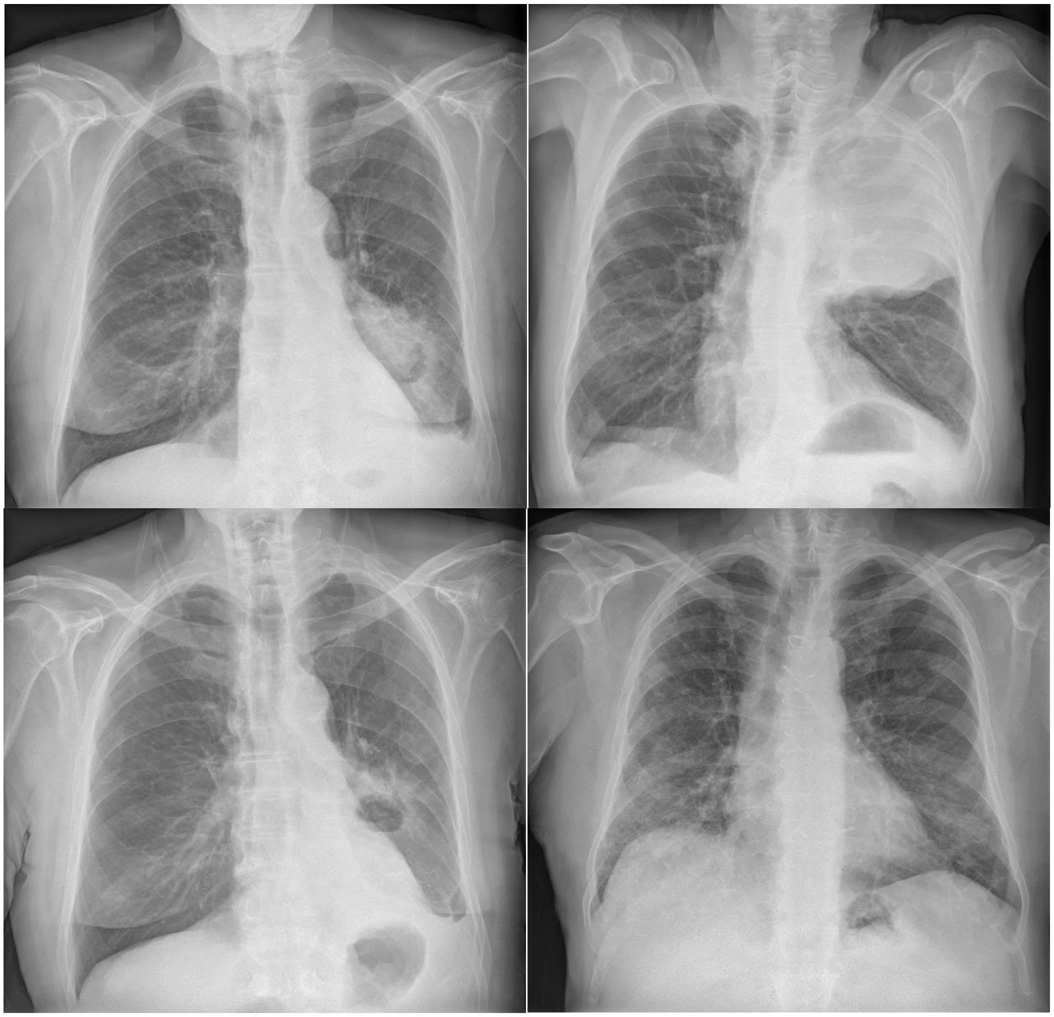


Figure 3. Example chest X-ray images from VIS evaluation dataset

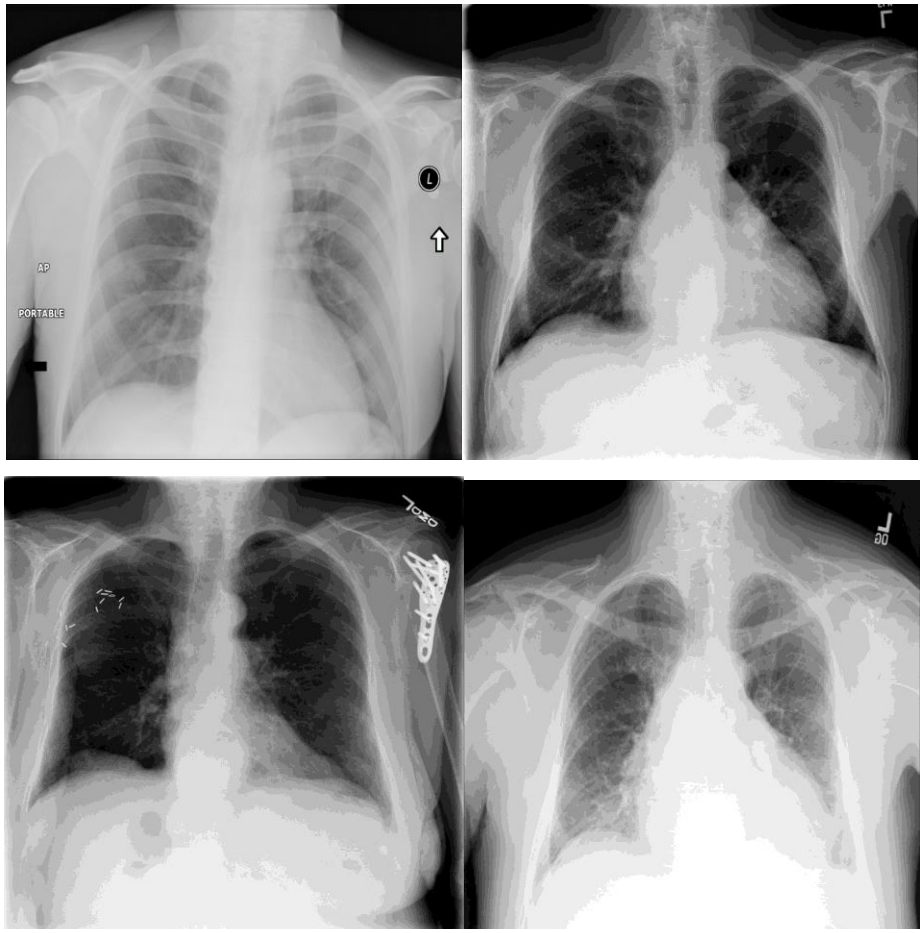


Figure 4. Examples from NIH evaluation dataset

What is significant of note in the NIH dataset is that many images contain digital overlay annotations (see image on the top left in Figure 4) and other objects which overlap with the chest region in the images. This is a significant confusion factor for any model analysing these images, particularly as no such effects were present in the VIS data used for training.

3.1.2.2 Evaluation of Chest X-ray Classification Model

The trained model being evaluated is based on the EfficientNetV2 with 512x512 grayscale image input size. Backbone is pretrained on ImageNet dataset, and then fine-tuned on VIS training data with task to classify cancer vs non-cancer images (healthy + other pathological condition).

For the datasets we applied the following performance measures:

- True positive (TP) rate or Sensitivity and true negative (TN) rate or Specificity, expressing the probability that positive findings (cancer) are correctly identified as such and probability that normal or other findings (non-cancer) are correctly identified as such, respectively. Ideal rates are 100%.

- False positive (FP) rate and false negative (FN) rate expressing the probability that negative findings (non-cancer: healthy or other findings) are incorrectly identified as positive (cancer) and probability that positive (cancer) findings are missed - incorrectly identified as such, respectively. Ideal rates are 0%.

- Area Under the Curve (AUC) applied to the ROC curve plot obtained by varying the classification threshold from one to the other extreme value, low to high. Ideal AUC score is 1 which means all the images were correctly classified, while a score of 0.5 signifies a random selection by the model.

- Equal Error Rate (EER) as the operating point (classification threshold value) at which false positive and false negative rates are the same. An ideal score for EER is 0, meaning ideal classification performance while an EER of 0.5 means a random selection.

- The F1 score is the harmonic mean of precision and recall and symmetrically represents both precision and recall. Ideal F1 score is 1.0, indicating perfect precision and recall, and the lowest possible value is 0, if either precision or recall are zero.

Due to their small numbers the data from other data providers, except VIS, were combined in the evaluation stage. Performance of the trained models on this data (AUTH+DISBA+UOA) provides a decent level of performance illustrated on the ROC curve in the Figure 5. Looking at a more balanced UOA data, that contains both positive and negative examples, performance is better, see ROC curve in Figure 6. Achieved AUC is 0.82 with EER of 0.22, suggesting that the other two heavily positive dataset skew the result somewhat. Relatively coarse evaluation on this data is evident in the blocky nature of the ROC curves shown in the plots.

Using only VIS data set, ROC curve presented in Figure 7, achieved Area-Under-the-Curve (AUC) measure is 0.79 with an Equal Error Rate (EER) of 0.29 and F1, at threshold of 0.5, score of 0.33. At this threshold Sensitivity is relatively low 0.2 but specificity is a perfect 1. At a more realistic threshold of 0.05 the F1 value increases to 0.73 with Sensitivity increasing to 0.59 at a decrease in Specificity to 0.86, a more realistic operating point.

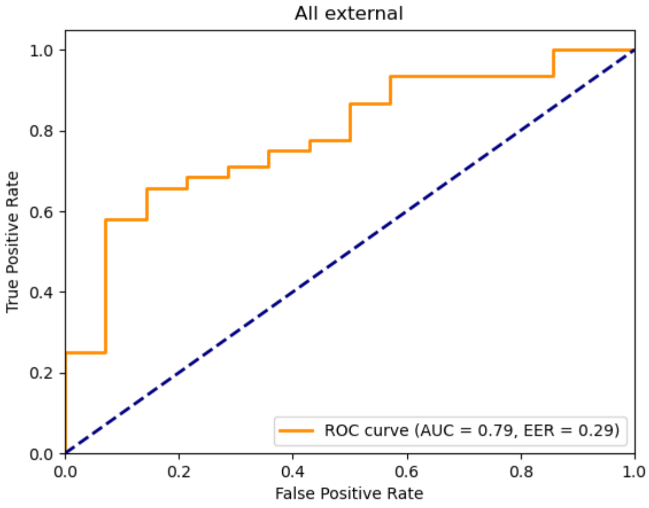


Figure 5: ROC plot for the Chest X-ray classification model on non VIS (AUTH+DISBA+UOA) INCISIVE data

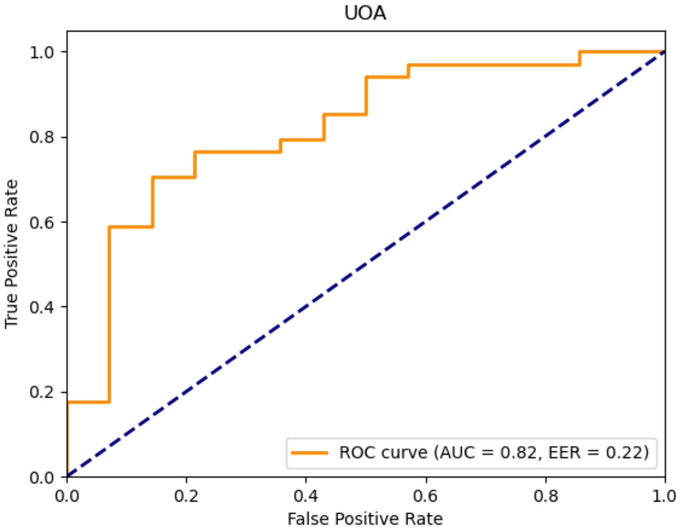


Figure 6: ROC plot of Chest X-ray classification model on the more balanced UOA dataset

Performance of the trained models is exceptional on VIS evaluation dataset images, as illustrated on the ROC curve in the Figure 7. Achieved Area-Under-the-Curve (AUC) measure is 0.96 with an Equal Error Rate (EER) of 0.09 and F1, at threshold of 0.5, score of 0.81. True positive rate of over 94% is a significant result, at a relatively low false positive rate of 9.4%.

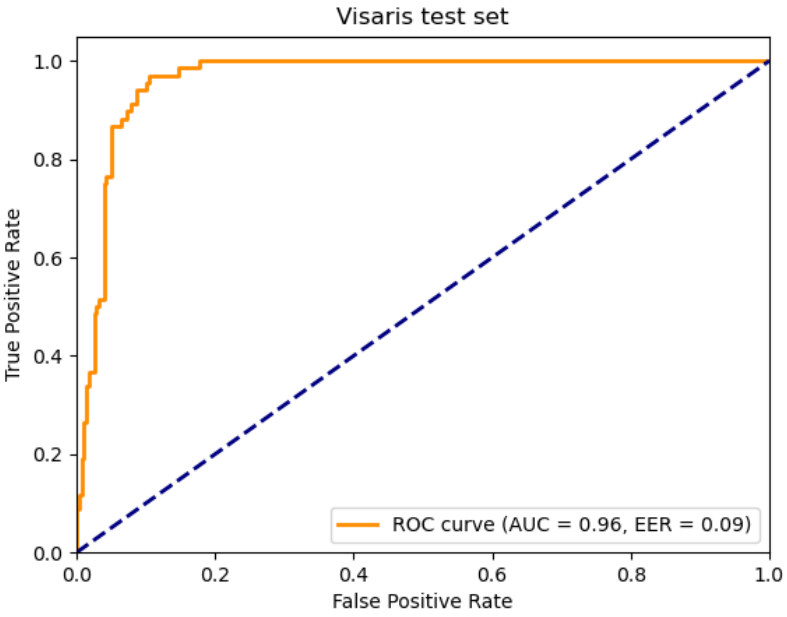


Figure 7. ROC curve of the model calculated on Visaris internal evaluation dataset

On the NIH dataset our trained model achieves AUC of 0.63, EER of 0.4 and F1 score (at threshold of 0.5) of 0.117. ROC curve calculated on the NIH dataset is illustrated in Figure 8. Note that difference in performance between 2 datasets can be explained by different appearance of Visaris and NIH dataset (different preprocessing). These results indicate a relatively low level of model's generalisation ability.

It is worth noting that the model had no exposure to the type of data contained in the NIH dataset or any other dataset during the training procedure. It is thus not unusual for such results to be achieved. Additionally, the evaluation test set contains a significantly higher number of images than the training set which was ~10,000 images.

Table 8 summarises the performance measures of the trained model. True positive rate (Sensitivity) and True negative rate (Specificity) for the Visaris and NIH datasets are quoted for the threshold value of 0.5 which is clearly not a suitable level in this case, as it results in an unacceptably low true positive rate, sensitivity while keeping false positives very low. In a more realistic application, the threshold would be decreased to favour sensitivity over false positive rate. A more realistic, lower threshold is used on the other INCISIVE data and UOA balanced datasets. Here it is obvious that false positive rates increase but so does Sensitivity which is the more important performance parameter.

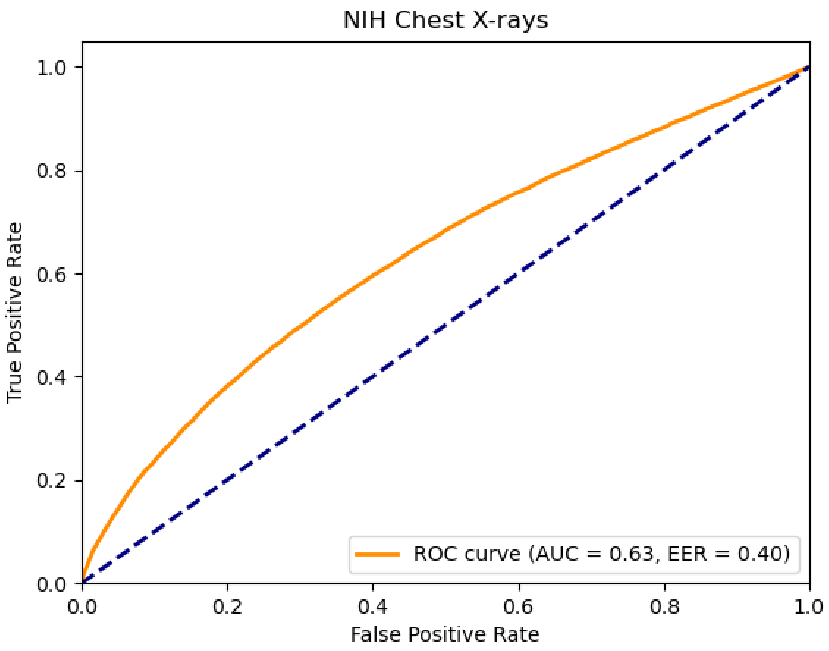


Figure 8: ROC curve of the model calculated on NIH dataset

Dataset Performance Measure	VIS (INCISIVE) Threshold = 0.5	AUTH+DISBA+UOA (INCISIVE) Threshold = 0.05	UOA (INCISIVE) Thr. = 0.05	NIH (External) Threshold = 0.5
True Positive Rate (Sensitivity)	94.1 %	59.2 %	61.8 %	6.5%
True Negative Rate (Specificity)	90.6 %	85.7 %	85.7 %	98.3%
False Positive Rate	9.4 %	14.3 %	14.3 %	1.7%

Dataset Performance Measure	VIS (INCISIVE) Threshold = 0.5	AUTH+DISBA+UOA (INCISIVE) Threshold = 0.05	UOA (INCISIVE) Thr. = 0.05	NIH (External) Threshold = 0.5
False Negative Rate	5.9 %	40.8 %	38.2 %	93.5%
AUC	0.96	0.79	0.82	0.63
EER	0.09	0.29	0.22	0.4
F1	0.81	0.73	0.74	0.117

Table 8: Summary of performance measures of the trained chest X-ray classification model on different evaluation datasets

3.1.2.3 Model Assessment and Maturity

The Visaris chest x-ray classification model was comprehensively evaluated on a large test dataset of images.

The results presented above show that the model is relatively well trained and can be applied to a specific narrowly defined set of images, processed using a specific pre-processing (same as that found in the VIS dataset). On the other INCISIVE datasets, the performance is varied. On the balanced UOA data, containing both positive and negative examples a decent level of performance can be reached. On the heavily positive sets from other data providers performance is lower.

The model however is not mature enough for wider and more general applications where input images are not of controlled appearance, where it provides a much lower level of performance. It appears that the model is overfit to the specific type of data. This can be resolved, and model performance improved by either:

- Retraining the model using additional data from other datasets in its training set to increase its exposure to other image appearances. This may include an increase in the size of the training dataset from ~10,000 currently to closer to 100,000, or

- Developing a pre-processing step that will harmonise the appearance of images originating from different devices and sources to a specific appearance.

3.1.3 CT scan lesion segmentation

This service performs segmentation of lesions in lung CT scans. A detailed description of the model design can be found in D4.3. Model training and validation was performed on INCISIVE data from 1 data provider (UoA). In this deliverable, we extend the validation of the model on new data, which come either from the same data provider or from a different one (AUTH).

Table 9 reports the number of samples (patient examinations) that were utilized to further validate the model.

Data Provider Name	Number of patients	Number of cases (examinations)
UoA	27	37
AUTH	57	140

Table 9. Number of patients and cases (CT examinations) used to further evaluate the model.

First, we evaluate the prioritization service, which assigns a low/high priority to the examination depending on whether any suspicious area was detected.

Dataset	F1-Score [%]
D4.3 (UoA)	100
UoA	100
AUTH	100

Table 10. Evaluation performance of lung CT scan prioritization service

As can be seen from Table 10, the performance of the prioritization service is very high and there are no false negative predictions.

Next, we evaluate the localization assistance service, which produces a list of the frames where a lesion was segmented. The results can be seen in Table 11.

Dataset	Sensitivity [%]	Specificity [%]
D4.3 (UoA)	64.7	94.7

UoA	50.6	93.5
AUTH	45.5	94.4

Table 11. Evaluation performance of lung CT scan localization service

Finally, we evaluate the lesion segmentation service on a pixel-level. The results can be found in Table 12.

Dataset	F1-Score [%]
D4.3 (UoA)	66.2
UoA	51.8
AUTH	39.9

Table 12. Evaluation performance of lung CT scan segmentation service

Overall, we observe that the performance of the services is lower on prospective data compared to the original test set, which the exception of the prioritization service. A possible explanation for this drop would be that there are inherent differences between these datasets and a more robust harmonization procedure is required. Therefore, the prioritization service showcases the highest maturity, while segmentation and localization assistance require improvements to increase their robustness before being used in a clinical setting.

3.1.4 Cancer Staging

The model was trained with TCIA open data, as previously described in D4.3. The INCISIVE data available (from all the data collection phases) were considered for testing, since they were not used in training.

In order to be used for testing, the data had to include a) CT images series (not FUSCT) at baseline for lung cancer, b) a tumor annotation in the folder (manual annotation at this stage), c) a lung mask available, as the output of the successful execution of the lung organ mask calculation within the pipeline , d) radiomics features available for the tumor and the lung lobes as the output of their calculation within the pipeline, e) accompanying clinical data, age and stage, f) accompanying vendor data, ie manufacturer, to be used for harmonisation.

After applying these criteria, in total 55 CT series in baseline were included in the test set, 12 of which belong to Class 0 (stage I or II) and 43 to class 1 (stage III or IV). Table 13 provides an overview of the data used for testing.

PROVIDER	retrospective	prospective	observational
AUTH	19	18	
UoA	12	4	3
Total:	31	22	3

Table 13. Number of annotated CTs per provider available for evaluation of the prediction model

Here, we need to highlight that the testing dataset was enriched compared with the one used in D4.3, where only 24 cases from one data provider were included (19 retrospective cases and 4 from the prospective study). In addition, data from more than one data provider were included, and thus for the harmonization both the vendor and the clinical site must be considered.

The final model was trained and finetuned on the whole TCIA dataset and tested with the INCISIVE data mentioned above. Changes include a modification in the normalisation of the *TuNo* features, now normalised to own lobe rather than average of two lobes, and a consequent slight modification of the selected features, small changes in the harmonisation process, and the thorough hyperparameter tuning of the final model. The performance results are depicted in Table 14.

	Accuracy [%]	Sensitivity [%]	Specificity [%]	Precision	F1	BA [%]
D4.3 (part of AUTH)	66.7 95% CI : [44.7 84.4]	62.5	75	83	62.5	68.8
INCISIVE data (AUTH+UOA)	76.4 95% CI : [62.9, 86.7]	69.77	100		82.19	84.88
INCISIVE data (docker integrated model)	73	65	100	100	79	83

Table 14. Evaluation performance of lung CT scan localization service. The positive class is the 'stage 3 and 4' class.

The docker integrated version of the staging AI service, currently utilizes an older version of the model. The updating of the model is an ongoing process, with the latest results being primarily in R. However, as our results presented in Table 14 indicate, when it is transferred to Python, the performance metrics remain close to those of the R version. Specifically, the differences observed are in accuracy, sensitivity and F1 score, where the R version slightly excels over the Python one. Thus, when the updated model is finished, the transition to the Python implementation and update of the docker version will be implemented without significant issues.

As can be seen in Table 14, the results are improved compared with those reported in D4.3, mainly in terms of specificity, that now predicts correctly all low stage cases. Sensitivity is also improved, up to 70%, but some stage 3 and 4 cases are missed. According to the feedback received by clinical experts, this was due to the fact that the model is more focused on the tumor information, and less focused on the nodules and metastasis dimensions of staging.

A small-scale analysis of errors took place. The characteristics of incorrectly classified patients are presented in Table 15. More details regarding the age distribution and the stage of the correctly and incorrectly classified patients are provided in Table 16 and Table 17, respectively.

Gender	Stage	Site	Age (65 yrs threshold)
8 male 21.05% of males	IIIA: 6, IIIB: 1, IVA:5, IVB : 1	site 1 (AUTH): 9 or 25% of site 1 cases	Elder: 9,
5 female **29.4% of the females**		site 2 (UoA): 4 , and 21.05% of site 2 cases	Middle-aged 4

Table 15 Patient characteristics for incorrectly classified cases.

Age statistics	Min	1st Qu.	Median	Mean	3rd Qu.	Max
correctly classified	43.00	55.50	64.00	63.88	71.75	82.00
Error cases	54.00	63.00	72.00	68.77	75.00	81.00

Table 16. Age distribution of correctly and incorrectly classified patients

Stage statistics	IA1	IA2	IA3	IB	IIB	IIIA	IIIB	IVA	IVB
Correct	1	4	4	2	1	3	9	11	7
Errors						6	1	5	1

Table 17. Patient stage of incorrectly vs correctly classified cases

With respect to manufacturers, the distribution of errors shows two prevailing manufacturers, but further relations between error and harmonisation problems could be investigated Table 18.

	GE	Siemens	Toshiba	Philips
Correct	10	22	3	7
Error	5	6	1	1

Table 18 Distribution of errors with manufacturers

A small-scale analysis was performed regarding fairness, with respect to age and gender.

In the training set, the distribution of ages is:

- Elder: 327
- Middle aged: 169

The gender distribution is:

- Female: 144
- Male: 352

Finally, the following stages are represented in the training set:

- I: 78
- IA1: 31
- IA2: 18
- IB: 25
- II: 36
- IIA: 4
- IIB: 16
- IIIA: 120
- IIIB: 165
- IVA: 3

Which are also summarized as:

- I: 152
- II: 56
- III/IV: 288

In the training set, the middle-aged and the females seemed to have a slightly higher accuracy, but no important model bias was found as regards age and gender, as indicatively shown in the Figure 9. A more extended analysis, and mitigation if needed, is currently ongoing.

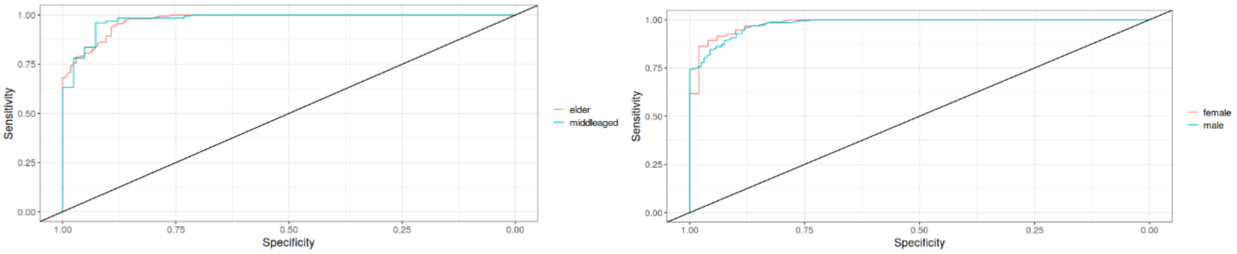


Figure 9. ROC curves per subgroup, (left) age subgroups, (right) gender subgroups

If a disparate_impact_remover is applied as a preprocessing step, age bias metrics are improved (Figure 10), especially the statistical parity, with a small decrease in the performance metrics, i.e. accuracy: 72.73, sensitivity: 67.44, specificity: 91.67.

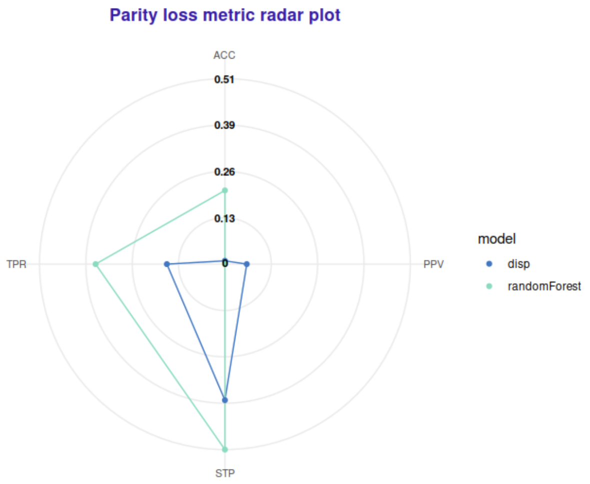


Figure 10 Radar plot comparing the various parity loss metrics in the initial model and the model using the data after preprocessing with disparate_impact_remover

3.1.5 Metastasis risk prediction

The metastasis prediction model was trained using open databases and in particular the NSCLC Radiogenomics database. For the initial testing of the model as reported in D4.3, the INCISIVE data that were available until July 2023 were used. The results that are reported in the current deliverable are based on the data that are available in the INCISIVE repository until October 2023. In addition, further improvements to the model parameters were made in order to improve performance. Table 19 provides more information regarding the data used in each phase of metastasis risk prediction model.

PROVIDER	Model training	Model testing (D4.3)	Model testing (D6.5)
Open data	111	0	0
INCISIVE data	0	14	23
Total:	111	14	23

Table 19 Number of cases used for each phase of the model development.

As also reported in D4.3, in order for the data to be eligible for model testing, specific requirements need to be met. In more details, the data for each case must include:

1. CT examination in baseline
2. Segmentation of the primary lung tumor in the CT baseline examination
3. Staging information in baseline (either TNM classification or overall staging)
4. The patient must not have metastasis in baseline, and
5. There should exist information related to the metastasis status in a 24month period after the baseline.

We recognized that the last requirement might limit the number of cases based on the information that is already available in the INCISIVE repository, and thus we asked specific data providers to provide additional information in case a requirement could not be met. Table 20 provides more information related to the data identified for each data provider and the data that were finally used.

The characteristics of the patients used for the evaluation are provided in Table 21.

The performance of the model based on the data that is reported above, is shown in Table 22. In that table, the respective performance metrics for each evaluation phase are also provided.

As mentioned in the previous section, the docker integrated version of the metastasis AI service, currently utilizes an older version of the model (as described in D4.3) and in the table, the evaluation was based on the data used from both data providers. In D6.5, the updated metastasis model that is implemented in R will be transformed in python and it will be integrated into the AI toolbox.

PROVIDER	Requirement 1+2	Requirement 1-3	Requirement 1-4	Requirement 1-5
AUTH	32	32	25	19
DISBA	19	0	0	0
UoA	34	22	11	4
Total:	88	54	36	23

Table 20 number of cases which fulfil the requirements set for the inclusion in the evaluation phase for the metastasis risk prediction model.

	stage	metastasis	Months of metastasis
AUTH	10 stage A, 9 stage B	5	12.2 ±2.8
UoA	1 stage A, 3 stage B	3	12.7±3.5
Total	11 stage A, 12 stage B	8	13.4±2.8

Table 21 Characteristics of the patients used during the evaluation phase. Class I refers to overall stage A or II while stage B for overall stage III or IV.

	Accuracy	Sensitivity	Specificity	Precision	F1	BA
D4.3 AUTH	78.6	75	70	88.9	82.4	77.5
D6.5 AUTH+UoA data	82,6 95% CI [61.2 95.1]	80	87.5	92.3	85.7	83.8
Docker Integrated model AUTH + UoA	78.3	75	80	66.67	70.6	77.5

Table 22 Performance metrics for the metastasis risk model using the data included in the previous evaluation phase (D4.3) and the current phase (D6.5].

