



DARE
DIGITAL LIFELONG PREVENTION

CODE NO. PNC0000002

Spoke 1 Deliverable

S1.D4.4

Report on developed computational models

This research is co-funded by the Ministry of University and Research within the Complementary National Plan PNC-1.1 "Research initiatives for innovative technologies and pathways in the health and welfare sector"
D.D. 931 of 06/06/2022, PNC0000002 DARE - Digital Lifelong Prevention



S1.D4.4 Report on developed computational models

| Deliverable information | |
|----------------------------|--|
| Spoke number and title | Spoke 1 - Enabling Factors and Technologies for Digital Prevention |
| WP number and title | WP 4 |
| Related task(s) | Task 4. 3 - Data Mining, AI, ML, Deep Learning, big data analytics |
| Lead beneficiary | UNIBO |
| Contributing beneficiaries | UNIPA, UNIBA, UNIPR, UNIROMA2, ENG, EXP, IRCCS AOUBO |
| Dissemination level | Public |
| Due date | 14/12/2025 |
| Actual date of delivery | 10/12/2025 |
| Author(s) | Federico Chesani (UNIBO), Antonella Carbonaro (UNIBO), |
| Contributors | Giulia Massimino (UNIBO) |
| Quality Assurance | Fabio Calefato (UNIBA), Eugenio Martinelli (UNIROMA2) |

Document history

| Version | Date | Author(s) /Reviewer(s) (Beneficiary) | Description |
|---------|------------|---|---|
| 0.1 | 15/11/2025 | Antonella Carbonaro (UNIBO) | Contribution collection |
| 0.2 | 19/11/2025 | Federico Chesani (UNIBO) | First draft |
| 0.3 | 1/12/2025 | Federico Chesani (UNIBO) | Added paragraphs on Computational models |
| 0.4 | 10/12/2025 | Federico Chesani, Antonella Carbonaro, Sabato Mellone (UNIBO) | Final draft for internal review |
| 0.5 | 12/12/2025 | Fabio Calefato (UNIBA), Eugenio Martinelli (UNIROMA2) | Revision |
| 1.0 | 15/12/2025 | Antonella Carbonaro, Federico Chesani (UNIBO) | Post-review revision and released version |

Disclaimer

This publication reflects only the author's views and the Funding Agency is not liable for any use that may be made of the information contained therein.

Table of contents

| | |
|---|-----------|
| Publishable summary | 6 |
| 1. Introduction | 7 |
| 1.1. How to read this Report..... | 7 |
| 1.2. Computational models developed within the project..... | 8 |
| 1.3. Health targets, and technical targets..... | 10 |
| 1.4. Development stages..... | 13 |
| 2. Computational Models | 15 |
| 2.1. Conversational Model for Analytical Data Exploration..... | 16 |
| 2.2. One Health Data Platform..... | 24 |
| 2.3. Personalised Environmental Clinical Risk Score | 30 |
| 2.4. 5-year Cognitive Decline Score..... | 32 |
| 2.5. Automatic CAP Recognition..... | 34 |
| 2.6. Physical Activity Index (PAI) | 37 |
| 2.7. MDS-UPDRS Classifier | 39 |
| 2.8. AI Framework for automatically revealing spatial-temporal-spectral EEG signatures. 42 | |
| 2.9. AI Framework for supporting the analysis of EEG-derived brain functional connectivity..... | 44 |
| 2.10. Workflow for trustworthy EEG decoding with deep neural networks..... | 46 |
| 2.11. Allograft Risk Score | 48 |
| 2.12. Automatic segmentation models to identify Region of Interests | 54 |
| 2.13. Model for Automatic Issue Classification..... | 57 |
| 2.14. Model for Emotion Recognition in software development | 59 |
| 2.15. Generic Augmentation of 3d neuroimaging data..... | 61 |
| 2.16. Framework for the automatic generation of regulatory documentation in AI-based medical software | 63 |
| 2.17. Identification of key factors causing intellectual disability in Down syndrome subjects..... | 65 |
| 2.18. Multilingual Medical Chatbot Based on Large Language Models..... | 68 |
| 2.19. Integrated Web Platform for Clinical Data Management and Medical Chatbot | 70 |
| 2.20. Automatic Extraction of Clinical Data from Reports and Population REDCap Databases..... | 72 |
| 2.21. Sleep Management Platform and Digital Support for Patient Health..... | 75 |
| 2.22. Artificial Intelligence Pipeline for the Automatic Classification of Periprosthetic Hip Fractures..... | 77 |
| 2.23. Automatic Detection of Noise in ECG Signals for Wearable Devices..... | 79 |

| | |
|---|-----------|
| 2.24. Computational Model for Breast Cancer Prevention and Diagnosis | 81 |
| 2.25. Cardiovascular Risk Assessment and Multimodal Data Integration..... | 83 |
| 2.26. A unified computational framework to describe individual dynamics, pairwise interactions, and high-order relationships within multivariate physiological data..... | 85 |
| 2.27. XAI for Histopathological Images | 88 |
| 2.28. Semantic Segmentation of gliomas on brain MRIs | 90 |
| 2.29. Glioblastoma Treatment Response Classification | 92 |
| 2.30. In Silico Trials to reduce the risk of hip fracture in fragile elders | 92 |
| 3. Bibliography of DARE-related publications..... | 95 |

Publishable summary

This Deliverable reports on the status of Computational Models developed up to Month 36 of the DARE project. Several Computational Models have been designed to address challenges in the health domain (as per the DARE goals), facing a huge diversity of tasks. This deliverable presents 30 different models, addressing a range of heterogeneous health-related issues, from more specific ones, such as the interpretation of EEG signals, to software frameworks that support common tasks like data management, filtering, and retrieval. For each Computational Model, a brief description is provided and, when relevant, links to the related scientific products are given.

1. Introduction

This Deliverable reports on the status of Computational Models developed up to Month 36. Within or in collaboration with Spoke 1 of the DARE project, several Computational Models have been designed to address a set of horizontal challenges, related to enabling tools and technologies, and vertical challenges related to specific clinical research topics within the pilot activities carried out in Spoke 2 and Spoke 3, facing a huge diversity of tasks. Indeed, the general health field is vast and presents a wide variety of problems to be tackled; as a consequence, a wide variety of methods, approaches, and tools are presented in this report. Hence, under the term “Computational Models”, we will include contributions that span fields like Software Engineering, Affective Computing, Medical Imaging Analysis, Decision Support Systems, Digital Twins, Data Collection and Management frameworks.

A common trait is shared among the models presented here: they might be directly related to DARE internal pilot projects, or rather they might address a more general issue about best software architectures, but all the models share the common goal of advancing automation, supporting experimental research, and enabling informed decision-making in a domain where human factors, data scarcity, and methodological constraints remain critical obstacles.

1.1. How to read this Report

This Deliverable enlists 30 different Computational Models that have been developed within the DARE project Framework. These models differ in several features, and although some traits are shared among several models, it would be very challenging to cluster them according to a specific criterion. The choice of any criterion would, in the end, result in an arbitrary decision.

To ease access to the many models, in this Section, we propose some tables that summarise the most important features. Aspects such as the health-related targets and the technical targets tackled by the models, which method is being exploited, which is the current development status, which data has been used, if the approach has been validated, and if it has already been published in scientific literature, are summarised in the following subsections.

Section 2 is entirely devoted to presenting each model with a deeper detail. For each one of the 30 models, a general introduction is provided, followed by a more technical section where details about techniques, the dataset used, and the results are presented. When a model has already been published, relevant publications (together with other bibliography) are reported at the end of each subsection: this would make it easier for readers interested in a specific model.

Finally, in Section 3, all the scientific publications related to the DARE project are listed. It is worth noting that 58 different papers are listed, thus proving the quality and quantity of results obtained within DARE.

1.2. Computational models developed within the project

As of the closing date of this deliverable, it is possible to report 30 different Computational Models. With the term “Computational Model”, we encompass a variety of applications that range from Risk Score Estimators to LLM-based human-machine interfaces, to transversal software platforms that provide common solutions to health-related frequent tasks, such as storing, retrieving, and filtering data. A list of the Computational Models can be found in Table 1.

Table 1: List of the Computational Models reported in this Deliverable

| # | Contributors | Model Name |
|----|--------------------------------|---|
| 1 | Engineering | Conversational Model for Analytical Data Exploration |
| 2 | Exprivia | One Health Data Platform |
| 3 | | Personalised Environmental Clinical Risk Score |
| 4 | | 5-year Cognitive Decline Score |
| 5 | University of Parma | Automatic CAP Recognition |
| 6 | University of Parma | PAI Index |
| 7 | | MDS-UPDRS Classifier |
| 8 | University of Bologna | AI Framework for automatically revealing spatial-temporal-spectral EEG signatures. |
| 9 | | Ai Framework for supporting the analysis of EEG-derived brain functional connectivity |
| 10 | | Workflow for trustworthy EEG decoding with deep neural networks |
| 11 | University of Rome Tor Vergata | Allograft Risk Score |
| 12 | University of Roma | Automatic segmentation models to identify Region of Interests |
| 13 | University of Bari | Model for Automatic Issue Classification |
| 14 | | Model for Emotion Recognition in software development |
| 15 | | Generic Augmentation of 3d neuroimaging data |
| 16 | | Framework for the automatic generation of regulatory documentation in AI-based medical software |

| # | Contributors | Model Name |
|----|--------------------------------|---|
| 17 | IRCCS AOUBO S. Orsola, Bologna | Identification of key factors causing intellectual disability in Down syndrome subjects |
| 18 | University of Parma | Multilingual Medical Chatbot Based on Large Language Models |
| 19 | University of Parma | Integrated Web Platform for Clinical Data Management and Medical Chatbot |
| 20 | University of Parma | Automatic Extraction of Clinical Data from Reports and Population REDCap Databases |
| 21 | University of Parma | Sleep Management Platform and Digital Support for Patient Health |
| 22 | University of Parma | Artificial Intelligence Pipeline for the Automatic Classification of Periprosthetic Hip Fractures |
| 23 | University of Parma | Automatic Detection of Noise in ECG Signals for Wearable Devices |
| 24 | University of Palermo | Computational Model for Breast Cancer Prevention and Diagnosis |
| 25 | University of Palermo | Cardiovascular Risk Assessment and Multimodal Data Integration |
| 26 | University of Palermo | A unified computational framework to describe individual dynamics, pairwise interactions, and high-order relationships within multivariate physiological data |
| 27 | University of Palermo | XAI for Histopathological Images |
| 28 | University of Palermo | Semantic Segmentation of gliomas on brain MRIs |
| 29 | University of Palermo | Glioblastoma Treatment Response Classification |
| 30 | University of Bologna | In Silico Trials to reduce the risk of hip fracture in fragile elders |

1.3. Health targets, and technical targets

Each model reported in this deliverable aims at a particular medical target; at the same time, such a goal can be viewed also from a technical perspective. It is worthy then to compare these two dimensions.

As a general technical consideration, the majority of the models can be grouped into four main classes:

- “Human-computer interfaces”, where a need for intelligent software able to better understand human-made documents and data is needed. Typical applications take care of ingesting natural language-based documents, but also to interact with human operators and help them to access the data; It should not surprise that Large Language Models (LLM) are the preferred solution;
- Signal analysis, in particular with respect to physiological signals such as EEG, ECG, but also signal collected from wearables; here the adopted solutions exploit both Deep Learning, as well as classical Machine Learning methods;
- Image segmentation and Recognition; typical applications are the recognition of regions with specific characteristics, starting from diagnostic images such as, for example, MRI; Deep Learning techniques seem to be preferred, although interesting result are obtained when mixing them with more standard classification techniques built on top of feature extraction algorithms.
- Classification tasks, towards diagnosis/prediction; classical Machine Learning and Statistical Learning techniques appear to be often preferred.

It is worthy to mention that a few models provide general software architectures supporting common process like, e.g. the management of health-related data, but also the best practices to adopt when dealing with AI-based solutions applied to the health field. Table 2 summarizes the health-related and technical targets of the considered models, highlighting their medical focus, intended technical objectives, and the technologies adopted for their implementation.

Table 2: Health and technical targets

| # | Model Name | Medical Target/Disease/Pathology | Technical target | Adopted technology |
|---|--|---|---|--|
| 1 | Conversational Model for Analytical Data Exploration | General query over domain-specific databases and general medical ontologies/standards | Transform natural language queries into SQL queries compliant with some target DB Schema. | LLM + RAG |
| 2 | One Health Data Platform | General-purpose Health- related data platform, with support for synthetic data; Current implementation focused on Senile Dementia | Integration of heterogenous data sources, together with support for synthetic data. | Pipeline-shaped architecture for ingestion of heterogenous data sources; Synthea™ for synthetic data generation. |
| 3 | Personalised Environmental Clinical Risk Score | Clinical Risk Score linked to environmental, pollution and social data about the subject | Adoption of existing state-of-the art approaches for estimating the Risk score. | Not yet committed; under considerations: Clustering, Autoencoders. |

| # | Model Name | Medical Target/Disease/Pathology | Technical target | Adopted technology |
|----|---|--|---|--|
| 4 | 5-year Cognitive Decline Score | Estimation of the evolution of cognitive abilities over time. | Adoption of existing state-of-the-art approaches for estimating the Risk score. | Not yet committed; under considerations: LSTM, Tree Ensemble (Gradient Boosting), Survival models (time-to-event). |
| 5 | Automatic CAP Recognition | Recognition of Cyclic Alternating Patterns (CAP) in EEG Signals | Time-related signals analysis and segmentation | Deep Learning |
| 6 | PAI Index | Novel Physical Activity Index (PAI) to quantify the daily physical activity. | Adoption and validation of data coming from off-the-shelf, commercial wearable devices; definition of an index, with proper thresholds and semantics. | Analytic formula |
| 7 | MDS-UPDRS Classifier | Parkinson Disease; evaluation of PD severity. | Classification task | Deep Learning (CNN and LSTM) |
| 8 | AI Framework for automatically revealing spatial-temporal-spectral EEG signatures. | EEG analysis | Time-related signals analysis, segmentation and classification. | Deep Learning |
| 9 | AI Framework for supporting the analysis of EEG-derived brain functional connectivity | EEG analysis | Time-related signals analysis, segmentation and classification. | Deep Learning (CNN, in particular) |
| 10 | Workflow for trustworthy EEG decoding with deep neural networks | EEG analysis | EE Time-related G signals analysis, segmentation and classification. | Deep Learning, with automatic hyperparameter search exploring algorithm. |
| 11 | Allograft Risk Score | Renal dysfunction in kidney transplantation | Regression task; anomaly detection | Partial Least Square Regression |
| 12 | Automatic segmentation models to identify Region of Interests | Vertebrae and metastatic Lesions identifications from CT; intervertebral discs identification from MRI | Image analysis and region identification | DL, mainly CNN. |
| 13 | Model for Automatic Issue Classification | General support to health-related systems | Text classification | LLM |
| 14 | Model for Emotion Recognition in software development | General Health | General Classification task | Clustering |
| 15 | Generic Augmentation of 3d neuroimaging data | Generation of T1-weighted MRI volumes | Data Augmentation | Diffusion Models |
| 16 | Framework for the automatic generation of regulatory documentation in AI-based medical software | Legal documentation generation and support for medical software devices. | Software Engineering support; MLOps. | MLOps |
| 17 | Identification of key factors causing intellectual disability in Down syndrome subjects | Down Syndrome, identification of factors involved in cognitive impairment worsening | Classification, regression. | Ensemble Trees (gradient Boosting); Shapely value analysis for explanation. |
| 18 | Multilingual Medical Chatbot Based on Large Language Models | Supporting physician in accessing and retrieving information | Multilingual LLM-based Chatbot | LLM |
| 19 | Integrated Web Platform for Clinical Data Management and Medical Chatbot | General human support in dealing with medical data storage; Lung Disease Prevention | Data entry automatization; automatic data extraction from natural language-based documents. | Software design; LLM integration |
| 20 | Automatic Extraction of Clinical Data from Reports and Population REDCap Databases | General data extraction from textual medical reports. | Automatic data extraction from natural language-based documents. | LLM |
| 21 | Sleep Management Platform and Digital Support for Patient Health | Software framework for collection and management of clinical data and textual information. | Software support to data collection and management, from sensors as well from human input. | Software design |
| 22 | Artificial Intelligence Pipeline for the Automatic Classification of | Hip fractures identification in radiographic images. | Region identification and classification in radiographic images. | Deep Learning |

| # | Model Name | Medical Target/Disease/Pathology | Technical target | Adopted technology |
|----|---|---|---|---|
| | Periprosthetic Hip Fractures | | | |
| 23 | Automatic Detection of Noise in ECG Signals for Wearable Devices | ECG signal analysis | Time-related signal filtering and analysis | ML and DL techniques |
| 24 | Computational Model for Breast Cancer Prevention and Diagnosis | Breast cancer diagnosis from extracted features and from MRI images. | Classification from extracted features; feature extraction from MRI images. | Tree ensemble; DL |
| 25 | Cardiovascular Risk Assessment and Multimodal Data Integration | Cardiovascular risk | Platform for collection of heterogenous data and fusion | Not applicable |
| 26 | A unified computational framework to describe individual dynamics, pairwise interactions, and high-order relationships within multivariate physiological data | General-purpose biomedical platform; classification of physiological and pathological states. | Data collection, feature extraction, classification. | Several different methods, adapting to the specific case. |
| 27 | XAI for Histopathological Images | Classification and Explanation generation for cancer diagnosis from tissue images. | Xai approaches for classification and explanation generation | XAI, several different techniques |
| 28 | Semantic Segmentation of gliomas on brain MRIs | Segmentation/classification of gliomas/glioblastomas in MRI images. | Image/region segmentation. | DL, CNN in particular. |
| 29 | Glioblastoma Treatment Response Classification | DDS supporting glioblastoma treatment prognosis | Classification | ML |
| 30 | In Silico Trials to reduce the risk of hip fracture in fragile elders | Prevention of hip fractures in frail elders | In Silico Trial | Biophysical Digital Twin |

1.4. Development stages

Although each model is usually developed following a tailored development process, it is worth considering the different maturity levels of each Computational Model. Some observations are mandatory:

- Not all the development processes started at the same time within the project's horizon: some models, strictly connected to pilots, have been investigated later than other models;
- Some models exploited existing dataset; other models instead have been trained on data freshly collected as part of the DARE project; clearly, Computational Models that were developed/trained starting from existing dataset might have reached a higher maturity level earlier than other models.

Table 3 reports the development stage of the Computational Models, detailing their current implementation status, the use of existing or newly collected data, and the approaches adopted for performance evaluation.

Table 3: Development stage of Computational Models

| # | Model Name | Development status | Use of existing data/Collection of new data | Performances evaluation |
|----|---|---|---|--|
| 1 | Conversational Model for Analytical Data Exploration | Prototype implemented; Proof-of-concept obtained; Initial validation performed. | Existing medical ontologies | Initial, qualitative evaluation of a specific LLM model. |
| 2 | One Health Data Platform | Prototype implemented; Proof-of-concept obtained; Initial validation performed. | Use of existing data; Generation of syntetic data | Initial, qualitative evaluation. |
| 3 | Personalised Environmental Clinical Risk Score | State-of-the-art analysis. | Use of the repository "One Health Data Platform" | Not yet in this stage |
| 4 | 5-year Cognitive Decline Score | State-of-the-art analysis. | Use of the repository "One Health Data Platform" | Not yet in this stage |
| 5 | Automatic CAP Recognition | Prototype implemented; Proof-of-concept obtained; Initial validation performed. | Use of existing repository "CAP Sleep Database" | Initial, qualitative evaluation. |
| 6 | PAI Index | Prototype implemented; Proof-of-concept obtained; Initial validation performed. | New data collected during the project. | Initial, qualitative evaluation. |
| 7 | MDS-UPDRS Classifier | Prototype implemented; Proof-of-concept obtained. | New data collected during the project. | Not yet in this stage. |
| 8 | AI Framework for automatically revealing spatial-temporal-spectral EEG signatures. | Implemented, validated. | Data already available | Yes |
| 9 | AI Framework for supporting the analysis of EEG-derived brain functional connectivity | Implemented, validated. | Data already available | Yes |
| 10 | Workflow for trustworthy EEG decoding with deep neural networks | Implemented, validated. | Data already available | Yes |
| 11 | Allograft Risk Score | Implemented, validated. | Data already available | Yes |
| 12 | Automatic segmentation models to identify Region of Interests | Implemented; Initial validation performed. | Data already available | Initial, qualitative evaluation. |
| 13 | Model for Automatic Issue Classification | Implemented, validated. | Data Already Available | Yes |
| 14 | Model for Emotion Recognition in software development | Implemented, validated. | Data Already Available | Yes |
| 15 | Generic Augmentation of 3d neuroimaging data | Implemented, validated. | Data already available | Yes |
| 16 | Framework for the automatic generation of regulatory | Implemented; Initial validation performed. | Not applicable | Initial, qualitative evaluation. |

| # | Model Name | Development status | Use of existing data/Collection of new data | Performances evaluation |
|----|---|--|--|----------------------------------|
| | documentation in AI-based medical software | | | |
| 17 | Identification of key factors causing intellectual disability in Down syndrome subjects | Implemented; Initial validation performed. | Use of Synthetic data; collection of real data ongoing | Initial evaluation |
| 18 | Multilingual Medical Chatbot Based on Large Language Models | Implemented; Initial validation performed. | Use of existing Medical Corpora | Initial, qualitative evaluation. |
| 19 | Integrated Web Platform for Clinical Data Management and Medical Chatbot | Implemented; Initial validation performed. | Not applicable | Initial, qualitative evaluation. |
| 20 | Automatic Extraction of Clinical Data from Reports and Population REDCap Databases | Implemented; Initial validation performed. | Use of data collected in a running medical trial, for validation purposes. | Initial evaluation |
| 21 | Sleep Management Platform and Digital Support for Patient Health | Implemented; Initial validation performed. | Not applicable | Initial evaluation |
| 22 | Artificial Intelligence Pipeline for the Automatic Classification of Periprosthetic Hip Fractures | Implemented; Initial validation performed. | Data already available | Initial, qualitative evaluation. |
| 23 | Automatic Detection of Noise in ECG Signals for Wearable Devices | Implemented; Initial validation performed. | Data already available | Initial, qualitative evaluation. |
| 24 | Computational Model for Breast Cancer Prevention and Diagnosis | Implemented, validated. | Data already available | Yes |
| 25 | Cardiovascular Risk Assessment and Multimodal Data Integration | Under implementation | Collecting data for future model definition | No |
| 26 | A unified computational framework to describe individual dynamics, pairwise interactions, and high-order relationships within multivariate physiological data | Implemented, validated. | Data collected within the project; data already existing | Yes |
| 27 | XAI for Histopathological Images | Implemented, validated. | Data collected within the project; data already existing | Yes |
| 28 | Semantic Segmentation of gliomas on brain MRIs | Implemented, validated. | Data collected within the project; data already existing | Yes |
| 29 | Glioblastoma Treatment Response Classification | Implemented, validated. | Data collected within the project; data already existing | Yes |
| 30 | In Silico Trials to reduce the risk of hip fracture in fragile elders | Implemented, validated. | data already existing | yes |

2. Computational Models

This section provides a more detailed view of the Computational Models developed within the DARE project. Thirty different Computational Models are reported here; for each model, an introduction is provided that states the context and goals, followed by some technical insight. The level of detail provided in each model description varies due to the significant heterogeneity of the targets, the technical solutions adopted, and the development stage of each model. “Mature” models will be described in more detail, while models in the early development phases may be simply sketched in terms of goals, development steps, and ideas.

