Proposal CSN5 research project

AIM_MIA

Artificial Intelligence in Medicine: focus on Multi-Input Analysis

Durata: 2025-2027
Area di Ricerca: Interdisciplinare
Responsabile nazionale: A. Retico (PI)
Unità partecipanti: BA, BO, CA, CT, FE, FI, GE, LE, LNS, MI, PI, PV

Abstract

PREAMBLE. Artificial intelligence (AI)-based solutions have become pervasive in the field of Medicine due to the broad opportunities they can offer, including enhanced diagnostics, personalized treatments, and improved patient care. However, to fully realize this potential, several challenges must still be addressed. Given that it is understood that the road to bringing an AI-based support tool from the laboratory to the patient's bed is very long and laborious, it is worth underlining that even the previous phase of research and development of these tools still requires a great effort from the scientific community. AI-based clinical support tools are often very narrow in scope, lack transparency and are not trustworthy.

RESEARCH OBJECTIVES. The **AIM_MIA** project will focus on a main methodological open issue related to the development of AI-based tools for medical data analysis: **1) mining multi-input data**, i.e. providing new analysis strategies to take advantage of the complementary information encoded in data coming from heterogeneous sources. To make progress in this field also requires addressing the following co-occurring key challenges: **2) handling incomplete/missing/limited datasets**; **3) developing a dedicated data and IT platform** for secure data management, linked to adequate computing resources. To achieve these goals, sharing data and knowledge within a broad scientific community (**networking**) will be a fundamental ingredient.

METHODS. Three work packages will be devoted to address the scientific issues enumerated above. In **WP1** advanced AI-based solutions to analyze relevant data regarding the health status of individuals (including demographic information, medical images acquired with different modalities, clinical scores, etc.) will be developed and validated. **WP2** will be focused on the implementation of the technical solutions for data curation, missing data imputation, data augmentation, sample balancing etc., in order to extract as much information as possible from the available datasets which in most real-word cases are incomplete, limited or unbalanced.

> The growing availability of public data repositories will ensure the feasibility of this project. The data will be organized and shared among the project collaborators via a dedicated data platform to be developed in **WP3**, which will rely on INFN computing resources. The continuous collaboration with clinical experts, relevant associations in the medical research field, and connections with other research projects funded by INFN or external Institutions will be managed in **WP4**, which is dedicated to scientific networking.
> **EXPECTED RESULTS.** A three-year research activity carried out by a research group with deep experience on these topics will certainly lead to a significant advancement in the strongly connected open issues described above. In terms of measurable performance indicators, the collaboration is expected to deliver several **scientific publications** in relevant journals (>5 per year). The **data platform** and **software repository** will constitute a tangible and reusable deliverable of the project.

Proposal

State of art

Precision medicine is an innovative approach that aims to tailor disease prevention and treatments to the specificity of each single person, by taking into account the individual differences in genes, environment, and lifestyles. To make precision medicine a reality, various areas of intervention need to be improved (Denny and Collins, 2021): collecting data from large longitudinal samples of subjects, taking into account diversity and including minorities; improving the availability of data from the electronic health records (EHR) and genetic tests for research purposes; acquiring also parameters related to the environment and lifestyle of the subjects; archiving and preserving data in compliance with privacy and ethical requirements; **exploiting big data collections to develop Artificial Intelligence (AI)-based models to support subjects-specific diagnostic and treatment pathways**.
Regarding the last point in particular, the scientific community began developing AI-based systems for the automatic analysis of medical data starting in the 1980s, when the first AI-based decision support tools for automated reading of mammograms and chest X-rays were developed (Giger 2008).
Jumping forward 4 decades to arrive directly at the present day, academic research has been developing several AI-based tools to support clinical workflows (Litjens 2017), and some solutions are also commercially available as CE-marked products (van Leeuwen 2021). Nonetheless, it has been highlighted that the performance of current AI algorithms to support diagnostic imaging evaluated in the R&D stage is difficult to maintain in clinical use,

i.e. most AI-based tools lack generalization capability (Park 2022). Moreover, it has been pointed out that AI algorithms are typically developed for a single specific task, e.g. to efficiently detect certain types of image abnormalities or a specific pathological condition. Their *too narrow scope* prevents their effective use in clinical workflows. AI-based systems would be more useful if they could predict more meaningful clinical endpoints, such as malignancy of lesions, need for treatments, patient survival (Oren 2020). To widen the scope of AI algorithms, they need to be fed with as much information as we have on the health status of subjects.

The **integration of the complementary information** encoded in omics data, EHR, imaging data, clinical data, phenotypic information and lifestyle of subjects is expected to increase the understanding of human health and disease conditions, and finally to allow personalized preventive, diagnostic, and therapeutic strategies. The potential benefits that similar approaches would bring to the oncology field have been highlighted recently (Lipkova 2022). Two main **challenges** have to be addressed: the collection of a significant amount of such data for a large population; the development of **advanced analysis strategies to mine heterogeneous data** to exploit the complementary information they encode. It has been agreed in recent years that scientific data should be collected and managed according to the **FAIR guiding principles** (Wilkinson 2016), according to which data should be **F**indable, **A**ccessible, **I**nteroperable and **R**eusable. Medical data collection for research purposes is however a world-wide issue that is beyond the scope of this proposal to be fully addressed. INFN has been developing computing infrastructures for enabling research in several fields of fundamental physics, and this expertise, which is extremely useful in medical data analysis (Retico 2021), is being extended to medical research also thanks to the projects funded by the PNRR. However, the recent review paper by Acosta *et al.* (2022) states «... we are far better at collating and storing such data, than we are at data analysis.» Therefore, since the development of **innovative strategies to analyze data from heterogeneous sources**, including solutions based on AI, falls very well within the expertise and the research interests of our research team members, the AIM_MIA proposal will be focused mainly on this challenge.