As shown in Table 22 the performance is higher than the one observed using the dataset from D4.3, and this can be attributed from the fact that further improvements were made on the model parameters and but also to fact that more data were used during the testing process. However, the fact that more data from class 1 (metastasis group) were added in the dataset, (only 4 cases were considered as having metastasis during the 2 years period) lead to a more specific model, which means that the model is able to predict the metastasis more accurately (from 70% to 87.5% specificity). This finding can help the clinical experts to adapt the treatment plan accordingly.

3.2 Breast cancer

3.2.1 Breast cancer final AI service overview

The final breast cancer AI services integrated in the INCISIVE AI platform are presented in Figure 11. These services provide the support for two modalities mammography (MG) and magnetic resonance imaging (MRI).

Breast Cancer services:

1. **MMG prioritization service** - a classification service which aims to flag MMG image as high priority in cases when there is an indication of the presence of suspicious lesions. For each MMG image labelled as high priority, the service offers explanation as visualization of the breast tissue areas considered as suspicious.
2. **MMG localization service** - automatically detect and localizes suspicious lesions in MMG images. The service provides presentation of uploaded MMG images, visualization of its basic characterizes and localization of suspicious lesions marked with bounding boxes.
3. **MMG lesion segmentation service** - automatically localizes suspicious lesion and contours all the pixels in MMG images that constitute the lesion. The service provides presentation of uploaded MMG images, visualization of its basic characterizes and the input image with contours enclosing potentially suspicious lesions.
4. **BIRADS classifier** service - analyses a MMG image and predicts the BIRADS score. This service is complemented with an XAI explanability of model predictions based on GradCam.

5. **Breast Density classification service** - This service analyzes a MMG image and predicts if the breast is dense, informing whether the tissue requires an ultrasound follow-up. The classification is binary non dense (A, B) and dense (C, D). This service is complemented with an XAI explanability of model predictions based on GradCam.
6. **MRI lesion localization** – relies exclusively on the first post-contrast sequence to:
   a) mode prioritization - classification that indicates the presence of lesions in breast MRI.
   b) mode localization - localization of slices where the lesion is present.
   c) mode segmentation - performs bounding box detection of lesions on the localized slices.

Breast Cancer pipelines:

1. **Mammography pipeline -** This pipeline is a serial connection of MMG prioritization and if applicable MG lesion localization and segmentation, BIRADS classifier with XAI and Breast density classifier with XAI.
   Medical report - combines the predictions of the AI models with additional spatial analysis and outputs a diagnostic report at the end of the MG pipeline.

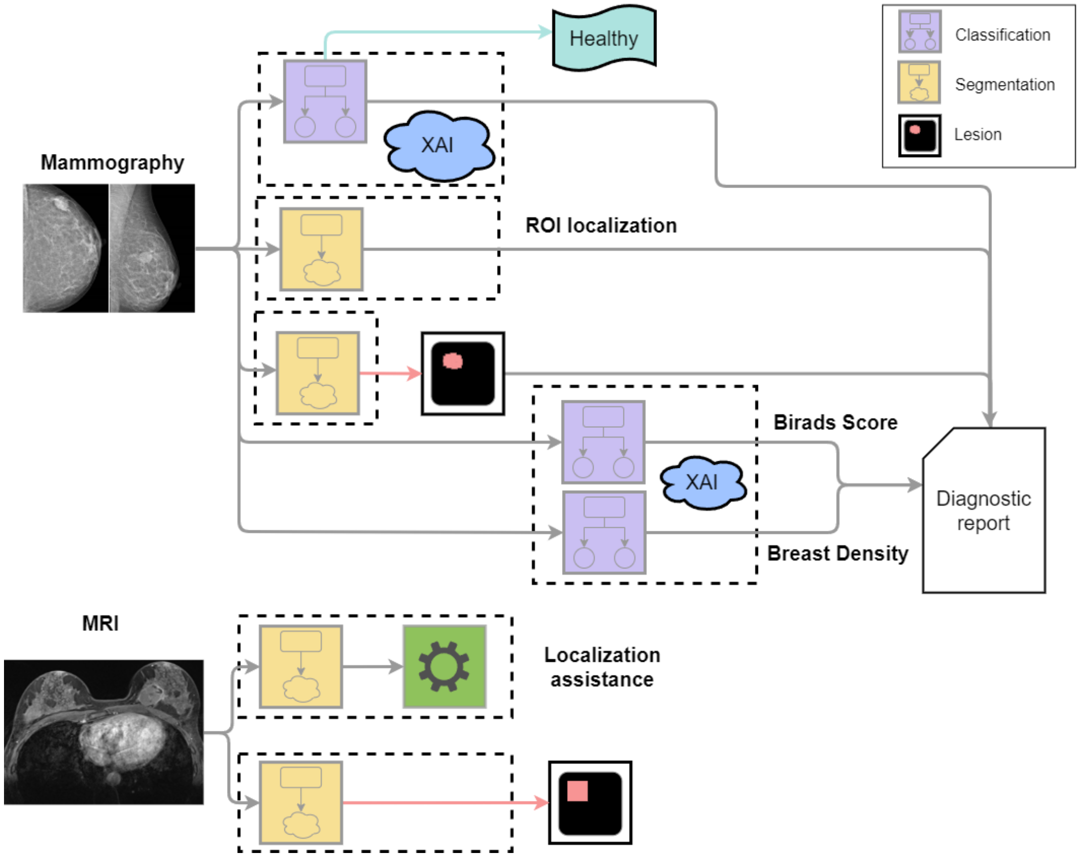


Figure 11: Overview of INCISIVE Breast Cancer AI Services

3.2.2 INCISIVE breast cancer repository images available for model evaluation

3.2.2.1 Mammography images (MMG)

Evaluation of the trained models is performed on the collected observational data, as well as on all retrospective data not used in the development of the final models reported in D4.3. Subset of retrospectively collected data uploaded after M32 has as well been used in this evaluation process in order to improve the estimates of the generalization (prediction) errors made by the final models. The set of images selected for evaluation was additionally careful inspected to avoid inclusion of specimen images, or images with breast implants. Table 23. presents the count of the collected number of patients who have available MMG examinations (studies) across the breast cancer data partners (AUTH, UNS and HCS).

MG		
Partner	Number of annotated patients	Number of annotated studies
AUTH	5	9
UNS	55	76
HCS	598	601
TOTAL	**658**	**686**

Table 23.: Number of curated INCISIVE MMG studies available for evaluation.

The data includes both diagnostic (i.e., baseline) and, where available and applicable, follow-up MMG examinations, during treatment or after surgery. The follow-up screenings can not be considered as representative results, given that the initial models did not learn from these cases, as they were initially trained only with baseline images. This is in accordance with the envisioned usage of these services, providing initial diagnostic information for a patient before the potential recovery journey. As with the initial set used for model creation, the MMG images were annotated with possible suspicious, malignant or benign lesions and those without. These annotations can be used to form a binary classification problem for a prioritization service, while the full annotations are utilized for the lesion localization and segmentation services in breast mammograms. The total number of individual MMG images, from patients reported in Table 23, is presented in Table 24, where the clear division between baseline (diagnostic) images and follow-up, time points images has been made. It is worth noting that the majority of images, 1436, are diagnostic images useful for the evaluation.

In order to evaluate the prioritization service for breast cancer mammography, we include information on the number of MMG images where the presence of any lesion (malignant/benign/suspicious) has been noted by medical professionals, and number of images without any documented and annotated lesions among the evaluation set of MMG images. This information will be used for binary classification in the prioritization process, where images with any lesion detected should be in the high priority class, whereas the other images should be classified as no-lesion images (low priority). Table 25 summarizes the number of images with and without lesions in the breast cancer mammography evaluation data set, both in baseline/diagnostic time point and in patient follow-up images.

PROVIDER	BASELINE	FOLLOW-UP
AUTH	13	8
HCS	1205	6
UNS	218	22
Total:	1436	36

Table 24. Number of annotated observational INCISIVE MMG images per provider available for evaluation of detection and segmentation models

PROVIDER	LESION		WITHOUT LESION	
	BASELINE	FOLLOW-UP	BASELINE	FOLLOW-UP
AUTH	13	7	0	1
HCS	940	6	265	0
UNS	153	12	65	10
Total:	1106	25	330	11

Table 25. Numbers of the individual MMG images with and without lesions in baseline and follow up available in mammography evaluation set

3.2.2.2 Magnetic resonance images (MRI)

For purpose of evaluation of the breast MRI lesion localization service we have extracted a subset of MRI images from the DUKE dataset (Saha, et al., 2018) (used for the model development). The randomly selected subset of ~9% of data, or in total 100 studies, has been withdrawn from the DUKE dataset (with total 922 MRI studies) before the model training/validation and testing for the final model development, as presented in D4.3. These 100 studies are thus an unseen observational test set for this service, preserved in case no breast MRI images are uploaded/or they do not contain the first post-contrast sequence used by the model.

3.2.3 Mammography normal/suspicious classification

The models developed for the prioritization service, which classifies mammographic images as normal (no lesions detected) or suspicious (lesion detected), have been designed with the aim of fast, automated processing of images obtained during daily clinical routine. All such models have been developed based on the YOLOv5 detection architectures of varying size. The training was performed on a combined dataset, using both annotated MMG images from open databases and annotated MMG images collected by data providers of the INCISIVE consortium. For computational reasons, during training, validation and testing, all images were proportionally reduced to 640 pixels in height.

Due to heterogeneity of the available dataset, two types of detection models were developed. In the first type, raw (downscaled and down-sampled) images were presented to the model in hope that the deep neural architecture will be able to automatically extract features regardless of the image content. In the second type, a custom-made pre-processing procedure was developed and applied to each image before it is presented to the model. The purpose of the pre-processing was to harmonize image appearance and level their dynamic range, ensuring it effective intensity range utilization across all the MMG images, and finally simplifying the task of feature extraction later performed by the backbone of the YOLO network. A detailed explanation of the network architecture, properties of the available datasets, various models obtained during training, and specifics of the pre-processing procedure were described with more details in D4.3.

As a reminder, the *model names* were selected in such a way that their important properties can be deduced from the name. Each model is named as **M**-**P**-**y**, where:

- **M** refers to the dataset used during training (**In** = just INCISIVE data, **InOD** = INCISIVE data + data from selected Open Datasets),
- **P** indicates whether pre-processing was used during training (**N** = No, **Y** = Yes), and
- **y** refers to YOLOv5 model size (**n** = nano, **s** = small).

Table 26 outlines classification performance of a selection of different characteristic models obtained during model development and presented in D4.3 (measured on the INCISIVE retrospective test set used for error evaluation upon the model development with the patients unseen by the models during training or validation). The Table 26 shows classification performance for three different characteristic confidence levels: 25%, 50%, and 75%. Confidence threshold of 25% (the one performing the best, and the one used for the deployment) implies that an image will be labelled as suspicious if at least one lesion is detected with confidence equal to 0.25 or higher. Choosing lower values for the confidence is reasonable not only because the performance indices are highest for that case, but also because false negative results are considered to be significantly more severe compared to errors of the false positive type. The performance indices used in D4.3 and reported in Table 26 are Precision (P), sensitivity (S) i.e. recall and F1 measure.

MODEL NAME	TEST SET PREPROCESSED	25%			50%			75%		
		P	S	F1	P	S	F1	P	S	F1
In-N-n	NO	0.66	0.76	0.71	0.90	0.41	0.56	1.00	0.06	0.12
	YES	0.58	0.81	0.67	0.80	0.45	0.57	0.92	0.07	0.14
In-N-s	NO	0.73	0.69	0.71	0.88	0.44	0.59	0.94	0.14	0.25
	YES	0.58	0.73	0.64	0.67	0.50	0.58	0.87	0.17	0.28
In-Y-n	NO	0.73	0.47	0.57	0.97	0.18	0.30	1.00	0.01	0.01
	YES	0.62	0.71	0.66	0.88	0.36	0.51	1.00	0.03	0.05
In-Y-s	NO	0.86	0.29	0.43	0.93	0.23	0.37	0.97	0.20	0.33
	YES	0.79	0.51	0.62	0.88	0.43	0.58	0.93	0.36	0.52
InOD-N-n	NO	0.76	0.61	0.68	0.86	0.32	0.46	0.93	0.04	0.08
	YES	0.59	0.77	0.67	0.79	0.41	0.54	0.95	0.06	0.11
InOD-N-s	NO	0.76	0.61	0.67	0.86	0.35	0.50	0.96	0.08	0.14
	YES	0.60	0.75	0.67	0.78	0.47	0.58	0.89	0.11	0.19
InOD -Y-n	NO	0.81	0.41	0.55	0.97	0.23	0.37	1.00	0.02	0.03
	YES	0.70	0.63	0.66	0.94	0.35	0.52	0.86	0.02	0.04
InOD-Y-s	NO	0.92	0.26	0.41	0.94	0.20	0.33	0.96	0.17	0.29
	YES	0.83	0.48	0.61	0.87	0.42	0.56	0.94	0.36	0.52

Table 26. Overview of a selection of different models and their performance on INCISIVE test MMG set measured for the normal/suspicious classification task important for the prioritization service. The models integrated in the INCISIVE platform are highlighted in yellow. Performance for three characteristic confidence thresholds (25%, 50%, and 75%) are shown.

Based on the performance during testing, two specific models were *selected for integration in the INCISIVE AI platform: In-N-n, and InOD-Y-n.* Both are nano-sized models, however the first was trained solely from the data obtained as part of the INCISIVE project, while the second one is also pre-trained on the data from the open MMG datasets. The other difference is that the first model is trained using raw data, and the second one uses pre-processing. It is important to stress that the same input strategy (without pre-processing for In-N-n and with pre-processing for InOD-Y-n) is used in the implementations on the INCISIVE AI platform. For reference, we show performance for all three threshold values, although only threshold 0.25 is used in the platform. Performance of these two models on the observational data is listed in Table 27, while Table 28 provides comparison of the results obtained on the D4.3 test set and on the INCISIVE observational dataset (for the confidence threshold 0.25).

Overall, a slightly better performance has been achieved using the observational data, compared to the test data used to evaluate performance in the previous phase of the INCISIVE project. However, the results are comparable which confirms stable generalization (performance aligned with the estimated errors) of the obtained models.

MODEL NAME	TEST SET PREPROCESSED	25%			50%			75%		
		P	S	F1	P	S	F1	P	S	F1
In-N-n	NO	0.85	0.67	0.75	0.95	0.30	0.46	1.00	0.06	0.11
InOD-Y-n	YES	0.89	0.60	0.72	0.98	0.28	0.44	1.00	0.02	0.05

Table 27. Performance of the two selected models when tested on the INCISIVE observational data. Confidence threshold of 25% is used in the deployment.

MODEL NAME	TEST SET PREPROCESSED	D4.3 TEST SET			OBSERVATIONAL SET		
		P	S	F1	P	S	F1
In-N-n	NO	0.66	0.76	0.71	0.85	0.67	0.75
InOD-Y-n	YES	0.70	0.63	0.66	0.89	0.60	0.72

Table 28. Comparison of the performance obtained in the development test phase (reported in D4.3) and on the observational dataset for confidence threshold value equal to 0.25.

Table 29 shows specifically the number of true and false positive, and true and false negative detection for the observational dataset. The total number of images in this dataset was 1469, of which 1121 images contained lesions annotated by a radiologist, while 347 correspond to healthy patients.

MODEL NAME	PREPROCESSED	SUBSET	25%								
			POS	NEG	TP	TN	FP	FN	P	S	F1
In-N-n	NO	ALL	1122	347	753	131	217	367	0.85	0.67	0.75
		AUTH	10	7	8	3	4	2	0.67	0.8	0.73
		UNS	165	75	139	42	33	26	0.81	0.84	0.82
		HCS	945	266	607	169	97	338	0.86	0.64	0.74
InOD-Y-n	YES	ALL	1122	347	647	264	83	447	0.89	0.60	0.72
		AUTH	10	7	9	6	1	1	0.9	0.9	0.9
		UNS	165	75	111	66	9	54	0.93	0.67	0.78
		HCS	945	266	555	191	74	391	0.88	0.59	0.70

Table 29. Performance evaluation of two integrated models: confusion matrices. The total number of images 1469 (out of which 347 are patients with no lesions reported)

3.2.4 Mammography Lesion Localization

The service for lesion localization uses the integrated model which is the same convolutional network models based on YOLOv5 architecture as in prioritization service, with the only difference being that localization service offers bounding box localization of the identified lesions.

As before, the performance was monitored and compared using two sets of measures. The first are standard pixel-wise scores, such as precision (P), sensitivity (S) and F1 (Dice) score. As already reported in D4.2 and D4.3, we found that in human-in-the-loop scenarios pixel wise metrics are too pessimistic. Namely, in many cases in which pixel level metrics were relatively low, the cases were positively evaluated by practicing radiologists, as despite imperfect overlap the localizations were targeting relevant regions. More realistic performance measure, as by radiologist standards, can be achieved by the object-based metrics proposed in D4.2 and D4.3: precision (SP), sensitivity (SS), and F1 score (SF1) on the segment level. For completeness, we repeat their definitions, which are:

$$SP = \frac{1}{R}\sum_{i=1}^{R} 1_{x\geq\theta_r}\left(\frac{A_{s_i}}{A_{r_i}}\right),$$

$$SS = \frac{1}{L}\sum_{i=1}^{L} 1_{x\geq\theta_l}\left(\frac{A_{s_i}}{A_{l_i}}\right),$$

$$SF = \frac{2\, SP \cdot SL}{SP + SL},$$

where R is the total number of image segments detected as lesions by the system, L is the total number of lesions in a set, A_{r_i} is area of i-th detected surface, A_{l_i} is area of i-th lesion, A_{s_i} is are of spotted part of lesion belonging to i-th detected segment or lesion, θ_i $(i = r, l)$ thresholds levels, and $1_{x\geq\theta}(x)$ indicator function.

For completeness, in Table 30 we repeat the performance of a selection of trained models obtained during training on the INCISIVE test set. Models selected for deployment are highlighted in yellow.

MODEL	TEST SET	PIXEL-WISE			OBJECT-WISE ($\theta_{r,l}$ =50%)		
	PREPROCESSED	P	S	F1	SP	SS	SF1
In-N-n	NO	0.554	0.353	0.431	0.572	0.511	0.540
	YES	0.354	0.436	0.391	0.443	0.530	0.483
In-N-s	NO	0.598	0.302	0.401	0.599	0.451	0.514
	YES	0.383	0.376	0.380	0.525	0.496	0.510
In-Y-n	NO	0.263	0.194	0.223	0.587	0.280	0.379

	YES	0.553	0.376	0.448	0.570	0.491	0.528
In-Y-s	NO	0.779	0.153	0.256	0.769	0.199	0.316
	YES	0.751	0.261	0.387	0.748	0.373	0.498
InOD-N-n	NO	0.621	0.288	0.394	0.641	0.403	0.495
	YES	0.407	0.409	0.408	0.523	0.525	0.524
InOD -N-s	NO	0.634	0.289	0.397	0.626	0.398	0.487
	YES	0.465	0.371	0.412	0.577	0.522	0.548
InOD -Y-n	NO	0.730	0.200	0.314	0.669	0.287	0.402
	YES	0.628	0.334	0.436	0.650	0.462	0.540
InOD -Y-s	NO	0.877	0.156	0.265	0.784	0.189	0.304
	YES	0.697	0.262	0.381	0.781	0.350	0.483

Table 30. Lesion detection performance obtained on the retrospective INCISIVE test set during model development in D4.3. Highlighted rows correspond to the models selected for integration in the INCISIVE AI platform.

Table 31 shows comparative performance of the deployed models on the retrospective INCISIVE test data set (used in D4.3) and observational data. During testing (and integration) the model **In-N-n** was presented with raw MMG images (no pre-processing), while the model **InOD-Y-n** was fed with pre-processed images. More specific results, granulated per individual providers are shown in Table 32.

MODEL	DATA SET	PIXEL-WISE			OBJECT-WISE ($\theta_{r,l}$ =50%)		
		P	R	F1	SP	SR	SF1
In-N-n	TEST SET (D4.3)	0.554	0.353	0.431	0.572	0.511	0.540
	OBSERVATIONAL DATASET	0.455	0.483	0.468	0.482	0.474	0.478
InOD -Y-n	TEST SET (D4.3)	0.628	0.334	0.436	0.650	0.462	0.540
	OBSERVATIONAL DATASET	0.593	0.41	0.485	0.618	0.44	0.514

Table 31. Performance of the models integrated in the INCISIVE AI platform, measured on the observational data.

MODEL	TEST SET PREPROCESSED	PIXEL-WISE			OBJECT-WISE ($\theta_{r,l}$ =50%)		
		P	R	F1	SP	SR	SF1
In-N-n	ALL	0.455	0.483	0.468	0.482	0.474	0.478
	AUTH	0.618	0.588	0.602	0.643	0.636	0.640
	UNS	0.497	0.519	0.508	0.562	0.578	0.570
	HCS	0.433	0.463	0.447	0.460	0.451	0.455
InOD -Y-n	ALL	0.593	0.41	0.485	0.618	0.44	0.514
	AUTH	0.705	0.517	0.596	0.800	0.545	0.649

	UNS	0.767	0.325	0.456	0.793	0.430	0.558
	HCS	0.545	0.449	0.493	0.583	0.441	0.502

Table 32. Performance of the detection models integrated in the INCISIVE AI platform per each data provider.

The performance of both models on the two datasets was comparable, and the models achieved similar overall level of success. It can be notices that the object-wise measures are somewhat better when measured on the test data in D4.3 (e.g. SF1), while pixel-wise measures tend to be superior on the observational data (e.g. pixel-wise F1 score is better on the observational data for both models integrated in the INCISIVE AI platform).

As a discussion on small differences between test and observational performance of the two models is that the test set and the observational set, are significantly different in size and contain different ration of data coming from various INCISIVE data providers. During training, we observed (and reported in D4.3) that data coming from different providers tend to be quite distinct, and that the train models tend to perform differently (on average) when evaluated on data coming from different providers. There is another layer of complexity, as some of the data providers (HCS) has collected data from multiple hospitals, where some MMG images were of excellent quality, while some were scanned mammograms converted to dicom. However, our aim here is to provide performance as close as possible to the one that can be achieved for quite large span of quality levels that can be seen in clinical practice.

As an illustration of the performance of the trained models on the observational dataset, we show one MMG image, and comparatively show bounding boxes annotated by a trained clinitian (radiologist) in blue together with bounding boxes identified by the model in red (Figure 12). The image is shown after pre-processing, and that the predicted bounding boxes are generated by the **InOD-Y-n** model. It is important to stress that several suspicious regions appear in this image, and that they were all successfully identified by the model. The match between area identified by a trained radiologist and the area identifies by the model seems to be quite good.

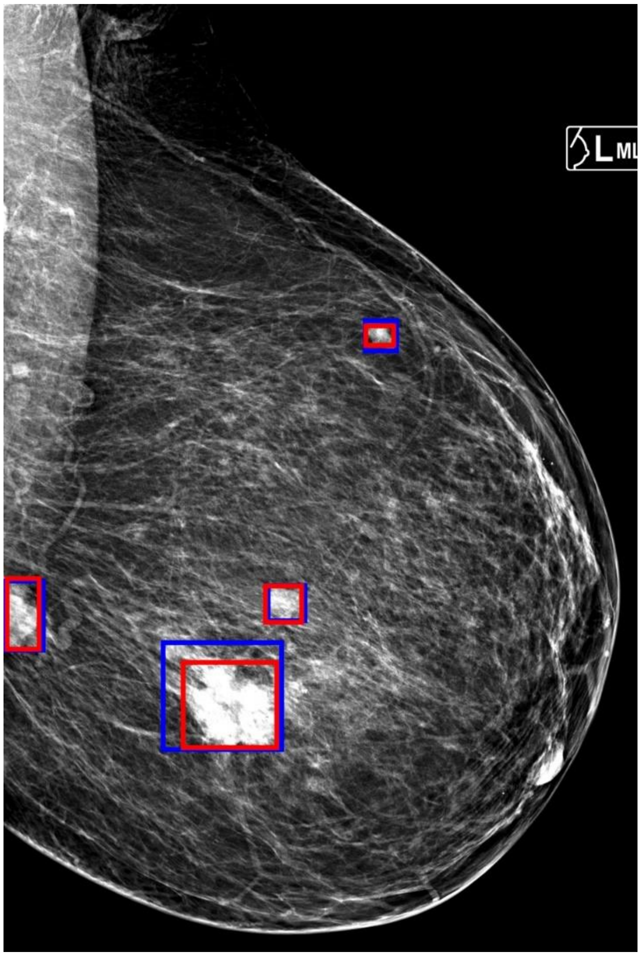


Figure 12. Comparative visualization of the radiologist annotations (blue) and annotations provided by InOD-Y-n model (red).

3.2.5 MRI Lesion Localization

During model development phase, we were faced by the lack of the MRI images in breast cancer and insufficient numbers to train a trustworthy model, for these reasons the DUKE MRI open dataset was used open dataset was used (Saha, et al., 2018), as reported in D4.3, The annotations provided by this dataset are 3D bounding boxes. However, our methodology was to approach the localization task on slice-by-slice manner, using object detection by means of YOLO networks (we used YOLOv5 architecture).

The observational study reassured us that, since MRI data are scarce, using significant quantity of readily available data from the DUKE dataset was a right decision, and that the obtained models can be easily refined in the future, once more INCISIVE MRI data becomes available. As explained in 3.2.2.2, 100 MRI studies were withdrawn prior to the model development and left for the fair evaluation of the final model during the observational study phase. An additional restriction is the use of the first post-contrast MRI sequence, where the lesions are most distinct (i.e. MRI examination with this type of sequence is a requirement for model usage).

During training, the data was prepared by taking the central annotated image slice (the central slice of the 3D bounding box), along with two annotated slices from each side shifted by a variable offset, since the lesion thickness is not consistent in the data set. In total up to three slices with lesions are selected for each patient. It is important to stress that, in some cases of smaller lesions, the annotation may be visible only in a single slice. This procedure was applied to all three data partitions: to the training, validation and test subset during model development. When talking about test sets in the development phase, the results in D4.3 are reported for:

- “reduced lesion test set” - the central annotated image slice (the central slice of the 3D bounding box), along with two annotated slices from each side.
- “lesion test set” - all image slices where a lesion annotation present
- “full test set” – all image slices in the MRI first post-contrast sequence.

We perform the same dataset split procedure to the observational data, for the sake of a fair comparison. The observational data was split into “reduced observational lesion dataset”, “observational lesion dataset”, and “full observational dataset”.

Three distinct detection models based on the YOLOv5 object detection architecture were developed in D4.3. For reference, we list here those models and report their performance on the D4.3 test sets used during model development in three different types of test sets. It is important to stress that the **-n**, **-s**, and **-m** suffixes of model names encode network size, where n stands for “nano”, **s** for “small”, and **m** for “medium”. Full details regarding model architectures and size can be found in D4.3.

The results on the “reduced lesion data set” taken from the development test set and from the observational dataset are presented in Table 33 and Table 34, respectively.

The results on the “lesion test set” taken from the development test set and from the observational dataset are presented in Table 35 and Table 36, respectively.

The results on the “lesion test set” taken from the development test set and from the observational dataset are presented in Table 37 and Table 38, respectively.

MDL.	PIXEL-WISE			OBJECT-WISE ($\theta_{r,l}$ =50%)		
	P	S	F1	SP	SS	SF1
D-n	0.436	0.571	0.494	0.646	0.745	0.692
D-s	0.47	0.536	0.501	0.695	0.657	0.675
D-m	0.746	0.39	0.513	0.845	0.514	0.640

Table 33. Performance of MRI lesion detection models on the reduced lesion test set. The models integrated in the INCISIVE AI platform are highlighted in yellow.

MDL.	PIXEL-WISE			OBJECT-WISE ($\theta_{r,l}$ =50%)		
	P	S	F1	SP	SS	SF1
D-n	0.49	0.63	0.55	0.64	0.76	0.70
D-s	0.47	0.58	0.52	0.70	0.72	0.71

Table 34. Performance of the deployed MRI lesion detection models on the reduced observational lesion set.

MDL.	PIXEL-WISE			OBJECT-WISE ($\theta_{r,l}$ =50%)		
	P	S	F1	SP	SS	SF1
D-n	0.529	0.448	0.486	0.665	0.559	0.608
D-s	0.590	0.385	0.466	0.740	0.491	0.590
D-m	0.786	0.254	0.384	0.874	0.340	0.490

Table 35. Performance of MRI lesion detection models on the lesion test set. The models integrated in the INCISIVE AI platform are highlighted in yellow.

MDL.	PIXEL-WISE			OBJECT-WISE ($\theta_{r,l}$ =50%)		
	P	S	F1	SP	SS	SF1
D-n	0.56	0.45	0.50	0.65	0.57	0.61
D-s	0.52	0.43	0.47	0.70	0.54	0.61

Table 36. Performance of the deployed MRI lesion detection models on the observational lesion dataset.

MDL.	PIXEL-WISE			OBJECT-WISE ($\theta_{r,l}$ =50%)		
	P	S	F1	SP	SS	SF1
D-n	0.208	0.448	0.284	0.230	0.559	0.326
D-s	0.258	0.385	0.309	0.374	0.491	0.424
D-m	0.515	0.254	0.340	0.590	0.340	0.432

Table 37. Performance of MRI lesion detection models on the full test set. The models integrated in the INCISIVE AI platform are highlighted in orange.

MDL.	PIXEL-WISE			OBJECT-WISE ($\theta_{r,l}$ =50%)		
	P	S	F1	SP	SS	SF1
D-n	0.25	0.45	0.32	0.26	0.56	0.36
D-s	0.28	0.43	0.34	0.37	0.54	0.44

Table 38. Performance of the deployed MRI lesion detection models on the full observational dataset.

It is obvious that model performance on the observational data is comparable to the performance on the test set, in fact our models tend to perform slightly better on the observational dataset. As expected, the same issues that were observed during testing, were also observed during observational study. In particular, the detector does not exploit the 3D image structure and interdependencies between image slices and does not correct for simple errors, such as skipping a lesion in one slice, and identifying it on the adjacent slices. Moreover, structures of similar intensities within the imaged body volume, such as heart and other tissue, sometimes trigger false positive detection. As noted also in D4.3, a possible solution for this problem could be additional breast segmentation prior to network input (which has to be done carefully as breast MRI serves for an improved estimate of disease dissemination) or detection post-processing (where adjacent slices could be used to revise positive or negative detections). Figure 13 presents an example of MRI localization service on several consecutive slices.