2.1. Conversational Model for Analytical Data Exploration

This model focuses on providing an intuitive and interactive medium for exploring — and potentially aggregating — complex clinical data, presented both in textual form and through graphical or diagrammatic representations. Specifically, it is about an intelligent conversational interface capable of understanding natural language requests, in both Italian and English.

The interface would ideally be able to interpret user intents expressed in natural language and translate them into executable queries, such as SQL statements for an OMOP CDV v5.4¹ compliant relational database (one of the healthcare data representation standard), or environmental time-series data currently stored in databases query-able through SQL-like protocols.

Depending on the interaction, the model could generate context-aware queries aligned with the clinical data model and return insights or visual summaries to facilitate data navigation and understanding.

A key challenge is the interpretation of clinical vocabularies and ontologies, since OMOP relies on complex concept tables that require semantic reasoning to map user terms to the correct entities.

To enable such capabilities, a Lightweight Language Model (LLM) — possibly an open-source model enhanced with Retrieval-Augmented Generation (RAG) — could be employed to support semantic understanding and contextual coherence.

In a more advanced stage, this approach might evolve towards an agent-based architecture, where dedicated agents or modules would handle tasks, such as: concept disambiguation and mapping to OMOP vocabularies, managing clarification dialogues when user requests are ambiguous or incomplete, data pre- and post-processing for visualization, and interfacing with heterogeneous database dialects.

This use case fits within the broader domain of **Generative AI for Business Intelligence**, aiming to enhance data accessibility, reasoning, and decision support through natural language interaction.

2.1.1. Technical Insight

The design focuses on defining a modular and extensible architecture that operationalizes the **Generative AI for Business Intelligence** scenario described above.

It aims to translate natural language requests into structured analytical queries through a Text-to-SQL pipeline enhanced with Retrieval-Augmented Generation (RAG) and domain-specific reasoning. The solution is not bound to a specific database implementation, yet it assumes that data are harmonized under a relational schema (e.g., OMOP for clinical data). A relevant requirement to support accurate query synthesis is that the database structure must expose metadata and

¹ OHDSI. (s.d.). *OMOP CDM v5.4*. Extracted in 2025 from OMOP Common Data Model: <https://ohdsi.github.io/CommonDataModel/cdm54.html>

descriptive information (such as table comments and column annotations), which the model can use during reasoning and schema interpretation.

The adopted design emphasizes modularity, interoperability, and explainability, providing a foundation for iterative enhancement, domain-specific fine-tuning, and future integration with agent-based or multimodal components. This architecture (see Figure 1) enables multi-turn, context-aware data exploration, while maintaining independence from specific database technologies or model providers — thus supporting future extensions such as fine-tuned LLMs or multimodal input channels.

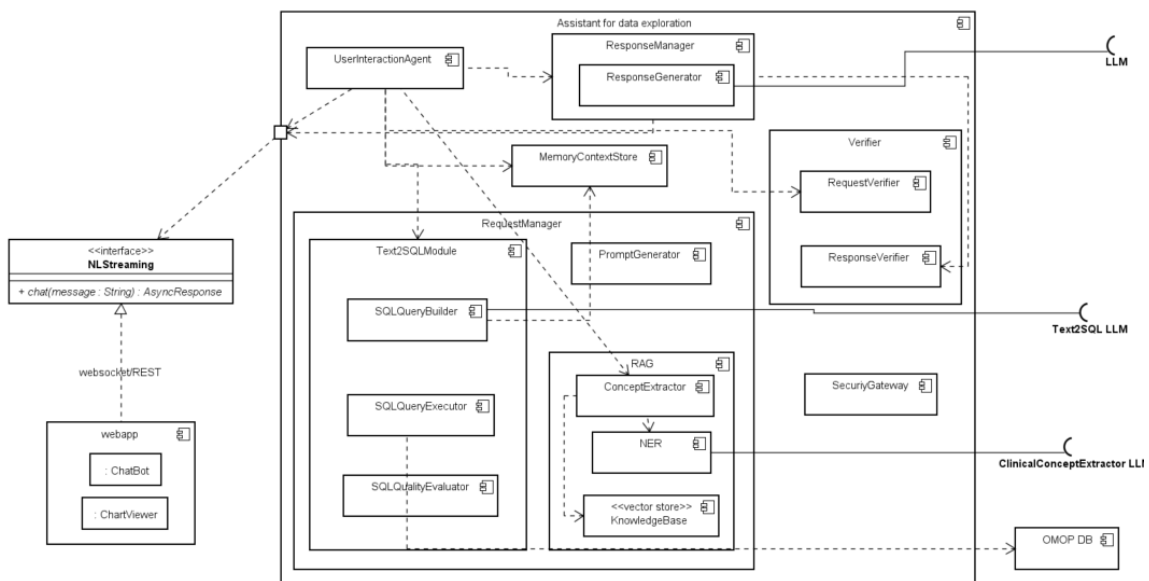


Figure 1: Architecture of the conversational system

The proposed architecture follows the structural and conceptual logic described in the **System Layer** of the *Trusted AI Reference Architecture* presented in (Lu, 2023). As such, several direct correspondences can be identified between the components of that model and those developed in the current system. The **UserInteractionAgent** represents the *Agent* described in the paper, acting as the central orchestrator of user interactions. The **KnowledgeBase** corresponds to the *Vector Database*, while the clinical dictionaries such as SNOMED and LOINC represent the *Internal Data Sources*. The **PromptGenerator** serves the same purpose as the *Prompt Patterns* component, dynamically building model instructions, whereas the APIs through which the system interacts with different language models (such as Llama CPP) embody the *Governance via AI APIs* mechanism. Finally, the models used for specific tasks map directly to the *Fine-tuned foundation models* of the paper. The **Verifier**, in turn, plays the same role as the *Verifier-in-the-loop*, ensuring responsible oversight over all AI-driven processes.

The system acts as an intelligent assistant for exploring clinical and environmental data stored in relational databases. At its core lies the **UserInteractionAgent**, which manages the conversational flow with the user. It interprets natural language queries, maintains context throughout the dialogue,

and decides when to request clarifications to refine ambiguous or incomplete questions. By mediating between the human user and the technical subsystems, the agent transforms general requests into precise operations and ensures that the conversation remains coherent and focused on the data available.

Supporting the agent is the **RequestManager**, a coordination component that handles the execution of each query. Within it, the **PromptGenerator** constructs adaptive prompts based on a library of reusable prompt patterns. Depending on the nature of the user's request—be it generating an SQL query, extracting clinical concepts, or producing a natural language response—the **PromptGenerator** assembles the appropriate structure, embedding relevant context and knowledge retrieved from the system's internal resources. This design allows the system to interact with different language models efficiently and with minimal human intervention, ensuring consistency and improving the interpretability of model outputs.

The **Text2SQLModule** is responsible for translating user intents into executable SQL queries. It is composed of three internal units: the **SQLQueryBuilder**, which relies on the Arctic Text2SQL model to generate the query syntax; the **SQLQualityEvaluator**, which verifies the syntactic correctness and evaluates the expected performance of the generated query; and the **SQLQueryExecutor**, which finally executes the validated SQL on the underlying database. This modular organization allows the assistant to autonomously transform high-level natural language instructions into efficient database operations, while maintaining control over accuracy, execution cost, and reliability.

Complementing the Text2SQL component, the system integrates a **RAG** module that provides semantic enrichment and concept normalization. The **ConceptExtractor** within this module interprets domain-specific terms appearing in user requests and resolves them into standardized codes. It relies on a dedicated **NER module**, implemented through a fine-tuned LLM, capable of identifying relevant entities in natural language text with contextual understanding, handling synonyms, abbreviations, and linguistic ambiguity. Once the entities are extracted, they are passed to a **VectorResolver** connected to a **FAISS-based vector store**, which retrieves the most semantically similar standardized codes from the **KnowledgeBase**. This process allows the system to understand phrases such as “bad cholesterol” and correctly associate them with the clinical concepts corresponding to LDL cholesterol, ensuring accurate semantic grounding in the clinical domain.

The **KnowledgeBase** itself functions as the vector database of the architecture, containing embeddings of standardized clinical dictionaries, multilingual synonyms. Its role is to support both retrieval and contextualization: it provides domain knowledge to the Text2SQL module and serves as a semantic map for the ConceptExtractor. The adoption of FAISS (FAISS - Facebook AI Similarity Search) enables high-speed similarity search and scalable management of large embeddings collections, which is crucial for real-time interactive exploration.

Ethical and functional oversight can be handled by the **Verifier**, which embodies the *Verifier-in-the-loop* concept proposed in the Trusted AI model. It is composed of two elements: the **RequestVerifier**, which ensures that incoming user requests are legitimate, privacy-compliant, and ethically acceptable, and the **ResponseVerifier**, which evaluates the generated output to confirm that it is contextually accurate and responsibly phrased. This continuous verification process maintains the trustworthiness of the system and enforces the principles of responsible AI. Currently, the supported verification is limited to elementary controls, but it could be expanded by introducing LLM-as-a-judge in particular for the generated responses.

Once the data is retrieved and validated, the **ResponseManager** takes charge of composing and formatting the response. It works in tandem with the **ResponseGenerator**, which produces human-readable summaries or structured outputs based on the query results. Depending on the type of result, the system can generate natural language explanations for scalar values, structured JSON objects for visual representation in charts or dashboards, or clarification prompts when the query is ambiguous. This approach ensures that the output is always in a form suitable for both human understanding and graphical rendering in the client interface.

To preserve conversational context, the **MemoryContextStore** records user interactions, including past queries, clarifications, and responses. This allows the assistant to refer to previous exchanges, maintain continuity across sessions, and build a persistent reasoning context.

The **SecurityGateway**, meanwhile, enforces data protection and compliance rules, authorizing queries, masking sensitive information, and maintaining access logs in accordance with clinical data regulations.

All these modules interact with specialized LLMs through standardized APIs, such as **Llama CPP**, which serve as the governance layer for AI model invocation. The system employs multiple fine-tuned models, each optimized for a distinct task. This division of responsibilities ensures scalability, maintainability, and clarity in the system's operation, as each model can evolve independently without requiring changes in the overall architecture.

The following activity diagram better illustrates the main internal interactions of the modules that are triggered whenever an end-user submits a request through the conversational interface.

We can see in **Figure 2. Activity diagram of architecture** how user's inputs are processed by the Natural Language Understanding (NLU) Layer, which through the RAG layer, queries an external knowledge base to obtain the corresponding clinical codes. Once these codes are retrieved (or immediately, if no diseases are specified), the RequestManager becomes responsible for further understanding user's intents and for generating the SQL query through the Arctic-Text2SQL-R1-7B model.

To accomplish this, it adds the request to the conversational memory, fetches previously summarized dialogues to increase contextual awareness, and retrieves the database schema to provide the LLM only with the necessary tables.

Next, the **PromptGenerator** constructs the prompt to be submitted to the Text-to-SQL model. It uses the best-prompt defined in the model's documentation (Yao, 2025) and enriches it with the summarized dialogues, the clinical codes extracted via NER, and a structured set of guidelines and instructions to ensure that the model produces the expected response format.

The resulting query is then validated by the **SQLQueryValidator**, which verifies both the syntactic correctness with respect to the SQL dialect specified in the prompt and the suitability of its performance for the intended use case. If any validation criterion is not satisfied, control returns to the **PromptGenerator**, which refines the context or introduces additional constraints as needed.

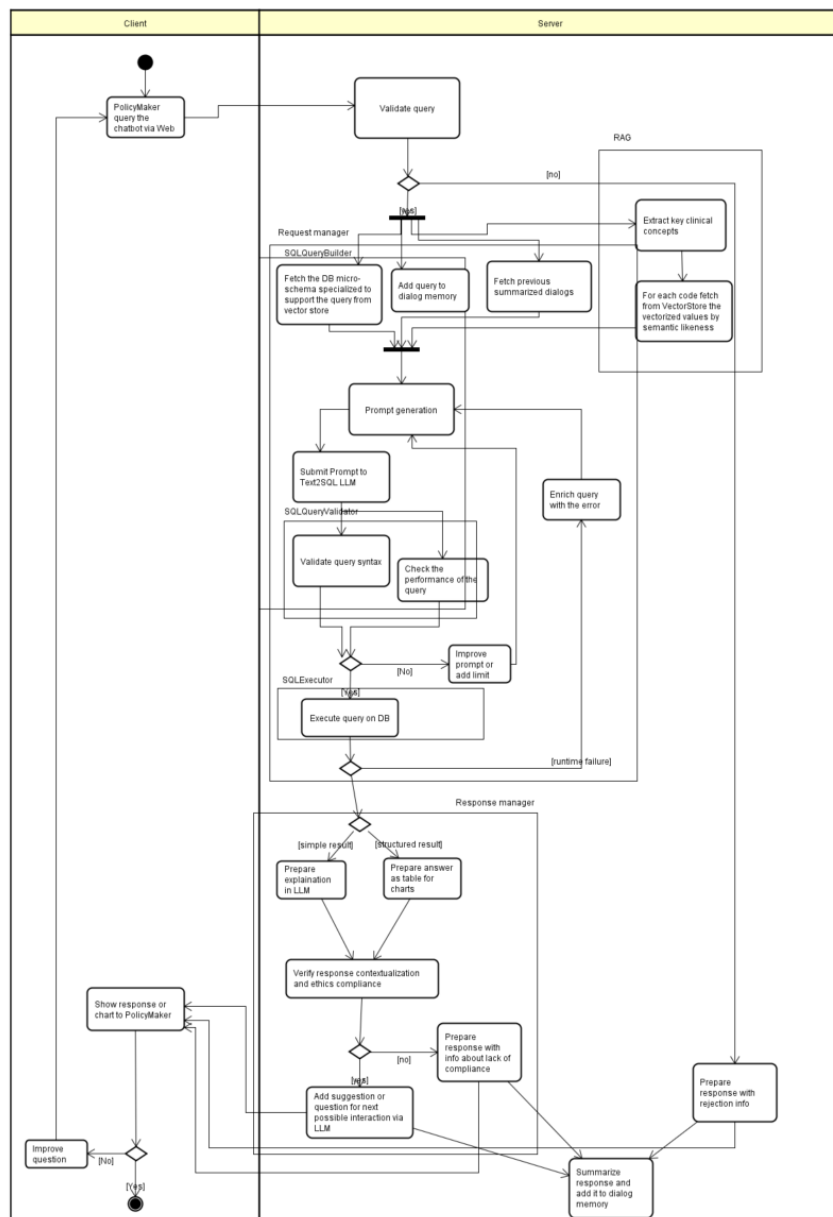


Figure 2. Activity diagram of architecture

If the query passes all validation checks, the **SQLExecutor** executes it against the database, monitoring the workflow and capturing any runtime errors. Should a failure occur, the error details are incorporated back into the prompt, and the process returns to the **PromptGenerator** for refinement.

Now the obtained results are handled by the **ResponseManager**, which formats the data in a structured format, so that they could be used to render charts, or in case of scalar results - by leveraging another LLM - it generates a human-friendly text. The prepared response is validated by checking both its level of contextualization and its compliance with ethical constraints. If these criteria are not met, the system generates a custom response indicating the lack of compliance; otherwise, the response is enriched with a set of possible suggestions on how the user may continue the interaction.

The last step consists of displaying the response to the user, who may then choose to refine the request with a new query or simply accept the provided output.

Within the scope of the present use case, a research activity has been conducted and is still ongoing to identify the most suitable Text-to-SQL model for the proposed application, which is intended to serve as the core engine of the system. An initial analysis was conducted on the open-source LLM deployable on-premises LLM model **Arctic-Text2SQL-R1-7B** (Yao, 2025) specifically trained to translate natural language expressions into SQL queries. The goal was to assess its performance through a simple, repeatable test that can be easily reproduced by anyone. The model was executed using Docker Model Runner (Docker, s.d.), a Docker-based tool that enables local and straightforward deployment of AI models. The Text-to-SQL model was quantized using a 4-bit quantization scheme (Q4) through llama.cpp (Gerganov, n.d.), allowing the deployment of significantly lighter version compared to the original Arctic-Text2SQL-R1-7B model, with only a minor trade-off in accuracy.

The Text-to-SQL model has been used and evaluated by following a 3 steps approach:

1. Definition of reference database schema;
2. Design of a set of natural language requests, progressively increasing in complexity, employed as benchmarks inputs to query the model. An example of question that can be managed by the model is: *"List the name of the products belonging to the beverage category in alphabetical order"*.
3. Application of the "best prompt" for Arctic-Text2SQL-R1-7B, to ensure optimal and comparable execution conditions.

The output generated by the models has been qualitatively assessed, focusing on:

- The syntactic and semantic correctness of the generated SQL queries with respect to the specified SQL dialect.

- The semantic alignment of the generated query with the intent expressed in the NL input.

The model produced a correct and contextually relevant SQL response after applying the NL request through the optimal prompt. Specifically, Arctic-Text2SQL-R1-7B generated the output in 25.8 seconds — 20.7 seconds for prompt processing and 5.1 seconds for token generation — producing a total of 37 tokens. This corresponds to approximately 15 tokens/s during the prompting phase and 7 tokens/s during generation, which is consistent with typical performance in local CPU-based on-premises environment.

During further testing, it was observed that the model's response time is directly affected by the size and token count of the input provided. When using more complex and structured database schemas, execution times increased significantly, highlighting the model's sensitivity to input complexity. In these cases, GPU-based setups proved essential to achieve acceptable inference performance and maintain response times within a valid operational range.

The performance results obtained from the model, combined with industry-standard benchmarks, led us to select Arctic-Text2SQL-R1-7B as our reference model for the Text-to-SQL module in our architecture, responsible for transforming user NL requests into executable SQL queries to query the database.

The current prototype still presents several limitations that need to be addressed through further research and experimentation.

First, the **data extraction process** remains a major challenge. Although the Text-to-SQL pipeline can produce syntactically correct queries, the model does not always fully grasp how data are structured or related within the database. This becomes particularly evident when queries involve complex joins, implicit relationships, or multiple heterogeneous data sources. In practice, the lack of a deeper understanding of the database schema sometimes prevents the model from retrieving results that are both accurate and clinically meaningful. In this context, there is also a scalability issue related to the size of the database schema as the bigger it is the more context tokens are used so impacting the overall performances. It could be improved by applying schema reduction techniques.

Second, the current design does not yet support **multi-turn dialogue management**. The system can interpret isolated user requests but lacks memory mechanisms for maintaining conversational context across turns. As a result, it cannot yet handle follow-up questions, clarifications, or comparisons with previous queries — all of which are key to creating a truly interactive and adaptive conversational experience. Moreover, the performance of the systems can also be further improved by enriching the prompts internally used as to increase the efficiency of the various selected LLM models. Such models, in the future will undergo further evaluations and, possibly, be replaced by upcoming open-source alternatives.

In addition, the **normalization of clinical concepts** and the **disambiguation of vocabulary** are still only partially managed, meaning that some medical terms or user expressions are not yet consistently mapped to the correct standardized concepts. The mapping of natural language expressions to standardized clinical concepts still relies on heuristic matching rather than fully automated semantic reasoning.

Lu, Q. Z. (2023). Towards responsible AI in the era of ChatGPT: A reference architecture for designing foundation model-based AI systems. *arXiv preprint*. doi:arXiv:2304.11090

Yao, Z. S. (2025). Arctic-Text2SQL-R1: Simple Rewards, Strong Reasoning in Text-to-SQL. *arXiv preprint*. doi:10.48550/arXiv.2505.20315

2.2. One Health Data Platform

The evolution of health, environmental and territorial information systems has produced a large amount of data over time, but this data is often fragmented into heterogeneous silos (Borowicc, 2024). This fragmentation is now one of the main barriers to the creation of analytical and predictive models consistent with the One Health approach, which requires the ability to correlate information from different domains to describe the health status of the population and territory in a unified manner, integrating analysis and prediction in the health and environmental fields.

The One Health Data Platform has been designed to overcome these limitations by providing a scalable and interoperable infrastructure capable of: (a) collecting and harmonising heterogeneous data; (b) ensure semantic, temporal and geographical consistency; and (c) generate integrated datasets for training artificial intelligence models and simulating complex scenarios for predictive and preventive purposes.

As a case study, **senile dementia** was selected. This disease was chosen because of its clinical complexity and the scientific evidence documenting the relationship between air pollution and cognitive decline (Qiu, 2023; Byeon, 2022; Peters, 2019; Wilker, 2023; Oliveira, 2024). In addition, several recent studies have demonstrated the effectiveness of applying artificial intelligence techniques in the predictive and diagnostic fields for this type of neurodegenerative disorder (You, 2025; Battineni, 2022; Newby, 2023; Camacho, 2025). Although these studies have been conducted mainly in international contexts, this project focuses on the Veneto Region, an area which, according to monitoring data from ISPRA (Istituto Superiore per la Protezione e la Ricerca Ambientale, Higher Institute for Environmental Protection and Research), has some of the highest levels of air pollution in Italy.