Project objectives

Objective 1. Mining multi-input data

The investigation of the reports of **different diagnostic tests and clinical information** available for a patient **allows medical experts to define an integrated clinical profile to outline the most appropriate care path**. **This task could be supported by AI**, through the design of data analysis models that accept heterogeneous input, thus allowing multidimensional analysis. However, it is not straightforward to develop models capable of analyzing digital data acquired with different instruments, given their high variability in format, dimensionality and informative content. **The goal of Objective 1 is to develop and validate AI-based analysis pipelines that can handle a combination of heterogeneous data sources**, which may include medical images, diagnostic tests, phenotypic and genetic data.

Objective 2. Handling incomplete/missing/limited datasets

In the context of multi-input data analysis, **managing incomplete data and ensuring sample balance are two critical challenges**. Sensor failures, manual data entry errors, data corruption, or the absence of a specific clinical test for some patients can strongly affect multi-modal data collection. Excluding patients with missing/incomplete entries may strongly reduce the dataset size and also lead to a selection bias. **Objective 2 is focused on the development of strategies for data curation, imputation and augmentation** to avoid severe reductions of the sample, and to compensate for unbalanced or small samples.

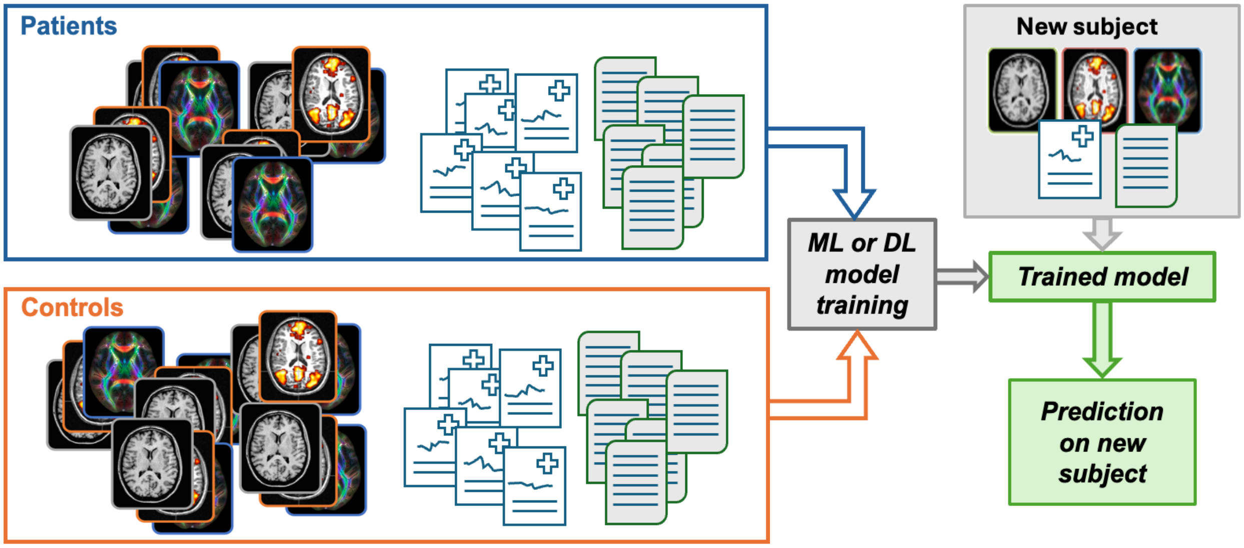
Objective 3. Developing a dedicated data and IT platform

To enable the development of the analysis strategies mentioned in Objectives 1 and 2, the **availability of large multi-modal data samples** is needed. To ensure effective access to data shared within the collaboration, including raw data and processed information, **Objective 3 is devoted to the development of a dedicated data platform**, integrated with computing resources, and **compliant with the FAIR principles and the GDPR regulation**.

Research methodology

Algorithms to mine multi-input data

The advancement of precision medicine, strongly depends on the possibility to combine data coming from different sources and to develop reliable AI models. Depending on the clinical question, a patient may undergo more than one different imaging modalities, such as CT, MRI or PET scans, different clinical tests, possibly genetic tests, and all these exams provide digital data with different types (numerical arrays, text) and dimensions. Even when we consider a single modality (e.g. the MRI with its multi-parametric nature), we may have to analyze more than one 3D volume of data. Moreover, complementary information, such as clinical or genetic data, have to be integrated and the correlation with the clinical outcome should be modeled. Due to the specific nature of data acquired in medicine (i.e. the



quantitative, semi-quantitative, contrast-based, structural or functional data, etc.), appropriate preprocessing is often needed to prepare and harmonize data before applying AI methods.

The best method to merge multi-modal data cannot be defined *a priori* as well as the best combination of input data, as the implementation of multi-modal algorithms is strictly bound to the specific available data and research question. Many different approaches can be explored to achieve this goal. Different fusion strategies will be investigated, such as early fusion, joint fusion and late fusion (Huang 2020) as well as more advanced AI approaches such as Vision Transformers (ViTs) (Pei 2023). Another interesting approach could be the use of Convolutional Neural Networks (CNNs) to extract features to be used in a Transformer (Li 2023) or, in case the problem needs to rely on large portions of images, the integration of radiomic features computed on those large areas with fine-grained features extracted instead by a ViT and/or a CNN.
A selection of case studies that will be addressed during the project, for which the data are already available by the AIM_MIA research team (see table of available datasets below) includes, for example, neuroimaging studies to integrate anatomical and functional characteristics in subjects with Autism Spectrum Disorder (ASD) or to perform virtual biopsies in subjects with glioblastoma.

Preliminary results by the research group on multi-input medical data analysis:

- Integration of features from structural and functional MRI images (Saponaro 2024)
- Integration of radiological images and clinical data (Lizzi 2024)
- Integration of radiomic and dosiomic features (Piffer 2024)

Methods for dealing with incomplete/unbalanced/limited datasets

Since medical image datasets are generally small, preserving all available information and filling in missing information is critical to developing reliable algorithms. This WP will focus on developing traditional and advanced data imputation techniques that can handle complex dependencies in multi-input data, and exploring novel data augmentation methods to generate high-quality synthetic data. Regarding the former, statistical methods for data imputation and tree-based algorithms with surrogate splits will be studied. As regards data augmentation, both traditional and advanced generative deep learning strategies will be applied. Traditional data augmentation consists in applying transformations such as rotations, flipping, zooming and so on and produces an increase of data samples that are quite similar to the original ones, achieving a slight improvement in the diversity of the dataset. For this reason, deep learning methods for data generation can be explored. These methods include Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) and diffusion probabilistic models. Moreover, another way to enlarge datasets can be the application of super-resolution techniques to transform low resolution images, e.g. DWI MRI, into high resolution ones.