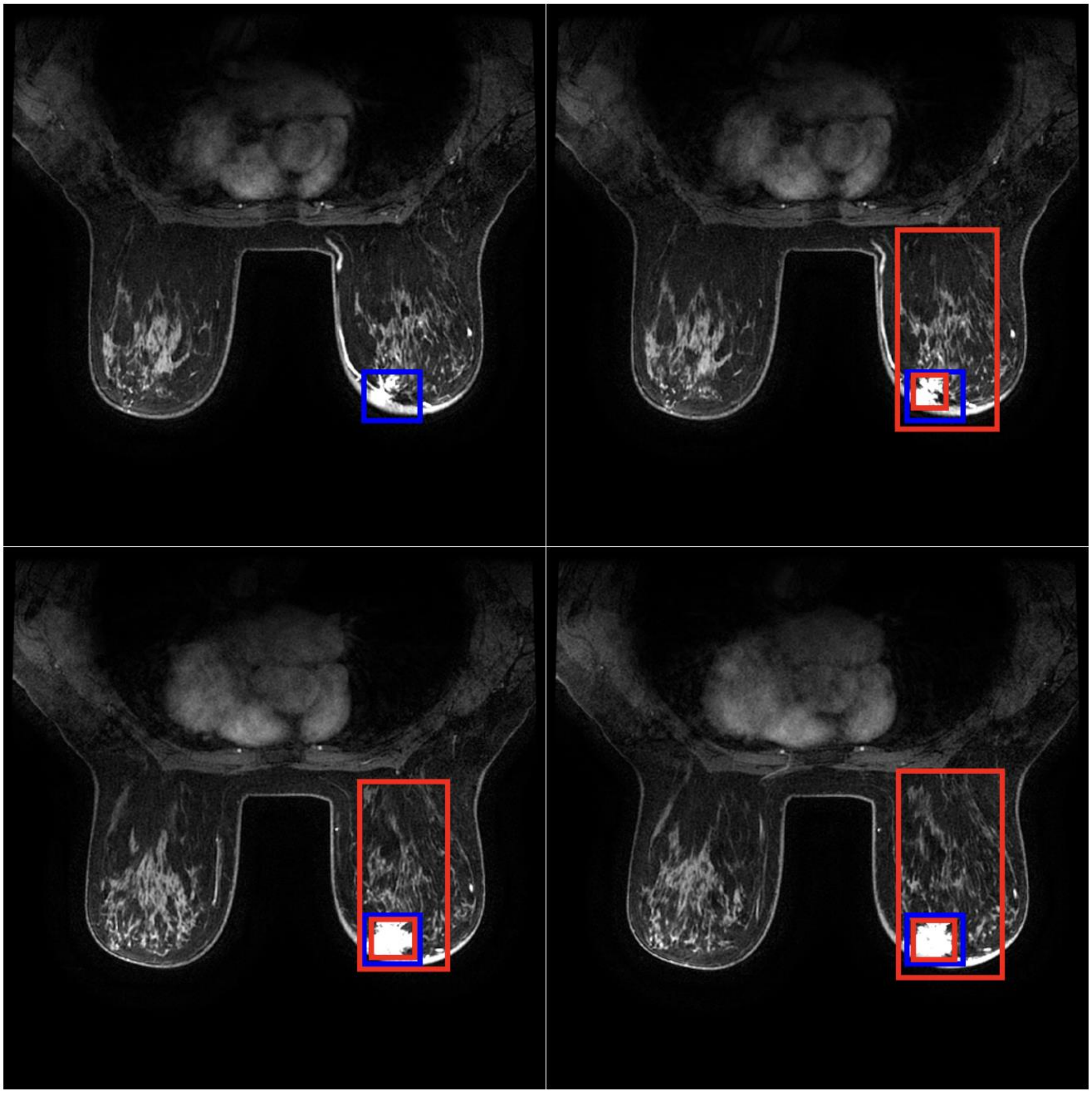


Figure 13. Example of several slices taken from a single observational MRI scan. The original annotations are drawn in blue, while the detections produced by the model are shown in red.

3.2.6 Mammography Lesion segmentation

In the observational study, model evaluation is restricted only on to the models integrated in the INCISIVE AI platform for the mammography lesion segmentation service, i.e. SegFormer models. Based on the results presented in the deliverable 4.3, SegFormer MiT-B3-G1/2 and MiT-B4-G2 have been selected for integration and use in the pilot studies. The results obtained on the dataset of MG images collected during the INCISIVE prospective data collection studies (Table 24) are presented in Table 39. The obtained F1 scores on pixel lever are similar to the initial scores obtained on the retrospective INCISIVE test set used for fair performance evaluation during the development phase. The F1 score on pixel level dropped for 5.5 % in case of MiT-B4-G2 model and 0.9 % in case of MiT-B3-G1/2, which indicate that the error estimation using the standard pixel level metrics was good during the model development phase, despite the INCISIVE retrospective MMG test set used in D4.3 is significantly smaller than full observational MMG data set used in this deliverable. Moreover, the training of the model has been performed using only with the images where lesion is present, as this is the segmentation model, while the results reported in Table 39 are average on all available images, regardless of the lesion presence.

The manual analysis shows examples of successful segmentation cases (Figure 14), where the predicted lesion (in red) successfully identifies the lesion marked by the expert annotation (in blue). The examples shown in Figure 15: Examples of false positives (A) and (C) and false detections (B) and (D) results for MiT-B3-G1/2 in the INCISIVE observational Figure 15 A) and C) imply the false positive cases, where the model predicts the suspicious lesions not annotated by the experts, while in panels B) and D) model has missed the annotated lesion, while the prediction marked by the model identifies other suspicious lesions.

The performance results averaged over all observational study images, irrespective of the confirmed presence of lesions, are summarized in Table 39. Moreover, we include the performance analysis in Table 40 when the model is presented only with images that contain the lesions to be segmented. This is more aligned with the scenario of the potential model usage, as the segmentation model in the breast cancer pipeline, is to be used once the presence of lesion has been detected.

SEGMENTATION MODEL	Test set	PIXEL-WISE			OBJECT-WISE ($(\theta_{r,l} = 50\%)$)			OBJECT-WISE ($(\theta_{r,l} = 10\%)$)		
		P	S	F1	SP	SS	SF1	SP	SS	SF1
MiT-B3-G1/2	AUTH	0.845	0.731	0.784	0.765	0.529	0.626	0.765	0.765	0.765
	HCS	0.483	0.517	0.500	0.417	0.474	0.444	0.571	0.565	0.568
	UNS	0.558	0.466	0.508	0.563	0.405	0.471	0.642	0.591	0.615
	ALL	**0.495**	**0.509**	**0.502**	**0.444**	**0.464**	**0.454**	**0.584**	**0.571**	**0.577**

MiT-B4-G2	AUTH	0.850	0.715	0.777	0.765	0.529	0.626	0.765	0.765	0.765
	HCS	0.500	0.506	0.503	0.470	0.443	0.456	0.558	0.547	0.553
	UNS	0.549	0.450	0.495	0.585	0.377	0.459	0.623	0.575	0.598
	ALL	**0.513**	**0.491**	**0.502**	**0.490**	**0.434**	**0.460**	**0.570**	**0.554**	**0.562**

Table 39. Lesion segmentation results obtained using the SegFormer models based on MiT-B3 and MiT-B4 evaluated on the INCISIVE observational data.

SEGMENTATION MODEL	Test set	PIXEL-WISE			OBJECT-WISE ($\theta_{r,l} = 50\%$)			OBJECT-WISE ($\theta_{r,l} = 10\%$)		
		P	S	F1	SP	SS	SF1	SP	SS	SF1
MiT-B3-G1/2	AUTH	0.845	0.731	0.784	0.765	0.529	0.626	0.765	0.765	0.765
	HCS	0.522	0.592	0.555	0.524	0.596	0.558	0.717	0.710	0.713
	UNS	0.729	0.437	0.546	0.791	0.569	0.662	0.902	0.830	0.865
	ALL	**0.578**	**0.535**	**0.555**	**0.565**	**0.591**	**0.578**	**0.743**	**0.727**	**0.735**
MiT-B4-G2	AUTH	0.850	0.715	0.777	0.765	0.529	0.626	0.765	0.765	0.765
	HCS	0.604	0.526	0.563	0.593	0.559	0.575	0.704	0.690	0.697
	UNS	0.809	0.362	0.500	0.810	0.523	0.636	0.863	0.797	0.829
	ALL	**0.658**	**0.466**	**0.546**	**0.625**	**0.553**	**0.587**	**0.727**	**0.706**	**0.717**

Table 40. Lesion segmentation results obtained using the SegFormer models based on MiT-B3 and MiT-B4 evaluated on the INCISIVE observational data containing annotated lesions.

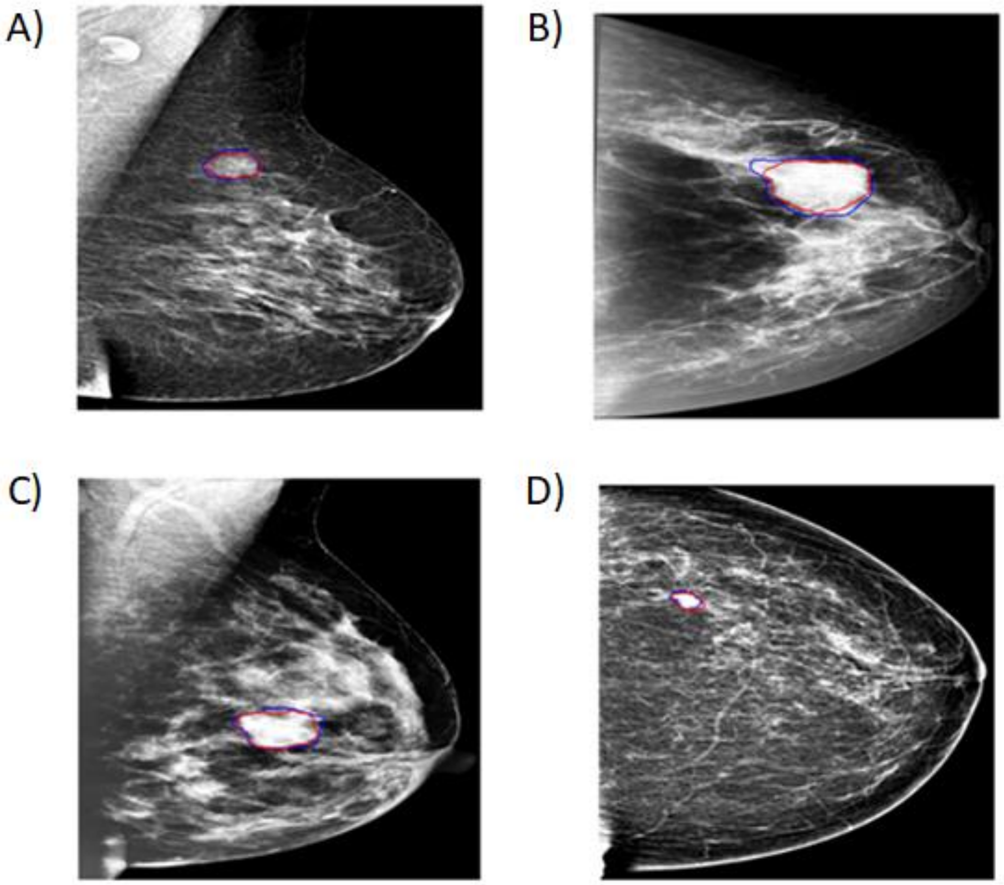


Figure 14: Some examples of successful segmentations for MiT-B3-G1/2 in observational dataset.

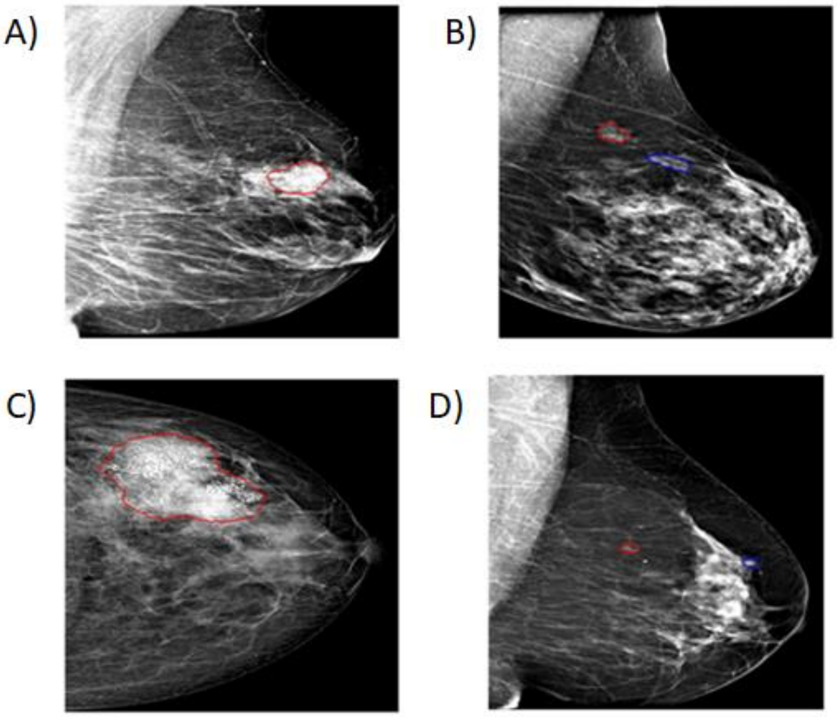


Figure 15: Examples of false positives (A) and (C) and false detections (B) and (D) results for MiT-B3-G1/2 in the INCISIVE observational dataset.

3.2.7 Mammography BIRADS classification

During the project's lifetime, the annotation standards, regarding the BIRADS classification, have been slightly changed. A BIRADS score per breast is now given, compared to an overall score in the retrospective data. In addition, after consultation with the medical professionals, BIRADS 0 and 6 were excluded from the possible model predictions. Therefore, retraining the model utilizing mammograms from both HCS and UNS retrospective datasets, has resulted in 71.2% F1 score. The performance of the model on UNS and HCS prospective datasets is reasonably reduced (55.2% and 59.4 accordingly), considering the heterogeneity of data and the detailed differentiation of the BIRADS scale (Table 41). The 100% score achieved on AUTH dataset is not indicative, since it is a small set of 6 images. Given the decrease in performance in the prospective data, the maturity of the model is not as high as expected, and possible examinations of different harmonization approaches may improve performance in the future.

Dataset	# Images	F1-Score [%]
HCS + UNS retro	3254	71.2
UNS	61	55.2
AUTH	6	100
HCS	319	59.4

Table 41: Breast BIRADS classification performance comparison

3.2.8 Mammography Breast density classification

As it is stated in D4.3, the Breast density classification model has been trained using 1700 mammograms from the UNS and reached 80% F1 score on the test set. For the inference purposes, prospective data from UNS, AUTH and HCS was utilized. As shown in Table 42 the model performs well in all cases, though, it presents reduced effectiveness in UNS and HCS cases. It is assumed that this occurs due to the heterogeneity and imbalanced distribution of the data, since such a performance drop is reasonable and expected. On the other hand, the model seems to perform perfectly on the AUTH prospective dataset with F1 score that equals 100%. Though, it is not an indicative result, considering that the dataset contains only 3 patients and 6 corresponding images. Therefore, additional investigations may be required to increase the model's maturity.

Dataset	# Images	F1-Score [%]
D4.3 (UNS)	1700	80
UNS	61	64.2
AUTH	6	100
HCS	319	58.4

Table 42: Breast density classification performance comparison

3.3 Prostate Cancer

3.3.1 Prostate cancer final AI service overview

The final set of services for prostate cancer has been developed for MRI as a main diagnostic imaging modality (Figure 16).

Prostate cancer services are:

1. **Prostate gland segmentation** - performs delineation of the prostate gland using T2W Axial View MRI in DICOM format.
2. **MRI lesion segmentation** – in the prostate gland based on T2W Axial View MRI in DICOM format using following services:
    a. **prioritization service** – classifies a MRI scan and assigns a low/high priority to the patient depending on whether any findings were identified
    b. **localization service** – identifies the slices of the scan where the model predicted lesions
    c. **segmentation service** – performs annotation of the potential findings in each located slice
3. **ISUP score classification service** utilizes T2W and DWI images from MRI examinations to deliver a binary classification of prostate cancer (clinically significant prostate cancer (csPCa) – class 1, and clinically insignificant prostate cancer (cisPCa) - class 0). The user is provided additional explanation of the classifier decision through XAI Lime or Anchors.

Prostate Cancer pipeline:

MRI pipeline – uses T2W and DWI images from MRI and runs all MRI related services – prostate gland segmentation – lesion segmentation and ISUP score classification with XAI user feedback.

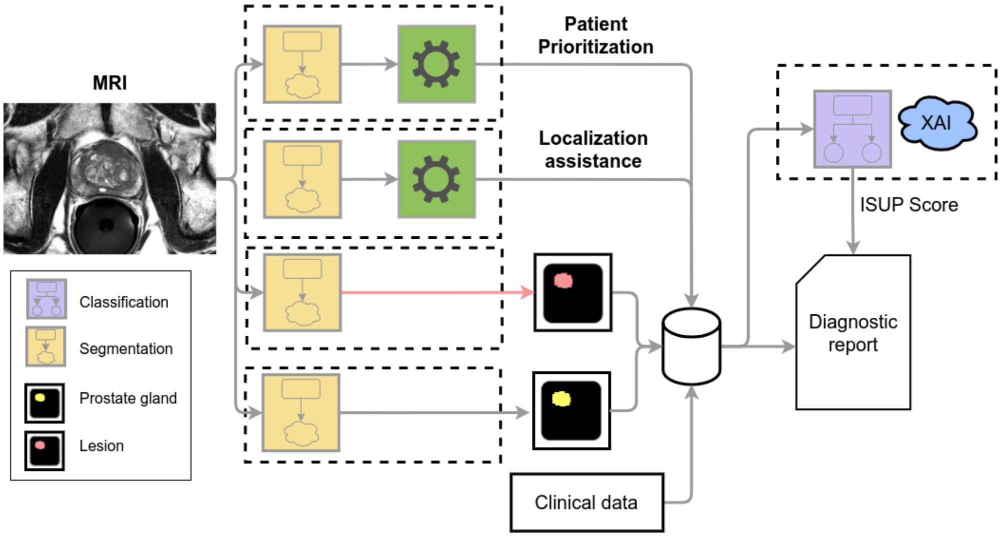


Figure 16: Overview of INCISIVE Prostate Cancer AI Services

3.3.2 Prostate Gland Segmentation

The first model developed for prostate cancer in INCISIVE is the delineation of the prostate gland area. Since the INCISIVE annotation guidelines did not include annotation of the prostate gland, the model was developed and tested on open data. As reported in D4.3, the dataset was split on an 80%-10%-10% ratio, to keep a percentage of the dataset for validation in this deliverable. The performance of the model on the 10% test set (100 MRI scans) can be found in Table 43. It is evident that the performance remains high, and that the maturity of the model is at an acceptable level.

Dataset	F1-Score [%]
D4.3 (PICAI)	85.2
D6.5 (PICAI)	83.7

Table 43: Prostate gland segmentation performance comparison

Due to the fact that gland segmentation masks were not part of INCISIVE's data collection, a quantitative assessment of the model's performance is not possible. However, the gland segmentation model is used as a pre-processing step in the lesion segmentation service to isolate the prostate gland area. Therefore, the model has been used with INCISIVE data, and an example of its performance can be found in Figure 17.

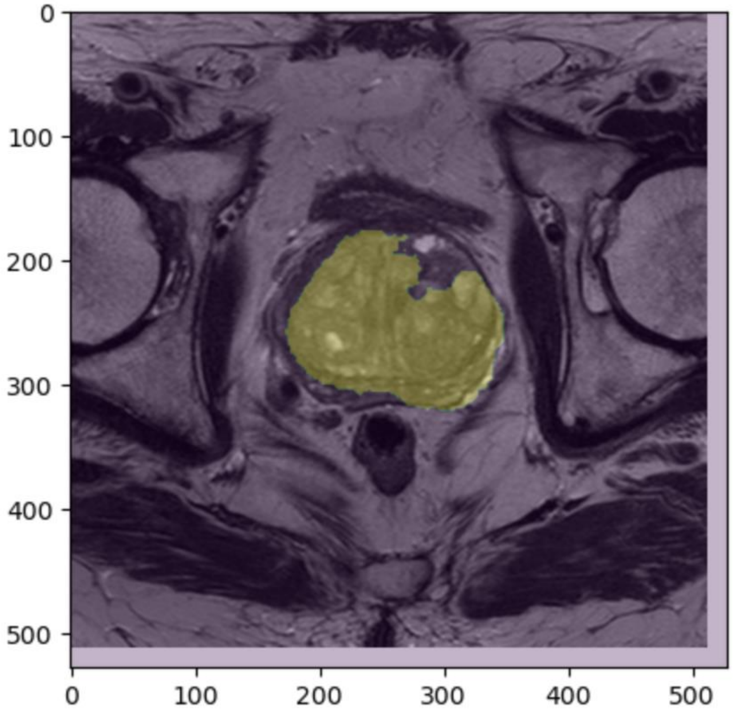


Figure 17: Prostate gland segmentation of an INCISIVE T2W MRI scan.

3.3.3 Lesion Segmentation

This service performs segmentation of lesions in Prostate MRI scans. It is the second model regarding the prostate case and utilizes the prostate gland segmentation model to produce proper inputs. In this section, this model is evaluated as a separate service. A detailed description of the model architecture can be found in D4.3. Model training and validation was performed on INCISIVE data from 1 data provider (UoA). In this deliverable, we extend the validation of the model on new data, which come either from the same data provider or from a different one (AUTH).

Table 44 reports the number of samples (patient examinations) that were utilized to further validate the model.

Data Provider Name	Number of patients	Number of cases (examinations)
UoA	44	44
IDIBAPS	50	50

Table 44. Number of patients and cases (MRI examinations) used to further evaluate the model.

First, we evaluate the prioritization service, which assigns a low/high priority to the examination depending on whether any suspicious area was detected.

Dataset	F1-Score [%]
D4.3 (UoA)	100
UoA	88.8
IDIBAPS	65.4

Table 45. Evaluation performance of Prostate MRI scan prioritization service

As can be seen from Table 45, the performance of the prioritization service is very high in the first two cases, while it is lower in the last case.

Next, we evaluate the localization assistance service, which produces a list of the frames where a lesion was segmented. The results can be seen in Table 46.

Dataset	Sensitivity [%]	Specificity [%]
D4.3 (UoA)	100	100
UoA	95.5	97.3
IDIBAPS	70.5	96.2

Table 46. Evaluation performance of Prostate MRI scan localization service

Finally, we evaluate the lesion segmentation service on a pixel-level. The results can be found in Table 47. Figure 18 exemplifies some lesion segmentation cases from the INCISIVE database.

Dataset	F1-Score [%]
D4.3 (UoA)	87
UoA	72.9
IDIBAPS	14.3

Table 47. Evaluation performance of MRI scan segmentation service

As we observe, the performance of the model on the UoA prospective data is well enough, comparing it to its performance on the retrospective data. Even the pixel-wise segmentation service achieves an F1-score of 72.9%. Some more qualitative results could occur by examining the images in Figure 18. Images in the left column depicts the ground truth annotations, while the images on the right present the predictions of the model. In the first case, it seems that the model performs a good segmentation of the lesion, while in the other two cases the model fails to detect the lesions.

According to the results presented in the aforementioned tables, the decline in the performance of the model, while it is evaluated on the IDIBAPS dataset, is substantial. Following further examination on this, we assume the following reasons for this behaviour:

1. The IDIBAPS dataset contains images with different, much higher pixel intensities than the UoA retrospective/prospective data;
    a. UoA – min: 0, max: 2674
    b. IDIBAPS – min: 0, max: 6373
2. The IDIBAPS' style of annotation differs a lot from the UoA's approach, and this seems to limit the model's potential. In particular, the IDIBAPS masks usually annotate a wider area, including the surrounding region in the mask.

Taking the above into consideration, the maturity of the model is high only for UoA patients, whereas for transferability to other data providers/vendors, additional harmonization techniques are required.

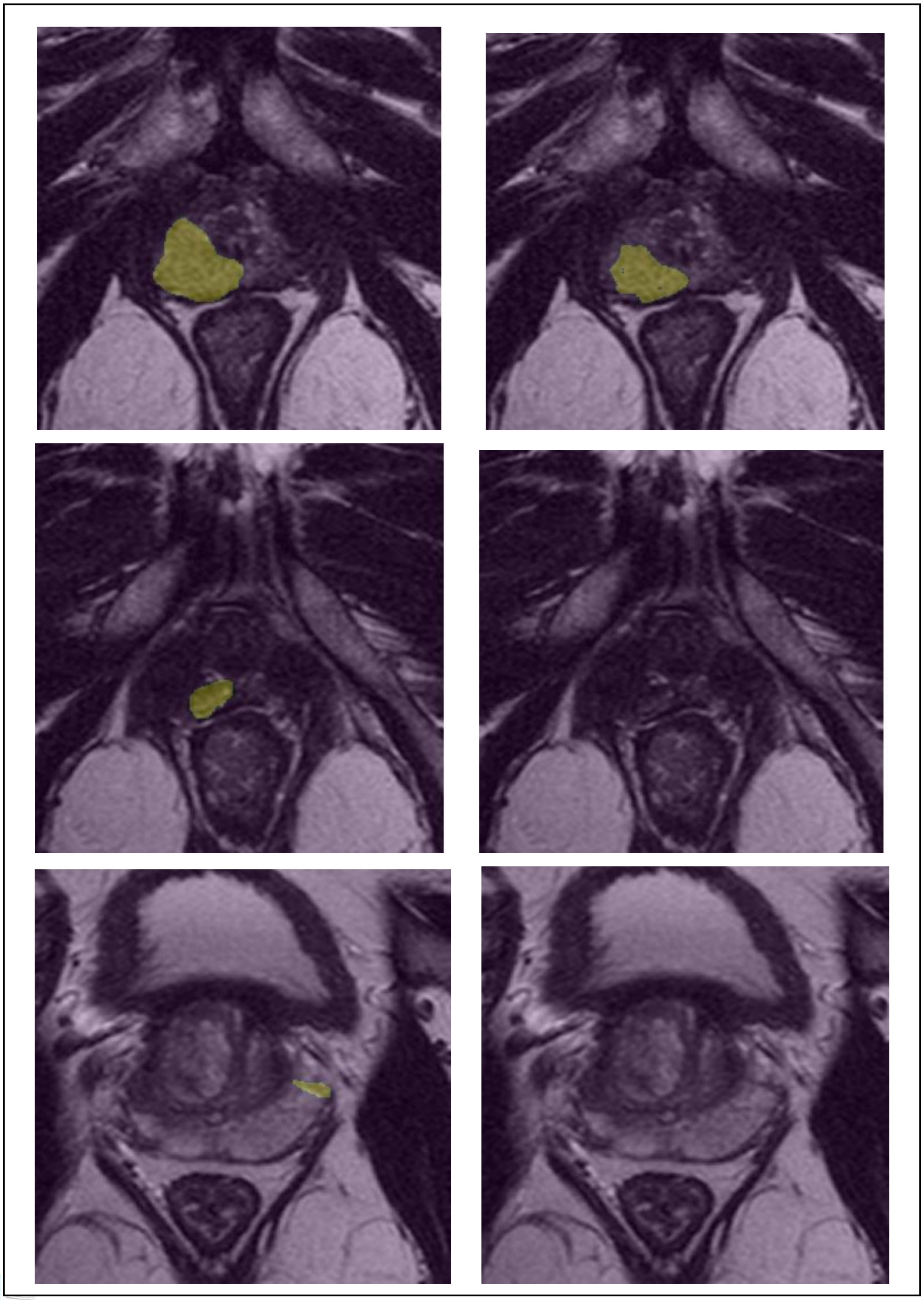


Figure 18. Some examples of prostate lesion segmentation cases for UOA patients. The left column depicts the original image and the ground truth, whereas the right one presents the model predictions.

3.3.4 ISUP Score Classification

The prostate ISUP classification model is based on the analysis of the whole prostate gland instead of requiring the accurate segmentation of suspicious lesions. The whole pipeline is based on the combination of radiomics features from the prostate gland with patient's clinical characteristics and is ordered to identify the presence of clinically significant (csPCa) of clinical insignificant prostate cancer (cisPCa).

The data utilized in developing the model were obtained from the Prostate Imaging: Cancer AI challenge (Anindo Saha, 2022), specifically from the public training and development dataset. This dataset comprises imaging data from 1500 patients suspected of clinically significant prostate cancers, identified through elevated levels of PSA or abnormal Digital Rectal Exam (DRE) findings. Among them, 1075 patients were classified as cisPCa, and 425 were diagnosed with csPCa. The MR images include T2-weighted (T2W) images (in axial, sagittal, and coronal views), Diffusion-weighted images (DWI) in the axial view, and Diffusion Coefficient maps (ADC) in axial view.

In addition to the imaging data, annotation files are also provided, covering lesion segmentation and segmentation of the entire gland. Concerning lesion delineation, human experts performed annotations for a limited number of patients (<17%), while AI-based annotations of lesions are available for all patients. As for whole gland annotation, segmentation was exclusively based on AI algorithms and applied to all patients. Finally, patients' clinical characteristics were also available, including age, prostate volume, Prostate Specific Antigen (PSA), and PSA density.

More details regarding the model development have already been provided in D4.3.

Regarding the data needs, specific requirements must be met in order to be used for the model testing in order to be eligible for the testing of the model. The requirements are listed below:

1. Images from T2W and DWI from the same area
2. Segmentation file of the whole prostate gland that fits both image series
3. The images must be acquired during diagnosis
4. PSA and age are available.
5. Gleason score or ISUP score is available.

Considering the requirements above, in INCISIVE three clinical sites provided data from prostate cancer (UoA, IDIBAPS and GOC). The data from multiparametric MRI examination are provided and a set of different folders included images acquired using different pulse sequences (included T2W and DWI). The information regarding the type of pulse sequence used is typically included in a specific DICOM tag of the images (Series Description (0008,103E)). However, after the anonymization step that was defined in INCISIVE, the content of this DICOM tag was removed making the identification of the appropriate images challenging. Fortunately, one clinical site (UoA) provided only two folders of images, one for T2W and one for DWI, placing the annotation file of the lesions into the folder containing the T2W images. This approach made it possible to indirectly identify the types of images that are provided. In total, 90 patients were eligible for testing (42 from retrospective study and 48 observational).

The main obstacle towards their use is the unavailability of whole prostate gland segmentation for all those patients, which delays the testing of the ISUP classification model using INCISIVE data until the AI-based whole gland service is implemented by a partner of the project.

For this reason, until the completion of the aforementioned service, further ideas were explored with the existing dataset.

Specifically, we tried to investigate other approaches in order to improve the performance of the model. In this respect, based on the prostate gland segmentation, two additional prostate areas we defined: (1) the periphery of the prostate which was defined as the area covering 4mm inside the prostate borders and (2) the core of the prostate which covers the remaining area of the gland and radiomic analysis was applied in both areas. Figure 19 depicts the areas that were considered for the radiomics analysis.