The Veneto Region also has a Diagnostic Therapeutic Care Pathway (PDTA) dedicated to dementia², which served as a methodological reference for modelling the simulated clinical processes, together with regional maps of senile dementia³, which provide up-to-date information on the geographical distribution of patients. To collect data on lifestyles and socio-economic factors, datasets from the PASSI d'Argento surveillance system⁴ were used, which are a resource for the behavioural and social characterisation of the elderly population.

2.2.1. Technical Insight

The **One Health Data Platform** architecture is organized into a **pipeline** that, starting from the generation and acquisition of data sources (through the “One Health Scenario Engine”), proceeds through the ingestion and validation modules, up to the storage and correlation of data within the

² Veneto Region – PDTA Dementia: <https://demenze.regione.veneto.it/pdta/il-documento/>

³ Veneto Region – A map for dementia: <https://demenze.regione.veneto.it/>

⁴ PASSI d'Argento: <https://www.regione.veneto.it/web/sanita/passi-dargento>

central repository. The architectural design followed the principle of separation of functional levels, which provides for a clear distinction between the level of data collection and generation, the level of processing and standardization, and the level of persistence and correlation. This approach allows the elements dependent on the simulated context to be isolated and ensures the replaceability of synthetic data with real data, where available, while preserving the reusability of the platform.

The “*One Health Scenario Engine*” is the entry point of the architecture, aggregating synthetic, environmental and socio-demographic data to generate consistent and customizable virtual populations.

The data produced in this way is processed in the “*Data Ingestion & Processing Engine*”, which performs harmonization, standardization, validation and geolocation.

The harmonised data then flows into the “*One Health Data Repository*”, a scalable repository based on the **HAPI FHIR** open-source project⁵, which ensures full semantic interoperability between the different information domains and forms the information core of the platform.

The logical diagram in **Figure 3** illustrates the main components of the architectural layer, the details of which are described in the following paragraphs.

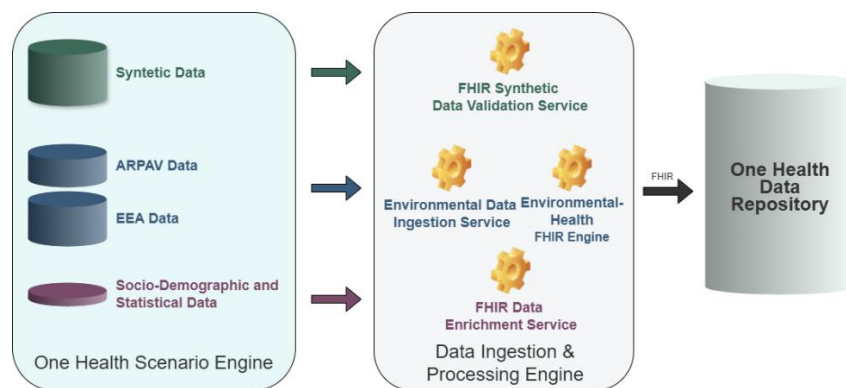


Figure 3: Logical diagram of the proposed architecture

Generation of synthetic Clinical Data

In the absence of real clinical data, due to restrictions related to privacy and ownership of health data, an approach based on the generation of synthetic clinical data using the open-source **Synthea™** platform⁶ was adopted. Synthea™ is a virtual patient generator that produces realistic, comprehensive clinical data consistent with international healthcare standards. The generation engine is based on probabilistic models and predetermined clinical pathways, defined in the form

⁵ HAPI FHIR: <https://hapifhir.io/>

⁶ Synthea™ - Open Source Synthetic Patient Generator: <https://synthea.mitre.org/>

of disease modules that describe the temporal progression of pathological conditions, diagnostic procedures, therapeutic interventions, etc.

Synthea™ allows the default generation of a synthetic population residing in the United States. Therefore, to make the model applicable to the Italian context, the **Synthea International** version⁷ was also used, which allowed the use and localization of certain datasets to support generation, such as the list of locations or healthcare providers.

Attention was then focused on the specific simulation model for dementia, calibrated to the territorial context of the **Veneto Region**, through a series of configuration and customization interventions. These configurations have made it possible to generate over 4.8 million virtual patients, of whom approximately 1.4% suffer from dementia. Synthea uses fixed deterministic and probabilistic rules, which in this context have been set up to comply with regional epidemiological statistics. The data obtained in this way maintain semantic and statistical consistency with real data, while not containing any information that can be traced back to individuals. In fact, according to the EU Artificial Intelligence Act, synthetic data can be considered “non-personal data”. The data was generated in accordance with the **HL7 FHIR** standard⁸, which uses international coding standards such as **LOINC**⁹ and **SNOMED CT**¹⁰, ensuring full semantic interoperability of the data and thus providing a solid basis for the development and validation of AI models.

Collection of environmental and social data

Real environmental data were retrieved and integrated from the **Veneto Regional Agency for Environmental Prevention (ARPAV)** and the **European Environmental Agency (EEA)** relating to air quality and, in particular, the atmospheric pollutants that most influence dementia (NO₂, PM_{2.5}, PM₁₀), as well as socio-demographic and behavioral data from sources such as the **National Institute of Statistics (ISTAT)** and the national surveillance system dedicated to the elderly population called **PASSI d'Argento**.

The environmental data were acquired through two complementary methods:

- **through public APIs** provided by the European Environmental Agency (EEA), which allow access to harmonised datasets in parquet format relating to air quality and atmospheric pollutants. Two datasets were used: the E1a dataset, containing validated data from 2013 to 2025, and the “E2a” dataset, which contains non-validated data that is constantly transmitted

⁷ synthea-international: <https://github.com/synthetichealth/synthea-international>

⁸ HL7 FHIR Standard: <https://hl7.org/fhir/R4/>

⁹ Loinc: <https://loinc.org/>

¹⁰ Snomed: <https://www.snomed.org/>

- **through a direct request logged with ARPAV**, which provided data relating to monitoring stations in the provinces of Belluno and Rovigo, as data for these provinces are not included in the previous datasets.

In addition to environmental data, socio-demographic and territorial context data were collected and integrated to provide a realistic representation of the regional population. This data includes the distribution of the resident population by age, gender and local health authority (ULSS), as well as estimates of the prevalence of dementia in the Veneto Region, derived from official sources (ISTAT and the Veneto Region). Well-being, lifestyle and social vulnerability indicators from the PASSI d'Argento system were also considered, such as smoking habits, alcohol consumption, levels of physical activity and sedentary lifestyle, nutritional status assessed through body mass index (BMI) and eating habits, etc.

Data harmonisation and standardisation

With regard to environmental data, datasets, originally available in different formats and structures, have been harmonised and standardised through the use of a shared data dictionary, which defines the format, unit of measurement, domain of permissible values and semantic relationships between variables for each attribute. Two modules were implemented to carry out this operation: the first module, called "*Environmental Data Ingestion Service*", deals with ingestion of data and conversion towards the project environmental csv; the second module, called "*Environmental Health FHIR Engine*", receives the pre-processed data and, using standard resources (Location, Device and Observation), represents it according to the FHIR standard and defines its reference coding domains.

Correlation algorithms

Several correlation algorithms have been used to enrich the data. In relation with the patient exposure to pollutants, temporal aggregation of the environmental data has been performed: environmental data values, initially available as daily averages, were aggregated on a quarterly basis, calculating the quarterly arithmetic mean of the concentrations recorded by each monitoring station for each pollutant. Spatial correlation between between patients and environmental monitoring stations was carried out as well, in the "*Environmental Health FHIR Engine*" module using a geospatial approach based on the *k-nearest neighbours (kNN) algorithm*, with parameter $k=3$.

For each individual, the geographical coordinates of residence (latitude and longitude) were considered and, by constructing a spatial index (KD-Tree) containing the positions of all the monitoring stations, the three closest monitoring stations were identified. A representative value of the level of exposure to pollutants was then calculated for these three monitoring stations using a

weighted average inversely proportional to distance (Inverse Distance Weighting, IDW), according to the formula:

$$V_{i,p,t} = \frac{\sum_{j=1}^k w_j \cdot mean_{c_j,p,t}}{\sum_{j=1}^k w_j} \text{ con } w_j = \frac{1}{(d_{i,j} + \varepsilon)^\alpha}$$

where:

- $V_{i,p,t}$: estimated exposure level for patient i , for pollutant p , in period t ;
- $mean_{c_j,p,t}$: quarterly average concentration of the pollutant detected by monitoring station c_j ;
- $d_{i,j}$: geodetic distance between patient i and monitoring station j ;
- $\alpha=1$: spatial attenuation parameter;
- ε : correction value (1 m) to avoid division by zero.

This method allows for a robust estimation of the average environmental exposure for each individual, favoring the closest monitoring stations and reducing the effect of outliers due to individual measurements.

With respect to the integration of socio-demographic data, it should be noted that by their very nature this data cannot be traced back to individual persons, but are distributed according to population classes, typically defined by factors such as age and gender. Data from official sources were organised into population clusters, providing statistical information on education levels, percentage of elderly population, sedentary lifestyle rates and risky lifestyles (smoking, alcohol consumption, physical inactivity).

These data were represented in the FHIR repository as population classes using standard resources (Group and Observation), thus ensuring interoperability with other project data. This task was carried out as part of the “*FHIR Data Enrichment Service*” module. The correlation between these aggregated data and the individual patient will be estimated directly by artificial intelligence algorithms.

Finally, the One Health Data Platform has been enriched with the Repository: the variables obtained in this way were normalised and correlated with each other, generating structured data by individual, time period and information domain (clinical, social, environmental) and made available within the CDR. From these data, it will be possible to create matrices for subsequent statistical analysis and for training predictive models, allowing multidimensional correlations and risk indices to be estimated. These matrices will be able to integrate variables from the clinical, environmental and socio-demographic domains, describing for each synthetic individual the main cognitive

indicators (MMSE, GPCog), the average quarterly concentrations of atmospheric pollutants and the clustered socio-territorial characteristics (population density, level of education, elderly population rate).

Battineni G. (2022). Artificial Intelligence Models in the Diagnosis of Adult-Onset Dementia Disorders: A Review. DOI: [10.3390/bioengineering9080370](https://doi.org/10.3390/bioengineering9080370)

Borowic S. L. (2024). Heterogeneous Data Integration: A Literature Scope Review. DOI: [10.5220/0012551000003690](https://doi.org/10.5220/0012551000003690)

Byeon H. (2022). Screening dementia and predicting high dementia risk groups using machine learning. DOI: [10.5498/wjp.v12.i2.204](https://doi.org/10.5498/wjp.v12.i2.204)

Camacho, M., et al. (2025). Low-cost predictive models of dementia risk using machine learning and exposome predictors. <https://doi.org/10.1007/s12553-024-00937-5>

Newby D., et al. (2023). Artificial intelligence for dementia prevention. DOI: [10.1002/alz.13463](https://doi.org/10.1002/alz.13463)

Oliveira M. (2024). Geospatial analysis of environmental atmospheric risk factors in neurodegenerative diseases: a systematic review update. DOI: [10.1186/s13643-024-02637-7](https://doi.org/10.1186/s13643-024-02637-7)

Peters R., et al. (2019). Air Pollution and Dementia: A Systematic Review. DOI: [10.3233/JAD-180631](https://doi.org/10.3233/JAD-180631)

Qiu X. et al. (2023) Association of Long-term Exposure to Air Pollution With Late-Life Depression in Older Adults in the US. DOI: [10.1001/jamanetworkopen.2022.53668](https://doi.org/10.1001/jamanetworkopen.2022.53668)

Wilker E.H. (2023). Ambient air pollution and clinical dementia: systematic review and meta-analysis. DOI: [10.1136/bmj-2022-071620](https://doi.org/10.1136/bmj-2022-071620)

You J. (2022). Development of a novel dementia risk prediction model in the general population: A large, longitudinal, population-based machine-learning study. DOI: [10.1016/j.jeclinm.2022.101665](https://doi.org/10.1016/j.jeclinm.2022.101665)

2.3. Personalised Environmental Clinical Risk Score

The DARE partner Exprivia currently has a Decision Support System (DSS) designed to support healthcare professionals in their decision-making processes through the representation and execution of clinical observation models (pathways). The system is capable of analysing a patient's clinical condition in real time and comparing it with the reference model, verifying the consistency between the observed data and the workflow planned for the management of the condition. Currently, the DSS decision-making logic is based on the Camunda process engine, which uses rules defined by DMN (Decision Model and Notation) tables for the automatic assessment of clinical conditions and the generation of operational suggestions.

A project's objective is to enhance the decision-making component of the DSS by integrating models based on Artificial Intelligence algorithms in order to provide predictive and personalised support for clinical. In the pilot's reference scenario, based on an analysis of the scientific literature and an examination of the available multidimensional features, a **Personalised environmental clinical risk score** has been the focus of research activities. The goal is to define a composite score that integrates individual exposure levels to air pollutants and socio-economic variables.

This Computational Model is still in its early stage of development. In particular, the dataset for training and evaluation has been selected based on the other Computational Model developed by the partner, i.e. the One Health Data Platform (see Section 2.2).

2.3.1. Technical Insight

In the case of the **clinical-environmental risk score**, while recognising a potential temporal dependence between environmental exposure and health status, as exposure to environmental factors can vary over time and influence health status, it has been chosen at this stage to treat the problem as **time-independent**. This choice is also motivated by the fact that, at present, the location of patients is only available as their place of residence. It is not possible to accurately track individual movements and the resulting temporal variation in exposure to different pollutants. Furthermore, since there are no reference labels directly describing the relationship between clinical status and level of environmental exposure, it is not possible to set up supervised training. An unsupervised exploratory approach was therefore adopted, aimed at identifying latent patterns and homogeneous risk groups based on the integration of clinical, environmental and socio-economic features.

Currently, for the development of the risk score, the following approaches are under investigation:

K-Means. The k-means algorithm is one of the best known and most widely used clustering methods. It works by dividing the data into a predetermined number of groups, called clusters, seeking to minimize the distance within clusters and maximise the distance between different clusters. It is a simple and very fast algorithm, particularly suitable for large datasets. However, the

need to decide the number of clusters in advance can be a limitation, as can its sensitivity to outliers and its tendency to find only fairly regular and spherical groups. Therefore, if the data contains clusters with complex shapes or noise, this method may not be the best choice.

Hierarchical Clustering. Hierarchical clustering, on the other hand, takes a more structured approach. It constructs a hierarchy of groups, represented by a tree structure called a dendrogram, which shows how the different clusters are progressively grouped together. This allows you to see the similarity relationships between data at different levels of granularity, providing a more comprehensive view of the underlying structure. On the other hand, hierarchical clustering is computationally more expensive than k-means and difficult to scale on very large datasets. Furthermore, once a grouping has been made, it is not possible to reassign points to different clusters in subsequent stages, which can reduce flexibility in the analysis.

DBSCAN. A very different approach is that of DBSCAN, which relies on point density to identify clusters. This method identifies areas with a high concentration of data and separates these areas from those with low density, which are considered noise or outliers. DBSCAN is therefore very effective at finding irregularly shaped groups and isolating anomalous data. However, the definition of the parameters that regulate density requires particular attention, because the quality of the clustering depends on it. Furthermore, if the density of the data varies greatly between clusters, the algorithm may encounter identification difficulties.

Autoencoder-based approaches. Finally, autoencoders belong to the family of deep learning algorithms and are very powerful as they allow for dimensionality reduction while maintaining the informational richness of the data. They consist of a neural network that learns to compress the original data into a more compact latent representation and then reconstruct it. This ability makes autoencoders extremely useful for discovering hidden and non-linear patterns in data, also facilitating the identification of complex correlations between environmental and social factors that influence health. However, they require a lot of data for training, are less immediately interpretable than traditional clustering algorithms, and require higher computational resources.

2.4. 5-year Cognitive Decline Score

The DARE partner Exprivia is extending its Decision Support System (DSS) with a third Computational Model (see Sections 2.2 and 2.3 for the other two Computational Models). This third model focus on a **5-year cognitive decline score**, which estimates the evolution of cognitive abilities over time, based on clinical variables and cognitive tests such as **MMSE** (Mini-Mental State Examination) and **GPCOG** (General Practitioner Assessment of Cognition).

2.4.1. Technical Insight

From a methodological point of view, it has been chosen to exploit the One Health Data Platform (Section 2.2). This choice, in addition to simplifying data management and maintaining semantic consistency between variables, allows **to preserve the possibility of evolving towards multi-task learning** approaches, which can be explored in subsequent phases of the project.

The **cognitive decline score** was considered **time-dependent**, as the progression of cognitive deterioration is closely linked to the patient's clinical evolution over time and to changes in associated factors (clinical, social and environmental). The availability of measurements on the time axis, such as MMSE and GPCog scores, allows targets to be constructed by analysing their temporal variations, thus providing a quantitative reference for training AI models.

The approach selected is supervised, with the aim of predicting the trajectory of cognitive decline based on the available multidimensional variables. The following approaches are being considered and undergoing preliminary tests:

LSTM (Long Short-Term Memory). LSTM is a type of recurrent neural network designed to learn and recognise long-term dependencies in sequential data, thanks to an internal structure with memory cells and several “gates” that control the information to be retained or forgotten. It is therefore very effective for temporal or sequential data such as physiological signals or cognitive time series. It effectively handles long-term dependencies and is excellent for sequential temporal data. On the other hand, it requires a lot of computational power and data for training. It can be complex to interpret how it works and there is a risk of overfitting on small datasets.

XGBoost (eXtreme Gradient Boosting). Gradient Boosting is an ensemble algorithm that builds predictive models by combining many weak decision models in sequence, progressively improving residual errors. It is widely used for classification and regression problems. It has high predictive accuracy and is robust to various types of data, as well as being good at interpreting important features. On the other hand, it is sensitive to noise in the data and can be slow to train. It is also complex to tune its parameters and there is a risk of overfitting if not adjusted properly.

Cox regression. Cox regression is a semi-parametric model that estimates the effect of variables on the risk of an event. It is an interpretable model, well established in the medical field, and allows for correction for confounding factors. On the other hand, it assumes proportionality of risks over time and limits the modelling of complex non-linear effects.

Random Survival Forest. Random Survival Forest is a non-parametric ensemble method based on the use of decision trees. This approach is particularly effective in handling data with non-linear relationships and complex interactions between variables, without the need to assume the proportionality of risks typical of other models such as Cox. Furthermore, it is characterised by being a robust and flexible method. However, compared to the Cox model, it is less interpretable and has greater computational complexity.

2.5. Automatic CAP Recognition

This Computational Model has been defined by the research group (Ciampolini-Matrella) of DIA (Dipartimento di Ingegneria e Architettura) of the University of Parma, to support the Pilot Project directed by prof. Marcello Maggio (Somnus-DARE) in relation with the University Interdepartmental Center for Sleep Disorders (so called CMS - Centro di Medicina del Sonno “Mario Terzani”) of the University Hospital of Parma.

The activity regards the **development of an algorithm for the automatic recognition of Cyclic Alternating Pattern (CAP) using electro-encephalon-graph (EEG) tracing.**

CAP detection is crucial for assessing sleep health because it reveals the microstructure of non-Rapid Eye Movement (NREM) sleep, reflecting the brain's balance between stability and activation. Unlike traditional sleep staging, which only classifies broad phases, CAP analysis identifies subtle fluctuations that indicate sleep quality and stability. A normal CAP pattern supports restorative sleep, while an elevated or disrupted CAP rate signals instability linked to conditions like insomnia, sleep apnea, epilepsy, and neurodegenerative diseases.

Thus, CAP serves as a sensitive biomarker for diagnosing sleep disorders, evaluating treatment efficacy, and understanding the brain's restorative processes during sleep.

2.5.1. Technical Insight

The EEG signal is acquired by recording the brain's electrical activity from the scalp using electrodes that detect voltage fluctuations generated by synchronized neuronal activity in the cerebral cortex. This is the starting point for detecting CAP segments. Indeed, CAP alternates between:

- Phase A – brief activation (micro-arousals, EEG shifts, muscle tone changes);
- Phase B – stable background sleep.

Detecting Phase A of the CAP is challenging because it involves identifying brief, subtle EEG activations within non-REM sleep that vary greatly across individuals and conditions. These activations are highly heterogeneous in amplitude, duration, and frequency content, making it difficult to define consistent thresholds for detection. The transition between background (Phase B) and activation (Phase A) is often gradual rather than abrupt, further complicating automatic segmentation. Moreover, Phase A can include different subtypes (A1, A2, A3), each with distinct EEG features and physiological correlates, requiring fine-grained feature extraction and classification. Noise, artifacts such as muscle activity, eye movements, and inter-individual variability in sleep EEG patterns add additional complexity, often leading to low agreement between human scorers and reduced accuracy in automated detection systems.

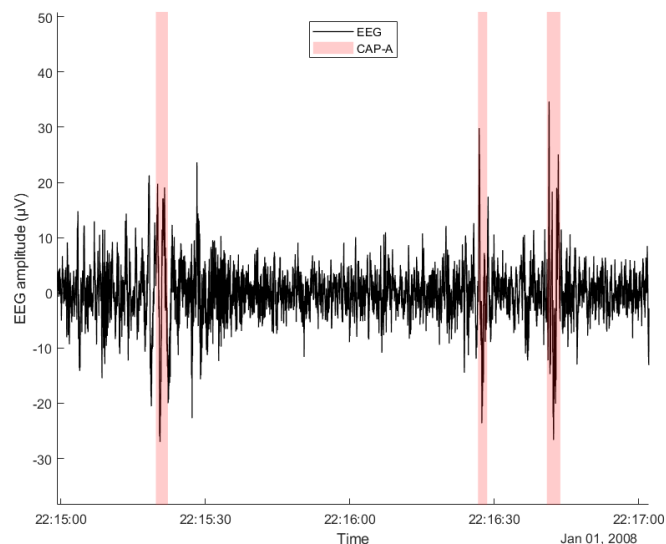
To face these technical challenges, robust software was developed during this period to detect Phase A of CAP based on the Terzano rule (Smerieri, 2007). The Terzano rules define the standard criteria for CAP and Phase A detection in non-REM sleep to ensure consistent CAP scoring as follows:

- CAP consists of a recurring sequence of Phase A (activation) followed by Phase B (background) during NREM sleep.
- Phase A is identified as a transient EEG activation lasting 2–60 seconds, characterized by clear changes in frequency or amplitude compared to the preceding background activity.
- To qualify as CAP, two or more consecutive A–B sequences must occur with inter-phase intervals (B phases) for shorter than 60 seconds.
- If the interval between A phases exceeds 60 seconds, the pattern is no longer considered CAP but a separate activation event.