In addition, medical images can suffer from a partial lack of information, and the missing regions of the images could be provided by inpainting methods. Possible causes of these difficulties may be due to artifacts or positioning outside the MR field of view or, in CT, incomplete projection data due to sparse sampling or data truncation. Architectures like Context Encoder Network or GAN will be explored.

Preliminary results by the research group on incomplete/unbalanced/limited datasets:

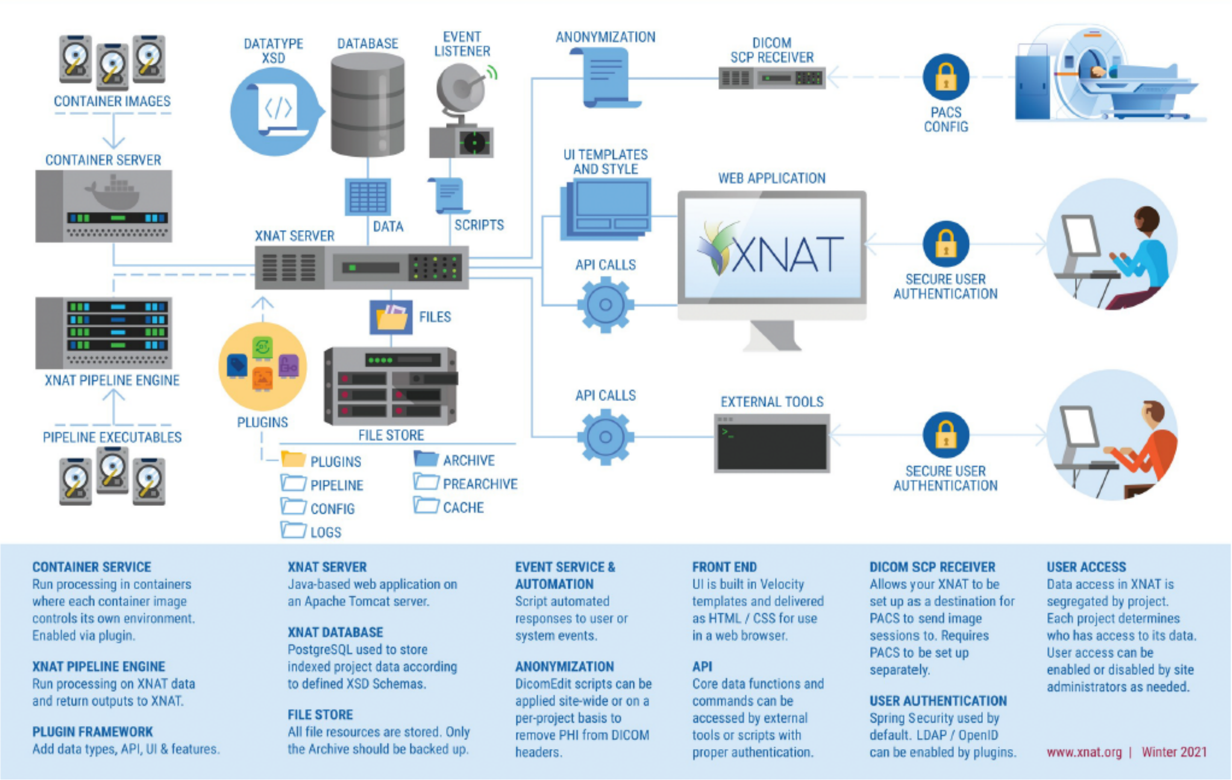
- Systematic review of methods to tackle the small data problem (Piffer 2024)
- Mixed use of public and proprietary data collections (Ubaldi 2021) or merging multiple data samples (Lizzi 2022) in AI model training
- Transfer learning approaches in brain imaging (Fiscone 2024)

Data platform

An IT data platform will be developed on top of XNAT (https://www.xnat.org/), a fully open-source informatics software platform, designed for medical imaging-based research; it includes features suitable for managing medical imaging data and associated metadata (e.g. uploading and viewing images, organizing the metadata, sharing and downloading the data). The user interface allows the exploration of multimodal, multidimensional, and heterogeneous datasets through sorting and filtering. The configuration of a XNAT-based platform is highly customizable, also in terms of user permissions, and new data types can be created. The data platform will be compliant with the FAIR data principle and GDPR regulation, even though in the first instance the platform will be used to store public data only. The plan is to link the platform to adequate INFN computing resources so that the analysis tools can run on the stored data and the processed information can be stored back within the same platform. This unified structured platform will help in sharing source data, analysis tools, and output data within the collaboration.

Preliminary results by the research group on data platform development:

- A GDPR compliant platform was developed within the ARIANNA project (Retico 2017), https://arianna.pi.infn.it/it
- A platform to foster the development of clinical AI-based support tools was designed in collaboration with AIFM (Retico 2021)



CONTAINER SERVICE
Run processing in containers where each container image controls its own environment. Enabled via plugin.

XNAT PIPELINE ENGINE
Run processing on XNAT data and return outputs to XNAT.

PLUGIN FRAMEWORK
Add data types, API, UI & features.

XNAT SERVER
Java-based web application on an Apache Tomcat server.

XNAT DATABASE
PostgreSQL used to store indexed project data according to defined XSD Schemas.

FILE STORE
All file resources are stored. Only the Archive should be backed up.

EVENT SERVICE & AUTOMATION
Script automated responses to user or system events.

ANONYMIZATION
DicomEdit scripts can be applied site-wide or on a per-project basis to remove PHI from DICOM headers.

FRONT END
UI is built in Velocity templates and delivered as HTML / CSS for use in a web browser.

API
Core data functions and commands can be accessed by external tools or scripts with proper authentication.