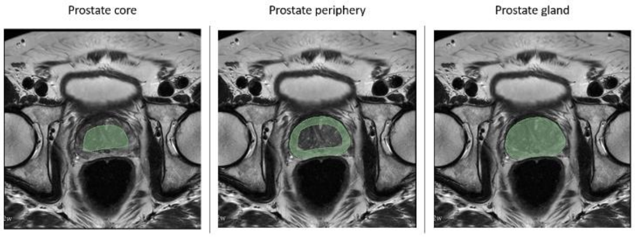


Figure 19 Prostate areas that were used for the extraction of radiomic features.

The whole pipeline is already described in D4.3 and it is summarized in Figure 20 below

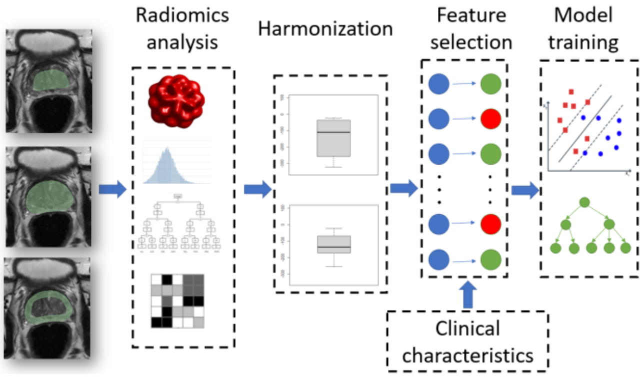


Figure 20 Schematic overview of the approach applied for the development of the prostate staging model

In brief, the pipeline is based on the extraction of radiomic features from different parts of the prostate gland, the harmonization of the features based on the vendor that was used for the image acquisition, the feature selection and finally the training of the model. For the harmonization of the data based on the vendor machine used, we applied the COMBAT pipeline (Fanny Orlhac, 2022). For feature selection, we initially removed the outliers using PCA outlier detection, with hotelling T2 test and on SPE/DmodX criteria, followed by the exclusion of the readiomics features which are not statistically significant. (<0.05). In addition, the remaining features were checked for correlation and those which presented correlation higher than 95% were excluded from the analysis. For the remaining features, Repeated Elastic Net Technique (RENT) was used (with C=0.2, l1=0.7) (Anna Jenul, 2021). Finally, based on the results of D4.3, we focused on two models, the first was Support Vector Machines (SVM) with linear kernel and the second one was Balanced Random Forest (BRF). Of note, Deep Learning classifiers for tabular data were also extensively explored but did not succeed in surpassing the SVM/BRF performance. In total, four different models were tested, with two of them using the radiomics from the whole prostate gland (case1) and two combining features from different parts of the prostate (case2), with SVM and BRF.

Following the data preprocessing phase, which involved removing the cases with missing clinical information or erroneous segmentation files leading to the inability to extract radiomics features, we included 1413 patients (cisPCa = 1002; csPCa = 411) for the case1 model and 1395 patients (cisPCa = 989; csPCa = 406) for the case2 model. Among these, 80% had data from SIEMENS.

For each segmentation mask, 1132 radiomics features were computed, resulting in a total of 2264 and 4528 features utilized for case1 and case2, respectively. Additionally, age and PSA were among the selected features for both cases. The radiomics features underwent harmonization for vendor batches, as illustrated in Figure 21 showing the first two principal components of the harmonized dataset.

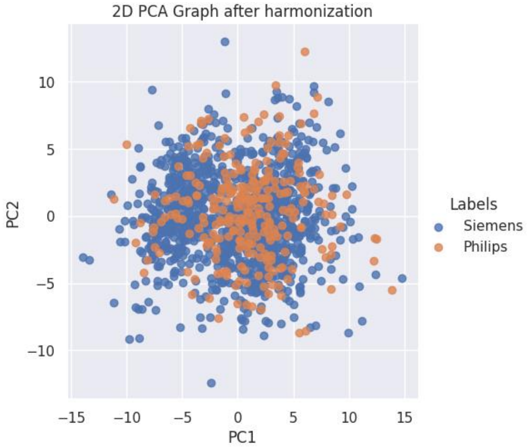


Figure 21 PCA after harmonization of radiomics features from images acquired using different vendor scanners

After the estimation of the statistically significant features and the exclusion of highly correlated features, the feature set for case1 and case2 was 351 and 749 features, while the final feature sets after RENT feature selection included 42 and 54 features respectively.

Table 48 includes the performance metrics for the two models, after 5 runs, using different radiomics features, the first using the radiomics from the whole gland and the second from both the periphery and the core. As observed the SMV classifier presents the highest performance, compared with the BRF. However, while the performance results are very close in all metrics, using the radiomics from the whole prostate gland seems to outperform the consideration of different prostate areas.

	SVM	*BRF*	*SVM*	*BRF*
Sensitivity	79±3	78±5	72.9±3	76.6±4
Specificity	79±2	72±3	80.5±1	74.2±2
Precision	61±2	54±2	60.6±2	54.9±3
Accuracy	79±1	74±4	78.3±1	74.3±2
Balanced accuracy	79±1	75±2	76.7±2	75.4±3

Table 48 Classification performance for cases 1 and 2 for SVM and BRF classifiers

In Figure 22, the 10 most important features (as estimated by permutation importance) for the two cases using the SVM classifier are depicted. As observed, in *case1*, 7 radiomics features extracted from DWI, 1 from T2W images, age and PSA were the most important features, whereas the respective list in *case2* includes 6 features from DWI, 3 from T2W and age. DWI texture features are more relevant, while in terms of the areas that were considered as more informative, among the 9 radiomics features, 6 features were extracted from the periphery or the prostate gland and only 3 from its core.

Furthermore, as mentioned on the dataset website, the segmentation of the gland was based on AI model which presented higher accuracy in the images from SIEMENS. Our hypothesis was that the accuracy of the segmentation file can affect the accuracy of the ISUP classification model, and harmonisation may blur the statistics of the radiomics features. For this reason, we further investigated the effect of vendors on the model development process, and we proceeded to the implementation of vendor specific models using the radiomics from the two areas of the prostate gland. As a result, we developed two models for SIEMENS and two models for PHILIPS scanners. Table 49 provides the performance results while for comparison purposes we included on the table the respective performance metrics when both vendors are used.

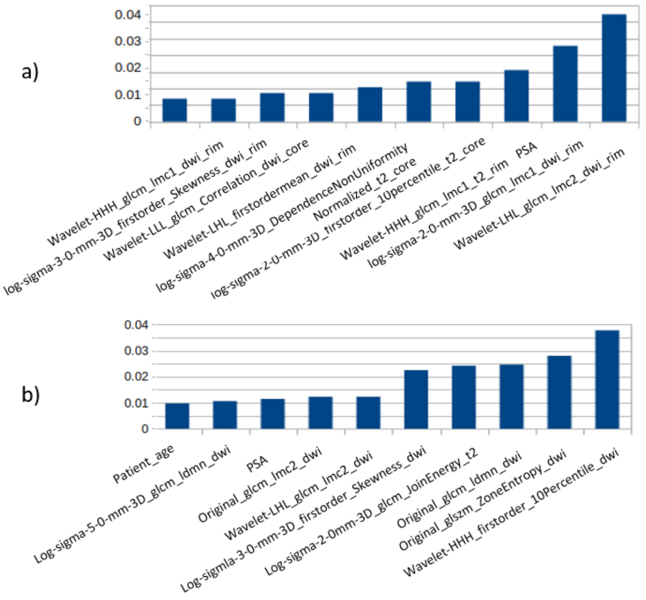


Figure 22 The most important features for the SVM classifier in case1 (a) and case2 (b). the suffix of each feature appeared in the x-axis denotes whether the feature was extracted from DWI (dwi) or T2W (t2) and in the case2, whether it was related to the core or the periphery (rim) of the prostate.

	SIEMENS		PHILIPS		Both vendors	
	SVM	BRF	SVM	BRF	SVM	BRF
Sensitivity	77.8±6	71.8±5	76.5±3	78.9±3	72.9±3	76.6±4
Specificity	74.4±5	70±6	79.5±2	75.5±5	80.5±1	74.2±2
Precision	56.5±3	50.5±4	60.5±2	57.3±4	60.6±2	54.9±3
Accuracy	75.8±2	70.5±4	78.6±2	77.74±3	78.3±1	74.3±2
Balanced accuracy	76.1±1	70.9±3	80±2	77.2±2	76.7±2	75.4±3

Table 49 Performance metrics for the vendor specific models (SVM and BRF) using the radiomics features from the core and the periphery of the prostate gland, using data

As observed, in all cases, the SVM classifier presented the higher Balanced Accuracy which is comparable to the model that was implemented using data from both vendors. On the other hand, BRF has an advantage in terms of Sensitivity. Regarding the most important features for the vendor specific models, it was found that for the SIEMENS-based model only 11 radiomics features were found to be important with the 6 of them derived from DWI and the rest from the T2W images. With regard to the prostate gland, 5 of them were calculated in the prostate gland core and the rest from its periphery zone. Regarding the PHILIPS-based model 41 features were found to be important, including 21 radiomics featured extracted from the DWI and 18 from T2W, patient age and PSA. From the list of the radiomics features, 25 were calculated in prostate periphery zone and the only 14 form the prostate core. These results were submitted as a paper to ISBI24 conference.

As a next step, we started the investigation of the masks that were recently made available by the whole gland segmentation service of INCISIVE. There are currently two challenges we are working on: A) the correct transformation of the mask files that were produced by the service (png file format) to a nifti file format, which is expected by the ISUP classification model. B) The generation of mask(s) that can be applied in both image series (T2W and DWI) since the segmentations are based on the T2W images and they must be transformed properly so that they can be used for the segmentation of the DWI, as well. In Figure 23 below we depict an example of a few slices that include the prostate gland and the segmentation mask overlayed.

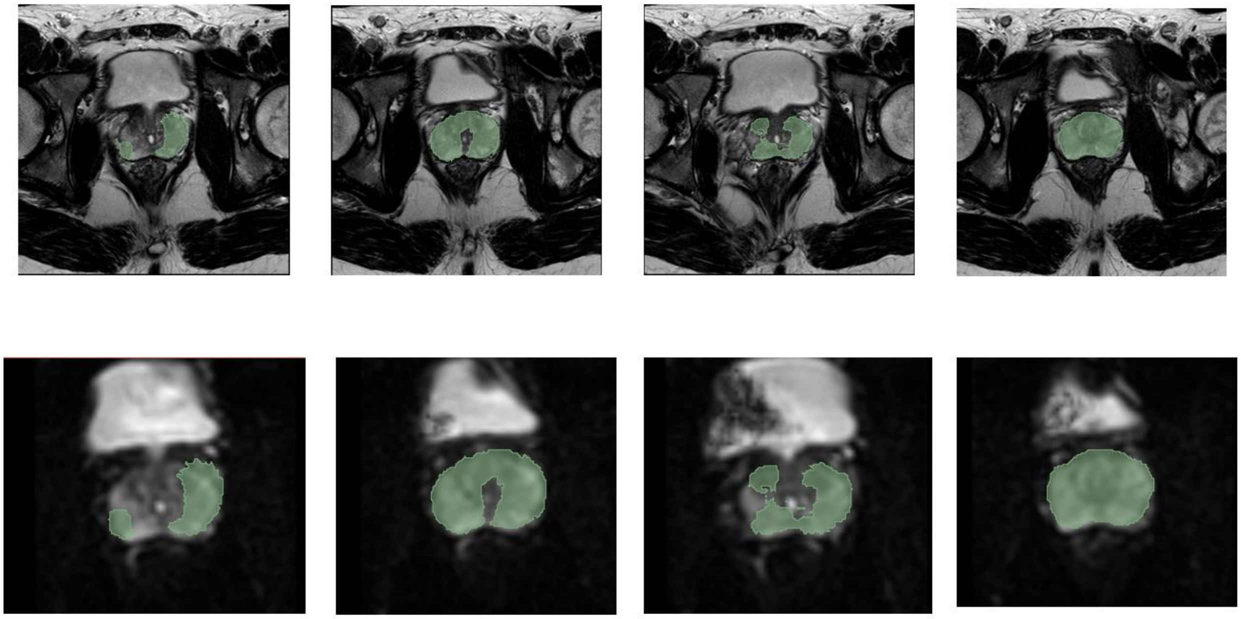


Figure 23 sample images from T2W (upper) and the respective DWI (bottom) and the segmentation masks that are prodiced by the INCISIVE whole gland segmentation service.

The mean maximum Gleason score for the UoA patients that will be used for the testing is 4.79±2.7, with the 68% of them to be considered as ISUP = 1 (cisPCa) and the rest as ISUP>1 (csPCa). The performance of the model based on those patients will be documented in D6.6.

3.4 Colorectal cancer

3.4.1 Colorectal cancer final AI service overview

Services for colorectal cancer have been developed for MRI and histopathological images (Figure 24).

Colorectal cancer services are:

1. **MRI lesion segmentation** – is based on T2W Sagittal view MRI scan in DICOM Format using following services:
   a. **prioritization service** – classifies a MRI scan and assigns a low/high priority to the patient depending on whether any findings were identified
   b. **localization service** – identifies the slices of the scan where the model predicted lesions
   c. **segmentation service** – performs annotation of the potential findings in each located slice
2. **Histopathological image analysis service** –performs the analysis of digitalized hematoxylin and eosin (H&E) stained samples from patients with colorectal cancer. It extracts morphological features and cell proportions from the H&E images, and predicts therapeutic responses based on progression free survival (PFS). Colorectal H&E slides should be in SVS format, and output of this service is a category 'Low' if the PFS is less than or equal to 2 years or 'High' if the PFS is greater than 2 years

There are no pipelines for colorectal cancer.

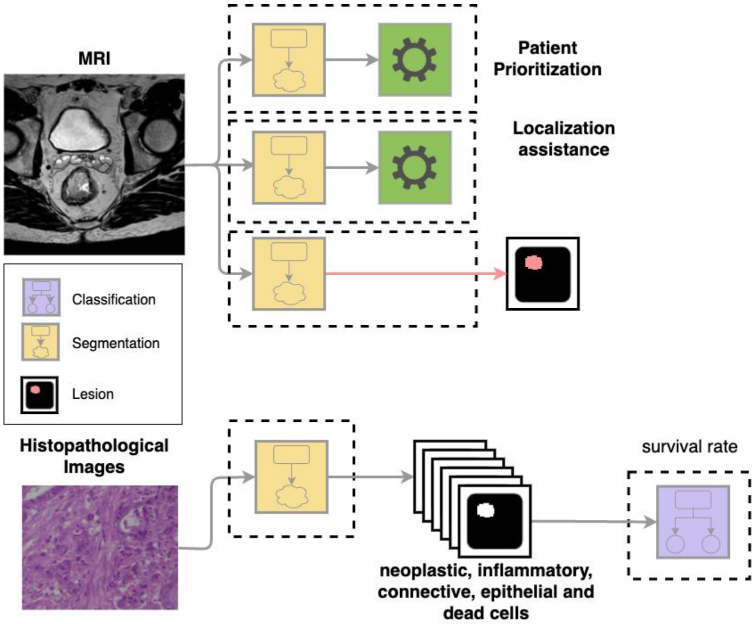


Figure 24: Overview of INCISIVE Colorectal Cancer AI Services

3.4.2 Lesion Segmentation

The lesion segmentation model was developed on T2W sagittal view MRI sequences. Unfortunately, the prospective data collection contains only a few MRI scans (approximately 25), out of which only two T2W sagittal view sequences were detected. Therefore, the validation of the developed model on prospective data, as well as the assessment of its maturity, are not meaningful.

3.4.3 Histopathological Image Analysis

3.4.3.1 Introduction

This model is based on a Deep Learning (DL) AI engine for the automated segmentation of cell types from histopathological images, which has been described in detail in Deliverable 4.2. In brief, the developed HEIP (H&E Image Processing) pipeline processes hematoxylin & eosin (H&E) stained images as input and provides re-processing segmentation and identification of different cell types as well as feature extraction, which is the basis for the prognostic model on survival prediction. The HEIP pipeline is a very versatile spatial model since it was trained using >205,000 cell annotations from several cancer types. The HEIP software has been made compatible with the INCISIVE infrastructure and is integrated into the INCISIVE AI Engine/Service.

Validating DL models based on histopathological images presents several challenges due to the unique nature and complexity of medical imaging. The key challenge is the limited availability of high-quality annotated data, particularly when expert pathologists are required to create ground truth labels. This limited availability of labelled data can make it challenging to train and validate deep learning models effectively, as they often require large data sets for optimal performance. To address these challenges, researchers and practitioners in the field of histopathological image analysis often employ techniques such as transfer learning, data augmentation, and cross-validation.

For the purposes of the blind evaluation, we carefully selected two openly available validation sets on colorectal cancer: the PanNuke (Gamper, et al., 2020) and the Lizard data (Graham, et al., 2021). Both are comprehensive datasets including digitalized histopathological images from different tissue types, as well as the corresponding annotated nuclei with their class labels. As an example, for usage in different cancer types, we also present the validation in a high-grade serous ovarian cancer (HGSC) cohort.

3.4.3.2 Cross-validation on the PanNuke colorectal cancer data set

PanNuke is a multimodal open-access dataset that includes five different tissue types: pancreas, lung, kidney, esophagus, and colorectal. This diversity of tissue types allows researchers to explore and develop models for various types of cancer. Another key advantage of the PanNuke dataset is that it includes manually annotated regions of interest within each image, delineating different cell types and structures, including nuclei, epithelium, connective tissue, inflammatory cells, and dead cells. These annotations can serve as ground truth data for model training and evaluation (Gamper, et al., 2020).

We selected a 3-fold cross-validation approach for the evaluation of HEIP. Cross-validation helps ensure that the model's performance metrics are reliable and not overly dependent on the specific data split into training and testing sets. It provides a more accurate estimate of a model's performance on unseen data compared to a single train-test split. The approach ensures that the training and validation datasets are always strictly independent, while also using all data for the training. In other words, when data are limited, cross-validation allows to maximize the utility of the dataset by repeatedly partitioning it into training and testing sets. Yet another advantage is that cross-validation provides a more comprehensive view of a model's performance by averaging results over multiple subsets of the data. This reduces the risk of drawing conclusions based on a single, potentially unrepresentative data split.

For the first validation, we selected only the subset of PanNuke data that contained colorectal cancer samples. The dataset was organized in folds, fold 1 contains 478 colon patches, fold 2 contains 468, and fold 3 contains 494 colon patches. In total 1,440 patches were used. For our 3-

fold cross-validation of the HEIP model, we trained on 3 folds, basically resulting in three slightly different models, as follows:

1) model 1 was trained with PanNuke fold 1 and fold 2 and validated with fold 3
2) model 2 was trained with PanNuke fold 1 and fold 3 and validated with fold 2
3) model 3 was trained with PanNuke fold 2 and fold 3 and validated with fold 1.

This approach ensures that the training and the validation set are always truly independent from each other.

The model provides instance segmentation which is a challenging computer vision task where the goal is to not only detect objects in an image but also segment each object into individual instances. Existing metrics, like TP, TN, FP, FN, sensitivity, specificity, and precision can only either evaluate semantic or instance segmentation separately, are therefore not used for instance segmentation tasks. Benchmark metrics commonly used to evaluate the performance of instance segmentation models are segmentation quality (SQ), detection quality (DQ) and panoptic quality (PQ). These image analysis performance metrics are defined as follows (Kirillov, He, Girshick, Rother, & Dollar, 2019):

$$DQ = \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}$$

$$SQ = \frac{\sum_{(p,g)\in TP} \quad IoU(p,g)}{|TP|}$$

$$PQ = DQ \times SQ = \frac{\sum_{(p,g)\in TP} \quad IoU(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}$$

where *TP*, *FP* and *FN* denote the true positive, false positive and false negative, respectively. *IoU* denotes the intersection-over-union. The reported validation metrices in Table 50 represent the benchmark metrics averaged over all three validation folds (the mean for the three validation benchmarks along with the standard deviation).

The validation results for HEIP are similar compared with other instance segmentation methods validated on the PanNuke dataset. Gamper and colleagues report PQ values for colorectal cancer between 0.3 and 0.5 for cell types and between 0.3 and 0.4 for colon specifically (Gamper, et al., 2020).

Cell Type	PQ		SQ		DQ	
	mean	SD	mean	SD	mean	SD
neoplastic	0.440	0.047	0.705	0.044	0.554	0.055
inflammatory	0.403	0.02	0.621	0.016	0.479	0.022
connective	0.391	0.025	0.659	0.031	0.489	0.029
epithelial	0.481	0.008	0.731	0.017	0.602	0.011

Table 50 HEIP pipeline accuracies determined with cross-validation, based on 1440 patches of the colorectal set of the of PanNuke dataset.

3.4.3.3 Cross-validation on the Lizard colorectal cancer data set

For the second, independent validation, the Lizard dataset was used (Graham, et al., 2021). The Lizard dataset is the largest known available dataset for nuclear instance segmentation and classification including annotated nuclei of only colorectal cancer samples with their associated class labels.

Since the Lizard data set contains some colorectal images from the PanNuke dataset, those samples were excluded to create a validation set that is truly independent of the validation on PanNuke. The Lizard data set contains some cell types that have not been included in the original training of HEIP, namely neutrophil, plasma, and eosinophil cells. Therefore, it was necessary to fine-tune the HEIP model to recognize all new types of cells annotated in the Lizard dataset. Then the fine tune was done as 3-fold cross-validation as described in the previous section.

The dataset was organized in folds, fold 1 contains 1980 colon patches, fold 2 contains 2190, and fold 3 contains 1945 colon patches. In total 6115 patches were used. Again, we employed the following steps:

1) model 1 was fine-tuned with Lizard fold 1 and fold 2 and validated with fold 3
2) model 2 was fine-tuned with Lizard fold 1 and fold 3 and validated with fold 2
3) model 3 was fine-tuned with Lizard fold 2 and fold 3 and validated with fold 1

The values of the benchmark metrics averaged over all validation folds (the mean for the three validation benchmarks along with the standard deviation) are shown in Table 51.

Cell Type	PQ		SQ		DQ	
	mean	SD	mean	SD	mean	SD
connective	0.38	0.03	0.70	0.05	0.48	0.04
eosinophil	0.06	0.02	0.11	0.04	0.08	0.03
epithelial	0.48	0.07	0.69	0.07	0.62	0.09
lymphocyte	0.46	0.11	0.73	0.12	0.53	0.12
neutrophil	0.02	0.01	0.05	0.04	0.03	0.02
plasma	0.15	0.08	0.36	0.14	0.17	0.09

Table 51. HEIP pipeline accuracies determined with cross-validation, based on 6115 patches of the Lizard dataset.

Our validation results for HEIP are similar compared with other nuclei segmentation methods validated on the Lizard dataset. Graham and colleagues report PQ values between 0.158 and 0.559 depending on the cell type (Graham, et al., 2021).

3.4.3.4 Cross-validation on HGSC

In addition to the validation on colorectal cancer samples, we also showed the usability of the HEIP pipeline in different cancer types by validation on two high-grade serous ovarian cancer (HGSC) patient sample datasets. All data originated from the Turku University Hospital and was annotated by a specialized pathologist.

The first validation dataset was the CellTypeValidation dataset which contains 20 ROIs extracted from H&E images of 19 HGSC tissue samples. On this dataset, HEIP showed a PQ of 0.75, DQ of 0.88, and SQ of 0.85. In accordance with the validation on colorectal cancer images, for HGSC images, the best performance was observed in detecting neoplastic and epithelial cells, whereas the detection of connective and inflammatory cells was lower.

The second validation dataset was the TumorSiteCellValidation dataset that contained 36 ROIs located at the interface between tumour and stroma tissue of 18 H&E images of HGSC patients, including different tumour anatomical sites, like omental, peritoneal, and tubo-ovarian tumours. On this dataset HEIP achieved a PQ of 0.72, DQ of 0.80, and SQ of 0.90 for detecting neoplastic nuclei in omental tumours.

The complete HEIP model, including its training and validation on HGSC samples was recently published in the Journal of Pathology Informatics (Ariotta, ym., 2023).

3.4.3.5 Summary

Our validation results for HEIP are similar compared with other instance segmentation methods. This indicates that HEIP is a reliable tool for the segmentation and annotation of different cell types from cancers, including colorectal cancer. HEIP is particularly well-suited for identifying and distinguishing lymphocyte and epithelial cells in colorectal cancer. This is important as the lymphocyte abundance is clinically meaningful. It is important to note that eosinophils, neutrophils, and plasma are quite scarce in the samples.

3.4.4 Survival Rate Prediction

The basis of the survival model is the DL model HEIP, which extracts features from histopathological images that are used in survival prediction. As the survival model is dependent on the quality of the features extracted by HEIP, our validation efforts focused on the segmentation performance of HEIP as described above.

For the development of our Survival prediction model, we used The Cancer Genome Atlas (TCGA, https://www.cancer.gov/ccg/research/genome-sequencing/tcga), specifically the TCGA-COAD dataset, which contains clinical and histopathological data from patients with Colorectal Adenocarcinoma. The survival prediction model was developed through a rigorous 10-fold cross-validation. Furthermore, as detailed in Deliverable 4.3, the model underwent further scrutiny by validation with a subset of TCGA samples that were not used in the training of the model. In D4.3, we present the prediction accuracy of our model using this independent validation dataset as this better reflects the model performance.

The challenges of validating DL models based on histopathological images are stated in 3.4.3.1. Validating survival prediction models presents further specific challenges as not only high-quality histopathological images but also the accompanying survival data are necessary. Typically, survival data sets in medical research are limited in size, and there are concerns regarding data quality and completeness with long follow-up times, as well as data protection issues.

Validating the models using an external data set is crucial to assess its generalizability, obtaining external datasets is next to impossible. For the purposes of the blind evaluation, we therefore chose to apply the Survival prediction model on a set of 70 randomly selected colorectal cancer samples from the TCGA data, which were not used before (neither in the training nor the test phase). Survival times from the TCGA-COAD dataset were binarized into good (progression-free survival (PFS) > 2 years) and poor (PFS <= 2 years).

The results of the evaluation can be found in Table 52.

Predicted/ True	Poor	Good
Poor	27	13
Good	16	14

Table 52: Confusion matrix for H&E image-based colorectal cancer survival prediction on an additional set of TCGA data.

On this additional set of TCGA data, the Survival prediction model achieved an accuracy of 59%, a sensitivity of 63%, a specificity of 52%, and an F1-score of 65%. These results are similar to the results presented for the test set in D4.3, as is to be expected. The accuracy, sensitivity, specificity, and F1-score are only slightly lower compared to the results on the test set in D4.3. The results have been discussed and put into the context of the existing literature in much detail in D4.3.

3.5 Analysis of discrepancies in model prediction and annotations

The observational study protocol includes analysis of model errors whenever discrepancy exist between the model predictions and annotations/labels set by the participating HCP within DPs during data collection process. The reasons for this discrepancy can be twofold:

- Models have been developed based on limited amount of data and limited/selected information on patient data, technical teams are aware that in many cases the prediction errors are indeed due to model imperfection and a need to increase and diversify the knowledge base with more data sources and involve more cases in the training process.
- erroneous inputs and human errors are possible in the time-consuming data collection process, both during clinical data collection and image annotation. The letter is even more pronounced in the demanding task of image annotation, where using new software tool and different annotation labels on the same image, can lead to accidental errors in this labour-intensive process.

In the process of model development, INCISIVE technical teams have been aware of the possibility of errors in input data/labels and thorough analysis of collected data in INCISIVE image repository has indicated that revision of expert annotations might be needed. As reported in D4.2 and in within this deliverable in 2.3, within INCISIVE project examination of the provided annotations in previous and all running data collection cycles has been done under the umbrella of T6.3/WP6.

For each cancer type evaluators and leading radiologist were nominated from the data provider providing mostly data for this cancer type: breast cancer (UNS), lung cancer (UoA), prostate cancer (IDIBAPS), and colorectal cancer (AUTH). The leading radiologist has prepared the listing of inconsistences, wrong usage of the annotation tool, and all sort of disambiguates, e.g. in cases when:

a) More labels were suggested, where certain phenomena could be treated under two labels (e.g., in breast mammography micro-calcifications could be labelled as micro-calcifications or as benign or malignant, depending on their nature)
b) If larger variability in number and presentations of lesions and other specific morphological changes is possible:
c) In some cancers and modalities, the annotation process required consistency in annotating all noticeable changes in an image, which was sometimes difficult to achieve under heavy workload
d) In some modalities annotation of only the largest lesion was the only feasible option and this lesion was described in the clinical template, while the presence of other changes was only indicated in the comprehensive clinical template.
e) If not straightforward, how are the borders of the lesion determined (radiologist had different approaches).

The procedures followed for all workshops, for each cancer types is described in the process in Figure 25. The main results achieved are **the new refined sets of guidelines for each cancer type,** that resolved many ambiguities that have been identified during the annotation process. Moreover, all image annotations have been double checked and corrected to conform to the new guidelines, forming a set of consistently annotated images.