These rules form the foundation for both manual and automated CAP scoring by providing precise timing and spectral criteria for detecting Phase A events.

The developed software was tested on the CAP Sleep Database¹¹, which was contributed to PhysioNet by the team of sleep experts of the Sleep Disorders Center of the Ospedale Maggiore of Parma, Italy. The software successfully detects the CAP-A events as shown in **Figure 4**.

Since the large data consumes long time of scoring, the Deep Learning model was developed to track this event with the aim to filter out all the parts, containing the CAP. The built DL model was able to detect whether the CAP exist in each 3 minutes with accuracy more than 85 %.



¹¹ CAP Sleep Database v1.0.0, *physionet.org*. <https://physionet.org/content/capslpdb/1.0.0/>

Figure 4. CAP-A detection by developed software

Moreover, heart rate per minute during the subtype of phase A was analyzed to compare with the non-CAP interval. From that, we can examine the difference of heart rate during CAP event with the normal state.

As shown in the Figure Figure 5, number of heartbeat per minute during CAP-A1, A2, A3 conduct large variations, while No CAP periods have stable number of heart rate. The phase A can occur very fast, with minimum of 2 s, but the heartbeat speed changes significantly during these short intervals.

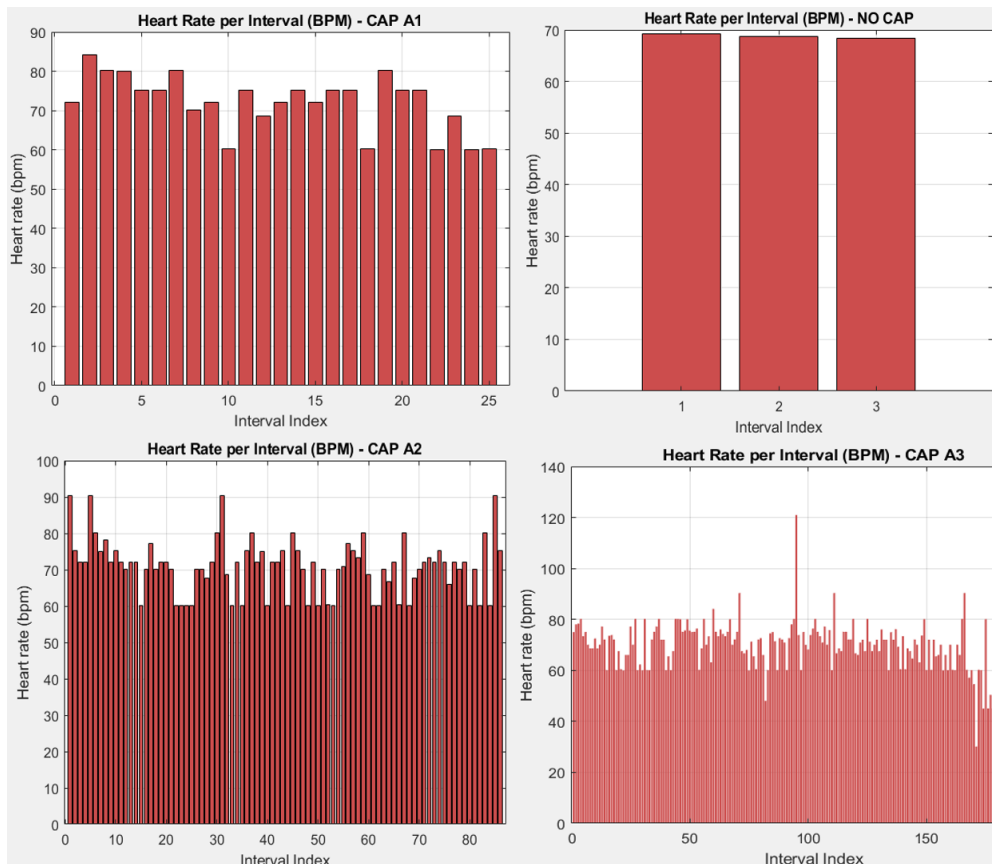


Figure 5 Heart rate statistic during CAP-A and NonCAP

Smerieri A., et al. (2007). Cyclic alternating pattern sequences and non-cyclic alternating pattern periods in human sleep. DOI: <https://doi.org/10.1016/j.clinph.2007.07.001>.

2.6. Physical Activity Index (PAI)

In the framework of DARE project, the Internet of Things (IoT) Lab research group (Ferrari *et al.*) at the Department of Engineering and Architecture of the University of Parma has contributed proposing computational algorithms for analysing data collected with wearable systems from both healthy subjects and patients affected by Parkinson's Disease (PD). Initially, the support targeted the Pilot Project directed by Prof. Bellafiore (UNIPA), due to their adoption of Garmin smartwatches, the same used by the IoTLab to collect daily information from the participants in the Pilot. Then, due to a technology change during the Bellafiore's Pilot project execution, the IoTLab moved its support to an “*in-house*” Pilot, collecting data from volunteers recruited at UNIPR.

In particular, in the context of continuous and long-term monitoring of the health status of healthy adults, it has been defined an innovative index to quantify the daily physical activity, referred to as **Physical Activity Index (PAI)**.

2.6.1. Technical Insight

The PAI relies on an IoT network, specifically developed to extract various health parameters collected by Garmin smartwatches. The PAI combines motion-related data in a strategical way, according to official regulatory guidelines and state-of-the-art references and is calculated as a weighted sum of five selected daily indicators, i.e., step count, climbed floors, intensity minutes, active minutes, and *Physical Activity Level (PAL)*, as pictorially shown in Figure **Figure 6**.

In details, the PAI is computed by first comparing each parameter value against a reference threshold corresponding to the minimum value of the parameter needed to be considered active, thus, obtaining partial percentages of activity for each considered parameter. All the thresholds are set by taking into account specific recommendations by international organizations, such as World Health Organization (WHO) and Food and Agriculture Organization of the United Nations (FAO) and other literature references. As a preliminary analysis, the PAI has been computed for 6 healthy adults over a 7-month period on a daily, and, then, the average value per week has been derived, as shown in Figure **Figure 7**, demonstrating the feasibility of the monitoring approach.

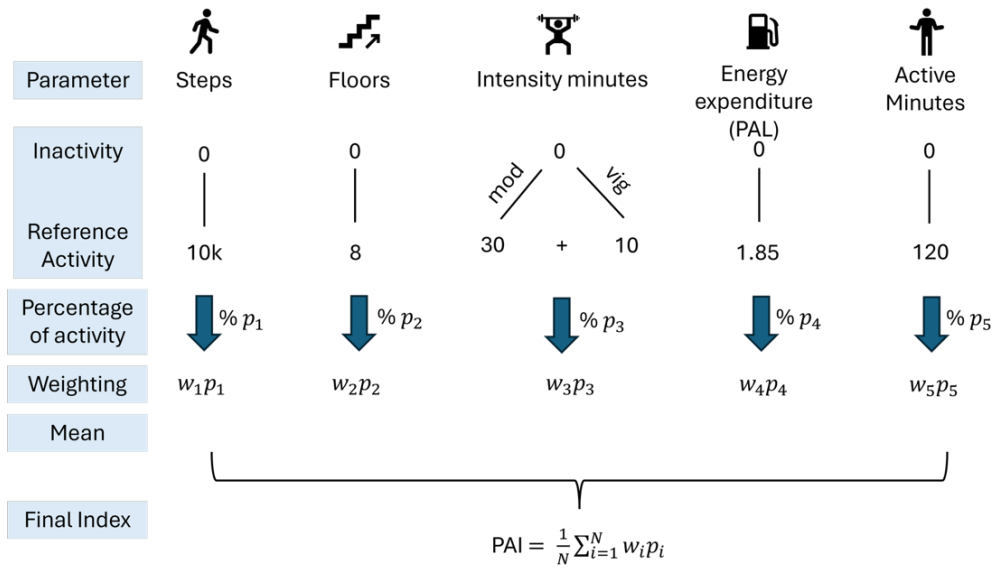


Figure 6. Generic representation of the PAI computation, considering five components of interest, N=5.

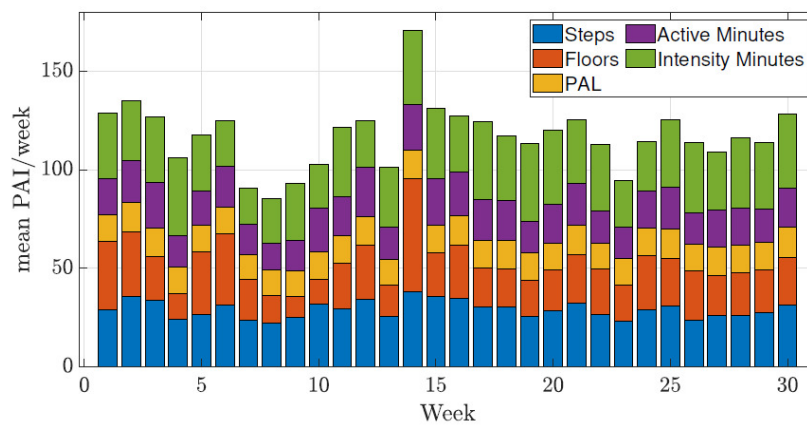


Figure 7. Average value per week of the PAI metric computed for one healthy adult over a 7-month period.

Mattioli, V. et al. (2025). Analysis of Daily Physical Activity by Garmin Smartwatches: A 7-Month Experiment. doi: 10.1109/ISMICT64722.2025.11059422.

2.7. MDS-UPDRS Classifier

The Internet of Things (IoT) Lab research group (Ferrari *et al.*) at the Department of Engineering and Architecture of the University of Parma has contributed to the developing of a second Computational Model. As for the first Computational Model (see Section 2.6), data collected with wearable systems from both healthy subjects and patients affected by Parkinson's Disease (PD) have been analyzed. Similarly to the previous model, data has been collected in a first phase within the Pilot Project directed by Prof. Bellafiore (UNIPA); at a later stage the IoT Lab moved its support to an “in-house” Pilot, collecting data from volunteers recruited at UNIPR.

In particular, the two-stage DL-based **UPDRS Evaluation model** (Goetz, 2007) aims to automatically classify the PD severity from gait signals acquired by a single-node wearable unit. As shown in the block diagram in Figure **Figure 8**, the MDS-UPDRS evaluation is carried out via two main modules, denoted as *Walk Detector* (WaDe) and *Clinical Classifier* (CliC), which have been then implemented as a Python script through the PyTorch library. In the following, an overview on both modules is provided.

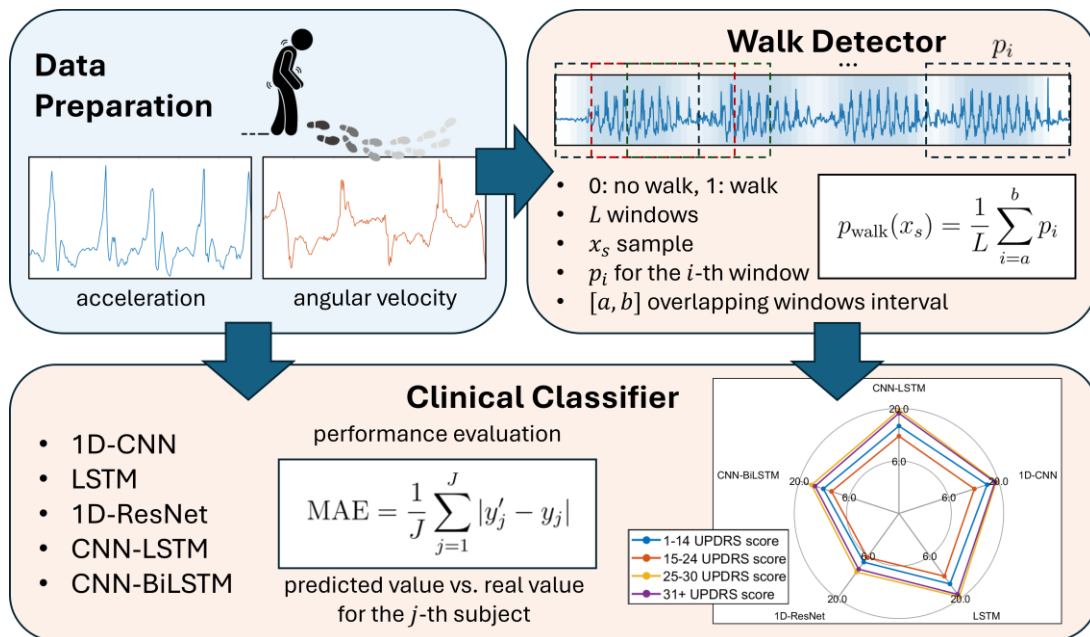


Figure 8. Designed two-stage DL-based UPDRS classification.

2.7.1. Technical Insight

The MDS-UPDRS Classifier is mainly composed of two components: the Walk Detector and the Clinical Classifier.

The Walk Detector (**WaDe**) module consists of a 1D-ResNet (He, 2015) specifically trained on motion signals acquired by a single IMU node positioned at the lumbar region of the involved participants. Then, as an output, WaDe the probability that a given window of the input signal represents a *walking* or *non-walking* phase (e.g., a static phase or transitions), thus enabling the automatic segmentation of the gait signal. In particular, signals are segmented into windows of 2 s with an interlacing factor of 75%, to capture at least one gait cycle per window. More in detail, the 1D-ResNet model is composed of one convolutional layer with 64 output channels, a Batch Normalization (BN) layer, a ReLU activation function, and three residual blocks composed of two convolutional layers with a kernel of size 5, a BN layer, a ReLU activation, and a dropout regularization layer. The residual blocks employ progressively increasing dilatation parameters of 1, 2 and 4, allowing the model to expand its receptive field and capture long-range correlations. The final stage of the 1D-ResNet architecture includes a global adaptive average pooling and a fully connected layer with dropout, projecting the feature vector from 64 to 2 output neurons corresponding to *walking* and *non-walking* classes.

The ClinicalClassifier (**Clcic**) module has been designed on the basis of the five DL architectures – namely: 1D-CNN, LSTM, 1D-ResNet, CNN-LSTM, and CNN-BiLSTM with attention mechanism – and processes the probabilities returned by WaDe aiming at classifying the PD severity (assigning more relevance to the walking phases during the training).

- The 1D-CNN model (Kiranyaz, 2021) is composed of three convolutional layers, each followed by a ReLU activation function and a *max pooling* layer after the second convolution. The extracted features are aggregated by an *adaptive average pooling* layer and passed to a *fully connected* layer projecting the output classes.
- The LSTM model (Hochreiter, 1997) consists of a single recurrent layer and bidirectional connections, followed by a *fully connected* layer.
- The 1D-ResNet is implemented with a convolutional block, followed by a BN layer, a ReLU activation function, and two residual blocks. Each residual block contains two convolutional layers with BN and skip connections. A *global average pooling* layer aggregates temporal features before the final classification head.
- The CNN-LSTM model (Ullah, 2024) combines a convolutional block, a ReLU activation function, and a *max pooling* layer, whose output is fed into a BiLSTM. The concatenated *forward* and *backward* hidden states are projected to the class output through a linear layer.
- Finally, the CNN-BiLSTM with attention mechanism (Shan, 2021) is implemented with two convolutional layers, followed by a GELU activation function, a BN and a *max pooling* layer to extract local features. Then, these features are processed by a BiLSTM, whose outputs are aggregated using a *self-attention pooling mechanism* that adaptively weights the most informative temporal segments.

- Goetz, C.G. et al. (2007), Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinimetric testing plan. doi: 10.1002/mds.21198.
- He, K. et al., (2015). Deep Residual Learning for Image Recognition. doi: 10.48550/ARXIV.1512.03385.
- Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. doi: 10.1162/neco.1997.9.8.1735.
- Kiranyaz, S. et al. (2021). 1D convolutional neural networks and applications: A survey. doi: 10.1016/j.ymssp.2020.107398.
- Shan, L. et al. (2021). CNN-BiLSTM hybrid neural networks with attention mechanism for well log prediction. doi: 10.1016/j.petrol.2021.108838.
- Ullah, K. et al. (2024). Short-Term Load Forecasting: A Comprehensive Review and Simulation Study With CNN-LSTM Hybrids Approach. doi: 10.1109/access.2024.3440631.

2.8. AI Framework for automatically revealing spatial-temporal-spectral EEG signatures.

A deep learning-empowered framework for characterizing EEG oscillations in the frequency, spatial, and temporal domains, based on the features learned by a fully interpretable convolutional neural network, has been designed. The Computational Model has been already successfully published (Borra, 2025a; Borra, 2025b; Borra, 2025c; Borra, 2024).

2.8.1. Technical Insight

The convolutional network learns a bank of bandpass filters to be applied to minimally pre-processed EEG. Then, frequency-specific spatial and temporal filtering allow the learning of the most salient spatial and time samples, separately for each frequency component. Finally, the framework processes the learned interpretable features to reveal meaningful EEG signatures (see **Figure 9**).

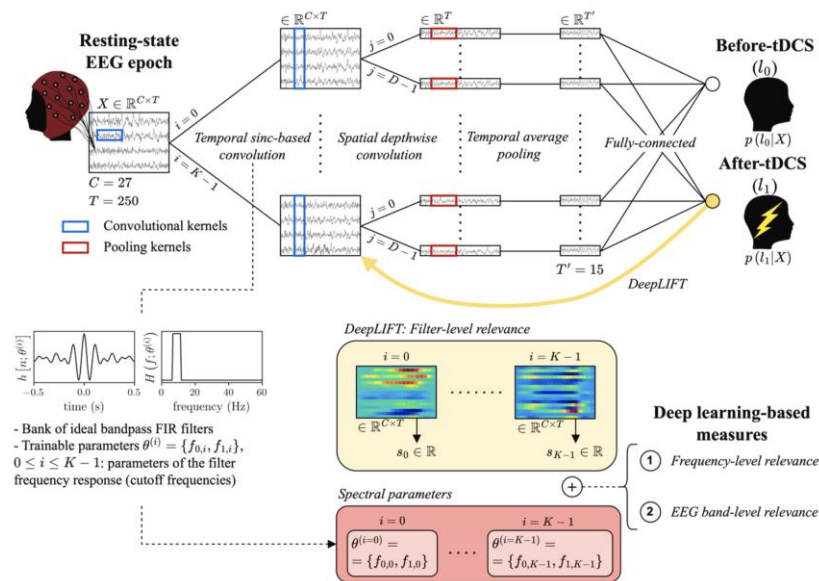


Figure 9: Explainable artificial intelligence approach for analyzing tDCS-based intervention in disorder of consciousness (DOC). An interpretable neural network classifies before-tDCS vs. after-tDCS resting-state EEG epochs recorded from DOC patients with a minimally conscious state diagnosis. The network architecture is illustrated at a high-level on top. The model learns interpretable frequency-domain features, estimating the spectral parameters of the optimal bank of bandpass filters (cutoff frequencies, light-red box) used for classification. An explanation technique (DeepLIFT, yellow arrow and box) is then applied to quantify the relevance of each bandpass filter (filter-level relevance) for predicting the after-tDCS condition. The derived relevance values, together with the interpretable spectral parameters of the network, are combined to derive deep learning-based measures sensitive to DOC intervention, revealing the frequencies (frequency-level relevance) and the EEG bands (band-level relevance) most modulated after the intervention.

The approach was tested on EEG data recorded in different tasks and populations: motor imagery in healthy participants (Borra, 2025b; Borra, 2025c), visual oddball in patients diagnosed with autism spectrum disorder (Borra, 2024), and resting state in patients with disorder of consciousness (Borra, 2025a). The proposed framework enables the characterization of brain oscillations in an automatic, optimal and end-to-end way, and is useful for boosting our comprehension of brain functions in healthy participants and in patients, tracking their neurophysiological/neuropathological modulations.

The developed model is available at: <https://zenodo.org/records/16156956>.

Borra D., et al. (2025a). Revealing EEG signatures of intervention in disorder of consciousness using artificial intelligence: methodology and feasibility. <https://doi.org/10.1016/j.cmpb.2025.109159>

Borra, D., & Magosso, E. (2025b). Unveiling multi-domain signatures of EEG oscillations using a fully-interpretable convolutional neural network. <https://doi.org/10.1016/j.cmpb.2025.109008>

Borra D., et al. (2025c). Automatically revealing multi-domain EEG signatures related to motor imagery using a fully-interpretable convolutional neural network. In press.

Borra D., et al. (2024). EEG Features Learned by Convolutional Neural Networks Reflect Alterations of Social Stimuli Processing in Autism. https://doi.org/10.1007/978-3-031-72341-4_9

2.9. AI Framework for supporting the analysis of EEG-derived brain functional connectivity

This Computational Model consists of a deep learning-enriched framework aimed at analyzing spectral directed functional connectivity. The model has been already successfully published in (Borra, 2025a; Borra, 2025b; Borra, 2024).

2.9.1. Technical Insight

The knowledge learned by an interpretable convolutional neural network – trained to discriminate brain states from functional connectivity – is used to define novel inflow and outflow measures, characterized for being non-linear, and for combining the information across brain regions and frequencies in an optimally discriminative way. Moreover, by explaining network decision with an explanation technique, the framework reveals the most relevant frequency contents and connectivity inflow/outflow. We applied our approach to EEG functional connectivity estimated at both the scalp and cortex level from healthy participants, during motor imagery tasks.