DICOM SCP RECEIVER
Allows your XNAT to be set up as a destination for PACS to send image sessions to. Requires PACS to be set up separately.

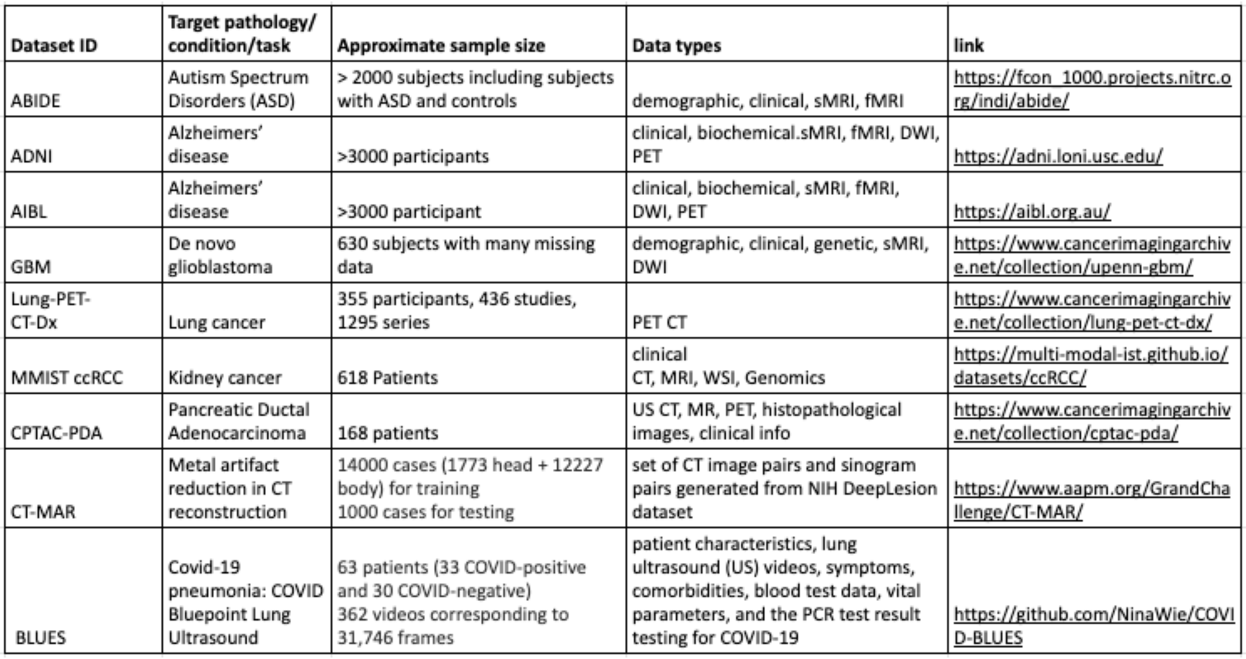
USER AUTHENTICATION
Spring Security used by default. LDAP / OpenID can be enabled by plugins.

USER ACCESS
Data access in XNAT is segregated by project. Each project determines who has access to its data. User access can be enabled or disabled by site administrators as needed.

www.xnat.org | Winter 2021

Data collection

Several multimodal datasets (see the table below) will be collected from publicly available data sources and organized in the AIM_MIA data platform.



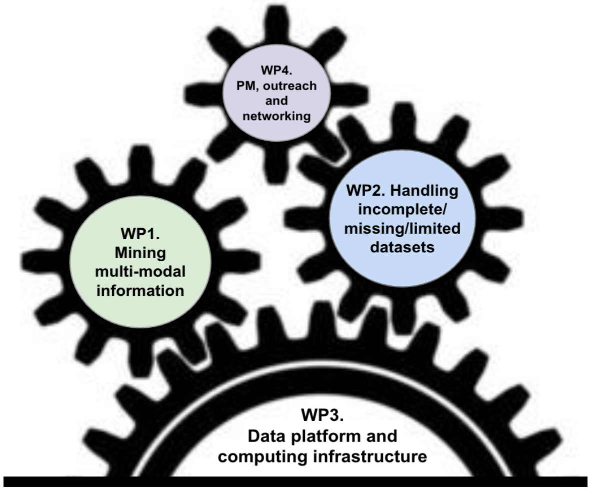
Dataset ID	Target pathology/ condition/task	Approximate sample size	Data types	link
ABIDE	Autism Spectrum Disorders (ASD)	> 2000 subjects including subjects with ASD and controls	demographic, clinical, sMRI, fMRI	https://fcon_1000.projects.nitrc.org/indi/abide/
ADNI	Alzheimers' disease	>3000 participants	clinical, biochemical.sMRI, fMRI, DWI, PET	https://adni.loni.usc.edu/
AIBL	Alzheimers' disease	>3000 participant	clinical, biochemical, sMRI, fMRI, DWI, PET	https://aibl.org.au/
GBM	De novo glioblastoma	630 subjects with many missing data	demographic, clinical, genetic, sMRI, DWI	https://www.cancerimagingarchive.net/collection/upenn-gbm/
Lung-PET-CT-Dx	Lung cancer	355 participants, 436 studies, 1295 series	PET CT	https://www.cancerimagingarchive.net/collection/lung-pet-ct-dx/
MMIST ccRCC	Kidney cancer	618 Patients	clinical CT, MRI, WSI, Genomics	https://multi-modal-ist.github.io/datasets/ccRCC/
CPTAC-PDA	Pancreatic Ductal Adenocarcinoma	168 patients	US CT, MR, PET, histopathological images, clinical info	https://www.cancerimagingarchive.net/collection/cptac-pda/
CT-MAR	Metal artifact reduction in CT reconstruction	14000 cases (1773 head + 12227 body) for training 1000 cases for testing	set of CT image pairs and sinogram pairs generated from NIH DeepLesion dataset	https://www.aapm.org/GrandChallenge/CT-MAR/
BLUES	Covid-19 pneumonia: COVID Bluepoint Lung Ultrasound	63 patients (33 COVID-positive and 30 COVID-negative) 362 videos corresponding to 31,746 frames	patient characteristics, lung ultrasound (US) videos, symptoms, comorbidities, blood test data, vital parameters, and the PCR test result testing for COVID-19	https://github.com/NinaWie/COVID-BLUES

In addition, Natural Language Processing techniques will be used to automate the search of publicly available datasets, and hence to expand the project data collection. Algorithms for scalable literature mining and document retrieval (e.g. ElasticSearch) will be adapted to the task and employed to mine and index publicly available corpora of scientific papers (e.g. PubMed) according to specific questions provided by experts. Deep Learning-based language models, e.g. BioBERT (Lee 2020), will be used to link questions and documents based on their semantic similarity.