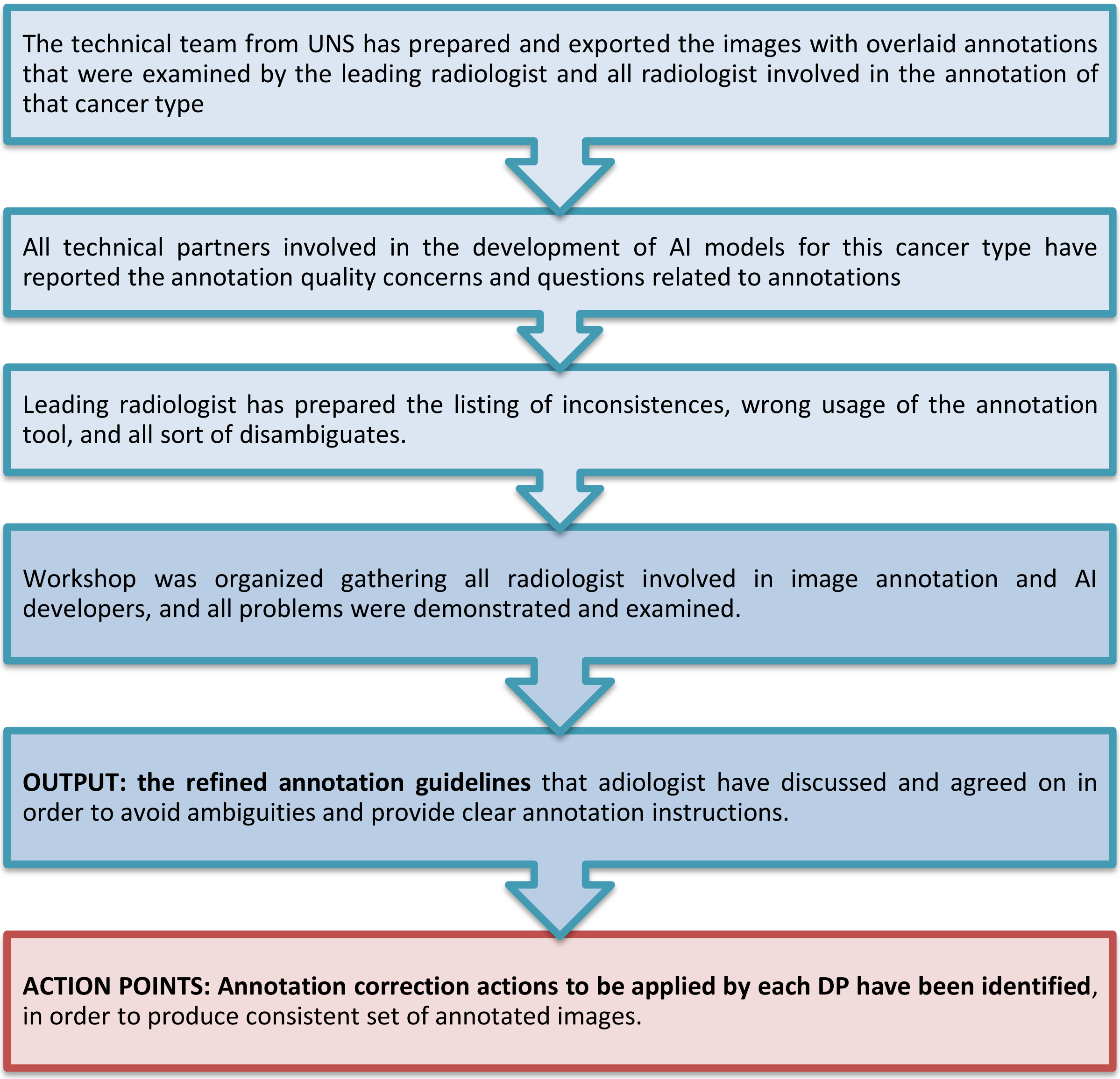
These corrective procedures have facilitated both AI training during the final models development phase, but as well contributed to an improved quality of annotations collected in the observational pilot study. For these reasons, annotations made by radiologists and other labels taken from clinical records (such as cancer stage, cancer type etc), can be considered trustworthy, due to the curation processes and quality checker control in the data preparation process.

However, to align with the observational study protocol, we have performed additional validation of radiologist annotations during the evaluation of the final AI models in the observational pilot study.

Having in mind all efforts already made in the manual review and correction of annotations by both the participating clinicians and technical supporting teams, we had no resources to rigorously manually examine all annotations again during the model evaluation. Additionally, the manual inspection process requires two external radiologists whose time is very limited. For these reasons we have selected two segmentation models and one classification model, for which the expert annotations/labels provided will be re-examined manually and compared to the model predictions. This type of models has been selected as they are the most sensitive to errors introduced in the annotation/data collection process, and in general being the most challenging to learn. Segmentation models learn directly from the radiologist annotations in the training process, and here both predictions and expert annotations have been evaluated in order to assess the percentage of residual annotation errors.

The selected models are: (1) suspicions lesion localization form MMG images in breast cancer, described in 3.2.4, (2) lesion segmentation in breast MMG images, described in 3.2.5; and (3) lung cancer staging based on clinical data and CT scans, described in 3.1.4.

The selection of two different models for breast cancer MMG images has been made since MMG annotation has proved as the most challenging in terms of number of labels available (thus higher possibility of error), inter-reader variability and number of corrections that were made. This additional validation step was needed in order to examine if the refined annotation guidelines for breast cancer MMG have resolved ambiguities and improved annotation quality. Lung cancer staging model is challenging as it includes both clinical inputs and image features and makes predictions with limited patient information (for some cancer stages more clinical information would lead to more accurate predictions).



The technical team from UNS has prepared and exported the images with overlaid annotations that were examined by the leading radiologist and all radiologist involved in the annotation of that cancer type

All technical partners involved in the development of AI models for this cancer type have reported the annotation quality concerns and questions related to annotations

Leading radiologist has prepared the listing of inconsistences, wrong usage of the annotation tool, and all sort of disambiguates.

Workshop was organized gathering all radiologist involved in image annotation and AI developers, and all problems were demonstrated and examined.

OUTPUT: the refined annotation guidelines that adiologist have discussed and agreed on in order to avoid ambiguities and provide clear annotation instructions.

ACTION POINTS: Annotation correction actions to be applied by each DP have been identified, in order to produce consistent set of annotated images.

Figure 25. The procedure followed for each annotation workshop: preparatory activities, refined guidelines as the main workshop output, and set of corrective actions for annotation corrections to confirm to the new guidelines

3.5.1 Aanalysis of Lesion Localization Service in Breast Cancer

The lesion localization service is mainly a supporting diagnostic service, designed for the low-cost mammography, with an aim of detection of any suspicious lesion in the breast tissue. The rigorous service evaluation against the radiologist annotation using pixel-based metrics has been reported in the section 3.2.4. We have amended these results with the adjusted set of metrics, object-based metrics, which are better suited when humans are in the loop, as in this scenario where the service is used as a decision support. Namely, this set of metrics does not favour exact pixel overlap, but rather at least 25% of overlap of predicted lesion surface with the annotated surface, as the main outcome should be to focus the attention of the radiologist on the potentially suspicious area.

In this section we present the error analysis results on the randomly selected subset of 457 MMG images, with the help of two external evaluators who examined all cases when prediction and existing annotation were not aligned. This should serve additional re-examination of annotation quality, and as well potentials of the model to detect some additional lesions, that might have been overseen in the annotation process.

For the lesion localization service, the randomly selected set of 457 images from all DPs involving both images with and without lesions, overall, 331 images with at least one lesion and 126 images without any lesion annotated, as summarized in Table 53. The potential errors include: omission of all lesions leading to FN, omission of at least one lesion, but detection of some lesions (TP), prediction of additional lesions besides the annotated lesions (TP) and prediction of one or more additional lesions in images where no lesions were annotated (FP).

In total some form of disagreement between predicted localizations and annotated lesions has been found in 101 out of 200 images from HCS, 98 out of 240 from UNS and 4 out of 17 from AUTH. Analysis of these disagreements in Table 53 indicates that omission of a lesion is more common model mistake, affecting approximately 1/3 of images with a lesion.

External radiologists review of the expert’s annotations provided with the images provided an overall high quality of annotations, where only in 8 images, out of 457 images some smaller lesions are omitted by the annotating radiologist. It is worth noting that manual annotation process is very tedious and time-consuming task, especially when radiologist is not accustomed to an annotation tool. The statistics observed in this set of images show the large improvement in the annotation consistency after the annotation workshop corrections process. In this subset of MMG images, no inadequate expert annotations have been found, all existing annotations indeed related to some suspicions lesion, while in 8 out of 457 images (1.75%) it happened that some of seemingly benign lesions have not been annotated by the expert.

Besides some successful examples, already shown in Figure 12 where quantitative model performance was expressed for large corpus of images, here we present some images where the model output deviates from annotations. It is worth noting that original annotations were contours, here converted into minimum escribing box, in order to match the detector box predictions. Figure 26 presents two examples where in panel A) the model correctly identifies the lesion, but the box alignment is not perfect; while in panel B) there is an example of a true positive, where only one lesion has been identified, while three lesions are missed. By the opinion of both radiologists involved in the project, and external experts, the cases such as the one in Figure 26 A) are true positive successful cases where, even though the boxes do not perfectly overlap, the model prediction conveys the correct information about localization of suspicious lesion.

DIFFERENT TYPE OF PREDICTION ERRORS									
Data provider	#MMG images			CORRECT DETECTION		AT LEAST ONE LESION MISSED		ADDITIONAL LESIONS PREDICTED	
	total	P	N	TP	TN	FN	TP	TP	FP
UNS	240	165	75	109 (66.06%)	66 (88.00%)	56 (33.94%)	28	5	9 (12%)
AUTH	17	10	7	7 (70.00%)	7 (100%)	3 (30%)	0	1	0
HCS	200	156	44	95 (60.89%)	31 (70.45%)	61 (39.10%)	20	10	13 (29.55%)
TOTAL	457	321	126	211	104	120	48	16	22

Table 53. Number of images in the evaluation of the MMG lesion localization service and types of errors observed

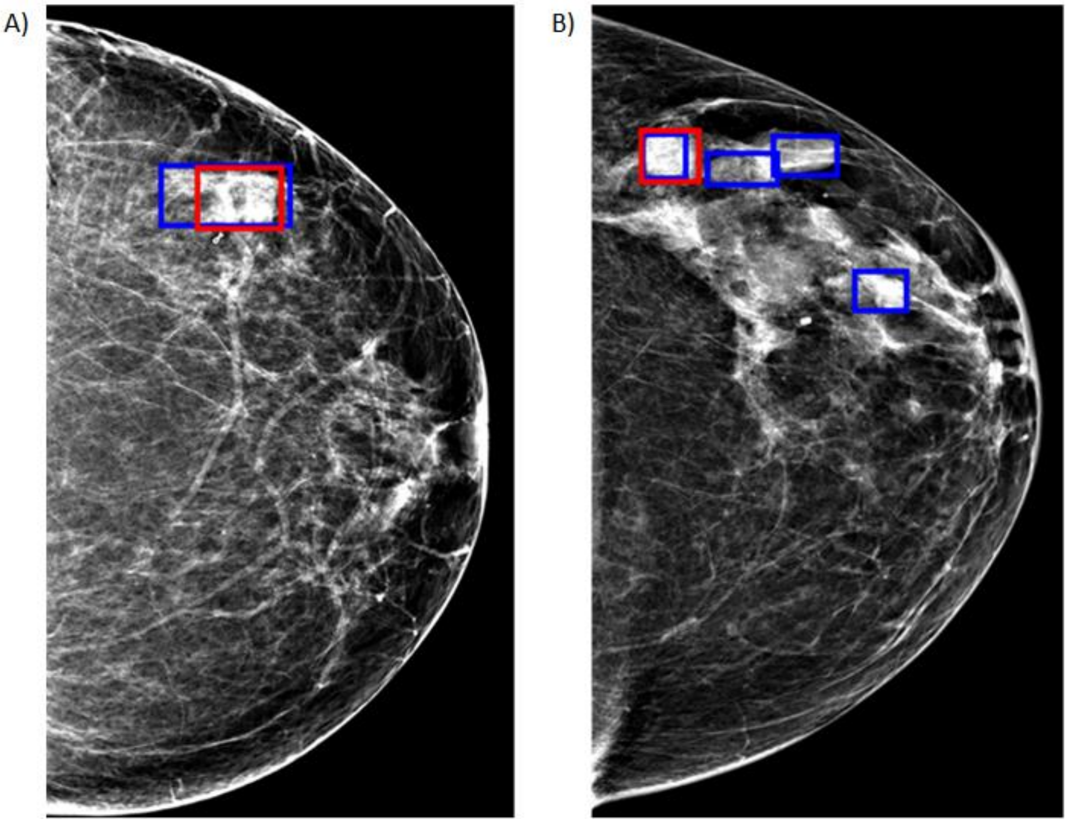


Figure 26. Some examples of discrepancies in the annotations (blue box embedding the annotator's contour) vs. model prediction (red box): A) the examples where both indicate the same region, but the area differs, which is not a critical issue, as the relevant area has been correctly identified; B) examples of missed regions, where model correctly identifies one of four annotated regions.

In Figure 27 two examples of cases where original annotated lesions are very small, and went undetected, but some additional suspicious lesions, according to external reviewers are identified by the model.

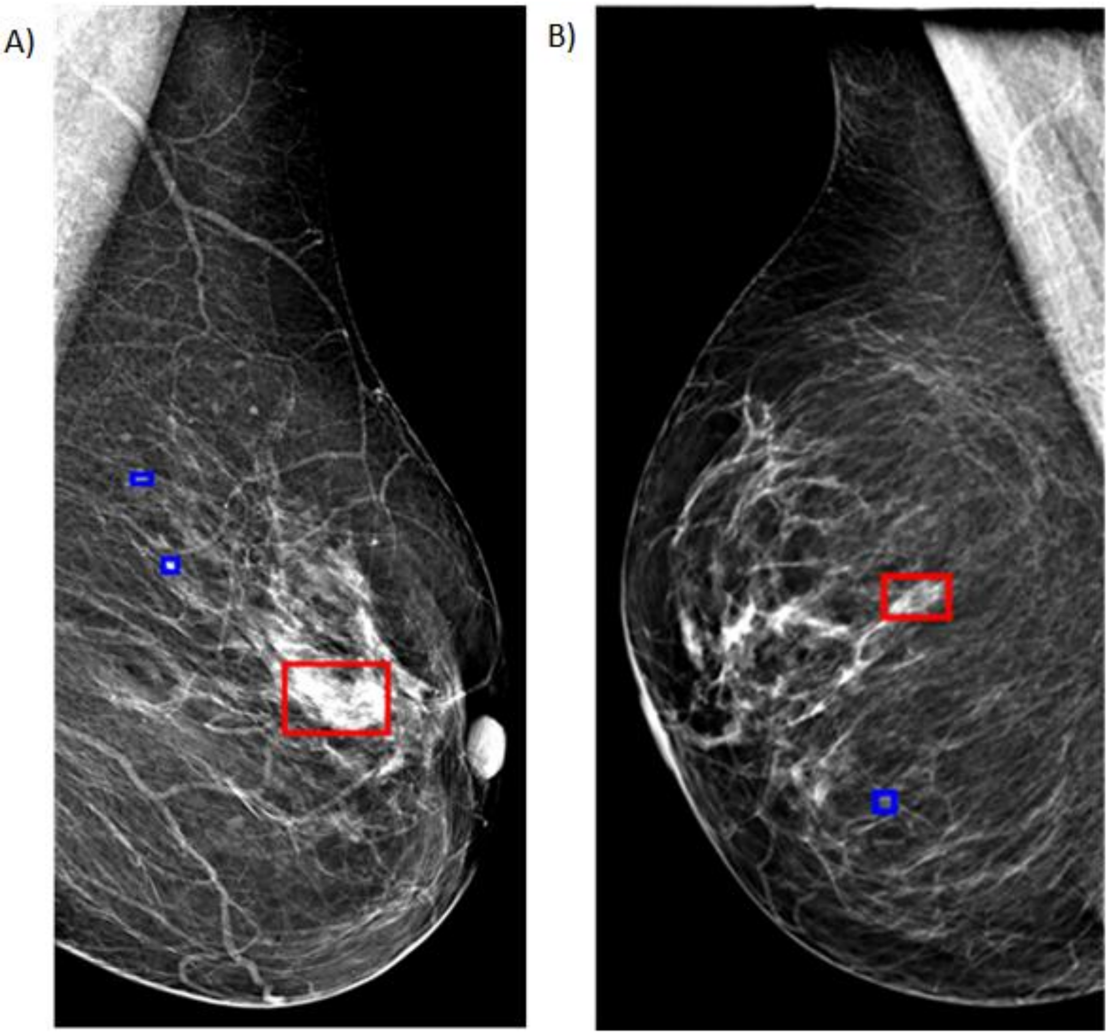


Figure 27. Examples where the model misses annotated lesions of very small surface, but still indicates some lesions that are suspicious, i.e. provides meaningful and useful output, and generated a true positive result.

3.5.2 Aanalysis of Lesion Segmentation Service in Breast Cancer

The segmentation service is based on the model MiT-B3-G1/2 trained on the images with any suspicious lesion, i.e. all lesions marked as "malignant", "benign" or "suspicious" were taken into account, without differentiation and presented as "suspicious" lesion during the training process. We have randomly selected 160 images from HCS, 101 from UNS and 17 from AUTH all with the presence of at least one annotated lesion (positive examples) in order to evaluate the segmentation model performance in identification and segmentation of the lesions present in the MMG images. The manual inspection of model prediction against the existing annotation has been performed by two external radiologists and the performance has been summarized in Table 54.

Data provider	#MMG images	ALL ANNOTATED LESIONS SEGMENTED	AT LEAST ONE ANNOTATED LESION MISSED		ADDITIONAL LESIONS SEGMENTED (among TPs)	
	WITH LESIONS	TP	TP	FN	SUSPICIOUS LESIONS	FALSE LESION PREDICTIONS
UNS	101	83 (82.17%)	18 (17.82%)	0	15 (14.85%)	17 (16.83%)
AUTH	17	13 (76.47%)	3 (17.65%)	1 (5.88%)	3 (23.08%)	0
HCS	160	117 (73.13%)	39 (24.38%)	4 (2.5%)	41 (25.63%)	18 (11.25%)
TOTAL	278	213	60	5	59	35

Table 54. Number of images in the evaluation of the MMG lesion segmentation service where the predictions were aligned with the annotations and when some type of discrepancy has been noticed. The percentages are provided with respect to the total number of images per DP reported in the second column.

Overall, the external reviewers have examined all images and expressed satisfaction with the correct segmentations produced. Among 278 randomly selected images from 3 different INCISIVE DPs in 213 images segmentation of all annotated lesions was successful. It is worth noting that not an ideal overlap between the annotated and predicted contour has always been noticed, but all lesions contoured by the medical professionals were identified and to a large extent segmented by the model. The reason why external reviewers adhered to lose acceptance values is large intra-reader and inter-reader variability (in 3 DPs more experts were involved in annotation process).

Discrepancies noticed between expert annotations and model predictions belonged to three categories not mutually exclusive: (1) at least annotated lesions were not segmented by the model; (2) expert annotation was inadequate; and (3) some additional lesions were segmented by the model. In cases when one or more annotated lesions were missed most often it was one of more lesions annotated, and in a very small percentage it led to false negative outcome, i.e. no lesion was predicted (Table 54). In all of these cases expert annotations referred to all suspicious regions. Some examples are provided in Figure 28 B) and C), where more lesions are present, and the model misses one in B) and three in the example under C). It can be noticed that more similar areas are present in the breast tissue, but no false detections by the model have been made. Figure 28 A) illustrates one of the rare cases (5 in 278 randomly selected images) where the expert annotation was inadequate, in this case as shifted from the correct placement, while in some other cases the annotation was not corresponding to the image, but probably to some other image. For these cases we have informed the DP and error correction process has been initiated. In cases when some additional lesions were detected (which can happen alongside with missed lesions) in many cases those were benign/suspicious lesions that were not annotated; Figure 29 illustrates some of these cases in panels A), B), C) and D). In panel D) a very challenging expert annotation has been successfully reproduced. However, even though some benign lesions predicted by the model have not been annotated, experts have annotated some other lesions suspected of malignancy as the top priority.

In some other cases of additional lesion detection, false predications have been made, as according to external radiologists, these were some areas where the breast tissue is dense, and image preprocessing has further amplified intensities. Those false predicted areas were usually small.

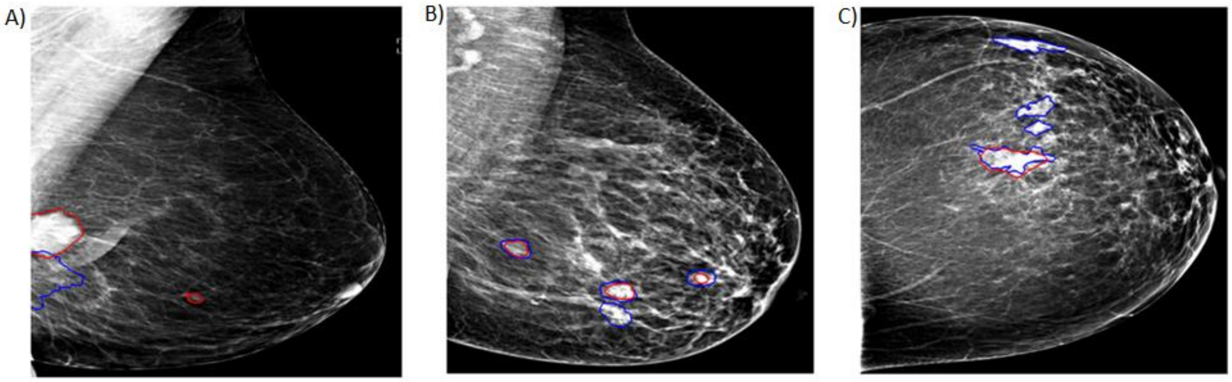


Figure 28. Examples of discrepancies between the segmenter model (red contours) and expert annotations (blue contours): A) expert annotations seam to bi misaligned with the lesion, while model predicts two suspicious lesions correctly; B) model predicts three out of four annotated regions; C) model predict one out of four annotated regions.

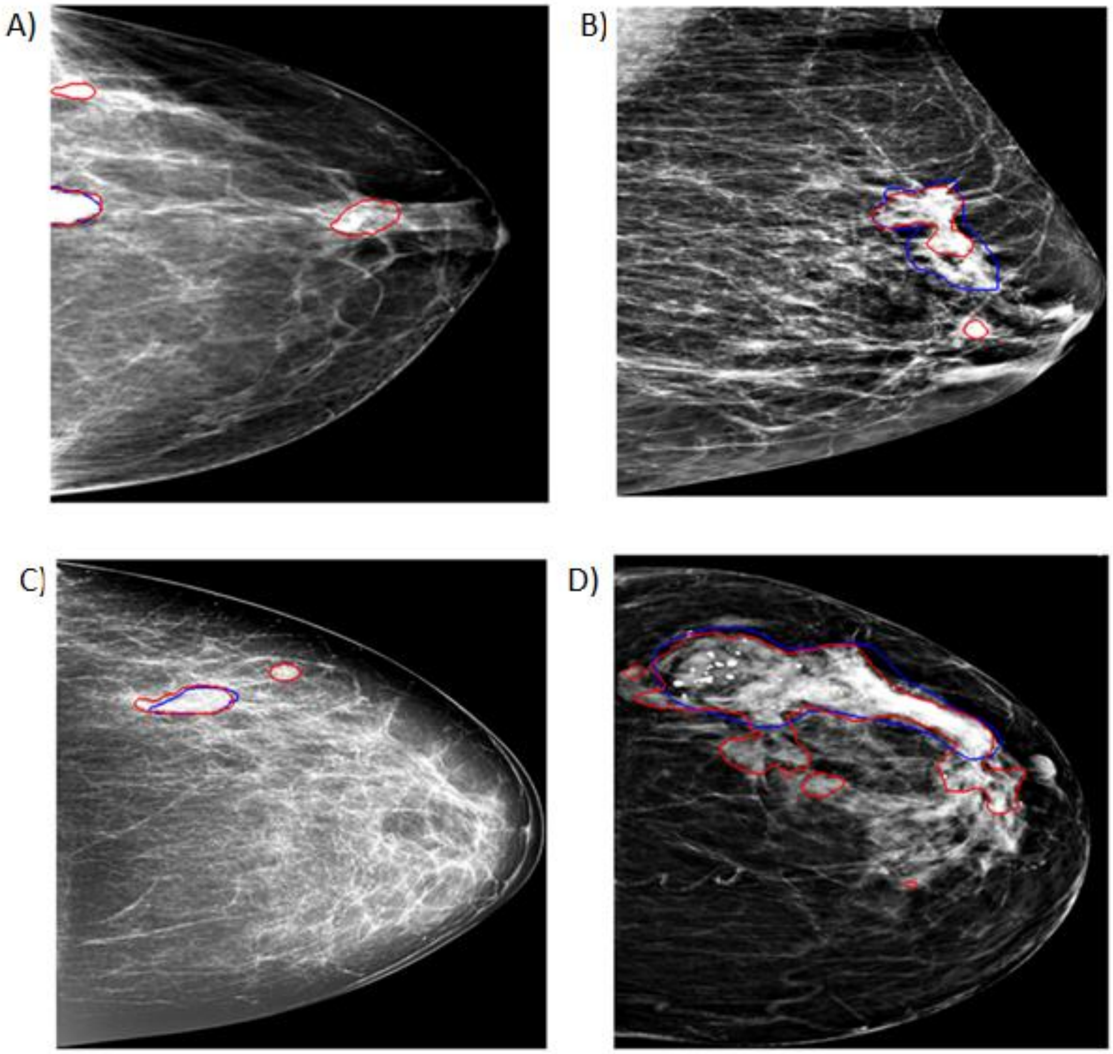


Figure 29. A), B) and C) examples where the segmentation model contours (in red) identify some additional suspicious lesions, that went unlabelled by un expert annotator; D) an example of the challenging large lesion successfully detected by the segmentation model (in red). Contours in blue are the expert annotations.

3.5.3 Analysis of Cancer Staging in Lung Cancer

As part of the evaluation of the cancer staging model, we also considered its evaluation by a clinical expert and thorough examination of prediction errors. In more details, a radiologist with expertise on lung cancer was asked to review a proportion of cases that was used for the evaluation of the model in order to perform the staging only by the inspection of the CT images but also to consider any additional information that is available for each patient (e.g. lab results or other than CT imaging examinations). The expert's staging results were then compared with the outcome of the classification model.

According to the cases used for the validation of the model, the ground truth stage for 11 patients was I or II (class 0) and for the rest (44) the stage was III or IV (class 1). Based on the model prediction, all the patients with class 0 were classified correctly while all the errors were related to the patients with ground truth class 1 (Table 17).

For the clinical evaluation of the model 10 cases were randomly selected from the patients assigned to class 1. Those cases included cases where the prediction was made correctly as well as cases where there is a misclassification based on the clinical characteristics available in the clinical template of the INCISIVE repository. All the cases were from AUTH.

patientID	Model prediction	Ground truth	Clinical evaluation			
			T	N	M	Overall staging
001-0000246	0	1	1	2	1B	IVA
001-0000254	1	1	4	3	1C	IVB
001-0000237	0	1	1	2	0	IIIA
001-0000251	1	1	Tx	3	0	IIIB
001-0000248	1	1	4	2	0	IIIA
001-0000008	1	1	4	2	0	IIIB
001-0000013	0	1	2a	3	0	IIIB
001-0000026	0	1	2	2	1	IVA
001-0000253	1	1	3	2	1	IVA
001-0000244	1	1	3	2	0	IIIB

Table 55. Clinical evaluation of lung cancer cases. Tx means Tumor that is proven histopathologically but cannot be assessed or is not demonstrable radiologically or bronchoscopically.

After discussing and analysing the errors with the support and feedback from clinical experts (Table 55), it was suggested that the model weakness is to predict the cases that are stage III or IV due to local or distant metastasis. In previous studies CT radiomics biomarkers were proposed for staging (Yu, 2019). CT radiomics biomarkers were also proposed for the prediction of mutations (like EGFR mutation status (Digumarthy SR, 2019)) that relate to metastasis. In order to enhance this part and give more focus on the metastatic properties of the tissue, ongoing efforts test the introduction of additional features (related to the position of the tumor) that could increase the chance for infiltration, and the properties of tissue around the tumor and in the organ perimeter, and in following steps extending the focus to include nodules, etc.

4 Discussions

The observational study has been conceived as a prospective study aiming for an objective evaluation of the INCISIVE AI services using additionally collected data (i.e. data unseen by the models behind AI services). The models have been developed using INCISVE data collected in the retrospective data collection stage and open data, aligned with the INCISIVE needs. This data has been used for model training, model selection and model evaluation using different performance metrics, as described in D4.3. All other data collected prospectively by all DPs could have been used for the blind evaluation of the developed models, integrated in INCISIVE platform behind AI services.

The observational study has been undertaken as planned within the parameters of the protocol, with adjustments where necessary. The main objective of the study – a quantitative evaluation of INCISIVE AI models on new data collected during the study has been achieved, and additionally new data have been collected and curated enriching the INCISIVE repository. During the study efforts have been invested to:

- inspect, re-evaluate and correct image annotations,
- improve the tools providing the quality check of adherence to guidelines related to structured and standardizes way of clinical data upload
- preserve the information on timeline and connections between the images and clinical data
- redefine and correct the accidental errors in folder structure for all image modalities in the repository
- development of the automated tools to control the folder structure, correct certain issues and generate error report as feedback to data provider with instructions on error correction.

In this way, we have managed to identify and correct to the largest extent errors due to human related factors.

Both development and evaluation of the models in the setting with multiple data providers, enriched with extensive open data sets, stressed the importance of data quality at the input and relevance of heterogeneous data volumes. The quantitative evaluation of models has indicated that stable and robust model performance in cases of models developed for image modalities having the largest representation within INCISIVE repository. In some more demanding tasks, such as disease characterization, we had to rely on fused imaging and clinical data and extensively analyse and revaluate different type of errors. These investigations have contributed set of directions for future developments beyond the project lifetime.

We have managed to evaluate 15 models for four cancer types, while we had no specific MRI sequences uploaded only for the colorectal lesion segmentation model. The number of cases to estimate generalization potential (performance) of the developed in most cases surpasses the minimum number of cases needed, as calculated during the protocol design (Table 1). For lung cancer 435 Xray and 255 CT scans all INCISIVE cases were used (681 needed), for breast cancer 1426 INCISIVE MMG cases and 100 MRI cases from open datasets (6 services were evaluated in total using 4641 samples), for colorectal cancer HEIP model was tested in 7555 image patches from open data set, while survival rate prediction model was tested on 70 patients from open data set , and prostate cancer services were tested on 184 MRI INCISIVE cases and 100 cases from open data set for prostate gland segmentation service.

The observational study has been initiated before its official start in M25, since it was considered that time is needed to recruit the patients, especially in those DPs where more cases have been foreseen. However, despite this risk mitigation measure, the data collection was slower than expected although this is not an uncommon issue is clinical studies of this kind. The main obstacles in the process were due to external reasons: pace of new patients' arrival in the participating hospitals, and patients' reluctance to engage in conversation related to participation in the study. Informing and consenting of the non-cancer participants and patients already under cancer treatment has also been done.

Overall, the experiences in all data collection stages imply that the process is time consuming and dependant on multiple external factors. Within INCISIVE we have experienced and overpassed different problems at all levels in data collection process: with the procedures related with ethical approvals, in patient recruitment, the availability and accessibility of patient data in different health information systems, the limited availability of clinicians/experts for annotation process and the time-consuming process of clinical data extraction. Additionally, many project related internal factors have interfered with the pace of the data collection in the observational study stage: lack of technical skills and support in some DPs, the need to transfer the INCISIVE repository amidst data collection from a temporary centrally based infrastructure to the final federated repository. Within INCISIVE all efforts have been directed towards overcoming difficulties and provision of all technical support to DPs: from detailed data collection guidelines to the infrastructure.