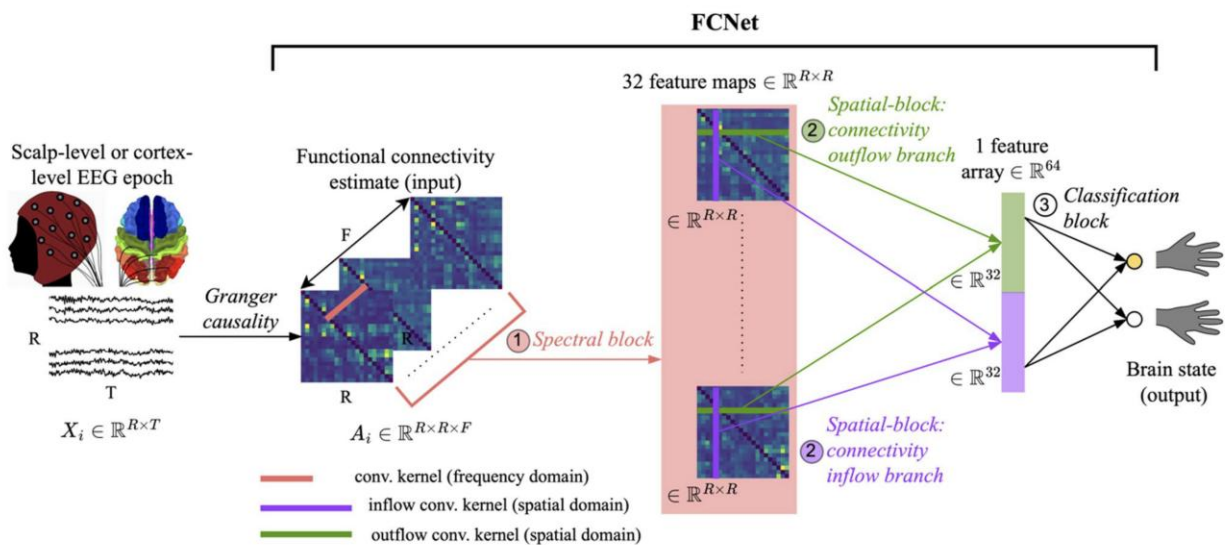


Figure 10: High-level graphical representation of the processing operated by FCNet blocks. FCNet comprises three main blocks: a spectral block, a spatial block (composed by two connectivity outflow and inflow parallel branches), and a classification block. Here we schematize the processing by reporting the output of each FCNet block while processing the input spectral functional connectivity $A_i \in \mathbb{R}^{R \times R \times F}$, computed from the EEG activity $X_i \in \mathbb{R}^{R \times T}$ at the scalp-level or cortex-level. The output is represented by the right-hand and left-hand motor imagery conditions.

Figure 10 illustrates the general architecture of this computational model. The framework reveals the spectral connectivity changes underlying motor imagery, and the network-based inflow/outflow measures captured connectivity changes with high strength and significance like graph theory measures (in degree, out degree, authority, hubness). Our framework is helpful to

elucidate the predictability of brain functional networks, and the most informative frequencies and connectivity inflow/outflows for the analyzed brain states.

Borra, D., & Magosso, E. (2025a). A deep learning-enriched framework for analyzing brain functional connectivity. <https://doi.org/10.1038/s41598-025-17635-5>

Borra D., et al. (2025b). A Compact Convolutional Neural Network for Decoding EEG Functional Connectivity: Application to Motor Imagery. https://doi.org/10.1007/978-3-031-82487-6_8

Borra, D., Ravanelli, M. (2024). Explaining Network Decision Provides Insights on the Causal Interaction Between Brain Regions in a Motor Imagery Task. https://doi.org/10.1007/978-3-031-71602-7_14

2.10. Workflow for trustworthy EEG decoding with deep neural networks

A comprehensive workflow for decoding EEG signals was developed, addressing the challenge posed by the introduction of many hyperparameters (defining data pre-processing, network architecture, network training, and data augmentation) and by the fluctuations of model performance due to the random initialization of model parameters. The model has been published in (Borra, 2025).

2.10.1. Technical Insight

The workflow employs automatic hyperparameter search exploring the hyperparameters characterizing the entire pipeline and includes multi-seed initialization for providing robust performance estimates. The architectural overview is shown in **Figure 11**.

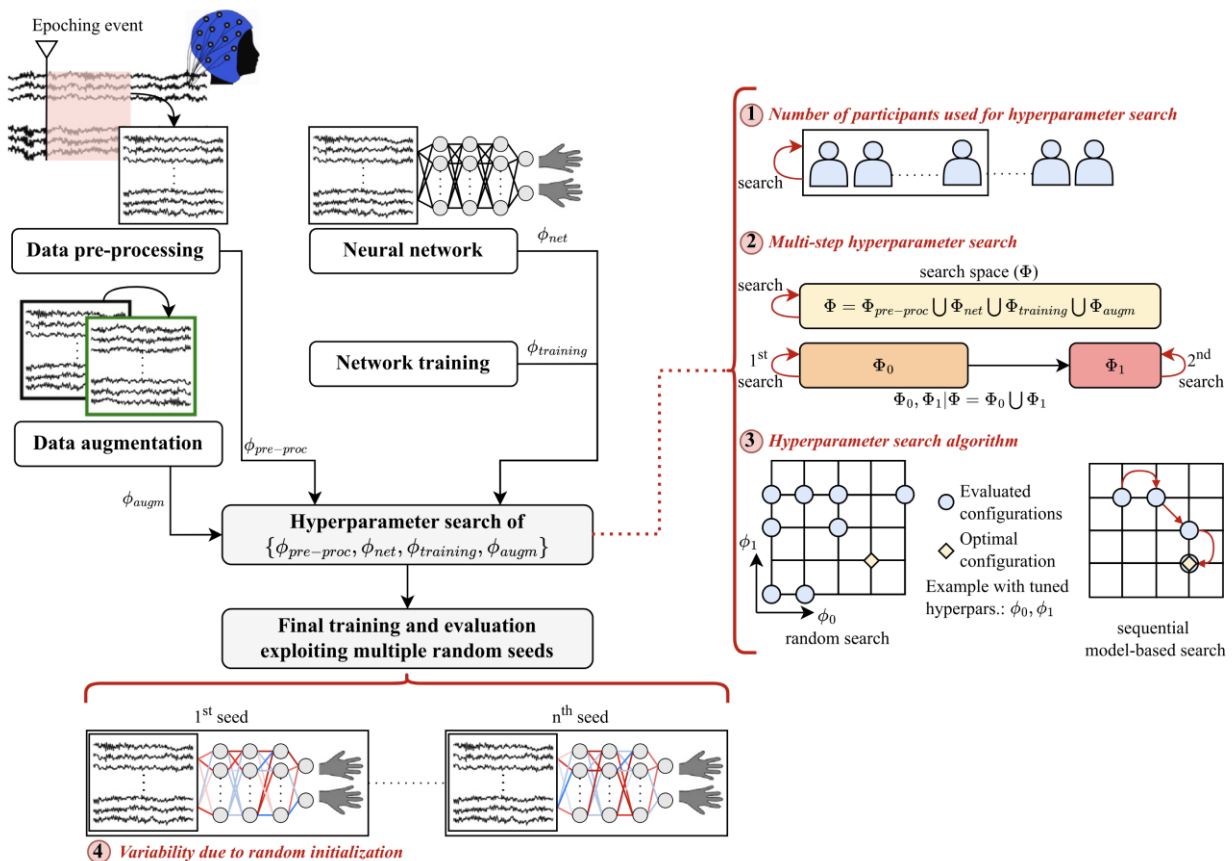


Figure 11: Scheme of the proposed decoding protocol and of the performed experiments (marked with red).

The approach was validated on 9 datasets about motor imagery, P300, SSVEP, overall including 204 healthy participants and 26 recording sessions, and on different deep learning models (Borra, 2025). Our workflow consistently outperformed baseline state-of-the-art pipelines, widely across datasets and models, and could represent a standard approach for neuroscientists for decoding EEG in a trustworthy and reproducible way.

Link to the developed workflow (included in the Python library SpeechBrain):
<https://github.com/speechbrain/benchmarks/tree/main/benchmarks/MOABB>.

Borra D., et al. (2025). A protocol for trustworthy EEG decoding with neural networks.
<https://doi.org/10.1016/j.neunet.2024.106847>

2.11. Allograft Risk Score

This Computational Model defines a new Allograft Risk Score (ARS) for early prediction of renal dysfunction in kidney transplantation (KT), based on a time-variant multivariate PLS model. The rationale behind the design of ARS is based on the hypothesis that, in successful KT, a predictive model can establish a consistent relationship between various predictors, including past Mid-Regional proadrenomedullin (MR-proADM) levels, and current creatinine levels. In the presence of an abnormal condition, such as suboptimal kidney function, this relationship becomes less predictable, increasing the model's prediction error. A higher prediction error indicates that the usual pattern no longer holds, signaling potential KT complications and resulting in a higher ARS for the subject. The study design is schematized in **Figure 12**.

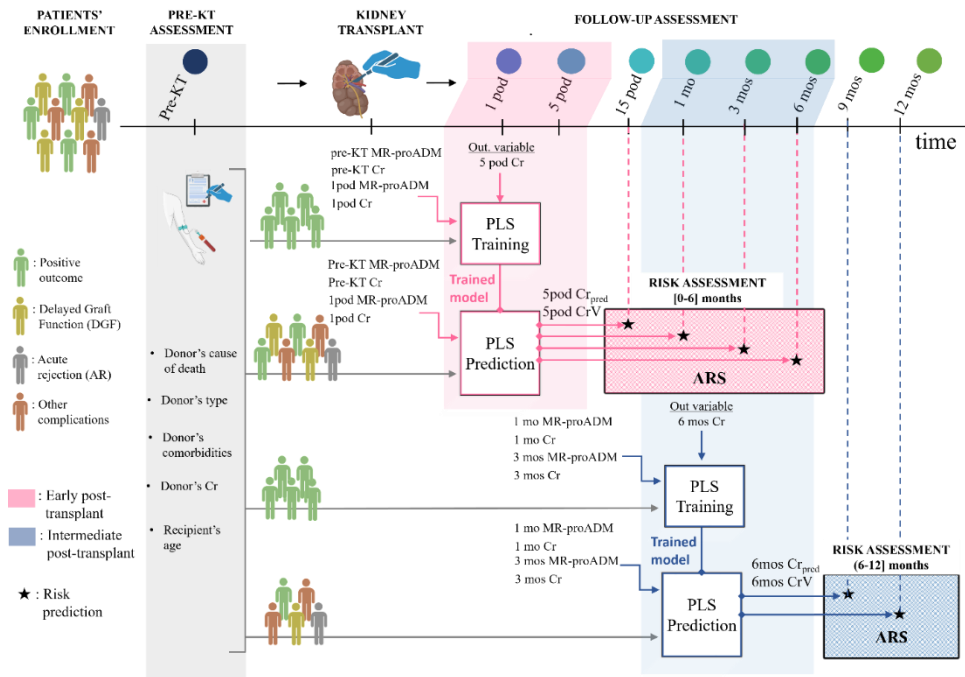


Figure 12: Schematic representation of the study protocol and derivation of the Allograft Risk Score (ARS). Patients' assessment was performed before kidney transplantation (pre-KT) and continued at multiple protocol times (tp) up to 12 months post-transplantation. Two reference PLS models for prediction of creatinine levels at time protocol $tp = 5$ POD and $tp = 6$ months are derived and trained with data relative to patients with positive outcomes (in absence of complications). On test subjects with unknown outcome, the Allograft Risk Score (ARS) was derived for risk assessment in the early (0-6 months) and intermediate (6-12 months) post-transplant phases, respectively, based on predicted creatinine levels (Cr_{pred}) and measured creatinine variations (CrV).

Abbreviations: AR, acute rejection; DGF, delayed graft function; KT, kidney transplantation; MR-proADM, MR-proadrenomedullin; POD, post-operative days; PLS, partial least squares.

2.11.1. Technical Insight

The **Allograft Risk Score (ARS)** was defined based on both the measured variations of serum creatinine (as indicator of the dynamics of the renal function) and on a deviation score from a reference profile. ARS is based on the information encoded in a multivariate predictive model performed on data from reference individuals who experienced KT with positive outcomes. Therefore, it encompasses the “desired” relationship between the values of n predictors and the temporal trend of serum creatinine. The more the test-individual's profile overlaps the reference ones, the lower the risk of transplantation failures.

Two reference Partial Least Squares (PLS) models for the prediction of creatinine value at $t_p = 5$ pod and $t_p = 6$ months have been trained, and are exploited for risk prediction in the early post-transplant phase [0-6 months] and in the intermediate post-transplant phase [6-12 months], respectively. Models are trained using subsets of reference subjects with positive transplant outcomes (control set in the absence of complications). During training, PLS regression (Rosipal, 2006) was used to model the relation between a set of predictors, x_1, x_2, \dots, x_N , and the predicted variable y_1 , i.e., creatinine value at $t_p = 5$ pod and $t_p = 6$ months. The N predictors include clinical data from both kidney donors and transplant recipients, as well as MR-proADM and creatinine values, cr , monitored at the two previous past protocol times. The ARS, is defined as the product of the creatinine deviation score (CDS) with the measured creatine variation (CrV):

$$ARS = CDS \cdot CrV = \text{median} |Cr_{pred}(t_p) - Cr_{meas}(t_p)| \cdot \frac{Cr_{meas}(t_p)}{Cr_{meas}(t_{p-2})}$$

The CDS quantifies the median deviation of the predicted creatinine values at

t_p , $Cr_{pred}(t_p) = \{cr_{pred_1}, cr_{pred_2}, \dots, cr_{pred_N}\}$, from $Cr_{meas}(t_p)$ obtained, in test, with $N = 100$ reference PLS models. The median operator has been chosen to make the ARS robust to outliers.

The higher the CDS difference is between the patient's profile and those having a positive transplant outcome, which indicates an increased risk of renal dysfunction in the future. Whilst the CDS term is a predictive estimate based on a set of indicators related to transplantation (i.e., donor's and recipient's data, past creatinine and MR-proADM values), the presence of the CrV factor normalizes the score based on the variations of the actual creatinine over the two past protocol time points.

Figure 13 illustrates the definition of ARS and its geometrical interpretation as the area of the rectangle defined by the x-y coordinates in the CDS vs CrV plot.

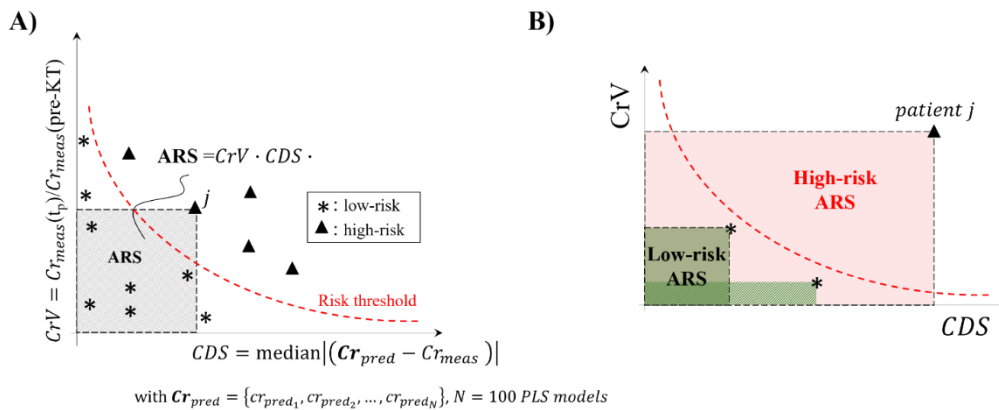


Figure 13: Definition of the Allograft Risk Score (ARS). (A) ARS is derived as the product of the measured creatinine variation, $Cr_{meas}(tp)/Cr_{meas}(pre-KT)$, with the absolute creatinine deviation score (CDS), which is obtained as the median of the absolute error of the creatinine predictive estimates over 100 reference PLS models. (B) Geometrical interpretation of ARS: ARS corresponds to the area defined by the xy coordinates, with $x = CDS$ and $y = CrV$. The higher the area, the higher the risk for graft dysfunctions.

Abbreviations: ARS, allograft risk score; CDS, creatinine deviation score; Cr, creatinine; PLS, partial least squares.

Internal cross-validation of the ARS was performed by repeating model training 100 with 80/20 random splits of the control group. We then evaluated the model responses both in test recipients with successful transplants (control) and in all the other recipients, including those who have experienced renal dysfunction. Performance of the PLS regression in cross-validation was assessed in terms of Mean Squared Error (MSE) for the four groups of interest: control patients in absence of complications, patients with miscellaneous complications, patients with DGF, and patients with AR. Patients were considered with recovered DGF when they interrupted hemodialysis sessions. Patients were assigned to the category with AR if they experienced histologically-proved AR within the prediction window. The predictive capability of individual variables and of the predictive scores was assessed in terms of Area under the Receiver Operating Characteristic (ROC) curve (Jong , 1992; Zweig, 1993). The optimal threshold, and corresponding sensitivity and specificity values, were determined at the point of the ROC curve closest to the vertex of coordinates (0,1). Statistical significance of the difference between the groups' means was assessed with one-way ANOVA across the 100 repetitions. For individual groups, multiple comparison against the control group was based on Dunnett's test. Analyses and figure generation were performed using Matlab R2024b.

The performance of the PLS regression is summarized in Table 4. MSE was computed for the control groups (in both training and test phases), and for test groups with other complications (AR, DGF or necessity of HD during the intermediate phase). As expected, MSE was lowest in the control group during training, and remained lower in the test controls compared with patients who developed complications, supporting the model's capacity as an anomaly detector. In the

intermediate post-transplant phase, MSE values were overall lower, with patients experiencing non-renal complications showing profiles closer to controls.

Table 4: Performance of the Partial Least Squares regression. F Average (and standard deviation) of mean-squared error in the creatinine predictive estimates obtained over the 100 repetitions of 20/80 cross-validation. P-values refer to one-way ANOVA comparing the MSE obtained across the 100 repetitions. For individual groups, multiple comparison against the control group is based on Dunnett's test. One-way ANOVA indicates means of MSE across the 100 repetitions with statistically significant difference ($p < 0.01^{}$) among the groups. The same holds for individual groups against the control group based on Dunnett's test ($p < 0.01^{**}$). Other complications include surgical (vascular, urological) and infective complications. Abbreviations: HD, hemodialysis; mos, months; MSE, Mean Squared Error.**

| | | Training | | | | Test | |
|--|----------------|--------------|-------------|---------------------|------------------------|-----------------|--------------|
| | | Control | Control | Other complications | Delayed graft function | Acute rejection | Total |
| Early post-transplant phase [<6 months] | n. of patients | 18 | 18 | 34 | 15 | 4 | 71 |
| | MSE | 2.04 (0.45) | 5.28 (3.70) | 9.88 (1.40) | 11.75 (2.04) | 7.28 (2.00) | 9.11 (11.15) |
| | p-value | - | - | $<0.01^{**}$ | $<0.01^{**}$ | $<0.01^{**}$ | $<0.01^{**}$ |
| | | Control | Control | Other complications | HD at 6-12 mos | Acute rejection | Total |
| Intermediate post-transplant phase [6-12 months] | n. of patients | 13 | 13 | 20 | 1 | 4 | 38 |
| | MSE | 0.01 (0.004) | 1.14 (0.10) | 0.26 (0.06) | 2.90 | 7.04 (1.37) | 0.98 (2.94) |
| | p-value | - | - | 0.61 | $<0.01^{**}$ | $<0.01^{**}$ | $<0.01^{**}$ |

Table 5 reports the predictive capabilities of individual input variables and the ARS in the early (15 POD, 1, 3 and 6 months) and intermediate (9 and 12 months) post-transplant phases. The ARS model demonstrated strong predictive ability, anticipating renal dysfunction up to 6 months in advance, with AUC: 0.81 [0.73-0.97] in the early phase and AUC: 0.98 [0.95-1.00] in the intermediate post-transplant phase.

Figure 11A-B shows the high sensitivity and specificity of ARS in both phases, as well as its capability to delineate patient risk profiles for post-transplant renal dysfunction. The most influential predictors, in terms of PLS weights, contributing to define the patient risk profile, were donor creatinine, donor comorbidities, and recipient MR-proADM and creatinine levels at POD 1 in the early phase and MR-proADM and creatinine level at month 3 in the intermediate phase (Figure 11C). Those variables are complemented, in the ARS scores, by creatinine values at POD 5 and months 6, respectively.