Project organization

The objectives of the project will be achieved by means of the realization of four interconnected work-packages (WPs), which are described below. For each WP and for each task, a short description of the planned work and the indication of the leaders and of the contributing research groups is provided.

The timeline of the project is shown in the Gantt chart below. Finally, a table with the list of the proposed Milestones and the corresponding deadlines is reported.


WP4.
PM, outreach and networking
WP2. Handling incomplete/ missing/limited datasets
WP1. Mining multi-modal information
WP3. Data platform and computing infrastructure

WP1. Mining multi-modal information

WP leader: A. Chincarini (GE)

Task1.1 Feature-based approach to multi-input analysis

Task leader: P. Oliva (CA); Participants: BA, GE, LE, LNS, PV

Several feature extraction techniques (including DL-based ones) will be set up to convert medical data into descriptive features to be processed with ML/DL approaches/

Task1.2 Integration of multi-parametric and multi-modal imaging data

Task leader: C. Talamonti (FI); Participants: GE, LE, LNS, PV

Automatic co-registration pipelines will be built for PET, structural and functional MRI for quantitative features extraction to develop ML approaches for diagnosis/prognosis and treatment monitoring of diseases.

Task1.3 AI solutions for heterogeneous data analysis

Task leader: M. Marrale (CT); Participants: CT, GE, LE, LNS, MI, PI, PV

Innovative approaches involving ViTs or multiplex networks (e.g. Heterogeneous Graph Learning for Multi-modal Medical Data Analysis) that integrate various types of image and clinical patient data, capturing the complex relationships between patients in a systematic manner, will be also explored.

WP2. Handling incomplete/missing/limited datasets

WP leader P. Oliva (CA)

Task2.1 Traditional approaches for data curation and augmentation

Task leader: G. De Nunzio (LE); Participants: CA, MI

Well established techniques for data augmentation ((flipping, resizing, cropping, brightness, contrast)) and dataset balancing will be integrated in the project SW repository. Traditional approaches for data imputation (e.g. k-means, NN) will be also delivered.

Task2.2 Medical Image Data Generation

Task leader: C. Testa (BO); Participants: FE, PV

Generation of highly-resolved diffusion MR images (DW images) by means of a NN. Investigation and development of generative models for medical image (radiological, CT, ultrasound images) augmentation. Evaluation of GAN augmented datasets (Inception Score, FID, visual inspection).

Task2.3 Data inpainting with CNN

Task leader: G. Di Domenico (FE); Participants: CA, CT

Development of deep learning methods for inpainting sinograms or volumes affected by metallic implant artifacts.

WP3. Data platform and computing infrastructure

WP leader(s): F. Lizzi (PI)

Task3.1 Definition of requirements and user roles

Task leader: C. Scapicchio (PI); Participants: BA, PV

Collection of the requirements for the platform and identification of which data and metadata should be stored. The data model will be defined together with the user roles to establish who can make what on which data.

Task3.2 Realization and maintenance of the data platform prototype

Task leader: A. Formuso (PI); Participants: BA, BO, CA, GE, PV

Technical configuration of the XNAT platform (project type, data model, use of plugins) by following the defined requirements and specifications. Maintenance of the platform supporting its usage by the partners.

Task3.3 Integration of data processing pipelines and output storing

Task leader: F. Sensi (GE); Participants: PI, CA

Integration of the AI-based analysis pipelines within the platform through the configuration of specific plugins available in XNAT to directly run the analysis on the data stored in the platform and directly upload and store the intermediate and final results in the platform.

Task3.4 SW organization and repository

Task leader: I. Postuma (PV); Participants: GE, PI, BO

Definition of a protocol for the storage of a dataset shared by multiple hospital partners and organization of the developed software in a well-organized repository.

Task3.5 Data collection

Task leader: N. Curti (BO); Participants: CA, PI

Collection of the data extracted from the datasets listed in the Table above and set up of NLP-based analysis to identify and other datasets of interest. Populating the data platform with the collected data both raw and pre-processed.

WP4. Project management, outreach and networking

WP leader: A. Retico

Task4.1. Project management and networking

Task leader: A. Retico; Participants: local group coordinators and WP leaders

Activity planning and continuous monitoring of progress progression fostering the cooperation and exchange of ideas among researchers from the different groups involved. Strengthening the network of researchers interested in the development of advanced analysis techniques for medical applications.