In INCISIVE we have been dealing with demanding disease characterization, covering several cancer types using different imaging modalities and different morphological presentations. These ambitious goals required collection of multitudes of different types of data from multiple sources, preferably perfectly aligned in time with the start of WPs that dealt with AI model development. What we have learned from the flow of the observational study is:

- data collection process has to be conceived and initiated as early as possible in the project lifetime
- technical support and basic technical skills related to data manipulation are needed in partners acting as DPs, as these skills facilitate smooth data collection progress
- data harmonization step is crucial for development of AI models with increased generalization capabilities, as image appearance is related to image acquisition protocol and quality of the equipment used, thus use of heterogeneous data sources is preferable in the training phase.
- Data curation and revision of the image annotation process, done during observational study, have been important in the final model development, and have significantly improved overall quality of the INCISIVE repository.
- having in mind duration of both the project and the observational study, focus on one or two cancer types might have resulted with more aligned data types and more mature models, as the smaller focus would allow more in-depth analysis.
- In this project we managed to overcome differences in understanding due to different backgrounds among medical and technical professionals. As technology spreads and determines the progress in medical field, it is relevant to work on mutual comprehension and technical education of HCP in using and understanding advantages and limitations of AI supporting tools and assistive technologies.

The outcomes of the observational study have shown that despite these difficulties, we have managed to deliver the set of AI services and the platform introducing the potentials of image-based AI models for different image modalities and different cancer types. The main bottleneck was slow paced of data and lack of some data types, such as lack of specific MRI sequences and histopathology images in the requested format, which had a largest impact on the colorectal cancer services. On the other side, in the observational study some DPs have provided predominantly one data type, which was lacking in the retrospective studies and for which the corresponding services could not be developed (such as PET/CT). For image types where the data was abundant, such as breast MMG, more in depth analysis was possible and more robust and mature models have been delivered.

5 Conclusions

The prospective observational study provides a quantitative performance assessment of models running behind AI services offered in the INCISIVE platform. The evaluation has been performed on the data collected in prospective INCISIVE studies, different to the data used for training, model selection and evaluation during the model development phase (as reported in D4.3). In certain cases, when possible, open data was used to increase the data volume and provide more realistic performance assessment.

The evaluation has showed similar or decreased performance in most of the models, and stressed the importance of image quality and harmonization of image appearance whenever services are envisioned to accommodate data from heterogeneous sources, which is the vision of the INCISIVE platform. Within observational study enormous efforts have been invested in data curation and annotation/label quality assessment and improvement.

Altogether the results of the study indicate that the AI services are ready for their external validation by HCPs in feasibility study, as initial limited validation with external experts done within observational study has showed their overall satisfaction with the performance and their understanding of advantages and limitations of machine learning/data driven supporting tools.

6 References

Anindo Saha, J. B. (2022). Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI: The PI-CAI Challenge (Study Protocol),.

Anna Jenul, S. S. (2021). RENT—Repeated Elastic Net Technique for Feature Selection. *IEEE Access, 9*, 152333-152346.

Ariotta, V., Lehtonen, O., Salloum, S., Micoli, G., Lavikka, K., Rantanen, V., . . . Hautaniemi, S. (2023). H&E image analysis pipeline for quantifying morphological features. *Journal of Pathology Informatics, 14*. doi:doi.org/10.1016/j.jpi.2023.100339

Digumarthy SR, P. A. (2019, Jan). Can CT radiomic analysis in NSCLC predict histology and EGFR mutation status? *Medicine (Baltimore), 98*(1), p. e13963.

Fanny Orlhac, J. J.-S. (2022). A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies. *Journal of Nuclear Medicine, 63*(2), 172-179.

Gamper, J., Koohbanani, N. A., Benes, K., Graham, S., Jahanifar, M., Khurram, S. A., . . . Rajpoot, N. (2020). PanNuke Dataset Extension, Insights and Baselines. *ArXiv*.

Graham, S., Jahanifar, M., Azam, A., Nimir, M., Tsang, Y.-W., Dodd, K., . . . Rajpoot, N. (2021). Lizard: A Large-Scale Dataset for Colonic Nuclear Instance Segmentation and Classification. *ArXiv*.

Kirillov, A., He, K., Girshick, R., Rother, C., & Dollar, P. (2019). Segmentation, Panoptic. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9404-9413.

Saha, A., Harowicz, M. R., Grimm, L. J., Kim, C. E., Ghate, S. V., Walsh, R., & Mazurowski, M. A. (2018). A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. *British journal of cancer, 119*(4).

Yu, L. T. (2019, 19). Prediction of pathologic stage in non-small cell lung cancer using machine learning algorithm based on CT image feature analysis. *BMC Cancer*, p. 464.

ANNEX 1. Pre-validation studies

This section describes the updates on results obtained during the period in which the collection of observational phase data took place (M37). Moreover, the Integrity of the data is also reported here for all collections of data (retrospective, prospective and observational). This is an ongoing work performed for all (meta)data collected. Results on the retrospective and prospective data have been reported in D6.2 and D6.3. The strategy followed (**Design and implementation**) is explicitly described in D6.1 and D6.2.

In general, the aim of pre-validation studies is to establish a framework towards data quality and identification of error in an effort to finally obtain harmonized datasets for further use in AI-services development. All analyses performed target error identification with respect to crucial aspects of model development such as bias (via fairness analysis) and privacy (e.g. anonymization) issues. These analyses could lead to data correction (information was sent to corresponding DPs) and /or problem-dependent selection of appropriate subsets that fulfill certain criteria.

A) Clinical Metadata Harmonization results and findings

In this subsection results on Data Quality with respect to *1. Completeness, 2. Validity 3. Consistency, 4. Integrity, and 5. Data Fairness* is presented.

1. *Completeness*:

As previously described, *Data Completeness* refers to the comprehensiveness or wholeness of the data. The results refer to the percentage of the records that are present/complete based on the definition rendering them as mandatory. The patients, in this case, are considered as records. The completeness analysis is performed during month 37 (see **¡Error! No se encuentra el origen de la referencia.**) for the amount of data already available in the INCISIVE repository and in this case observational data).

Data are also presented by Data Provider (**¡Error! No se encuentra el origen de la referencia.**

Case	Breast	Lung	Colorectal	Prostate
Overall (%)	49.66	44.58	35.59	54.14

Table A1_ 1 Overall Data Completeness results for observational data – Month 37

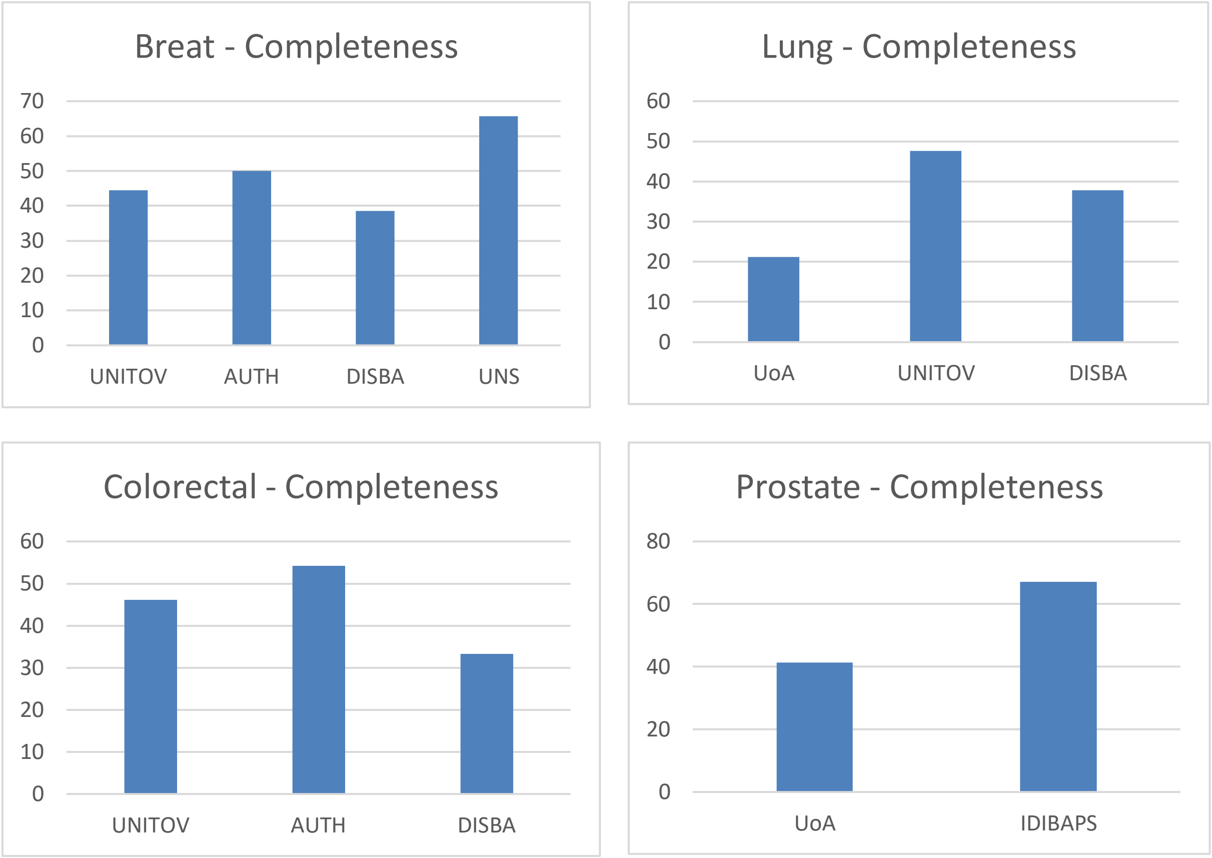


Figure A1_ 1 Data Completeness (%) results for each cancer type by DP.

Regarding the metric of case completeness, a correction needs to be made against what was reported in Deliverable 6.3, Annex 2: Pre-validation studies. In the measurements for retrospective and prospective studies, due to an error in the knowledge base, some fields of breast, lung and colorectal cases were not taken into consideration. The error was reported and fixed, so the correct measurements are depicted in the**¡Error! No se encuentra el origen de la referencia.**:

	Breast	Lung	Colorectal	Prostate
retrospective data Overall (%)	37.95	44.51	37.47	67.79

prospective data Overall (%)	45.97	41.16	42.88	63.87

Table A1_ 2 Data Completeness results (a) retrospective data; (b) prospective data – Month 32

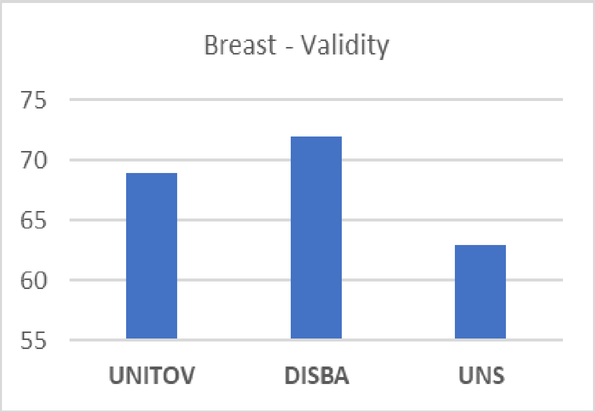
As shown in Table A1_1, completeness reaches the highest value which is in the case of prostate cancer (54.14 %). For lung and breast cancer the percentage of completeness is ~ 45-50%. When comparing with the retrospective or prospective phase, the results seem not to significantly improve indicating that there might be a difficulty in complying with the set instructions in collecting the clinical metadata or multiple sources of data (e.g. Health records, examinations performed elsewhere than the clinical site of the corresponding DP) to fill in the template during collection.

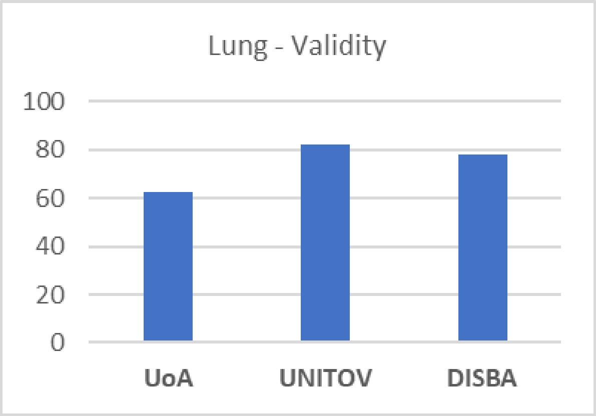
2. *Validity*:

Validity refers to how well data conforms to required value attributes based on specific rules set from the beginning of the study (format, allowable types and value ranges). The results refer to the percentage of records in which all values are valid (valid identified information). This metric is assessed during month 37 (**¡Error! No se encuentra el origen de la referencia.**), for the amount of observational data already available in the INCISIVE repository.

Patient	Breast	Lung	Colorectal	Prostate
Overall (%)	67.93	74.09	69.72	74.07

Table A1_ 3 Overall Data Validity results for observational data – Month 37.





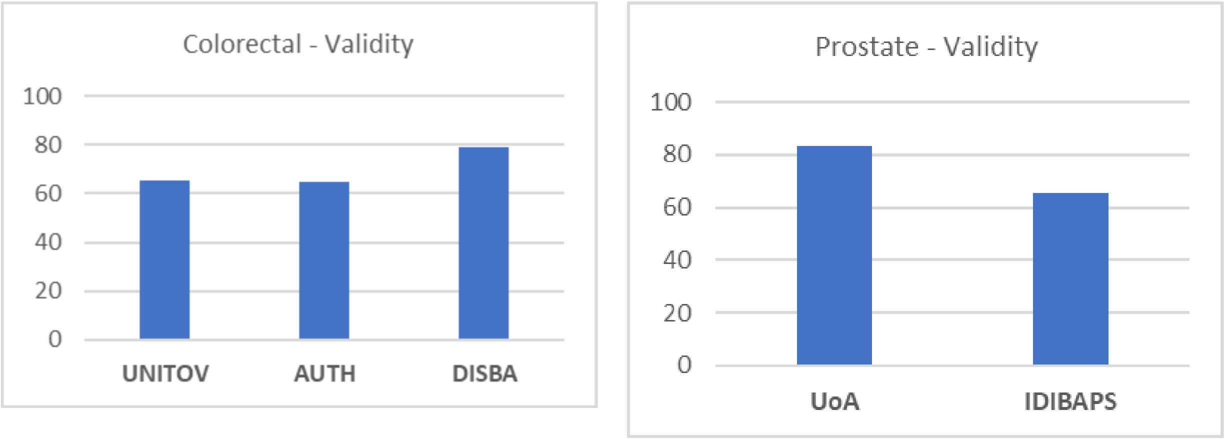


Figure A1_ 2 Data Validity (%) results for each cancer type by DP.

The results show that Validity percentage ranges from 67.93 to 74.09% which is slightly the same as in the case of prospective study further confirming the gradual familiarization and improvement of the data collection routine that it was observed when comparing the retrospective with prospective study results (TableA1_3 and FigureA1_2).

3. *Consistency*:

By the same token, data consistency refers to the process of keeping information uniform and homogeneous across different sites. The results refer to the consistency metric, which is the percentage of values that match across different records (**¡Error! No se encuentra el origen de la referencia.**).

Patient	Breast	Lung	Colorectal	Prostate
Overall (%)	75.14	54.05	40.57	-

Table A1_ 4 Overall Data Consistency results for observational data – Month 37.

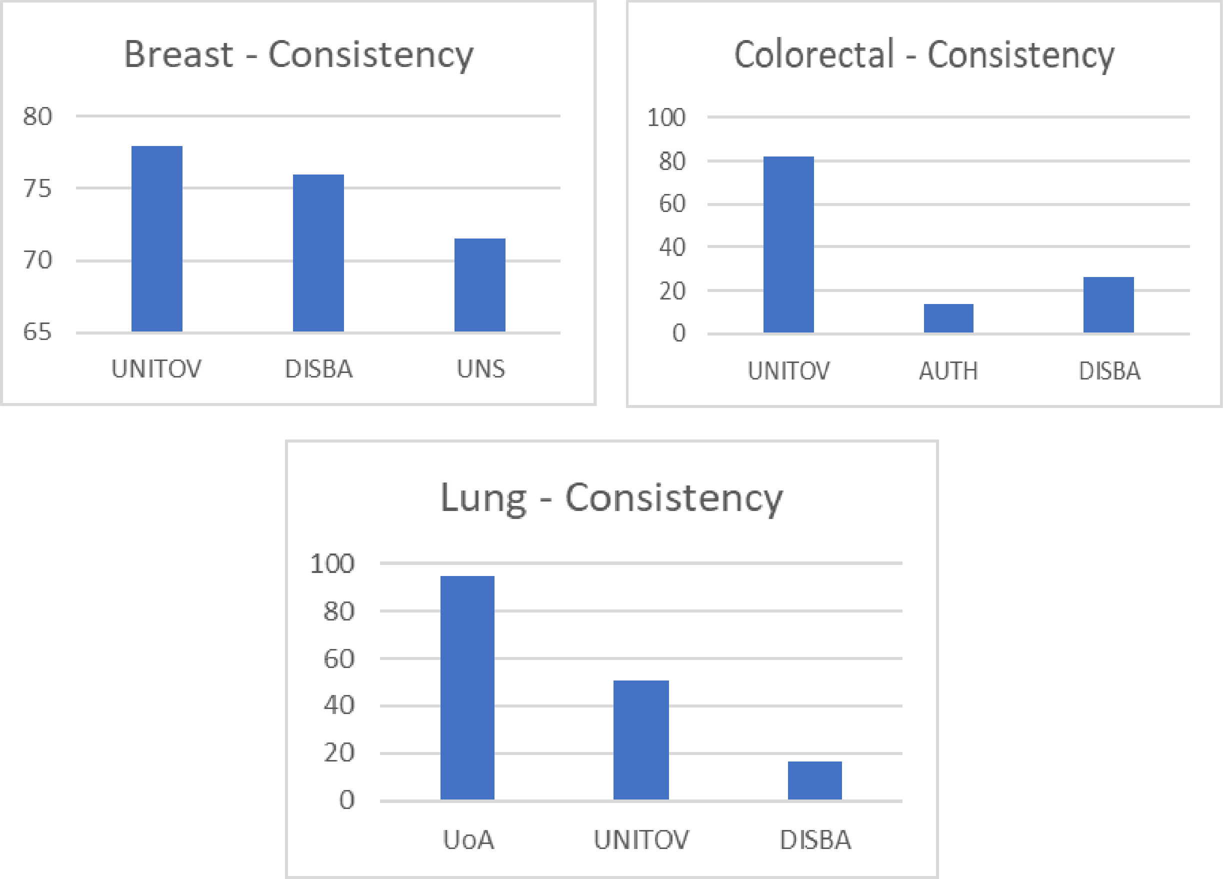


Figure A1_ 3. Data Consistency (%) results for each cancer type by DP.

The results show that consistency ranges from 40.57-75.14% with significant deviations among the DPs (**¡Error! No se encuentra el origen de la referencia.**).

4. *Data Fairness*:

Briefly, Data Fairness refers to data adequacy to be reliably combined in different ways and for various use cases. This metric is assessed for observational data for AI training and has been calculated for a) sex, b) grade, c) type and d) age presented below Tables A1_4-18) for each cancer type separately. Case: sex was not applied in the case of prostate cancer as all cases were male.

Breast

Age	*(20, 25]*	*(25, 30]*	*(30, 35]*	*(35, 40]*	*(40, 45]*	*(45, 50]*	*(50, 55]*	*(55, 60]*	*(60, 65]*	*(65, 70]*	*(70, 75]*	*(75, 80]*	*(80, 85]*	*(85, 90]*	*(90, 95]*
UNITOV	0	0	2	2	4	5	4	4	5	4	4	5	5	3	0
DISBA	0	0	0	0	2	4	4	3	0	2	1	2	2	0	0
UNS	0	1	1	0	2	4	3	3	2	4	4	2	1	0	0

Table A1_ 5 Data Fairness results, Case: Age, Cancer: Breast– M37.

Breast											
Cancer Type	*IDC*	*ILC*	*IPLC*	*IMC*	*IUC*	*IBC*	*MPT*	*SPC*	*HBOCS*	*UMN*	*DCIS*
UNITOV	22	1	0	0	0	0	0	0	0	0	0
DISBA	0	0	1	0	0	0	0	0	0	6	0
UNS	59	3	0	4	0	0	0	0	0	0	0

Table A1_ 6 *Data Fairness results, Case: Cancer type, Cancer: Breast– M37.*

Breast			
Cancer Grade	*1*	*2*	*3*
UNITOV	0	5	12
DISBA	0	2	5
UNS	9	51	6

Table A1_ 7 Data Fairness results, Case: Cancer grade, Cancer: Breast– M37.

Breast		
Sex	*Male*	*Female*
UNITOV	4	128
DISBA	0	23
UNS	0	35

Table A1_ 8 Data Fairness results, Case: Sex, Cancer: Breast– M37.

Colon															
Age	*(20, 25]*	*(25, 30]*	*(30, 35]*	*(35, 40]*	*(40, 45]*	*(45, 50]*	*(50, 55]*	*(55, 60]*	*(60, 65]*	*(65, 70]*	*(70, 75]*	*(75, 80]*	*(80, 85]*	*(85, 90]*	*(90, 95]*
UNITOV	0	0	3	2	1	2	4	3	5	5	5	5	5	4	0
AUTH	0	0	0	0	1	1	2	0	2	2	2	3	1	1	0
DISBA	0	0	0	0	1	0	0	1	2	1	2	0	2	0	0

Table A1_ 9. Data Fairness results, Case: Age, Cancer: Colorectal– M37.

Colon			
Cancer Grade	*1*	*2*	*3*
UNITOV	0	0	0
AUTH	11	9	1
DISBA	1	0	0

Table A1_ 10. Data Fairness results, Case: Cancer grade, Cancer: Colorectal– M37.

Colon					
Cancer Type	*Adenocarcinoma*	*Squamous cell carcinoma*	*Small-cell carcinoma*	*Large-cell carcinoma*	*Other*
UNITOV	0	0	0	0	0
AUTH	19	0	0	0	2
DISBA	1	0	0	0	0

Table A1_ 11. Data Fairness results, Case: Age, Cancer type: Colorectal– M37.

Colon		
Sex	*Male*	*Female*
UNITOV	79	61
AUTH	14	10
DISBA	6	6

Table A1_ 12. Data Fairness results, Case: Sex, Cancer: Colorectal– M37.

Lung															
Age group	*(20, 25]*	*(25, 30]*	*(30, 35]*	*(35, 40]*	*(40, 45]*	*(45, 50]*	*(50, 55]*	*(55, 60]*	*(60, 65]*	*(65, 70]*	*(70, 75]*	*(75, 80]*	*(80, 85]*	*(85, 90]*	*(90, 95]*
UoA	0	0	0	0	0	1	0	2	3	2	3	2	0	0	0
UNITOV	0	0	0	1	2	1	2	4	5	5	5	5	5	3	0
DISBA	0	0	0	0	1	0	2	1	0	2	1	4	1	0	0

Table A1_ 13. Data Fairness results, Case: Age, Cancer: Lung– M37.

Lung			
Cancer Grade	*1*	*2*	*3*
UoA	0	0	0
UNITOV	0	0	3
DISBA	0	0	0

Table A1_ 14. Data Fairness results, Case: Cancer Grade, Cancer: Lung– M37.

Lung					
Cancer Type	*Adenocarcinoma*	*Squamous cell carcinoma*	*Small-cell carcinoma*	*Large-cell carcinoma*	*Other*
UoA	5	0	1	0	0
UNITOV	0	2	0	0	1
DISBA	0	0	0	2	0

Table A1_ 15. Data Fairness results, Case: Cancer type, Cancer: Lung– M37.

Lung		
Sex	*Male*	*Female*
UoA	19	1
UNITOV	57	38
DISBA	8	6

Table A1_ 16. Data Fairness results, Case: Sex, Cancer: Lung – M37.

Prostate															
Age	*(20, 25]*	*(25, 30]*	*(30, 35]*	*(35, 40]*	*(40, 45]*	*(45, 50]*	*(50, 55]*	*(55, 60]*	*(60, 65]*	*(65, 70]*	*(70, 75]*	*(75, 80]*	*(80, 85]*	*(85, 90]*	*(90, 95]*
UoA	0	0	1	0	1	1	2	3	5	4	3	3	1	0	0
IDIBAPS	0	0	0	0	0	1	4	4	5	5	4	2	0	0	0

Table A1_ 17. Data Fairness results, Case: Age, Cancer: Prostate – M37.

Prostate					
Cancer Grade	*1*	*2*	*3*	*4*	*5*
UoA	0	0	0	0	0
IDIBAPS	34	29	20	18	5

Table A1_ 18. Data Fairness results, Case: Cancer grade, Cancer: Prostate – M37.

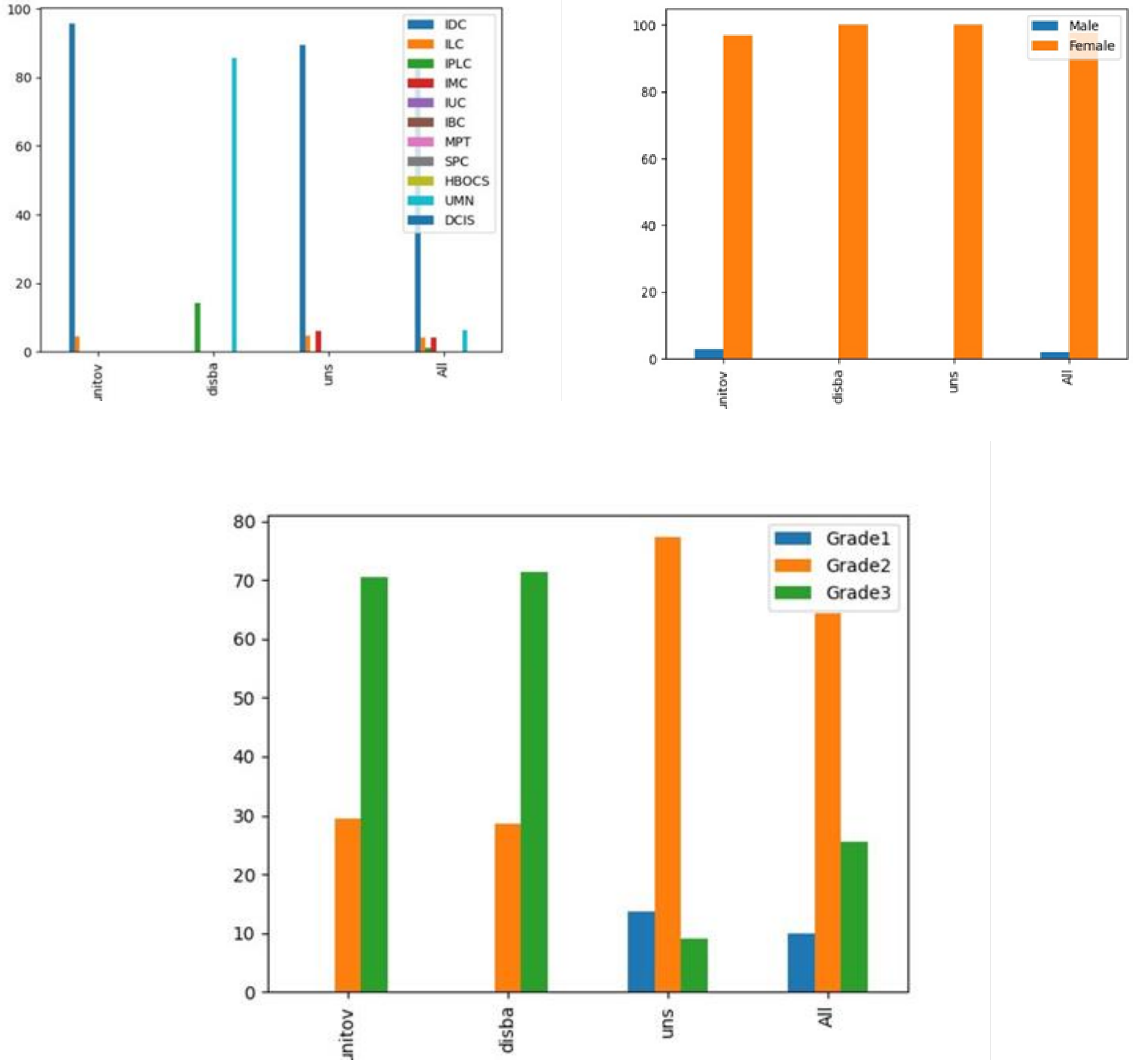


Figure A1_ 4 Representative graphs for Fairness results, Case: Cancer type, grade and sex for Breast cancer – M37.

Collectively, age groups in Breast cancer range from (30, 35] to (85, 90] similarly to previous studies. Most cases appear to have grade 2 (TableA1_7). Breast can also be diagnosed in men too, although it is considered as a women cancer (also observed during prospective study) (TableA1_8).

Age groups in Colon cancer range from (30, 35] to (85, 90] similarly to previous studies (TableA1_9). Case Grade, in colon cancer, seems to have equal distribution for all three grades (TableA1_10). Regarding cancer type in colon cancer, most cases appear with adenocarcinoma (TableA1_11). Regarding Sex, the results indicate that both male and female project the same number of records (Table A1_12).

Age groups in Lung cancer range from (35, 40] to (85, 90] similarly to previous studies (TableA1_13). Sex groups in Lung cancer range appear to be more evident in Male in the case of observational data collection (TableA1_16).

5. Integrity:

Integrity refers to the extent to which all data references have been joined accurately. The results are presented as percentage values. The metric that is used in assessing this dimension is the percent of data that is the same across multiple systems (Figure 5-7).