Table 5: Predictive capability of the Allograft Risk Score (ARS) in the early and intermediate phases after kidney transplantation. Average and 90% confidence intervals of AUC values are reported for individual variables (x_1, x_2, \dots, y_1) and for the proposed predictive scores, ARS. Results with average values higher than 0.70 are in bold. Abbreviations: ARS, allograft risk score; KT, kidney transplantation; MR-proADM, mid-regional pro-adrenomedullin; mo, month; mos, months; POD, post-operative day.

| Early post-transplant phase | | | | | | | | | | |
|-----------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-----|
| x_1 | x_2 | x_3 | x_4 | x_5 | x_6 | x_7 | x_8 | x_9 | y_1 | ARS |

| | <i>Recipient age</i> | <i>Donor type</i> | <i>Donor comorbidities</i> | <i>Donor cause of death</i> | <i>Donor Creatinine</i> | <i>Pre-KT MR-proADM</i> | <i>1 pod MR-proADM</i> | <i>Pre-KT Creatinine</i> | <i>1 POD Creatinine</i> | <i>5 POD Creatinine</i> | <i>ARS</i> |
|---|----------------------|---------------------|----------------------------|-----------------------------|-------------------------|-------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|
| [all] vs [renal dysfunctions] at 15 POD (n=71) | 0.51 [0.50-0.72] | 0.50 [0.50-0.62] | 0.62 [0.50-0.75] | 0.55 [0.50-0.77] | 0.60 [0.50-0.87] | 0.55 [0.50-0.66] | 0.76 [0.66-0.86] | 0.51 [0.50-0.66] | 0.65 [0.51-0.78] | 0.85 [0.80-0.95] | 0.87 [0.81-1] |
| [all] vs [renal dysfunctions] at 1 months (n=71) | 0.50 [0.50-0.67] | 0.52 [0.50-1] | 0.55 [0.50-1] | 0.59 [0.50,0.82] | 0.51 [0.50,0.79] | 0.50 [0.50,0.64] | 0.78 [0.73,0.88] | 0.63 [0.50-0.76] | 0.54 [0.50,0.61] | 0.71 [0.57-0.80] | 0.85 [0.75-0.94] |
| [all] vs [renal dysfunctions] at 3 months (n=71) | 0.59 [0.50-0.81] | 0.62 [0.50-0.81] | 0.55 [0.54-0.75] | 0.55 [0.50-0.72] | 0.67 [0.52-0.86] | 0.61 [0.50,0.90] | 0.62 [0.50,0.78] | 0.76 [0.74-0.85] | 0.59 [0.50,0.70] | 0.64 [0.50-0.76] | 0.81 [0.73-0.97] |
| [all] vs [renal dysfunctions] at 6 months (n=71) | 0.58 [0.50-0.94] | 0.56 [0.50-0.76] | 0.53 [0.50-0.63] | 0.50 [0.50,0.73] | 0.58 [0.50,0.84] | 0.52 [0.50,0.69] | 0.69 [0.53,1] | 0.75 [0.61-0.93] | 0.55 [0.50,0.69] | 0.60 [0.50-0.72] | 0.81 [0.72-0.91] |
| <i>Intermediate post-transplant phase</i> | | | | | | | | | | | |
| | x_1 | x_2 | x_3 | x_4 | x_5 | x_6 | x_7 | x_8 | x_9 | y_1 | <i>ARS</i> |
| | <i>Recipient age</i> | <i>Donor type</i> | <i>Donor comorbidities</i> | <i>Donor cause of death</i> | <i>Donor Creatinine</i> | <i>1 mo MR-proADM</i> | <i>3 mos MR-proADM</i> | <i>1 mo Creatinine</i> | <i>3 mos Creatinine</i> | <i>6 mos Creatinine</i> | <i>ARS</i> |
| [all] vs [renal dysfunction] at 9 months (n=38) | 0.53 [0.50-0.81] | 0.55 [0.50-0.87] | 0.63 [0.50-0.87] | 0.64 [0.50-0.86] | 0.52 [0.50-0.70] | 0.60 [0.50-1] | 0.89 [0.80-1] | 0.78 [0.66-0.93] | 0.84 [0.72-0.91] | 0.96 [0.92-1] | 0.96 [0.91-1] |
| [all] vs [renal dysfunction] at 12 months (n=38) | 0.52 [0.50-0.74] | 0.53 [0.50-0.76] | 0.57 [0.50-0.83] | 0.54 [0.50,0.75] | 0.51 [0.50,0.68] | 0.87 [0.81,1] | 0.78 [0.68,0.98] | 0.77 [0.55-0.99] | 0.81 [0.67,0.98] | 0.97 [0.94-1] | 0.98 [0.95-1] |

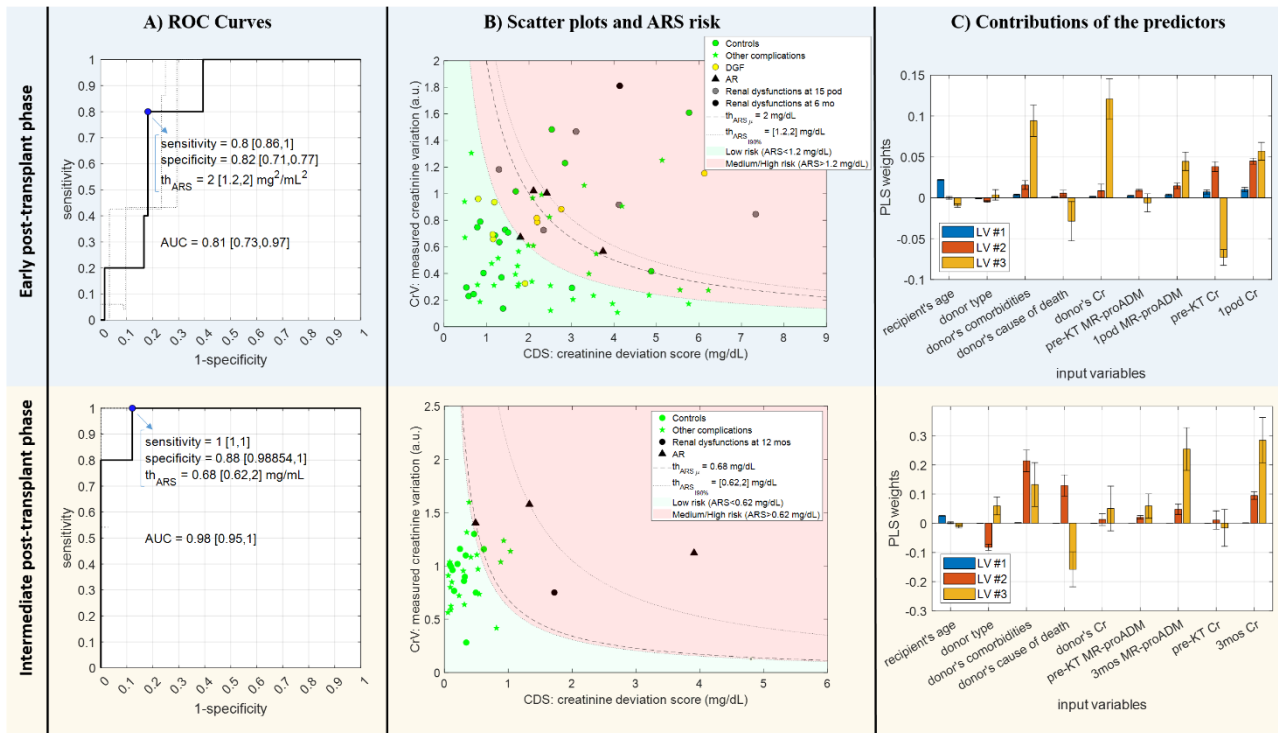


Figure 14: Discriminant capability of the Allograft Risk Score (ARS) in the early and intermediate post-transplant phases. The figure shows the capability of the ARS in the early (blu box) and intermediate (yellow box) post-transplant phases. (A) ROC curves for the risk scores obtained at 5 pod and 6 months after KT evaluated with reference to outcomes at 6 months (early post-transplant phase) and 12 months (intermediate post-transplant phase), respectively. AUC values are indicated together with sensitivity, specificity and threshold of the ARS (th_{ARS}) values obtained at the operating point closest to the vertex of coordinates (0,1), shown with a blue circle. Bootstrapping with 100 iterations was used to estimate the Studentized confidence intervals at the 90% confidence level. (B) Scatter plots of the creatinine deviation score (CDS) versus the measured creatinine variation, CrV. The dotted lines indicate the average (and 90% confidence intervals) ARS thresholds corresponding to the operating point closest to the vertex of coordinates (0,1) in the ROC curve which delimit the area at low risk (in green) from the area at higher risk (in red). (C) Contributions of the predictors to the latent variables of the reference PLS models.

Abbreviations: AR, acute rejection; DGF, delayed graft function; mo, month; mos, months; CDs, creatinine deviation score; Cr, creatinine, MR-proADM, mid-regional pro-adrenomedullin; POD, post-operative day.

Rosipal R, Krämer N. (2006). Overview and recent advances in partial least squares. doi:10.1007/11752790_2

Jong S. de, (1992). SIMPLS: an alternative approach to partial least squares regression. Chemom Intell Lab Syst. 1992;(18):251-263

Zweig MH, Campbell G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. doi:10.1093/clinchem/39.4.561

2.12. Automatic segmentation models to identify Region of Interests

This Computational Model exploits automatic segmentation approaches to identify the Regions of Interests (ROIs) for biomechanical Computational Models: vertebrae and metastatic lesions from CT images, and intervertebral discs (IVDs) from MR images. Additionally, vertebrae segmentation masks require the splitting into vertebral body (VB) and posterior elements (PE), while IVD masks require the separation of annulus fibrosus (AF) and nucleus pulposus (NP), if the disc degeneration allows it.

The work was carried out on public datasets, including: VerSe 2019 and VerSe 2020 (Large Scale Vertebrae Segmentation Challenge, available from OSF), Lumbar SPIDER (SPine segmentation: Discs, vERtebrae and spinal canal, available from Zenodo) for MRI images, and Spine-Mets-CT-SEG (Spine metastatic bone cancer: pre and post radiotherapy CT, available from The Cancer Imaging Archive) specifically for metastatic lesions.

2.12.1. Technical Insight

To segment vertebrae from CT and IVDs from MRI, deep learning models such as the U-Net architecture (a Convolutional Neural Network architecture specifically designed for image segmentation), the nnU-Net framework, the U²-Net architecture, and more recently hybrid architectures mixing U-Net and Transformers (Swin-U-Net, UNETR, Swin-UNETR) have been investigated. Models were evaluated by Dice metric and median Hausdorff distance (HD) averaged over classes. Then, the refinement of these masks into sub-ROIs was achieved via computational segmentation models exploiting image processing techniques, without the use of deep learning, to reduce the computational burden. Similar techniques were used to segment metastatic lesions given the annotation of the vertebra with the lesion.

Segmentation masks of vertebrae (**Figure 15**) and of IVDs (**Figure 16**) were obtained with deep learning models. Referring to vertebrae segmentation, some cases were segmented by the model in an accurate way (**Figure 15-a**), while in other cases the model assigned vertebral classes in an imprecise way, e.g. by assigning portions of different adjacent vertebrae to the same class (**Figure 15-b**). This variability in segmentation output masks leads to moderate values in metrics: 0.76 for test Dice metric, and 8.2 mm for median HD on the test set. The results refer to the use of nnU-Net. It is worth mentioning that cases where the model finds additional vertebrae to segment with respect to the reference mask make all labels shift, e.g. by +1, leading to larger values in HD. Referring instead to IVDs segmentation, model output masks were compared both in the MRI acquisition plane (**Figure 13-a**) and in out-of-plane view (**Figure 13-b**); the former helped in assessing the general shape of the IVD and its lower and upper surfaces along the inferior-superior direction, while the latter was useful to assess how good the model learned the IVD 3D shape. Out-

of-plane comparison was presented blindly to radiologists of the project, who generally preferred the model output to the reference mask. A test Dice metric of 0.79 and a median HD of 6.2 mm in the test set were obtained, using a U-Net model. Other architectures are currently being trained and/or evaluated for these two tasks.

Several approaches were developed for the separation of VB and PE in vertebrae segmentation. A first approach is based on calculating the distance map of points in the mask with respect to vertebral centroid along the anterior-posterior axis and adding constraints both based on the number of connected components and on grey-level intensity because of the different cortical bone content. In this approach, intensity distribution was modeled as a Gaussian mixture, and VB/PE separation was performed using an expectation-maximization algorithm. Separation of AF and NP was achieved by binary threshold on IVD mask. Metastatic lesions were segmented by combining binary threshold, flood-fill algorithm and morphological operations; segmentation was focused on the within-bone part of the lesion. For these steps, it was not possible to compute metrics as no reference segmentation was available.

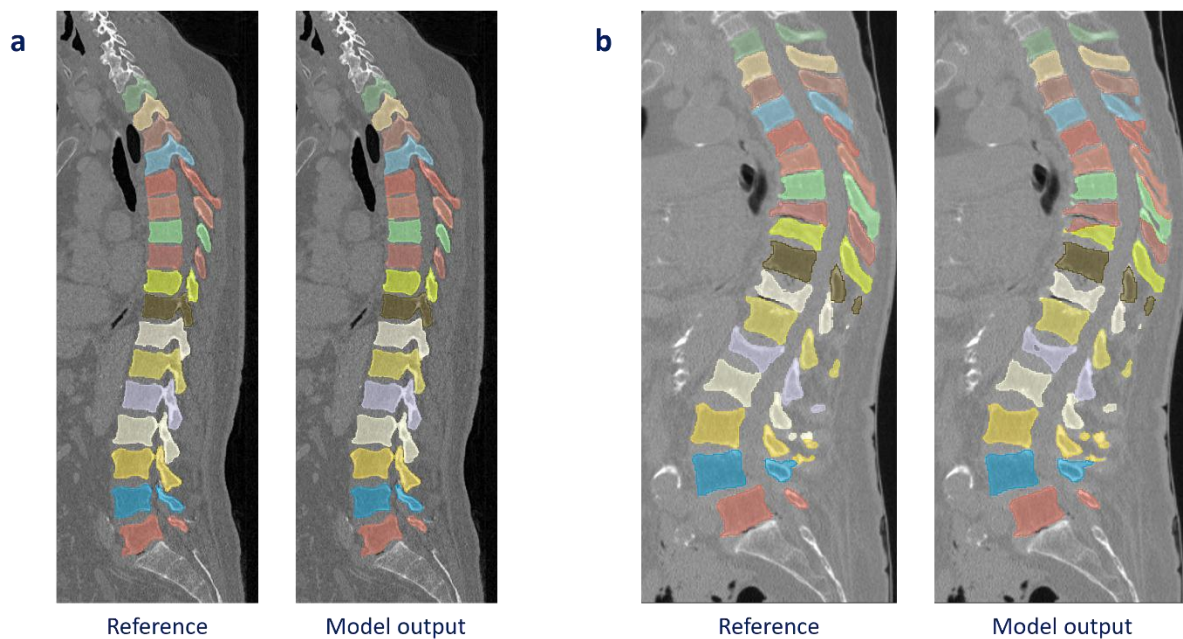


Figure 15: Visual comparison of vertebrae segmentation masks in two cases, a and b, in sagittal view. In both, reference segmentation mask is on the left while model output segmentation mask is on the right. Case a: the model correctly segmented vertebrae, identifying the same classes as in the reference. Case b: the model correctly identifies the classes of lumbar vertebrae but misidentifies three adjacent classes of thoracic vertebrae (starting from the first segmented vertebrae in the top of the image, classes 6-9 are not properly segmented).

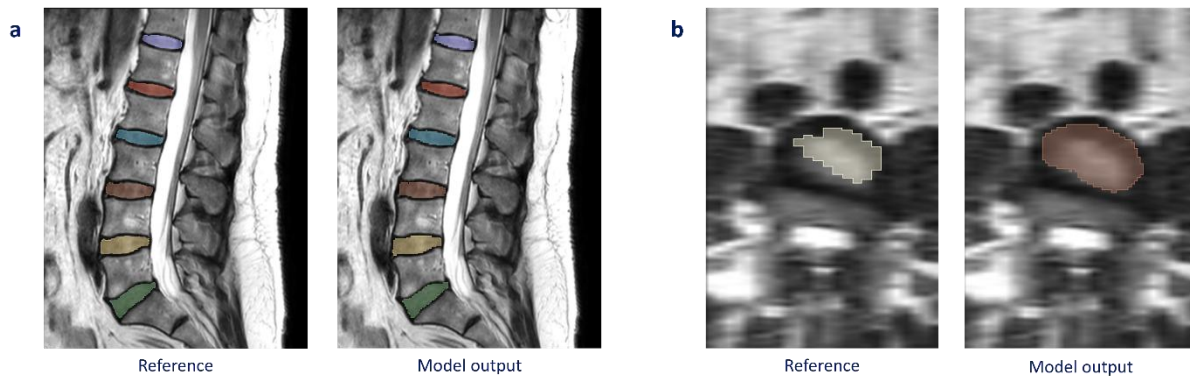


Figure 16: Visual comparison of IVD segmentation masks in two cases, a and b, in two views. In both, reference segmentation mask is on the left while model output segmentation mask is on the right. MRI were acquired in the sagittal plane. Case a: example of sagittal view in a T2W MRI acquired in the sagittal plane. Case b: example of axial view in a T2W MRI acquired in the sagittal plane.

2.13. Model for Automatic Issue Classification

This research effort addressed automatic issue classification, an essential task for distinguishing between bug reports, enhancement requests, and support issues.

2.13.1. Technical Insight

In Impact of Data Quality for Automatic Issue Classification Using Pre-trained Language Models, pre-trained transformer-based classifiers were adopted to automate this process, with specific attention to understanding how the quality of issue-tracking data affects model performance. In **Figure 17** the architecture of two different classifiers is presented.

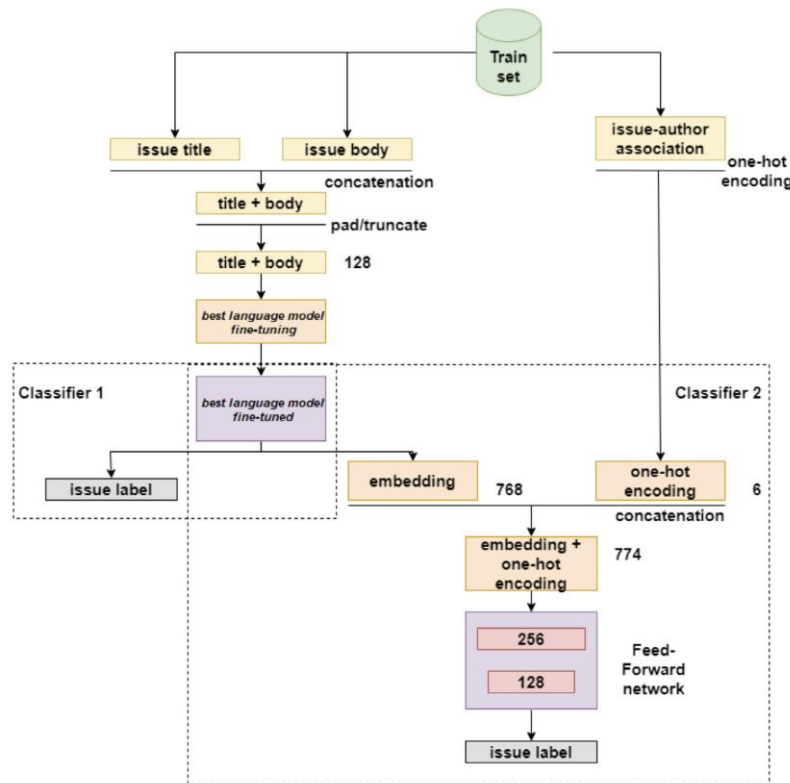


Figure 17: The two classifiers implemented for issue labeling.

Despite applying a variety of data quality filters, the experiments showed no substantial improvements in classification outcomes. The findings point to a deeper challenge—an underlying threat to construct validity in issue labelling practices—which limits the achievable performance of automated classifiers. This insight underscores how inconsistencies in ground-truth labels can become a fundamental bottleneck even when using state-of-the-art NLP models (Colavito, 2024).

Colavito G., et al. (2024). Impact of data quality for automatic issue classification using pre-trained language models. <https://doi.org/10.1016/j.jss.2023.111838>

2.14. Model for Emotion Recognition in software development

This Computational Model focuses on emotion recognition in software development, where timely identification of developers' emotional states can support productivity and well-being.

2.14.1. Technical Insight

In (Grassi, 2025), a novel modelling approach was introduced to mitigate the distortions caused by individual physiological differences. Using non-invasive biometric sensors, the method clusters developers according to their physiological profiles and trains cluster-specific emotion classifiers.

Figure 18 illustrates the adopted pipeline for constructing the model.

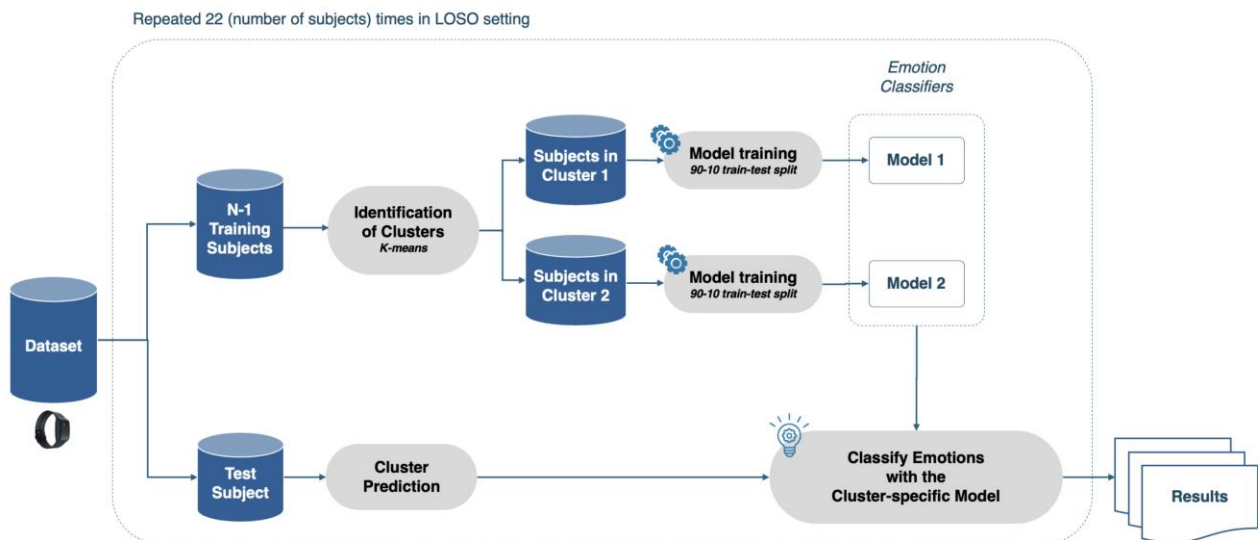


Figure 18: Pipeline of the cluster-based approach for train and evaluation of sensor-based emotion classifiers.