Task4.2. Collaboration with AIFM and other associations

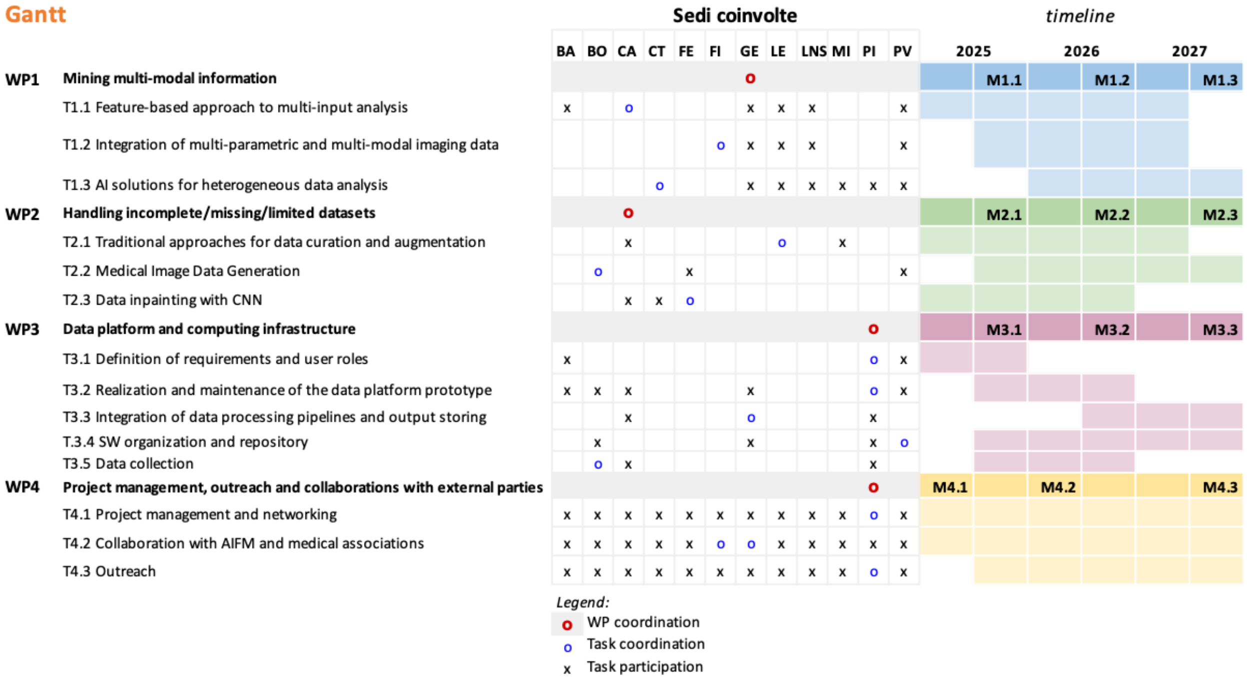
Task leaders: C. Talamonti (FI), A. Chincarini (GE); Participants: all groups

Continuous exchange of research ideas with AIFM and other associations active in the field of medical data analysis. Preparation of joint research proposals.

Task4.3. Outreach

Task leader: ME Fantacci (PI); Participants: all groups

Definition of an outreach program for the project, exploiting synergies and collaboration with AIFM, medical associations and collaborating Universities. Preparation of dissemination materials and participation in public events.



Gantt

		Sedi coinvolte												*timeline*					
		BA	BO	CA	CT	FE	FI	GE	LE	LNS	MI	PI	PV	2025		2026		2027	
WP1	**Mining multi-modal information**							o						■	M1.1	■	M1.2	■	M1.3
	T1.1 Feature-based approach to multi-input analysis	x		o				x	x	x			x	■	■	■	■	■	
	T1.2 Integration of multi-parametric and multi-modal imaging data						o	x	x	x			x		■	■	■	■	
	T1.3 AI solutions for heterogeneous data analysis				o			x	x	x	x	x	x			■	■	■	■
WP2	**Handling incomplete/missing/limited datasets**			o										■	M2.1	■	M2.2	■	M2.3
	T2.1 Traditional approaches for data curation and augmentation			x					o		x			■	■	■	■	■	
	T2.2 Medical Image Data Generation		o			x							x		■	■	■	■	■
	T2.3 Data inpainting with CNN			x	x	o								■	■	■	■		
WP3	**Data platform and computing infrastructure**											o		■	M3.1	■	M3.2	■	M3.3
	T3.1 Definition of requirements and user roles	x										o	x	■	■				
	T3.2 Realization and maintenance of the data platform prototype	x	x	x				x				o	x		■	■	■		
	T3.3 Integration of data processing pipelines and output storing			x				o				x					■	■	■
	T.3.4 SW organization and repository		x					x				x	o		■	■	■	■	■
	T3.5 Data collection		o	x								x			■	■	■		
WP4	**Project management, outreach and collaborations with external parties**											o		M4.1	■	M4.2	■	■	M4.3
	T4.1 Project management and networking	x	x	x	x	x	x	x	x	x	x	o	x	■	■	■	■	■	■
	T4.2 Collaboration with AIFM and medical associations	x	x	x	x	x	o	o	x	x	x	x	x	■	■	■	■	■	■
	T4.3 Outreach	x	x	x	x	x	x	x	x	x	x	o	x		■	■	■	■	■

Legend:
o WP coordination
o Task coordination
x Task participation

Deadline	Milestone	Description
31/12/25	M1.1	Definition of a generalizable AI-based pipeline that integrates features extracted from heterogeneous data
31/12/25	M2.1	Implementation of traditional data curation pipelines
31/12/25	M3.1	Identification of requirements, user roles and datasets to be included in the data platform, and creation of a shared SW repository

30/06/25	M4.1	Kick-off meeting organization and plan of collaboration with external parties
31/12/26	M1.2	Definition of a generalizable AI-based pipeline that integrated multiparametric or multimodal images
31/12/26	M2.2	Inpainting of metal artifact corrupted sinograms with NN
31/12/26	M3.2	Instantiation of a XNAT platform prototype and integration of data sets
39/06/26	M4.2	Identification of joint research or dissemination activities to promote the networking among researchers from different Institutions and associations
31/12/27	M1.3	Definition of an AI-based pipeline that integrates images and features extracted from heterogeneous data
31/12/27	M2.3	GAN generation of medical images
31/12/27	M3.3	Integration of data processing pipelines and output storing
31/12/27	M4.3	Organization of a workshop to discuss the results of the project with stakeholders

Description of the research group

The research team is constituted by INFN staff personnel (researchers and technologists), University staff members associated to INFN, and many young collaborators (PhD students, postdocs and young fellows) from either INFN or collaborating Universities. The main expertise and interests of each research group are described below. The involvement of each group in the WPs and tasks is highlighted in the previous section, whereas the use within the project of existing infrastructures and the collaboration with external entities and synergies with other projects are described in the following sections.

BA. Medical data analysis and managing of computing facilities.

BO. Biomedical data analysis (multi-omics, imaging, clinical information) through ML, AI and advanced statistical methods. Design and implementation of analysis algorithms (eg. classification, regression, dimensionality reduction, data imputation).

CA. Medical data preprocessing and analysis, simulations, imaging systems.

CT. Medical imaging data analysis (MRI, CT), deep learning in medicine, development of AI algorithms.