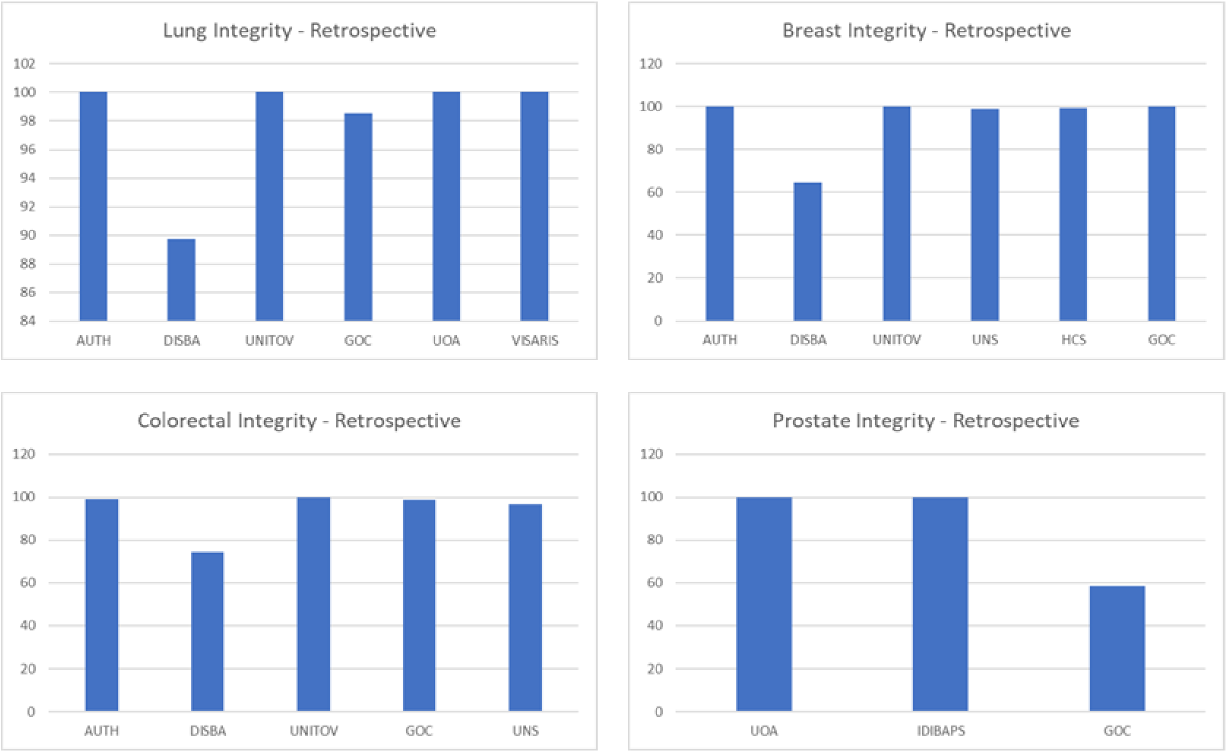
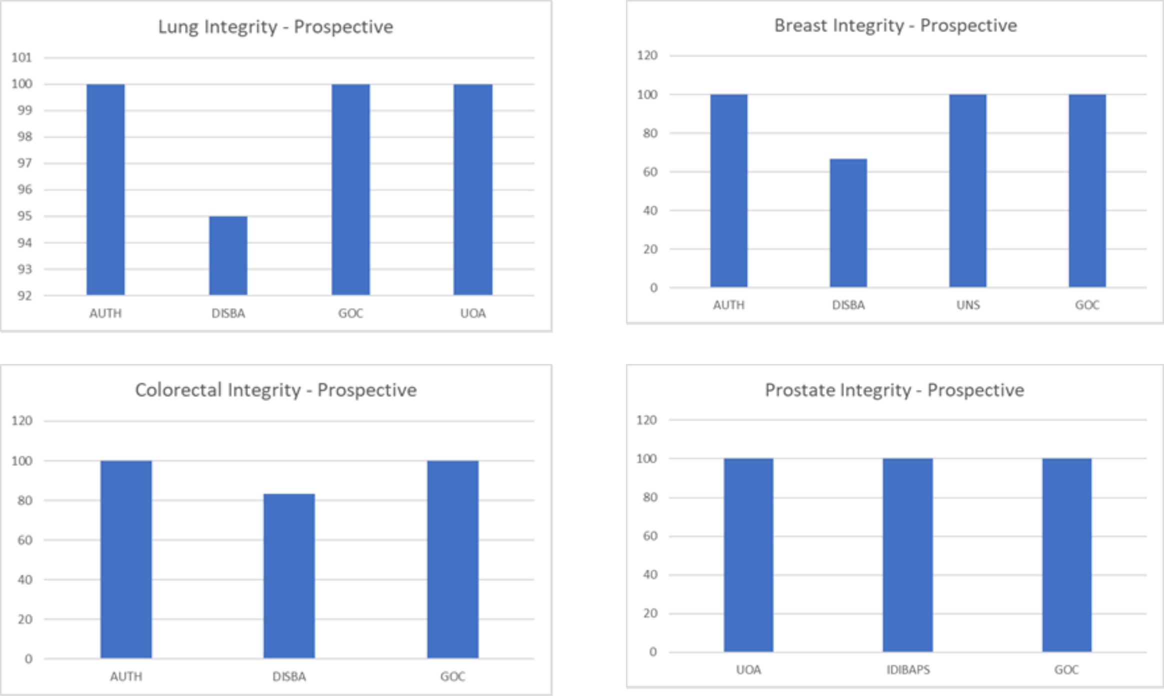


Figure A1_ 5. Integrity results (%) for retrospective phase for each cancer.



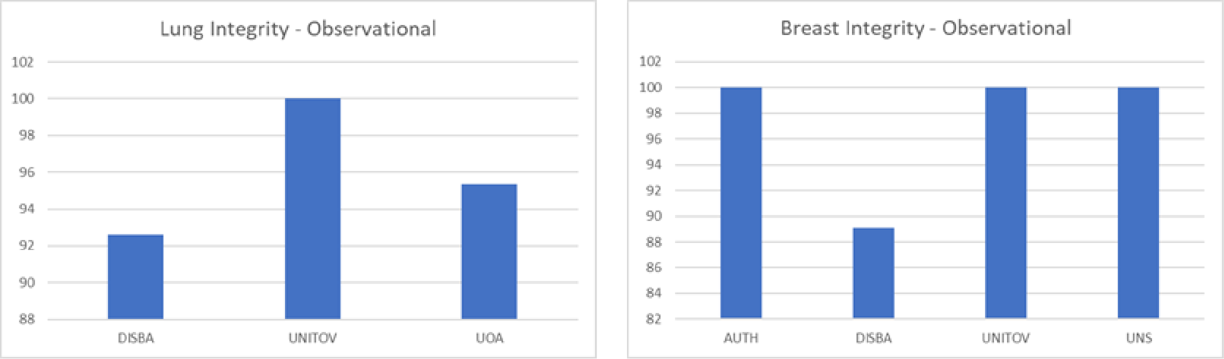
Lung Integrity - Prospective

Breast Integrity - Prospective

Colorectal Integrity - Prospective

Prostate Integrity - Prospective

Figure A1_ 6. Integrity results (%) for prospective phase for each cancer.



Lung Integrity - Observational

Breast Integrity - Observational

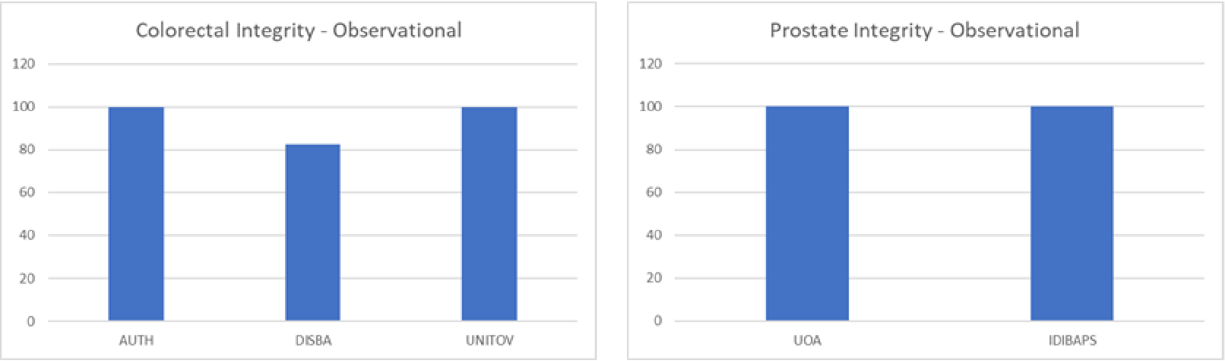


Figure A1_ 7. Integrity results (%) for observational phase for each cancer.

The results indicate that in all cases the integrity was 100% with a few deviations that do not exceed 40 %.

B. Image harmonization and quality results and findings

1. Annotation check results

Regarding the annotation check, this was performed by analyzing the labels from the segmentation masks, as they were previously defined during workshops dedicated to each cancer type. These labels are related to the cancer type but also to the imaging modality. For example, in mammography images, the data provider can characterize each lesion as (1) Benign, (2) Suspicious or Indeterminate, (3) Malignant, (4) Calcification, (5) Surgical clip or (6) Axial lymph node whereas on MR images for breast cancer the respective labels are related to (1) Benign, (2) Suspicious or Indeterminate or (3) Malignant. The Table below provides the information related to the labels for each cancer type and imaging modality (**¡Error! No se encuentra el origen de la referencia.**).

Cancer type	Imaging modality	Labels
Breast	MG	Benign Suspicious or Indeterminate Malignant Calcification Surgical clip Axial lymph node
	MR	Benign Suspicious or Indeterminate Malignant
	CT	Benign Suspicious or Indeterminate

		Malignant Macrocalcifications
	FusCT/PT	Benign Suspicious or Indeterminate Malignant
Lung	CT/FusCT/PT	Benign Malignant
	Xray	Suspicious Problematic
Colorectal	MR	Benign Malignant
	CT	Benign Malignant Lymph Node
	FusCT/PT	Benign Malignant
Prostate	MR	Benign Malignant

Table A1_ 19 Summary of the labels for each cancer type and imaging modality.

This analysis was completed only for AUTH data for all three phases of data collection (retrospective, prospective and observational) for each cancer type (breast, colorectal, lung).

Breast cancer

In the case of breast cancer, the results show that all annotations are correct (ROI and number of slices). This was performed for 174 annotation files for the current data in the incisive 2 repository (**¡Error! No se encuentra el origen de la referencia.**).

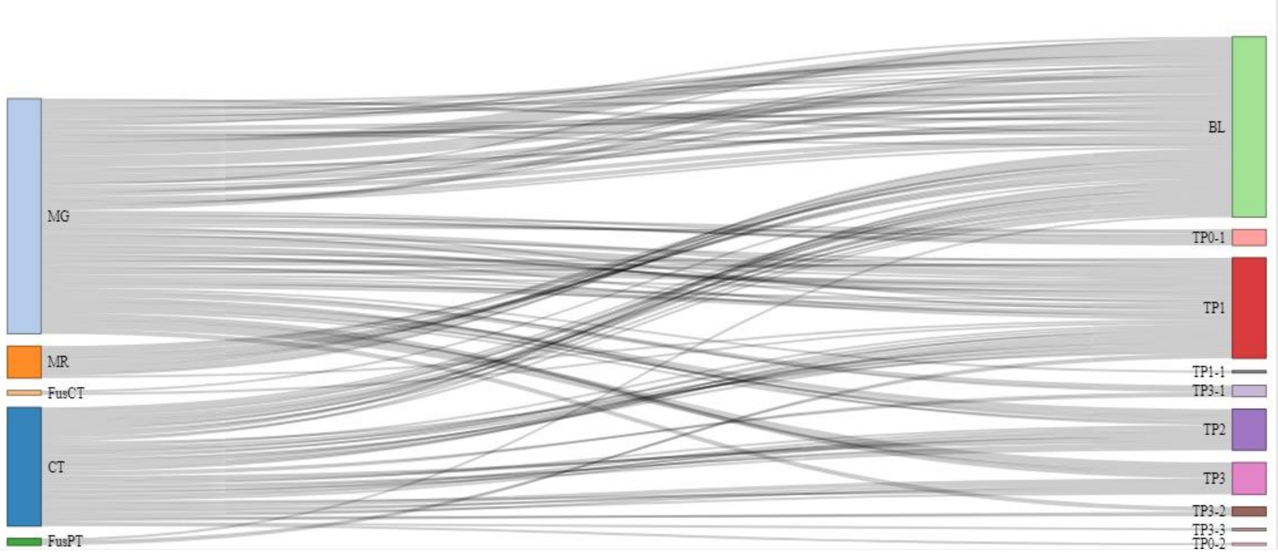
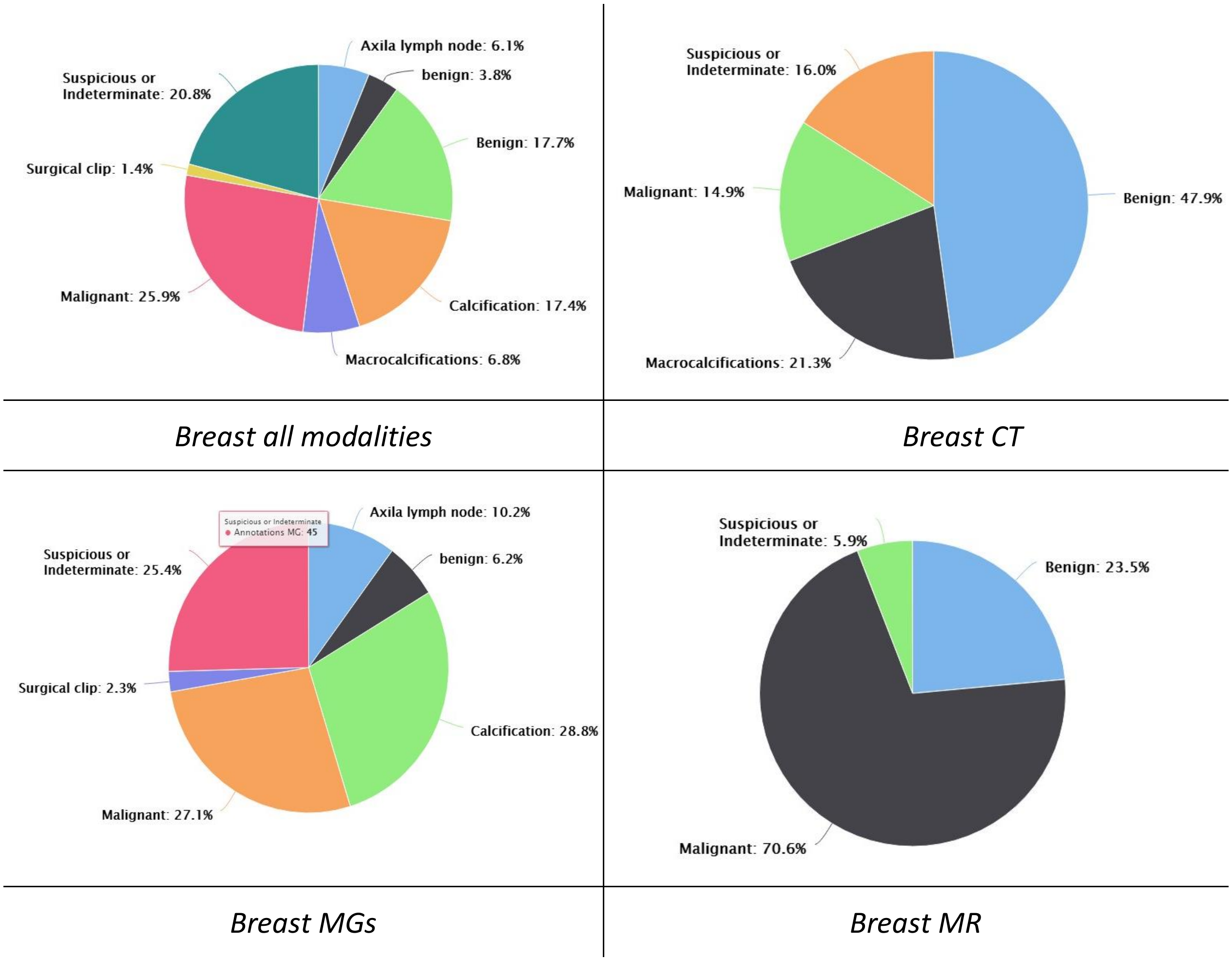


Figure A1_ 8. Annotations per timepoint for each modality in breast cancer. BL: baseline, TP-X, next timepoints

The types of annotation labels are presented in **¡Error! No se encuentra el origen de la referencia.**.



Breast all modalities

Breast CT

Breast MGs

Breast MR

Figure A1_ 9. Types of annotation labels per modality in Breast cancer.

For all the modalities taken into account the annotated regions/lesions involve suspicious or intermediate at 25.4%, Axial lymph nodes at 10.2%, benign at 6.2%, calcification regions at 28.8%, surgical clips at 2.3% and malignant at 27.1%. in the case of MR, the most found region was malignant (70.6%) whereas in the case of CT the most found region is benign (47.9%). Finally, in the case of mammographies all types of regions are found (FigureA1_9).

Lung Cancer

By the same token, in the case of lung cancer, 204 annotation files were checked. The results show that 8 annotations for PT (all that are for PT) are incorrectly placed in that folders and they appear correct only for the respective FusCT images. The remaining cases are correct (ROI and number of slices) (FigureA1_10).

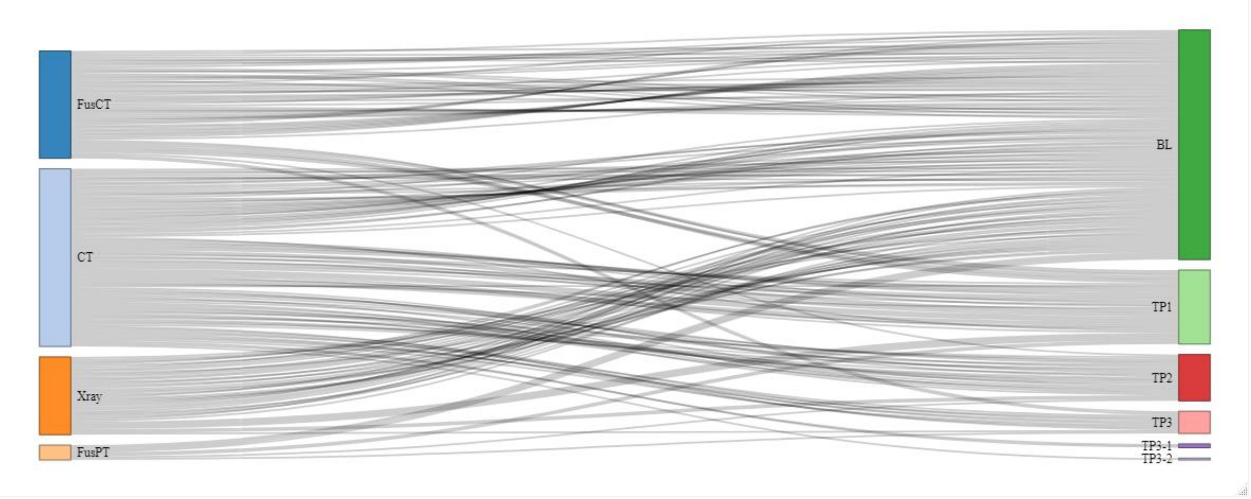
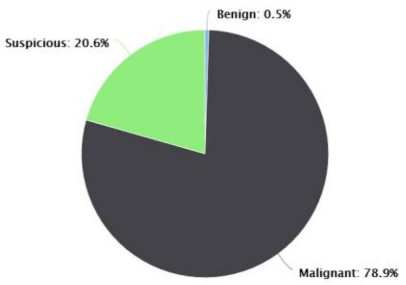
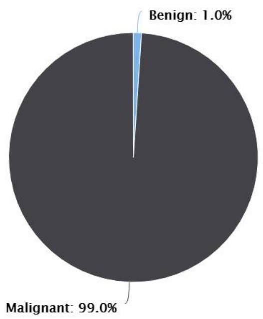


Figure A1_ 10. Annotations per timepoint for each modality in lung cancer.





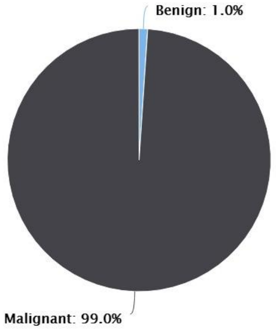
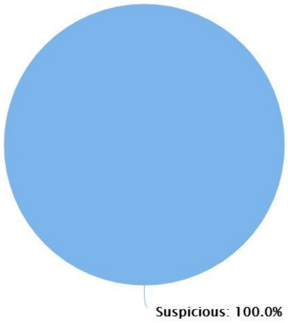
Lung all modalities	*Lung CT (96 annotations)*
	
Lung fusCT (58 annotations)	*Lung Xray (42 annotations)*

Figure A1_ 11. Types of annotation labels per modality in Lung cancer.

In the case of lung cancer, the regions found/annotated were malignant, benign and suspicious. In the case of CT and fusCT, malignant was the most common finding, whereas in the case of Xrays, all findings were suspicious (**¡Error! No se encuentra el origen de la referencia.**).

Colon cancer

Similarly, for the colorectal cancer, 97 annotation files were checked, and all were found correct (ROI and number of slices).

In colorectal cancer the regions annotated were malignant or benign for all modalities tested. In the case of MR and CT, the malignant region was 91.2 and 64.7% respectively whereas in the case of FusCT all regions were found as malignant. This is reasonable, as FusCT is typically performed for further investigation in malignant tumors (**¡Error! No se encuentra el origen de la referencia.**). This is reasonable, as FusCT is typically performed for further investigation in malignant tumors.

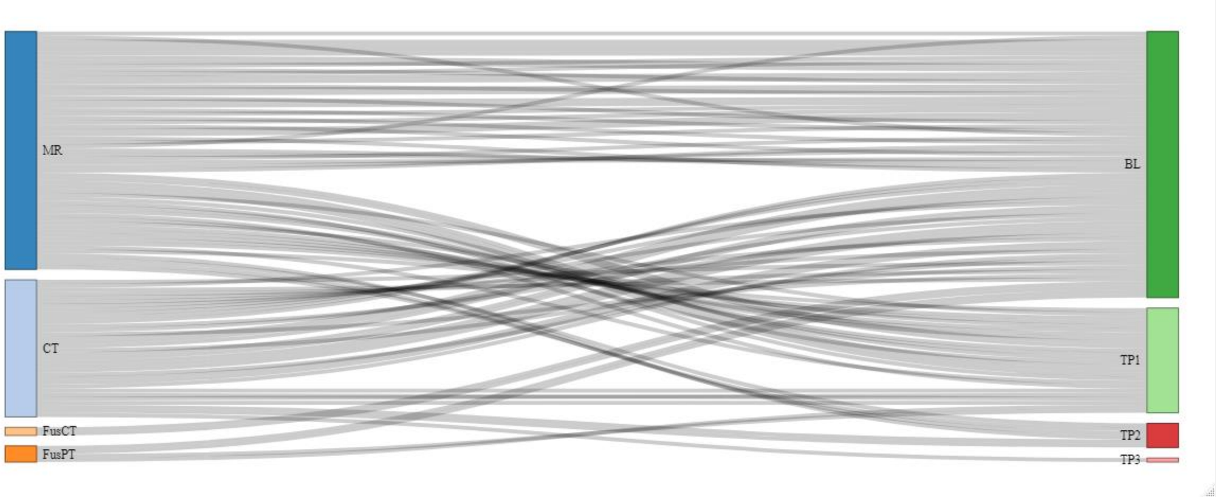
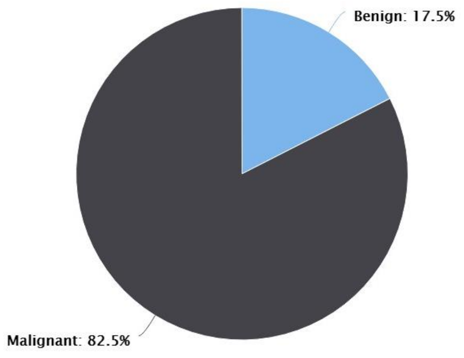
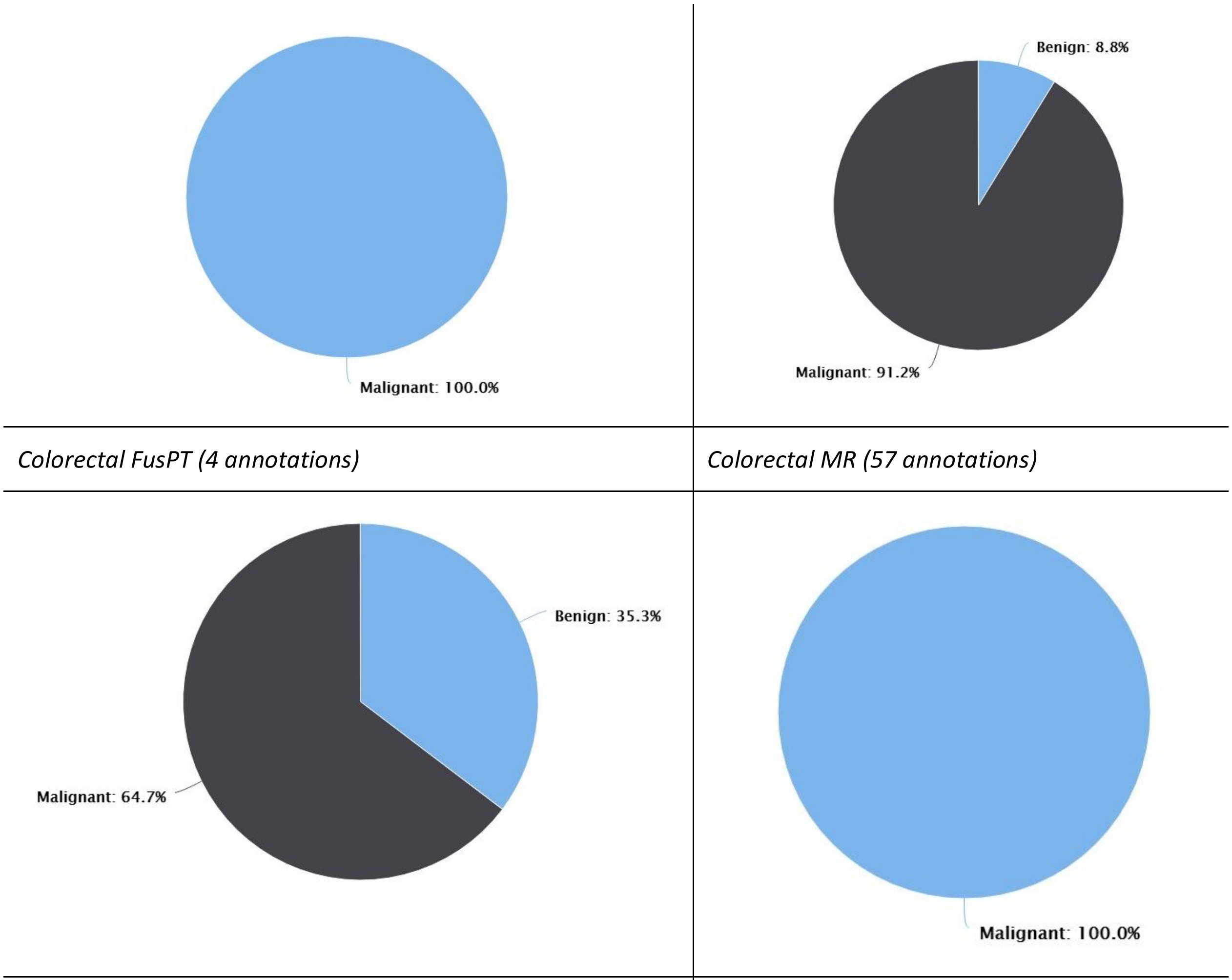


Figure A1_ 12. Annotations per timepoint for each modality in colon cancer.



Colorectal cancer all modalities (97 annotations)



Colorectal FusPT (4 annotations)

Colorectal MR (57 annotations)

Colorectal CT (34 annotations)

Colorectal FusCT (2 annotations)

Figure A1_ 13. Types of annotation labels per modality in Colon cancer.

2. Identification of DICOM reports

In this analysis, INCISIVE data were checked to identify images that contain DICOM reports. These reports are DICOM images that summarize some information of the imaging examinations. These images' headers can be de-identified by the de-identification tools, but they also contain burnt-in information written on the images' pixels, that can't be detected by the aforementioned tools. To that end, a tool was developed to detect such images in the data, and the analysis was performed indicatively in a part of the repository, reaching almost 75 % of the data. When this analysis was performed and for data that was checked, no DICOM reports were detected in the repository.

3. Image similarity detection

We implemented a process for identifying similar images in the data pool byte utilizing the *difPy* library, a Python tool developed for detecting duplicate or similar images within folders. Our primary goal was to analyze and identify similar images in order to enable a more effective search for potential duplicates. This experiment was executed on lung prospective and retrospective data for AUTH.

The difPy library enables this process by comparing images based on specified similarity grades ("low", "normal", "high") and providing detailed results, including the presence of duplicates, processing duration, and other relevant metrics. The similarity grades refer to the threshold of “mean squared error” (MSE), which is the metric employed in the tool implementation, to evaluate the results of the image comparison.

The approach we followed involved:

- Intra-directory Analysis: We first focused on comparisons within individual directories (each patient's examinations), grouping the results based on the patient's ID extracted from the directory names. This allowed us to determine the degree of similarity within each specific modality/timepoint folder.
- Inter-directory Comparisons: Next, we extended our analysis to comparisons between different directories i.e. patient’s different timepoint examinations based on the same type of modality. We grouped the data based on pairs of directories involved in each comparison, utilizing again the patients' IDs, their modality and timepoint. This approach provided insights into cross-directory image similarities.

For both intra- and inter-directory comparisons, we aggregated the data to quantify the total number of comparisons, the occurrence of similar images, and the distribution of these instances across different directories and directory pairs. The results reported below are for lung prospective and retrospective data for AUTH.

Intra-directory analysis:

The intra-directory comparison was conducted with a "high" similarity grade, corresponding to an MSE threshold of 0.1. This stringent criterion was chosen to ensure that only images with very high degrees of similarity were flagged as similar/ potential duplicates within the same folder.

Total number of directories (patients) examined	60	
Total number of individual folders compared	436	
Instances similar images were found	186	
Comparisons count grouped by modality: (total number of examinations / number of folders similar images found		
CT	237	35
FusCT	85	82
FusPT	69	69
Xray	45	0

In order to have a summarized view for each case separately, the percentage score of the similarity result was calcuclated I.e. the ratio of possible duplicate occurences (files flagged) to the number of comparisons performed (similarity percentage).

We observed a notably high number of similar images within the PET/CT scans (FusCT/FusPT). This phenomenon is characterized also by high similarity percentages for most of the cases involved. PET-CT scans involve multi-dimensional capturing techniques that result to metabolic and functional data that have a distinct clinical value. However, due to their low resolution -compared with other medical imaging modalities- and the decrease of heterogeneity in the spatial context of an organ –as reported in cases in the [1]- they require a more nuanced strategy for this kind of analysis, as suggested also by our results.

For the few cases in CT modality, and before we proceeded to verify these findings through a metadata-based analysis, we investigated some instances i.e. cases where similar images were detected in comparison between two directories, to examine if false positives appear also in this modality type. We selected cases with high similarity percentages to ensure that only the most relevant instances of potential duplicates were considered. Figure A1_14, illustrates two example series of CTs. In our analysis both were labeled as false positives, but the similarity percentages varied, with the first showing a 92.44% (across 119 slices) and the second 3.81% (across 105 slices).