Evaluated on a dataset comprising self-reported emotions and biometric signals collected during a Java programming task, the approach achieved notable improvements over non-clustered baselines. For valence detection, the model yielded substantial gains in precision (+33%), recall (+12%), and F1-score (+18%), and for arousal recognition, precision increased by 29% (see **Figure 19**).

| | Precision | Recall | F1 | Accuracy |
|--------------------------------------|-------------------|-------------------|-------------------|-----------------|
| <i>Valence</i> | | | | |
| Cluster-based approach | .60 | .68 | .59 | .70 |
| Baseline [13] | .45 | .61 | .50 | .67 |
| <i>Improvement over the baseline</i> | <i>+.15 (33%)</i> | <i>+.07 (12%)</i> | <i>+.09 (18%)</i> | <i>.02 (3%)</i> |
| <i>Arousal</i> | | | | |
| Cluster-based approach | .52 | .60 | .52 | .64 |
| Baseline [13] | .40 | .59 | .49 | .62 |
| <i>Improvement over the baseline</i> | <i>+.12 (29%)</i> | <i>+.01 (1%)</i> | <i>+.03 (7%)</i> | <i>.02 (3%)</i> |

Figure 19: Performance of the cluster-based approach for valence and arousal recognition compared with the baseline performance.

The classifier proved particularly effective in detecting negative high-arousal states—conditions potentially linked to stress or reduced well-being. While the results are promising, further data and model refinement are required before practical deployment (Grassi, 2025).

Grassi D., et al (2025). A Cluster-Based Approach for Emotion Recognition in Software Development. DOI: 10.1109/CHASE66643.2025.00034

2.15. Generic Augmentation of 3d neuroimaging data

This Computational Model addresses generative augmentation of 3D neuroimaging data, a domain where deep learning applications are significantly constrained by limited datasets, especially for rare neurodegenerative conditions.

2.15.1. Technical Insight

In (Mallardi, 2025) the use of 3D Denoising Diffusion Probabilistic Models (DDPMs) to generate synthetic T1-weighted MRI volumes is investigated. The model was trained on a multicenter dataset of healthy subjects and explored the balance between structural fidelity and variability. Quantitative evaluation using Maximum Mean Discrepancy showed that the generated images closely matched the distribution of real MRI data, while visual inspection indicated that global and local brain structures were preserved. Limitations remained in reproducing fine-grained anatomical details, but overall, the study showed that DDPMs represent a promising strategy for augmenting neuroimaging datasets and supporting downstream tasks such as segmentation or classification. This contribution established one of the most technically advanced generative pipelines within the project (Mallardi, 2025; Basile, 2024). **Figure 20** shows samples of real and synthetic MRI images, while **Figure 21** shows the denoising architecture adopted in this computational model.

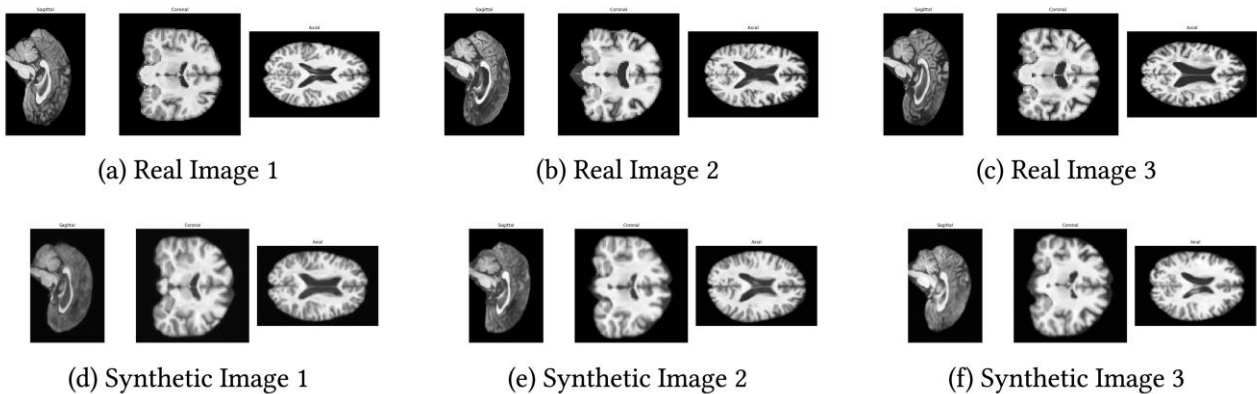


Figure 20: Comparison of real and synthetic MRI shows the diffusion model's anatomical accuracy

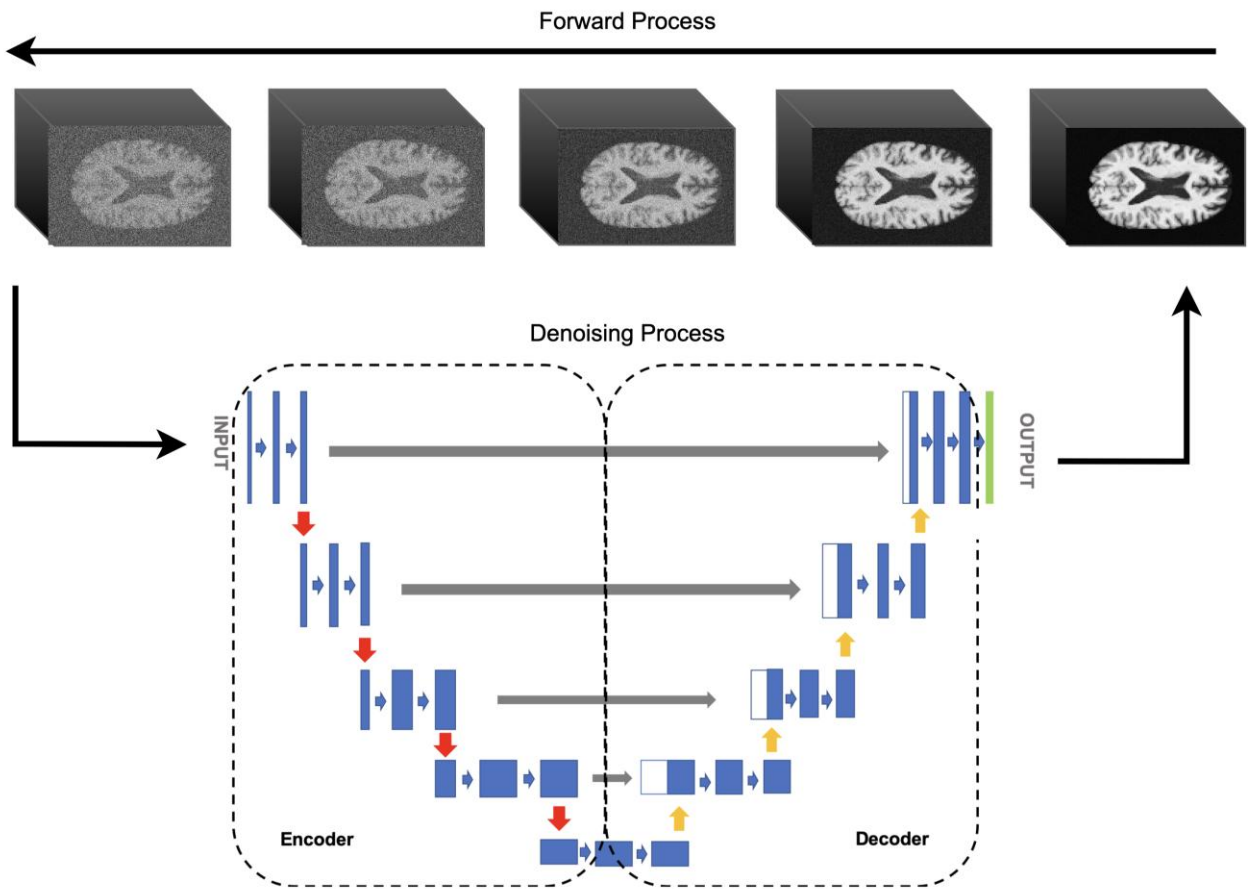


Figure 21: Architecture of the denoising diffusion probabilistic model (DDPM) adapted for 3D brain MRI synthesis

Mallardi G. et al. (2025). Diffusion Models for Neuroimaging Data Augmentation: Assessing Realism and Clinical Relevance. Journal of Medical Systems. Springer, 2025. (Under Review)

Basile A., et al. (2024), A Preliminary Study on Augmenting Neuroimaging data using a Diffusion Model. Proceedings of the 3rd AIxIA Workshop on Artificial Intelligence for Healthcare co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024) <https://ceur-ws.org/Vol-3880/paper24.pdf>

2.16. Framework for the automatic generation of regulatory documentation in AI-based medical software

This Computational Model provides a computational strategy for automating regulatory documentation in AI-based medical software, addressing a critical bottleneck in the Software as a Medical Device (SaMD) domain.

2.16.1. Technical Insight

MLOps-Driven Automation of Regulatory Documentation for AI-Based Medical Software proposed an approach that integrates MLOps principles—traceability, reproducibility, and continuous integration—directly into the development workflow. By linking documentation generation to artefacts produced during model development and execution, the system enables the production of consistent, audit-ready regulatory documentation with minimal manual effort. When applied to a representative healthcare AI project, the approach demonstrated its potential to streamline compliance processes and reduce the gap between rapid AI innovation and the formal requirements of medical device regulation. This contribution introduces a Computational Model designed to automate the generation of regulatory documentation within the DARE framework. (Rosmarino, 2025). The general architecture is shown in **Figure 22**.

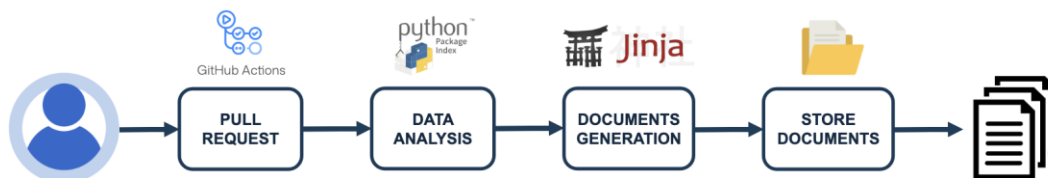


Figure 22: MLOps-Driven Automation of Regulatory Documentation for AI-Based Medical Software.

The five main stages can be roughly described as follows:

- **Pull Request:** When a developer opens a pull request on GitHub, a workflow configured via GitHub Actions is executed. This ensures that documentation is updated synchronously with code modifications.
- **Data Analysis:** The system extracts relevant metadata from the repository. This includes static code analysis, dependency extraction from configuration files, and querying the Python Package Index (PyPI) to retrieve licensing, versioning, and maintainer information.
- **Documents Generation:** The extracted metadata is passed to the templating component to generate structured documents. This stage produces two regulatory artefacts: the Software List and the SOUP List. Any unresolved or unverifiable fields are flagged for expert review.

- **Store Documents:** The generated documentation files are stored in a dedicated directory within the repository. These documents, along with auxiliary files such as historical software state snapshots and error reports, are committed and pushed to the pull request branch.
- **Final Output:** The result is a version-controlled, machine-generated set of documentation files, ready for inclusion in the technical file required by regulatory audits and certification procedures.

Rosmarino F., et al. (2025). MLOps-Driven Automation of Regulatory Documentation for AI-Based Medical Software. Proceedings of MLOps25, Workshop on Machine Learning Operations, co-located with the 28th European Conference on Artificial Intelligence (ECAI 2025), Bologna, Italy. CEUR-WS, 2025. (In press)

2.17. Identification of key factors causing intellectual disability in Down syndrome subjects

Down syndrome (DS) or Trisomy 21 is the most common genetic cause of intellectual disability. This Computational Model aims to understand the key factors that influence intellectual disability to create a care and treatment pathway for the prevention of worsening cognitive impairment in Down syndrome.

2.17.1. Technical Insight

The first step consists of building a database that includes personal, diagnostic, clinical, auxological and molecular data of children with DS. The second step will consist in the developing an AI-based algorithm, so as to support the large amount of collected data.

The aim is to ameliorate the diagnostic-therapeutic and care pathways for people with DS in order to prevent and control key aspects that have a worse impact on cognitive abilities. Specific objectives are: (a) to create a dataset of complex and mixed data of patients with DS using a FAIR method; (b) to include as many centers as possible in data collection; (c) to create machine learning models that can process data included in the dataset; (d) to highlight key aspects affecting the progression of intellectual disability in patients with Down syndrome; and (e) to create care pathways to prevent the worsening of cognitive impairment in these patients

From the methodological viewpoint, the integrated analysis of multiple data, not only with classical statistical approaches but also with Machine Learning approaches, innovatively addresses the need to integrate data of a heterogeneous nature to relax possible assumptions about the linearity of dependencies; such an approach may allow the generation of pathogenesis models and provide supports to improve diagnostic-therapeutic-care (PDTA) pathways aimed at people with pediatric Down syndrome.

In particular, the use of hybrid Informed Machine Learning methodologies, capable of combining sub-symbolic and symbolic elements, will enable the exploitation of existing medical knowledge (e.g., known causal effects) to reduce the amount of data required to train the models and improve their robustness. This will be achieved through neuro-symbolic, neuro-probabilistic and Machine Learning techniques with constraints.

The validation of the develop solution will be a key point to assess the Computational Models. To this end, an analysis pipeline for the identification of key correlates in a non-linear, multivariate setting has been set up. In particular, given a target feature (e.g. the Intelligence Quotient), the method allows to screen a dataset for other features that correlate with the target, either individually or synergistically. The pipeline works by learning a – typically non-linear – Machine

Learning model that attempts to estimate the target based on the features under screening. By their natures, Machine Learning are very effective at this type of task, though the relations they learn are not necessarily causal and incur the risk of overfitting, which therefore needs to be controlled. After a model has been learned, we employ state-of-the-art explainable AI techniques to interpret how the model is coming up with the estimate. While not sufficient for the identification of causal links, this process can speed up their discovery by highlighting complex correlations to find for a human researcher. In details, the designed pipeline is structured as follows:

1. Dataset preprocessing (training/test split, features scaling, etc.);
2. ML model learning, making use of hyperparameter tuning and k-fold cross-validation to control overfitting;
3. Accuracy evaluation for the target model;
4. Feature importance analysis using permutation importance scores and approximate Shapely values;
5. Automated feature selection via the Boruta method:

Two issues frequently arising in this kind of analysis are being addressed as well. Namely: 1) limited sample size; 2) importance attributions over correlated features. To address the first issue, a data augmentation strategy is being explored, based on the idea of *comparing the target values for pairs of examples (children with trisomy 21)*, rather than on directly estimating the target value for one individual. The underlying idea is that comparing pairs might be easier than estimating a specific value, and that pairs scale with the square of the original sample size. Therefore, this strategy can potentially lead to higher predictive accuracy (which translates into more reliable analysis results) and better data availability. The downside is that the approach requires more sophisticated analysis methods (e.g. specific Neural Network models) and its semantic is slightly different compared with the current, more straightforward, pipeline. This approach is currently under investigation.

The issue with importances correlated attributes is twofold: on the one hand, the dataset may include known correlations that or of little research interest, but that would be naturally picked up by a ML model (e.g. impact of age on specific metabolic concentrations, or on the unscaled IQ); on the other hand, confounders may lead a correlated feature to be deemed importance to the detriment of a true causal factor. To mitigate this issue, there is an attempt to develop an innovative approach that allow to construct synthetic representations of specific features that are de-correlated by construction with other, user selected, features. For example, the goal would be to obtain a representation of the creatinine blood concentration that is decorrelated from the age. Performing importance analysis on this new representation would allow one to measure the specific target correlation of creatinine, discounting the effect of age, without any artefact-prone grouping. This is a very challenging problem, for which we expect moderate progress by the end of the project, rather than a mature solution. That said, the issue is important enough to justify the research

effort. On this aspect, another challenge that occurs concerns the handling of missing data (Nan values) that can be used by the model as actual values useful for learning and predicting the target. The issue becomes more complicated when using a dataset that includes both numerical and categorical variables, such as ours.

From the technical viewpoint, the XGBoost approach has been chosen for the ML component, while Shapely value analysis and the Boruta algorithm have been chosen for the explanation analysis.

2.18. Multilingual Medical Chatbot Based on Large Language Models

This Computational Model consists of a multilingual medical chatbot designed to assist patients and healthcare professionals in managing basic clinical queries and consulting validated medical information.

The system is based on a pipeline of pre-trained language models (LLMs) arranged in sequence, ensuring natural interaction in Italian and clinically coherent responses that are terminologically and conceptually accurate.

The chatbot's logical workflow involves three main phases:

1. Translation of the user's question into English, using a pre-trained neural translation model (Transformer-based architecture).
2. Semantic processing and response generation by a medical LLM trained on clinical data, guidelines, and scientific literature. The model is prompted to respond as a qualified physician and aligned with the system's integrated Frequently Asked Questions (FAQ) database.
3. Reverse translation (English → Italian) of the generated response, producing an answer that is fluent, comprehensible, and free from linguistic ambiguity.

This modular approach combines the linguistic precision of translation models with the semantic depth and interpretive robustness of the medical LLM, resulting in contextually accurate and clinically informed conversations. The chatbot is also integrated within a clinical web application, extending its functionality: users can access an intuitive interface to submit questions, review previous interactions, and receive personalized guidance related to exams or reports uploaded to the platform.

2.18.1. Technical Insight

The system follows a multi-modular architecture, structured as a pipeline of neural models orchestrated through internal APIs and asynchronous connectors. The main functional modules are as follows:

- **Input Preprocessing Module:** Receives text in Italian, performs linguistic normalization (correction of spelling errors, punctuation, and medical abbreviations), and forwards the cleaned input to the translation model.
- **Translation Module (IT → EN):** Uses a pre-trained Transformer model (based on the Llama-7B architecture).

- **Medical LLM Module:** Implements a large-scale language model with approximately 4 billion parameters, trained on medical corpora, clinical guidelines, and biomedical datasets. The instruction prompt includes role conditioning (e.g., “Respond as a medical specialist”) and integrates a database of FAQs that guide answer generation.
- **Reverse Translation Module (EN → IT):** Applies a symmetric translation model to return the response in Italian while preserving medical terminology and meaning.
- **Postprocessing and Web Interface Module:** Handles text formatting, linguistic validation, and delivery of the final response to the user through the web application interface.

The entire system is implemented in Python, using FastAPI for orchestration and REST APIs for model communication. The computational infrastructure employs high-performance GPUs (CUDA) for inference, given the models’ computational complexity and large number of parameters.

Current limitations encountered so far are: (1) High computational complexity and high hardware resource consumption (CPU/GPU), limiting scalability in low-resource environments; and (2) dependence on translation quality, which may introduce minor terminological distortions in more complex clinical concepts.

Future work will focus on a number of aspects:

- Integration of an end-to-end multilingual model (eliminating intermediate translation steps) based on cross-lingual LLMs.
- Pipeline optimization using model quantization and distillation techniques to reduce response time and memory usage.
- Expansion of the training dataset with local clinical data and Italian-specific terminology to improve linguistic and cultural contextualization.
- Clinical validation of chatbot responses and usability testing in real-world healthcare settings.

2.19. Integrated Web Platform for Clinical Data Management and Medical Chatbot

As part of the screening project for lung disease prevention, an integrated web application incorporating artificial intelligence systems was developed to manage clinical data and facilitate patient interaction. The primary goal of the platform is to optimize the data flow derived from feasibility questionnaires and low-dose computed tomography (LDCT) exams, minimizing manual transcription errors and improving communication between patients and healthcare personnel.

The platform serves as a centralized management interface, capable of:

- Automatically receiving and organizing structured clinical data in REDCap.
- Populating clinical form fields directly from uploaded textual reports (PDF or free text).
- Allowing patients to view, edit, and download their own data and medical reports.
- Integrating a multilingual medical chatbot that responds to frequently asked questions and provides personalized assistance.
- Managing clinical appointments automatically, including WhatsApp notifications for reminders and updates.

This solution combines the capabilities of large language models (LLMs) with a scalable web-based infrastructure, creating a complete digital support system for population screening programs.

2.19.1. Technical Insight

The web application was developed as a full-stack system, following a modular architecture integrated with REDCap via secure REST APIs. The infrastructure is structured into three main layers:

1. **Frontend (User Interface).** Developed in React.js with dynamic components and a responsive design. Patients can: view and update their personal and clinical data; interact with the medical chatbot to receive instant answers and explanations about test results; manage and visualize their scheduled appointments; receive automated reminders through WhatsApp Business API integration.

Healthcare professionals access a separate panel that allows them to: review patient data; upload new reports; manage clinical schedules and appointments.

2. **Backend (Logic Management and AI Integration).** Developed in Python (Flask) with dedicated interfaces for communication with REDCap and LLMs. It includes: a report parsing module, automatically extracting relevant clinical information from uploaded documents (in

JSON format); a REDCap synchronization engine, responsible for populating or updating database fields automatically; and a scheduling service, managing appointments and sending notifications automatically.

3. **Database and Security.** The system relies on REDCap and Firebase for data storage and authentication. All transactions and access logs are recorded to ensure traceability and GDPR compliance. Authentication and access control are managed through Firebase, with differentiated permission levels for patients and healthcare staff.

The multilingual medical chatbot described in Section 2.18 was integrated into the platform, acting as a cognitive interface between patients and their clinical data. The chatbot workflow consists of:

1. Automatic translation of the patient's question from Italian to English (via a neural translation model).
2. Query submission to the medical LLM, trained on specialized corpora and clinical FAQs.
3. Response generation in English, followed by reverse translation into Italian.
4. Display of the response within the in-app chat interface.

This multi-model architecture ensures semantically coherent and accurate responses, although it introduces some computational overhead due to dual translation and GPU load. Despite this, the system has demonstrated high semantic accuracy and strong adaptability to the specific clinical context (LDCT exams, follow-up, eligibility for screening participation).

A limitation of the current development process stems from the fact that the doctor-patient interface still requires formal validation by the clinical team prior to large-scale deployment. As future work, the extension of the platform to additional clinical modules (e.g., cardiovascular, respiratory, and metabolic screening) will be pursued.

2.20. Automatic Extraction of Clinical Data from Reports and Population REDCap Databases

This Computational Model consists of a pipeline for the automatic extraction of structured clinical data from textual medical reports. Its main goal is to automate the population of a clinical database (REDCap), reducing the manual workload of healthcare personnel and ensuring greater consistency and quality of recorded data.

The entire process relies on a LLM specifically guided by a complex prompt that instructs the system to extract relevant information from medical reports according to the predefined structure of the target database. For each uploaded medical report, the system automatically generates the corresponding responses to the fields in the REDCap database, such as: Date of examination; ODI (phase/hour); BMI; Obstructive apneas (events/hour); and Heart rate (bpm).