FE. Medical imaging data analysis, deep learning in medicine, artificial intelligence in tomography.

FI. Medical imaging data acquisition, organization and analysis , clinical radiology and radiomic analysis.

GE. Data analytics, modeling, nuclear neuroimaging, advanced statistics.

LE. Medical data acquisition and analysis, development of AI algorithms, development of pre-processing algorithms (for missing data and for dataset balancing).

LNS. Medical image processing and analysis, in particular for non-invasive imaging techniques.
MI. Medical imaging systems, application of AI methods to clinical data
PI. Medical data acquisition and analysis, data platform developing, computing facility developing and maintaining, Data Protection Office team member (dpo@infn.it).
PV. Medical imaging data analysis (micro-CT, CT, MRI), deep learning in medicine, SW organization and repository, small datasets handling

Connections with external entities

A long-standing collaboration between the research teams involved in the project and the local clinical centers (University Hospitals and IRCCS) at each site is active. At the national level, a 5-year research framework agreement has been renewed in January 2024 between INFN and the Italian Association of Medical Physicists (AIFM). This agreement foresees the collaboration between the parties for the synergistic realization of common research objectives in the healthcare field. The collaboration can also leverage significant connections with medical associations thanks to the role appointed to some of its members. A few examples are :
- the European Alzheimer's Disease Consortium (EADC, https://eadc.online/) - A. Chincarini has been elected member of the Executive Board
-the Italian Association of Nuclear Medicine (AIMN, https://aimn.it/) - A. Chincarini has been elected member of the Neurological Study Committee.

Synergic projects funded in the last three years on connected research topics

A list of research projects on connected research topics active during the last three years and involving AIM_MIA members is provided below. The activity carried out in the INFN Data Cloud project (C3SN), which is connected to PNRR projects where a suitable and secure infrastructure to collect and analyze medical data is under development (EPIC platform), is definitely of interest for our collaboration. A continuous exchange of information and ideas with the INFN personnel involved in these projects is foreseen. Analogously, a collaboration is in place with the on-going AI_INFN project (CSN5), which is expected to be able to provide access to INFN cloud resources (including GPUs) via the AI_INFN platform.

Projects funded by INFN

- next_AIM [CSN5, 2022-2024], https://www.pi.infn.it/aim/
- AI_MIGHT [CSN5, 2022-2024]
- ML_INFN [CSN5, 2020-2023]
- AI_INFN [CSN5, 2024-2026]

Projects funded by Ministry of University and Research

- PNRR CN ICSC Spoke 8 [2022-2025]
- PNRR PE FAIR Spoke 8 [2023-2025]
- PNRR PE HEAL Spoke 2 [2022-2025]
- PNRR ECS THE Spoke 1 and Spoke 4 [2022-2025]
- PNRR ANTHEM Spoke 4 [2022-2025]
- PNRR POC-2022 [2023-2025]

Projects funded by Ministry of Health

- Piano Operativo Salute (POS), TELENEURART [2023-2026]
- Ricerca Finalizzata 2021, *Probing neuroinflammation in the prodromal stages of alpha-synucleinopathies. [2023-2025]*
- Ricerca Finalizzata, *AI algorithms to automate the Total Marrow (Lymph-node) Irradiation by VMAT optimization using WB-CT/MRI and synthetic WB-CT: AuToMI project. [2021-2024]*

Expected impact and transferability of the results

The main research objective of this proposal is to develop robust and effective analysis pipelines to make predictions about the health status of individuals, by extracting and combining via multi-input AI-based tools the complementary and heterogeneous information provided by different data sources (images, diagnostic tests, phenotypic and genetic data). A significant contribution to the advancement of this field of research is expected within this 3-year research project. As a measurable indicator of performance, we propose the number of published papers in peer-reviewed journals relevant in the field of medical data analysis, which we expect will exceed five per year in our project.

The transferability of AI-based tools into the Clinics deserves a dedicated effort. Should some multi-input analysis pipelines developed in this project demonstrate high performance, robustness and generalization capabilities such as to be attractive for commercial purposes, the INFN Technology Transfer office will be involved in investigating the feasibility of this possible exploitation. Otherwise, the multi-input analysis pipelines and the results obtained will be made available as open resources for the scientific community to promote the reproducibility of the research results, and the large-scale validation of the AI-based software tools.

In addition, the project includes the development of a dedicated data platform based on open-source software potentially of great interest to many similar projects that deal with heterogeneous medical data, and a prototype that could be replicated in other research contexts.

Finally, another significant impact of this project will be the research network that will be consolidated. It will certainly constitute a solid context for young people in which to experiment with new ideas, while more experienced researchers will contribute to bringing out and grounding innovative approaches in this very dynamic field of research.

References

Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. Nature Medicine, 28(9), 1773–1784. https://doi.org/10.1038/s41591-022-01981-2

Denny and Collins, Precision medicine in 2030—seven ways to transform healthcare. Cell 2021;184:1415–9. https://doi.org/10.1016/j.cell.2021.01.015.

Fiscone, C., Curti, N., Ceccarelli, M., Remondini, D., Testa, C., Lodi, R., Tonon, C., Manners, D. N., & Castellani, G. (2024). Generalizing the Enhanced-Deep-Super-Resolution Neural Network to Brain MR Images: A Retrospective Study on the Cam-CAN Dataset. ENeuro, 11(5). https://doi.org/10.1523/ENEURO.0458-22.2023

Giger ML, Chan H-P, Boone J (2008) Anniversary Paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM. Medical Physics 35:5799–5820. https://doi.org/10.1118/1.3013555

Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., & Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. Npj Digital Medicine, 3(1), 136. https://doi.org/10.1038/s41746-020-00341-z

Kim, S., Lee, N., Lee, J., Hyun, D., & Park, C. (2023, June). Heterogeneous graph learning for multi-modal medical data analysis. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 4, pp. 5141-5150). https://doi.org/10.1609/aaai.v37i4.25643

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Li, W., Zhang, Y., Wang, G., Huang, Y., & Li, R. (2023). DFENet: A dual-branch feature enhanced network integrating transformers and convolutional feature learning for multimodal medical image fusion. Biomedical Signal Processing and Control, 80, 104402. https://doi.org/10.1016/j.bspc.2022.104402