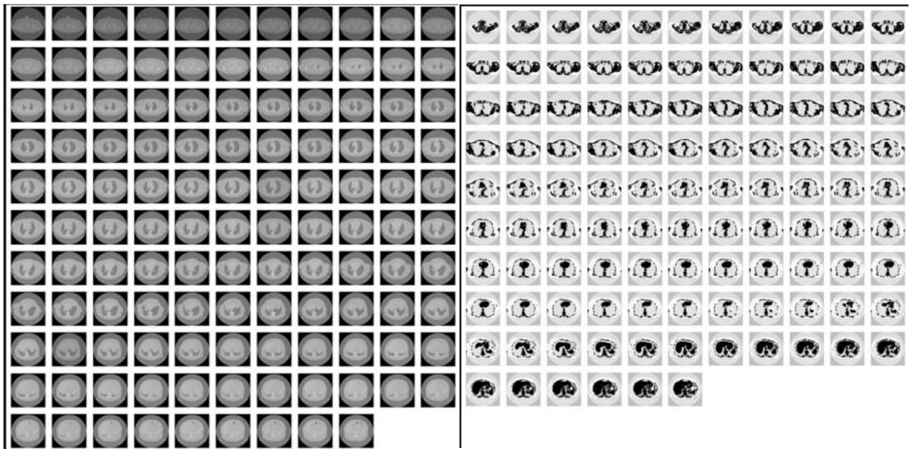


Figure A1_ 14. False positive of potential duplicate cases of CT examinations

Inter-directory analysis

For the inter-directory comparison, a "high" similarity grade was employed.

[1] Thomas, L. J., Huang, P., Yin, F., Luo, X. I., Almquist, Z. W., Hipp, J. R., & Butts, C. T. (2020). Spatial heterogeneity can lead to substantial local variations in COVID-19 timing and severity. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(39), 24180–24187. https://doi.org/10.1073/pnas.2011656117

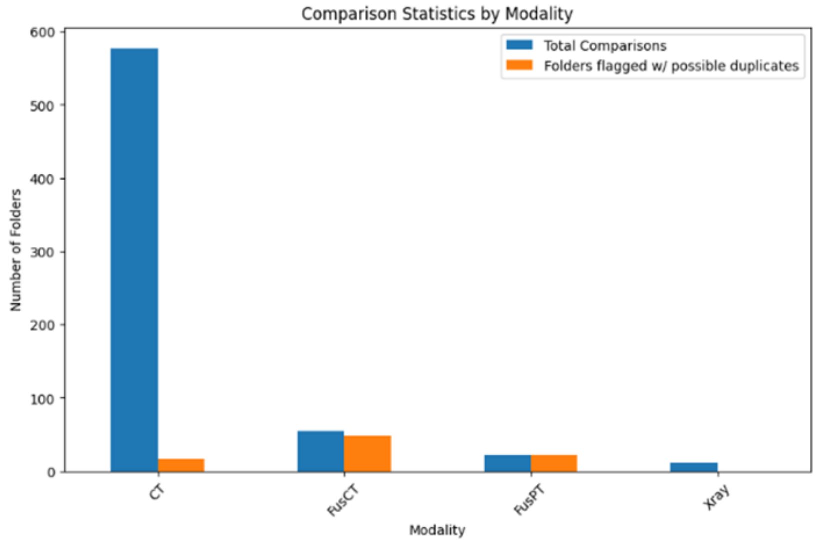


Figure A1_ 15. Results of inter-directory analysis

¡Error! No se encuentra el origen de la referencia. presents the number of total comparisons performed, and the instances of the directories where similar imageswere detected, are displayed by modality. This diagram reaffirms the issue reported previously in the intra-directory analysis for PET-CT scans. From the figure it is evident that CT scans were most frequently involved in the analysis, with 578 comparisons. Fusion CT (FusCT) and Fusion PT (FusPT) images followed with 54 and 22 comparisons, respectively, while X-rays had the least, with 11 comparisons. Similarly, to our intra-directory analysis, we decided not to further investigate the PET-CT scans. Instead, we focused on the CT modality and specifically on the timepoints aspect.

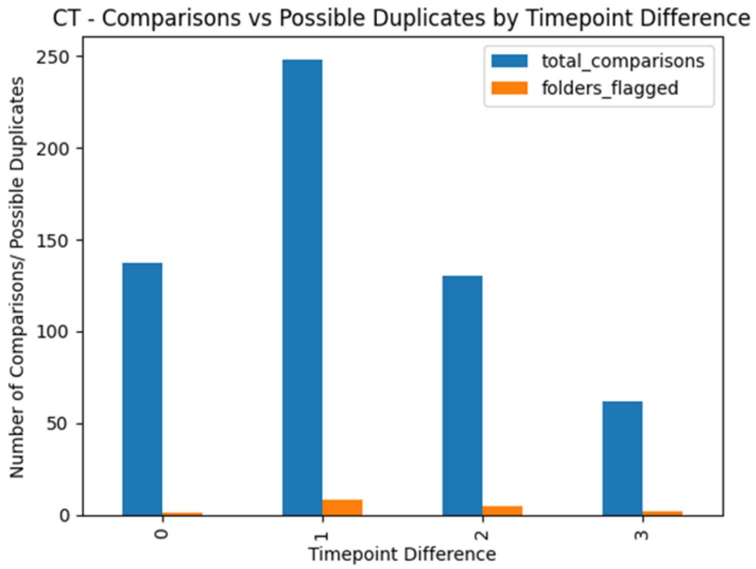


Figure A1_ 16. Results of similar/possible duplicate instances in CT modality, broken down by timepoint

¡Error! No se encuentra el origen de la referencia. illustrates the results of the inter-directory comparison based on the CT modality. The bar chart shows the number of comparisons between directories executed, and the number of the instances that were reported to have potential duplicates, categorized based on the time difference between the timepoints of the CT examinations. We observe that the category with the highest number of comparisons is the one where examinations were 1 timepoint apart and this is also the group with the most duplicate cases. This can be explained by considering that CT examinations close in time, such as follow-ups, may present a similar clinical status, which may lead to false-positive duplicate instances. Additionally, the standard protocols that are used for these types of examinations may amplify this phenomenon.

The experiment described so far, was repeated with a "normal" similarity grade, which corresponds to a higher MSE threshold. This less stringent criterion acknowledges the potential for slight variations between images from different directories or patient cases. The number of duplicate cases was higher both for the CT modality and the PET-CT scans, as anticipated. This can be attributed to the fact that the system flagged images as false positive duplicates due to the lenient similarity criteria selected.

Below, in Figure A1_17, we present an example of false positives, identified by our subsequent investigation. This case involves possible duplicates identified from a comparison between two series which are one timepoint apart. The similarity percentage recorded for this instance is 56%. In the images displayed, the tiles lined up on the left column are the same slice from the first series in the comparison, and the images on the right column are the slices that were flagged as possible duplicates from the second series.

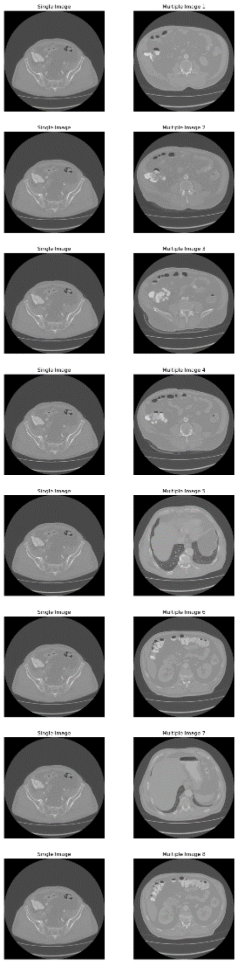
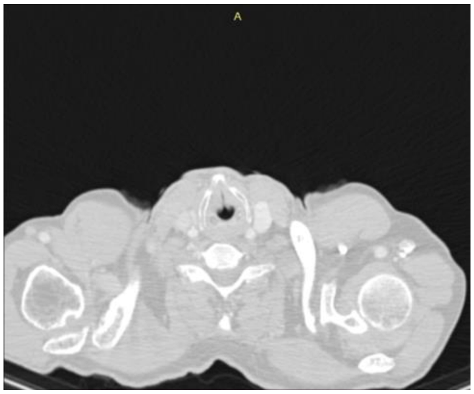
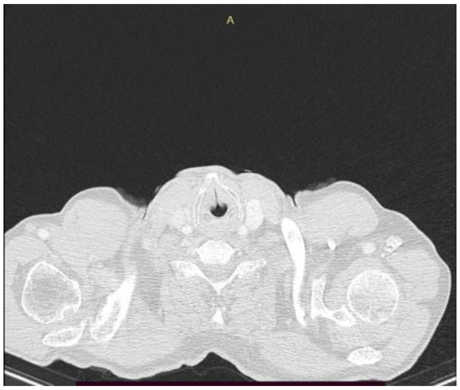


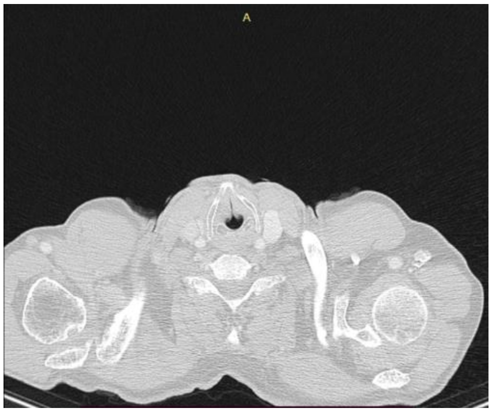
Figure A1_ 17. False positive example of a comparison between two series in the same timepoint

From the whole list of possible duplicates that were identified by the aforementioned approach, we randomly selected some cases to verify the existence of duplications in the repository. In one case it was found that the two series folders that were provided for the same patient, for the same timepoint and imaging modality (CT), were flagged as possibly duplicated. From a closer view we observed that the images were not duplicates however, they looked similar because they were acquired using different convolution kernels. The following example provide more information regarding the data that are included in the two series.

Folder: 001-000010_CT_TP1/Series-4	Folder: 001-000010_CT_TP1/Series-5
ROI: 512x512	ROI: 512x512
No of slices: 296	No of slices: 296 slices
Convolution kernel: B20f	Convolution kernel: B70f
slice thickness: 1.5	slice thickness: 1.5
Sample image depicted in Lung Window 	Sample image depicted in the default window 

In addition, similar images were found in the same patient in different timepoints, when the acquisition protocols were identical, and the shape of the human body did not change dramatically.

001-000020_CT_TP1/Series-5	001-000020_CT_TP2/series-10
ROI: 512x512	RO: 512x512
Slice thickness: 1.5	Slice thickness: 1.5
Convolution Kernel: B70f	Convolution Kernel: B70f
Number of slices: 296	Number of slices: 301

4. *Image Deduplication*

For the detection of potential duplicates, a complementary step was adopted. This step was based on the analysis of a specific DICOM tag named “media storage SOP /instance UID (“0002,0003). This attribute is unique for each DICOM image while during the anonymization process the hash function produces the same output for specific input. This approach allows us to identify the existence of duplicates, based on this DICOM tag, even after the anonymization step. In more details, for the “media storage SOP /instance UID” The root argument is a text string containing the UID root for the institution (for example, 1.2.840.4267.32.). The hash UID function creates a new UID by computing the MD5 hash of the existing UID, converting it to a base-10-digit string and prepending the root.

After the application of the aforementioned step in the whole lung cancer dataset provided by AUTH, no duplication was identified. This approach will be applied in the whole INCISIVE repository and the results will be documented in D6.6.

The whole approach for this analysis has revealed critical insights and highlighted areas that need further exploration. This is an ongoing work aiming to identify the most efficient way in image deduplication. One significant observation was the prevalence of false positives, especially in cases where entire patient folders appeared to consist of duplicate images. This underscores the complex nature of medical imaging data and the necessity for a more nuanced approach to duplication detection. As next steps the following are proposed:

- Selective Image Comparison: Instead of comparing entire series or folders, a more targeted approach might be appropriate in the case of detecting duplicates in a pool with medical image data. Comparing only specific slices or key images within a series could reduce false positives while still effectively identifying genuine duplicates.
- Image Preprocessing: The preprocessing of medical images, such as cropping or adjusting contrast, needs careful consideration. This step should be tailored to each specific modality case.

5. DICOM attributes

In the context of data quality and harmonization, DICOM tags were also analyzed for potential sources of error and uniformity of the datasets. In so doing, several attributes were analyzed. The selection of the present attributes was based on literature sources and clinical guidance by experts in the field (HCPs, radiologists).

The results are split for each cancer type by each DP in INCISIVE and the tables are oriented by each modality used. The results presented here have been obtained collectively for all three data collections (retrospective, prospective and observational) whereas the attributes studied are further split in categorical (Figures A1-18 & 19) and numerical groups (Figures A1 20 & 21).

The results are presented as median value with 1st and 3rd quartile.

In general, in the case of **categorical attributes**, the following were selected to analyze by each imaging modality: Convolution Kernel, Filter Type, Image Type, Manufacturer, Manufacturer Model Name, Scan Options, Timepoint, Fused Convolution Kernel, AngioFLag, Image Nucleus, In-plane Phase encoding Direction, Photometric Interpretation, Scanning Sequence, Scanning Variant, Software Versions, Count Source, Decay Correction, Field Of View Shape, Patient Position, Photometric Interpretation, Randoms Correction Method, Reconstruction Method, Series Type, and Units.

Indicative examples are presented in the following figures. The case of Convolution Kernel:

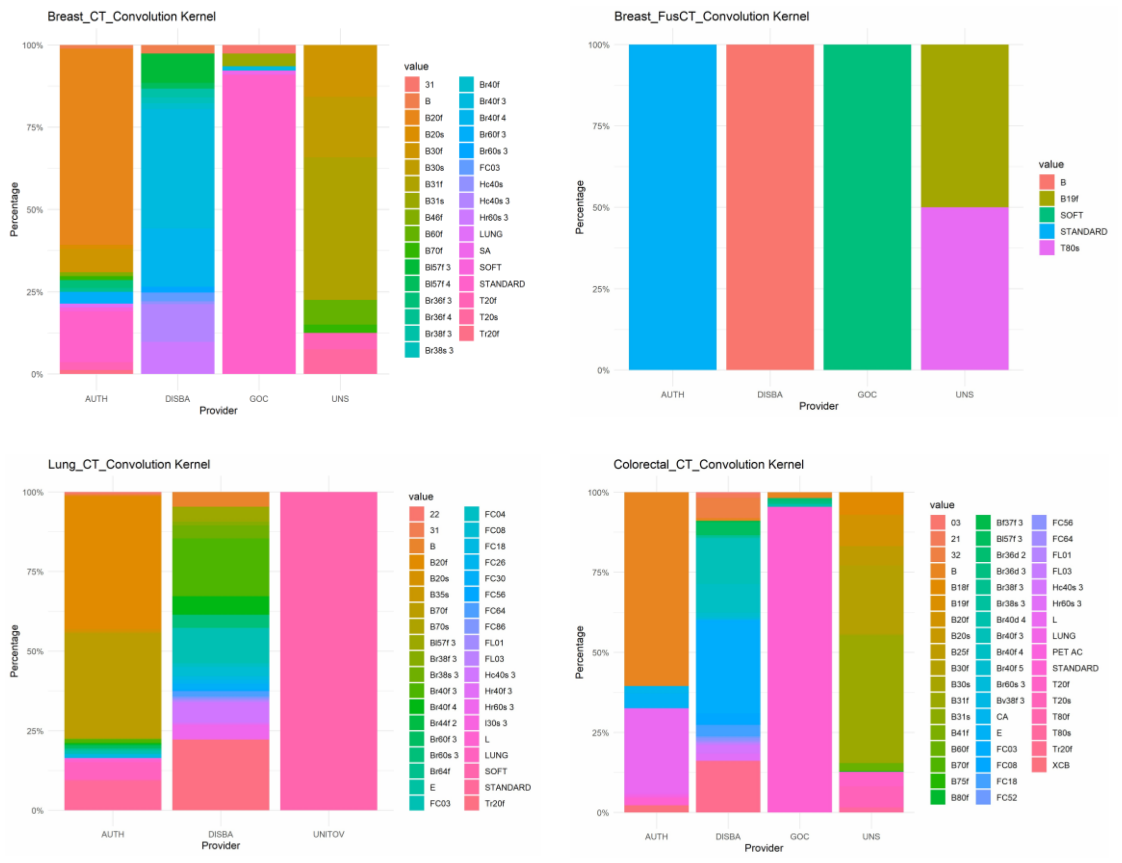


Figure A1_ 18. Comparison of the distribution of the Convolution Kernel value among different DPs, modalities or cancer type.

The results show that diversity in an attribute can occur not even among different DPs, but also when comparing the same modality in different cancer types or when comparing the similar modalities (CT with FusCT) even for the same cancer type.

The case of Manufacturer:

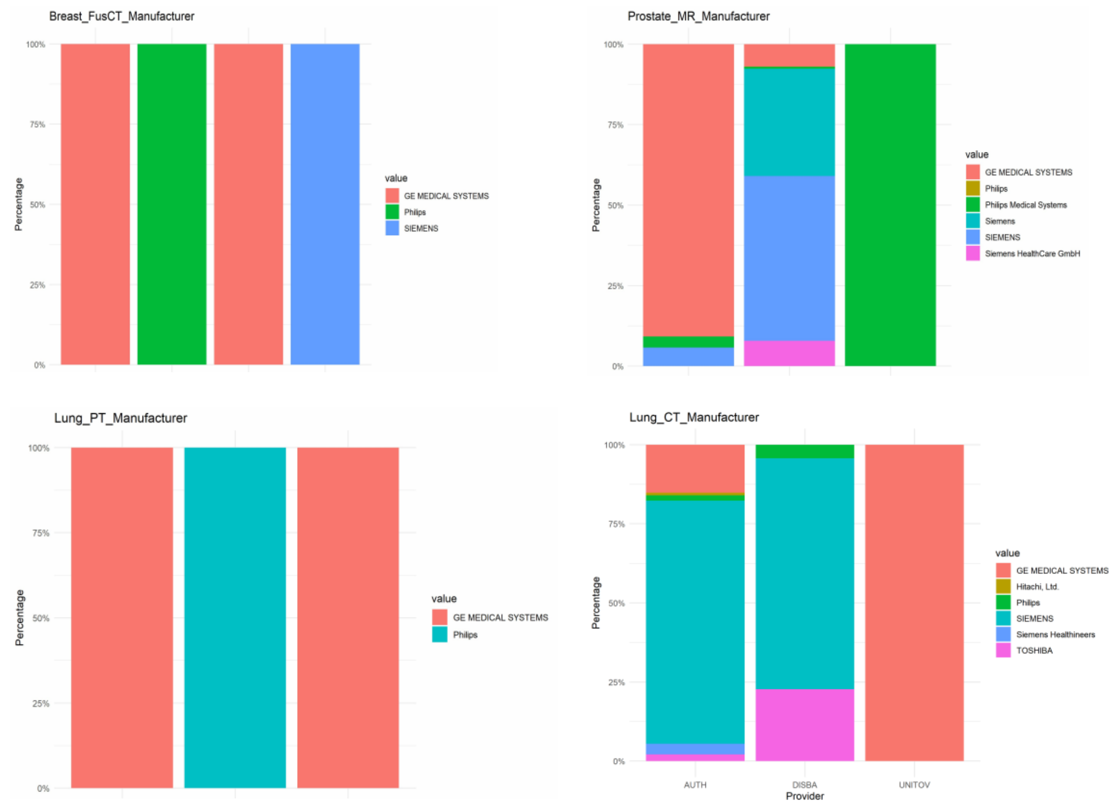


Figure A1_ 19. Comparison of the distribution of the Manufacturer value among different DPs, modalities or cancer type.

In the case of Manufacturer, it seems that there is lower diversity compared to other attributes indicating that is a more DP-oriented differentiation although this can also occur even among the same DP.

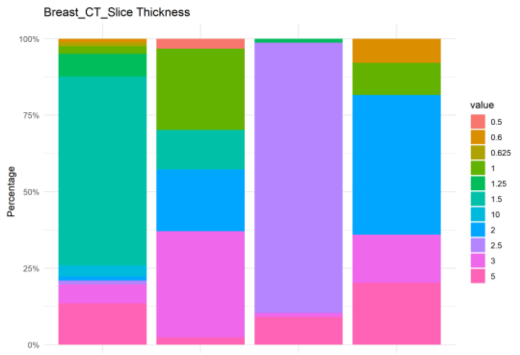
In the case of **Numeric attributes**, several attributes were analyzed, again, for each cancer type and by each modality used.

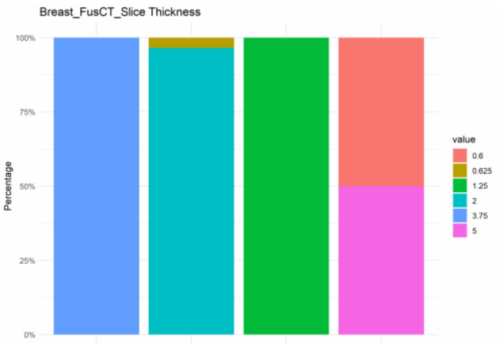
For instance in the case of Breast cancer PT, numImages, SliceThickness, ReconstructionDiameter, GantryDetectorTilt, ActualFrameDuration, Rows, Columns, PixelSpacing, RescaleSlope, EnergyWindowLowerLimit, EnergyWindowUpperLimit, RadiopharmaceuticalStartTime, RadiopharmaceuticalStopTime, RadionuclideTotalDose, RadionuclideHalfLife, RadionuclidePositronFraction, RadiopharmaceuticalStartDatetime, RadiopharmaceuticalStopDatetime, DecayFactor, DoseCalibrationFactor, ScatterFractionFactor, and DeadTimeFactor. Among these attributes ‘numImages’, ‘Exposure’ and ‘XrayTubeCurrent’ were found with highest diversity in values among the DPs. Same results were obtained also in the case of Fused CT. In the case of MR, ‘numImages’, ‘RepetitionTime’, ‘EchoTime’, ‘Number of Averages’, Imaging Frequency’, ‘Pixel Bandwidth’, ‘SAR’, and ‘Largest Image Pixel Value’ project the highest diversity among DPs. Similarly, in the case of PT images, ‘numImages’, ‘ActualFrameDuration’, ‘Rows’, ‘Columns’ and ‘Factor’ project the high diversity.

Similarly, in the case of Lung CT numImages, Slice Thickness, KVP, Pixel Spacing, Spiral Pitch Factor, reconstrusction Diameter, Rows, Columns, Exposure, FocalSpots, InstanceNumber, XRayTubeCurrent, Bits stored, and High bit were analyzed. In the case of Lung cancer, CT images were found to have higher diversity for ‘numImages’, ‘SliceThickness’, ‘reconstructionDiameter;, ‘Exposure’ and ‘XRayTubeCurrent’. Similarly, in the case of Fused-CT, ‘numImages’, ‘Slice Thickness’, ‘Exposure’, ‘InstanceNumber’, and ‘XRayTubeCurrent’ project the highest diversity. In the case of Lung PT images, ‘numImages project slight diversity as also ‘Slice thickness’ and ‘RadiopharmaceuticalStopTime’ and ‘DoseCalibrationFactor’.

In colorectal cancer, in the case of CT and Fused CT images, the results are the same as in the case of breast cancer CT and Fused CT images. In the case of MR images ‘repetitionTime’, ‘EchoTime’ and ‘PercentSampling’, ‘Pixel Bandwidth’, ‘FlipAngle’, ‘Rows’ ‘Columns’ and ‘LargestImagePixelValue’ were found to have the highest diversity. In the case of PT, ‘ReconstructionDiameter’, ‘Rows’, ‘Columns’, ‘Factor’, ‘RescaleSlope’ and ‘DoseCalibrationFactor’ project the highest diversity among DPs.

Representative examples are depicted in **¡Error! No se encuentra el origen de la referencia.** and **¡Error! No se encuentra el origen de la referencia.**. The case of Slice thickness:





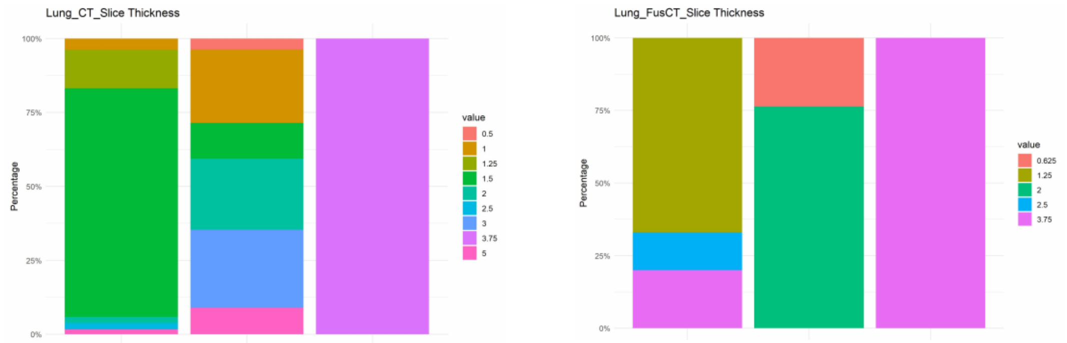


Figure A1_ 20. Comparison of the distribution of the Slice thickness value among different DPs, modalities or cancer type.

The case of Pixel spacing:

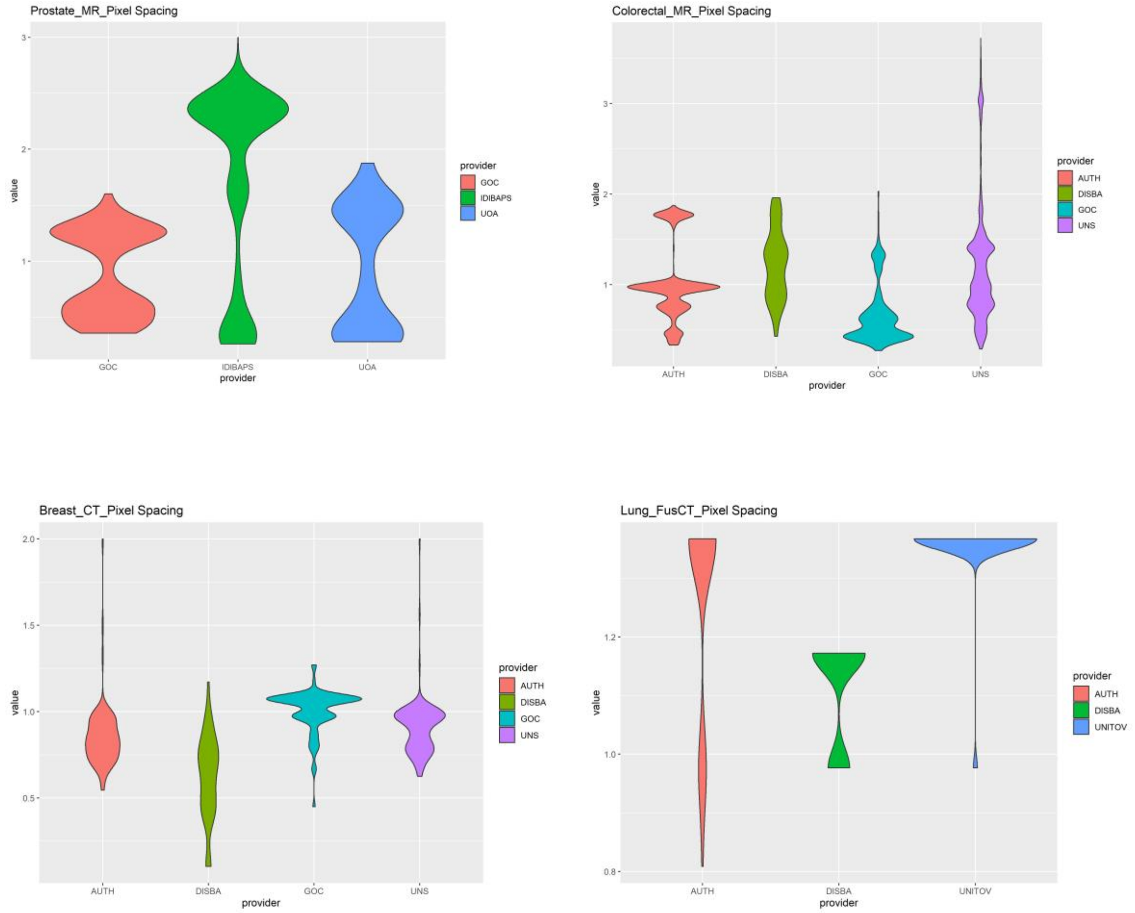


Figure A1_ 21. Comparison of the distribution of the Pixel Spacing value among different DPs, modalities or cancer type.

The median in each case is as follows in **¡Error! No se encuentra el origen de la referencia.**-23.

COLORECTAL MR	PARTNER			
Attribute	**AUTH**	**UNS**	**GOC**	**DISBA**
Pixel Spacing	1[0.8-1]	1[0.8-1.4]	0.6[0.4-0.7]	1.3[0.9-1.4]

Table A1_ 20. Median value with 1st and 3rd quartile in the case of Colorectal MR image type by DP.

PROSTATE MR	PARTNER		
Attribute	**UOA**	**IDIBAPS**	**GOC**
Pixel Spacing	0.9[0.3-1.5]	2.4[0.8-2.4]	0.8[0.6-1.2]

Table A1_ 21. Median value with 1st and 3rd quartile in the case of Prostate MR image type by DP.

LUNG CT FUSED	PARTNER		
Attribute	**AUTH**	**DISBA**	**UNITOV**
Pixel Spacing	1.4[1.4-1.4]	1.2[1-1.2]	1.4[1.4-1.4]

Table A1_ 22. Median value with 1st and 3rd quartile in the case of Lung FusCT image type by DP.

BREAST CT	PARTNER			
Attribute	**GOC**	**DISBA**	**AUTH**	**UNS**
Pixel Spacing	1.1[1-1.1]	0.7[0.4-0.8]	0.8[0.7-0.9]	1[0.8-1]

Table A1_ 23 Median value with 1st and 3rd quartile in the case of Breast CT image type by DP.

Slice thickness - Requirements analysis

Lung cancer

This was performed for 60 patients provided by AUTH.

In most patients, more than one series is provided. For all patients provided by AUTH, there was at least one image series per patient with slice thickness lower or equal to 1.5mm, except for one patient' with no image series covering this requirement. The threshold for the slice thickness was set by the AI developers and is related to CT images in lung cancer patients. For the series that did not cover the requirements (43 image series), the table below provides more details on the provided slice thickness.

Minimum	1st quantile	Median	Mean	3rd quantile	Maximum value
2	2.5	3.125	3.185	3.75	5

Table A1_ 23. Slice thickness for CT scan that did not have the minimum slice thickness

Breast Cancer

In breast cancer the slice thickness requirement differs based on the modality. In CT, the upper acceptable value is 5mm, for CT images and 3mm for MR. 27 image series do not cover the slice thickness requirements. However, all the patients included at least one series which covered the slice thickness requirement.

	>slice thickness (mean+/-std)	Wrong values	Missing value
CT	3 (10+/-0)	1	5
MR	17 (4.65=/-1.1)	1	0

Table A1_ 24. Slice thickness for CT and MR scans that did not fulfil the minimum slice thickness

Colorectal cancer

Slice thickness requirement was defined only for CT images (<6mm). All patients which had CT images covered the slice thickness requirement. However, more series are provided for some patients, in which the slice thickness was not able to be tested because this attribute was missing from the DICOM files (71 series), while for 11 series the value was not correct and for 5 the value was higher than 6mm.