Lipkova, J., Chen, R. J., Chen, B., Lu, M. Y., Barbieri, M., Shao, D., Vaidya, A. J., Chen, C., Zhuang, L., Williamson, D. F. K., Shaban, M., Chen, T. Y., & Mahmood, F. (2022). Artificial intelligence for multimodal data integration in oncology. Cancer Cell, 40(10), 1095–1110. https://doi.org/10.1016/j.ccell.2022.09.012

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Lizzi, F., Agosti, A., Brero, F., Cabini, R. F., Fantacci, M. E., Figini, S., Lascialfari, A., Laruina, F., Oliva, P., Piffer, S., Postuma, I., Rinaldi, L., Talamonti, C., & Retico, A. (2022). Quantification of pulmonary involvement in COVID-19 pneumonia by means of a cascade of two U-nets: training and assessment on multiple datasets using different annotation criteria. International Journal of Computer Assisted Radiology and Surgery, 17(2), 229–237. https://doi.org/10.1007/s11548-021-02501-2

Lizzi F, Brero F, Fantacci ME, Lascialfari A, Paternò G, Postuma I, Oliva P, Scapicchio C, Retico A (2024). A multi-input deep learning model to classify COVID-19 pneumonia severity from imaging and clinical data. IWBBIO, 1–12.

Oren, O., Gersh, B. J., & Bhatt, D. L. (2020). Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *The Lancet Digital Health*, *2*(9), e486–e488. https://doi.org/10.1016/S2589-7500(20)30160-6

Park, S. H., Han, K., Jang, H. Y., Park, J. E., Lee, J., Kim, D. W., & Choi, J. (2022). Methods for Clinical Evaluation of Artificial Intelligence Algorithms for Medical Diagnosis. *Radiology*, 1–12. https://doi.org/10.1148/radiol.220182

Pei, X., Zuo, K., Li, Y., & Pang, Z. (2023). A Review of the Application of Multi-modal Deep Learning in Medicine: Bibliometrics and Future Directions. International Journal of Computational Intelligence Systems, 16(1). https://doi.org/10.1007/s44196-023-00225-6

Piffer, S., Greto, D., Ubaldi, L., Mortilla, M., Ciccarone, A., Desideri, I., Genitori, L., Livi, L., Marrazzo, L., Pallotta, S., Retico, A., Sardi, I., & Talamonti, C. (2024). Radiomic- and dosiomic-based clustering development for radio-induced neurotoxicity in pediatric medulloblastoma. Child's Nervous System, 0123456789. https://doi.org/10.1007/s00381-024-06416-6

Piffer, S., Ubaldi, L., Tangaro, S., Retico, A., & Talamonti, C. (2024). Tackling the small data problem in medical image classification with artificial intelligence: a systematic review. Progress in Biomedical Engineering, 10, 22408–22418. https://doi.org/10.1088/2516-1091/ad525b

Retico A, Arezzini S, Bosco P, Calderoni S, Ciampa A, Coscetti S, et al. ARIANNA: A research environment for neuroimaging studies in autism spectrum disorders. Comput Biol Med 2017; 87. https://doi.org/10.1016/j.compbiomed.2017.05.017

Retico, A., Avanzo, M., Boccali, T., Bonacorsi, D., Botta, F., Cuttone, G., Martelli, B., Salomoni, D., Spiga, D., Trianni, A., Stasi, M., Iori, M., & Talamonti, C. (2021). Enhancing the impact of Artificial Intelligence in Medicine: A joint AIFM-INFN Italian initiative for a dedicated cloud-based computing infrastructure. Physica Medica, 91(October), 140–150. https://doi.org/10.1016/j.ejmp.2021.10.005

Saponaro S, Lizzi F, Serra G, Mainas F, Oliva P, Giuliano A, Calderoni S, Retico A. (2024). Deep learning based joint fusion approach to exploit anatomical and functional brain information in autism spectrum disorders. Brain Informatics, 11(1), 2. https://doi.org/10.1186/s40708-023-00217-4

Ubaldi, L., Valenti, V., Borgese, R. F., Collura, G., Fantacci, M. E., Ferrera, G., Iacoviello, G., Abbate, B. F., Laruina, F., Tripoli, A., Retico, A., & Marrale, M. (2021). Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples. Physica Medica, 90(September), 13–22. https://doi.org/10.1016/j.ejmp.2021.08.015

van Leeuwen, K. G., Schalekamp, S., Rutten,... van Ginneken, B., & de Rooij, M. (2021). Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. European Radiology, 31(6), 3797–3804. https://doi.org/10.1007/s00330-021-07892-z

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Additional documentation

Find attached the CV of the PI including a selection of 10 relevant publications in this research field.

Financial Information

Budget	Missioni	Inventario	Consumo	manutenzione	Licenze SW	Pubblicazioni	Totale
2025	43,5	2	19,5	2	1,5	10	78,5
2026	40	3	15	2	1,5	10	71,5
2027	40	3	15	2	1,5	10	71,5
Totale	123,5	8	49,5	6	4,5	30	221,5

Budget table

Details on financial requests for 2025

We made a request for the national INFN for accessing computing resources that are pivotal to carry out the AI-based analysis proposed by the AIM_MIA project. The request includes access to 2 NVIDIA A100 GPUs (with, possibly, 80 GB of VRAM) and 30 TB of fast storage which is necessary for the implementation of AI algorithms. Although medical image datasets are usually small, in fact, in multimodal analysis each subject may need a large amount of storage and VRAM especially for 3D images and genomic data.

Travel requests for project meetings and joint work sessions between partners were included. Consumable costs for maintenance and replacement of small IT devices (disks, monitors). Open access publication fees are requested when not covered by agreements with Publishers stipulated by INFN or by the collaborating Universities.

The costs of computational resources will be partially redirected on already existing INFN computing infrastructures (see also the requests in CALC5_TIER1).

Estimated financial plans for 2026 and 2027

The financial requests for the next two years of the project will not differ much from those of the first year.