The model therefore acts as an “intelligent semantic extractor”, capable of understanding natural medical language and converting it into structured and codified information.

After extraction, the data are automatically sent and synchronized with the REDCap database through secure APIs, and subsequently used to generate an interactive dashboard that allows researchers and clinicians to visualize aggregated patient data (clinical indicators, distributions, temporal trends, etc.).

The system is designed to operate in nightly batch mode, enabling automated and periodic data updates without requiring human intervention.

2.20.1. Technical Insight

The extraction pipeline was implemented in Python, following a modular design that facilitates integration into existing clinical environments. Its logical architecture consists of the following main modules:

1. **Report Acquisition and Preprocessing Module.** It imports files in PDF or plain-text format (with OCR extraction if necessary); it converts documents into .docx format for uniform processing.
2. **Semantic Extraction Module (LLM-based).** The core of the system is an instruction-following LLM, configured with a prompt describing the database schema. Then, the prompt guides the model to answer a series of predefined questions, each corresponding to a REDCap field (e.g., “What is the date of the examination?”, “What is the heart rate?”). Finally, the extracted answers are automatically converted into JSON format, ready for API transmission.

3. **REDCap Population Module.** The validated data are automatically sent to the REDCap platform via an authenticated REST API. Each field is tracked through unique report and patient identifiers, ensuring full data traceability.
4. **Visualization and Dashboarding Module.** A dashboard, developed in REDCap, allows dynamic visualization of the extracted clinical data. Charts and metrics are automatically updated with each nightly synchronization.

The infrastructure is optimized for execution on a dedicated server, with an asynchronous pipeline that supports parallel processing of multiple reports.

The system was evaluated on a sample of 30 real medical reports from a clinical screening project. The objective was to measure completeness and accuracy of data extraction compared with a manually built gold standard prepared by expert clinicians. The results are summarised as follows:

- Percentage of correctly extracted fields: 98%
- Average processing time per report: 30 seconds
- Mean semantic interpretation error: < 2% (mainly in reports with ambiguous or non-standard syntax)
- False positives in Boolean fields: 1.4%

From a qualitative standpoint, clinicians confirmed an excellent correspondence between textual content and structured output, achieving a manual entry time reduction of over 90%.

Since data updates occur automatically and asynchronously (nightly batch), computational latency does not represent an operational bottleneck. The achieved accuracy (98%) is considered more than sufficient for research and data management purposes, with selective supervision possible for ambiguous cases.

Few limitations are known:

- Extraction quality is strongly dependent on the structural consistency of reports, which may vary significantly between clinical centers.
- Some conceptual fields (e.g., “clinical recommendations”) require higher interpretive reasoning and may remain partially ambiguous.
- Processing large or numerous reports with the LLM entails significant computational costs.

Among future works, it is planned to take into consideration the implementation of a continual learning module, enabling the model to progressively adapt to the specific reporting style of each



clinical center. Also, the integration with automated clinical validation systems, for example by cross-checking extracted data with reference values or previous reports of the same patient, will be investigated in the future.

2.21. Sleep Management Platform and Digital Support for Patient Health

Sleep is an essential parameter for human health, strongly correlated with cardiovascular, metabolic, and neurological diseases. Despite growing awareness of sleep importance, systematic data collection on sleep quality and habits remains complex, fragmented, and highly dependent on clinical context. To address this challenge, a digital platform was designed in collaboration with the Sleep Center, aiming to:

- Continuously and systematically monitor subjective and objective sleep parameters;
- Collect information useful for clinical evaluation;
- Improve patient communication and management through digital tools;
- Promote the adoption of good sleep hygiene practices and greater therapeutic adherence.

The system enables comprehensive patient management, offering diary functionality, clinical questionnaires, report uploads, appointment management, and automated notifications, all synchronized with REDCap.

2.21.1. Technical Insight

The platform is implemented as a modular web app, developed in Flask (frontend) and Python (backend), with a microservices-oriented architecture and REST APIs. It integrates several components covering the entire sleep monitoring process:

1. Sleep Diary – Patients can record daily sleep information (e.g., bedtime and wake-up time, perceived duration, subjective quality (Likert scale 1–5), nocturnal awakenings). Data is saved in REDCap and displayed via a calendar, allowing both patient and doctor to track sleep quality over time.
2. Dream Diary – Optional module for recording dream content and associated emotions. Data, properly anonymized, can be used for qualitative and psychometric analyses or studies correlating dreams with sleep disorders.
3. Clinical Questionnaires – The system allows automatic sending of standardized questionnaires (e.g., PSQI, Epworth Sleepiness Scale, ISI). Questionnaires are single-entry; responses are recorded in REDCap and cannot be modified. Assignment can be configured by clinical staff according to follow-up needs.
4. Report Management – Doctors can upload clinical reports (e.g., polysomnography, actigraphy reports, overnight EEG), stored in the patient profile. Patients can view and download them but cannot modify them.

5. Sleep Hygiene – The platform includes a section dedicated to sleep hygiene guidelines, customizable according to the patient profile (e.g., insomnia, obstructive apnea, circadian disorders). Recommendations are provided dynamically.
6. Appointment Management and Notifications – The system includes an appointment management module, allowing patients to book visits, view upcoming appointments, and receive automatic WhatsApp reminders the day before. Clinical staff have a dashboard showing all scheduled appointments, questionnaire response status, and uploaded reports.

The platform is fully synchronized with REDCap, which serves as a structured repository for all clinical data. Integration occurs via authenticated REST API, automatically managing the patient record creation and updates; the questionnaire response uploads; and the data export for statistical analyses and longitudinal studies.

The current implementation exhibits a few limitations, namely the dependence on self-reported data, subject to perceptual bias; the lack of modules for direct acquisition from sensors (actigraphy, smartwatches, portable EEG); and a large-scale clinical validation has not yet been implemented.

Future work will be devoted to tackle a few challenges, like the integration with wearable devices for automatic collection of physiological parameters (heart rate, movement, oxygen saturation); the implementation of predictive modules for sleep disorder risk; the expansion of the system for long-term home monitoring and telemedicine interventions; and a multicenter clinical validation in collaboration with other sleep medicine centers.

2.22. Artificial Intelligence Pipeline for the Automatic Classification of Periprosthetic Hip Fractures

The goal of this Computational Model is to classify periprosthetic according to the Gruen using an AI-based deep learning pipeline, providing objective decision support for orthopedic surgeons. This model represents the first automated system capable of correlating radiographic classification with surgical and perioperative outcomes, paving the way for broader clinical adoption of AI-assisted preoperative planning.

2.22.1. Technical Insight

The system was designed as a multi-stage image processing pipeline, composed of four main modules:

1. Prosthesis Detection: automatically identifies the position and contours of the prosthetic stem, generating a reference mask for subsequent processing steps. It has been implemented using an object detection model (Single Shot Detector, SSD) with a VGG16 backbone, pre-trained on radiographic datasets.
2. Image Rotation and Normalization: it ensures correct alignment of the prosthesis with the horizontal axis by rotating and normalizing the input images.
3. Fracture Identification: it employs another object detection model (EfficientNet backbone) to localize the fracture along the femoral shaft, distinguishing between proximal, diaphyseal, and distal fractures.
4. Gruen Zone Classification: it determines the Gruen zone involved based on the anatomical position of the detected fracture relative to the femur.

All models were implemented in TensorFlow/Keras, trained on manually annotated images by expert orthopedic surgeons, and optimized through data augmentation (rotations, scaling, noise addition) to enhance generalization. The entire pipeline is orchestrated by a central control module that manages data flow, image normalization, and the integration of intermediate outputs into a structured final report.

As shown in Table 6, the multi-stage pipeline demonstrated high performance across all processing phases, confirming the effectiveness of the integrated approach.

| Module | Metric | Value |
|----------------------|--------|--------|
| Prosthesis detection | mAP | 99.56% |

| | | |
|---------------------------|-------------|-------|
| Rotation correction | Accuracy | 95.9% |
| Fracture identification | mAP | 86.2% |
| Gruen zone classification | Accuracy | 96.7% |
| Sensitivity (Gruen zones) | Sensitivity | 97.1% |
| Specificity (Gruen zones) | Specificity | 94.4% |

Table 6. Results of this Model

These results indicate robust performance in both detection and classification tasks, with a high correlation between automated and expert-annotated labels.

2.23. Automatic Detection of Noise in ECG Signals for Wearable Devices

The objective is to accurately differentiate between arrhythmias and noisy signals, to avoid compromising automatic arrhythmia detection in wearables, leading to false positives or undetected pathological events. The reference dataset included ECG signals manually annotated for noise presence, comprising examples from both clinical recordings and consumer wearable devices.

The project involved the development and evaluation of multiple machine learning (ML) and deep learning (DL) models for noise classification in ECG signals, aiming to balance diagnostic accuracy, computational efficiency, and memory usage.

2.23.1. Technical Insight

The overall architecture comprises four main components:

1. Feature Extraction – Four distinct approaches were tested: Raw ECG signal (waveform data); time-domain heart rate variability (HRV) features (mean, standard deviation of RR intervals, pNN50, RMSSD); combined time- and frequency-domain HRV features (power spectra, LF/HF ratio, spectral entropy); parameters derived from a genetic programming (GP) reconstruction algorithm, providing coefficients representing signal complexity.
2. Classification – Several models have been implemented and compared: Random Forest (RF); Support Vector Machine (SVM); Multi-Layer Perceptron (MLP); Gradient Boosting (GB); Convolutional Neural Network (CNN) for direct raw-signal processing
3. Post-Processing Module – Classification results are aggregated over temporal windows and used to dynamically filter noisy segments, preventing the storage or transmission of low-quality data.

The system can be embedded in wearable devices with a configurable activation threshold, determining whether a signal should be recorded or discarded. As showed in Table 7, the comparative analysis showed that:

- CNN models based on raw data achieved the highest accuracy but required significant computational and memory resources, making them less suitable for wearable implementation.
- Traditional models (RF, SVM) were lighter but slightly less accurate in complex noise scenarios.
- The Multi-Layer Perceptron (MLP) model, using combined time- and frequency-domain HRV features, represented the best trade-off between performance and efficiency.

| Model | Features | Acc | Sens | Spec | F1 | Memory/Time |
|-------|--------------------|-------|-------|-------|-------|-------------|
| CNN | Raw signal | 97.4% | 97.8% | 96.9% | 0.973 | High |
| MLP | HRV (time + freq) | 96.8% | 97.1% | 96.2% | 0.967 | Medium |
| RF | HRV (time) | 93.4% | 94.0% | 92.8% | 0.932 | Low |
| SVM | HRV (time + freq) | 94.8% | 95.2% | 94.0% | 0.945 | Medium |
| GB | Genetic parameters | 91.6% | 92.5% | 90.4% | 0.916 | Low |

Table 7. Results of this Model

The MLP model was thus selected as the primary candidate for wearable deployment, due to its optimal balance between accuracy, computational efficiency, and energy consumption. The final architecture enables real-time filtering of ECG signals, significantly reducing false alarms and optimizing on-device storage.

2.24. Computational Model for Breast Cancer Prevention and Diagnosis

This Model focuses on classifying medical images (mammographics, in particular) for diagnosis of breast cancer.

An initial review was performed to analyse the main shallow and deep learning classifiers employed for medical imaging. This study outlined how the choice of an algorithm must depend on data availability, computational resources, and the required degree of explainability. Traditional models such as SVM, Random Forest, and XGBoost were presented as effective for limited datasets and structured features (e.g., radiomics), while deep learning architectures offer automatic feature extraction from large-scale imaging data (Prinzi, 2024a). This review provided the conceptual foundation for the subsequent research activities.

2.24.1. Technical Insight

Concerning shallow learning methods, combined with an interpretable radiomic feature extraction, a key contribution to early breast cancer diagnosis involved the development of an interpretable radiomic signature for breast microcalcification detection and classification. Using handcrafted radiomic features extracted from mammographic regions of interest, multiple classifiers (SVM, Random Forest, XGBoost) were trained to distinguish between healthy, benign, and malignant tissue.

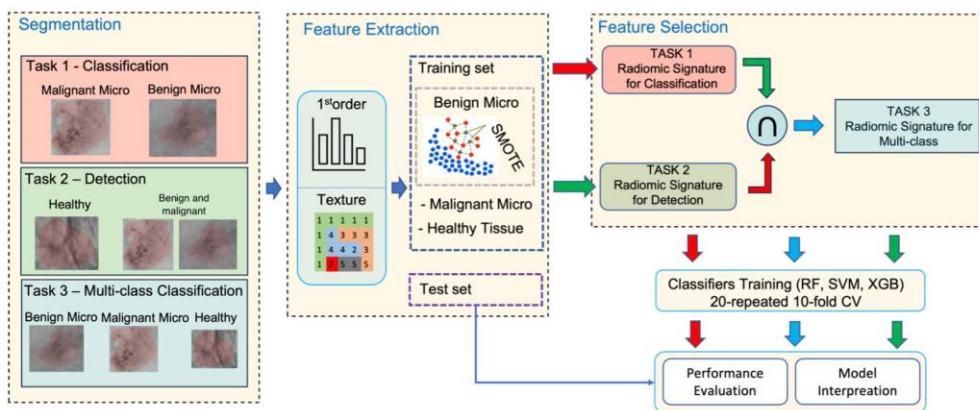


Figure 23: Overall Architecture. The segmented data were divided into healthy tissue and benign and malignant microcalcifications. The same training pipeline was applied for task 1 (malignant vs. benign microcalcifications) and task 2 (healthy tissue vs. microcalcifications). In particular, after the feature extraction process, SMOTE was applied to the benign microcalcification samples for data balancing. Several feature selection steps were employed to select the best signature for tasks 1 and 2. The intersection between the two signatures was used to train a multi-class model, which can simultaneously distinguish healthy tissue, and benign and malignant microcalcifications (task 3). The validation performance were computed using a 20-repeated 10-fold cross-validation strategy. Finally, the performance of the trained models were computed on the test set, and their introspection was performed.

Figure 23 shows the overall architecture. The models achieved AUC values up to 0.876, with feature importance analysis confirming the clinical relevance of descriptors such as GLCM Contrast

and FO Entropy. This work demonstrated how explainable machine learning can bridge the gap between algorithmic prediction and radiological interpretation, supporting both diagnostic accuracy and clinical trust (Prinzi, 2024b).

Regarding still radiomic feature extraction and classification in breast cancer, to exploit the temporal evolution of contrast enhancement in breast MRI, we designed models capable of analyzing entire Dynamic Contrast Enhanced MRI (DCE-MRI) sequences rather than isolated time instants. Two complementary strategies were developed: a Graph Neural Network (MUGI-MRI) that models each temporal instant as a node or layer in a multiplex network, aggregating radiomic information across time (Ceccarelli, 2024); and a multivariate time series classification framework using algorithms such as Rocket, MultiRocket, and Time Series Forest, achieving an AUC of 0.85 and balanced sensitivity/specificity (Prinzi, 2024c). Both approaches confirmed that integrating temporal information markedly improves diagnostic performance compared to static analysis.

The strategy to use shallow approaches combined with radiomic features, address the issue of training with limited data and moreover the interpretability issues. For this reason, to overcome the classic trade-off between model accuracy and interpretability, the Rad4XCNN method was proposed. This framework connects deep CNN representations with interpretable radiomic descriptors, providing quantitative and global explanations of model behavior. Rad4XCNN was tested on a multi-center dataset, showing high predictive accuracy while enabling clinicians to extract semantic insights consistent with known disease patterns. This hybrid approach represents a significant advance toward transparent deep learning models in oncology (Prinzi, 2025).

Ceccarelli F., et al. (2024). MUGI-MRI: Enhancing Breast Cancer Classification through Multiplex Graph Neural Networks in DCE-MRI. <https://doi.org/10.1109/IJCNN60899.2024.10650117>

Prinzi F., et al. (2024a). Shallow and deep learning classifiers in medical image analysis. <https://doi.org/10.1186/s41747-024-00428-2>

Prinzi F., et al. (2024b). Interpretable radiomic signature for breast microcalcification detection and classification. <https://doi.org/10.1007/s10278-024-01012-1>

Prinzi F., et al. (2024c). Breast cancer classification through multivariate radiomic time series analysis in DCE-MRI sequences. <https://doi.org/10.1016/j.eswa.2024.123557>

Prinzi F., et al. (2025). Rad4XCNN: A new agnostic method for post-hoc global explanation of CNN-derived features by means of Radiomics. <https://doi.org/10.1016/j.cmpb.2024.108576>

2.25. Cardiovascular Risk Assessment and Multimodal Data Integration

This Computational Model is still in the earliest stage of development, and targets cardiovascular diseases, focusing on data fusion and multimodal learning. In particular, the development is at its first phase, where the collection of a data set is ongoing. The first version of the dataset has been already released and published in (Prinzi, 2025).

2.25.1. Technical Insight

This research activity has begun with the acquisition and collection of a multimodal dataset to allow the investigation and development of predictive model for digitalized prevention. The **MultiD4CAD** dataset was designed and released as a comprehensive multimodal resource integrating Coronary CT Angiography (CCTA) imaging with clinical biomarkers for coronary artery disease (CAD) analysis. **Figure 24** shows the adopted flow for constructing the dataset, while **Figure 25** shows the clinical characteristics considered in the dataset, and their distribution.

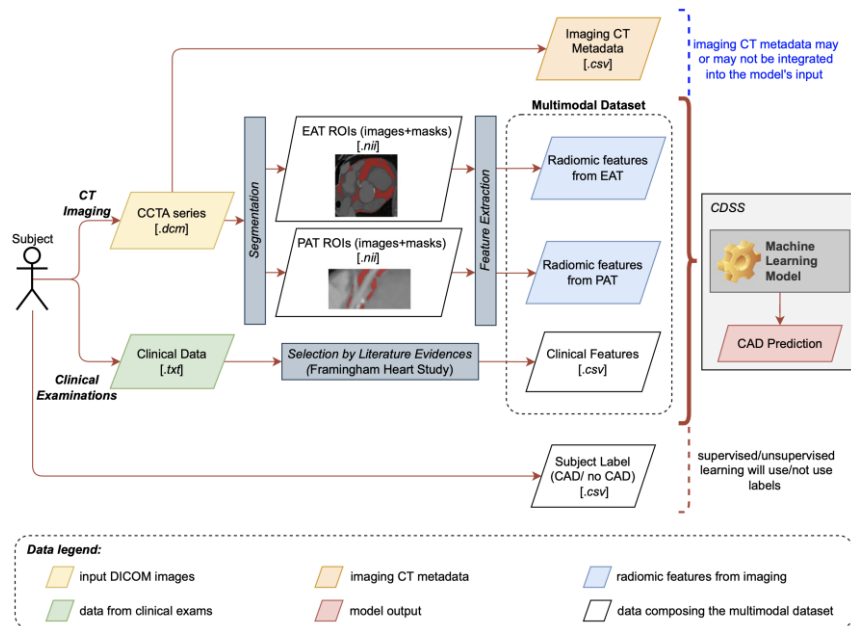


Figure 24: Overall flow diagram showing the processing chain implemented to obtain MultiD4CAD, the multimodal dataset proposed with this study. Acronym specification: CAD: Coronary Artery Disease; CCTA: Coronary Computed Tomography Angiography; CDSS: Clinical Decision Support Systems; CT: Computed Tomography; EAT: Epicardial Adipose Tissue; PAT: Pericoronary adipose tissue; ROI: Region Of Interest. File format specification: csv: comma-separated values; dcm: DICOM; nii: NIfTI; txt: text.

| | with CAD | without CAD | TOTAL (with CAD + without CAD) |
|---|-------------------|-------------------|--------------------------------|
| Samples | 78 | 40 | 118 |
| Women | 22 | 13 | 35 |
| Men | 56 | 27 | 83 |
| Age (mean \pm std.dev) [years] | 53.78 \pm 14.82 | 63.81 \pm 10.97 | 60.40 \pm 12.2 |
| BMI (mean \pm std.dev) [Kg/m ²] | 26.51 \pm 3.96 | 26.39 \pm 3.96 | 26.43 \pm 3.94 |
| Smokers (no/yes) | 40/38 | 19/21 | 59/59 |
| Hypertension (no/yes) | 35/43 | 25/15 | 60/58 |
| Diabetes (no/yes) | 63/15 | 34/6 | 97/21 |
| Familiarity (no/yes) | 35/43 | 22/18 | 57/61 |
| Hypercholesterolaemia (no/yes) | 51/27 | 25/15 | 76/42 |
| Obesity (no/yes) | 67/11 | 35/5 | 102/16 |

Figure 25: Distributions of the clinical characteristics considered in the study.

The imaging component includes validated Epicardial (EAT) and Pericoronary (PAT) adipose tissue segmentations –tissues known to reflect local inflammation and metabolic activity. This dataset will enable the development of supervised learning models for CAD outcome prediction, radiomics-based characterization, and deep learning studies on segmentation or classification.

Prinzi F., et al. (2025). MultiD4CAD: Multimodal dataset composed of ct and clinical features for coronary artery disease analysis. <https://doi.org/10.1038/s41597-025-05743-w>

2.26.A unified computational framework to describe individual dynamics, pairwise interactions, and high-order relationships within multivariate physiological data

This model focuses on information-theoretic methods for the analysis of complex physiological systems to create interpretable and data-efficient tools to quantify interactions among physiological variables. The research is aimed to provide a unified computational framework to describe individual dynamics, pairwise interactions, and higher-order relationships within multivariate physiological data.

This framework combines information-theoretic, data-driven approaches and time-series analysis through four integrated stages—feature extraction, feature selection, feature importance, and classification—transforming physiological signals into compact and discriminative representations.

2.26.1. Technical Insight

During the first stage, the research focused on developing novel information-theoretic measures and applying them to physiological data, forming the basis for subsequent feature-based analyses. Indices derived from the concepts of Granger causality and autonomy were implemented to characterize individual dynamics and causal interactions in physiological signals recorded from patients with coronary artery disease (Saputo, 2024a). In parallel, a non-causal measure of Mutual Information Rate was introduced to quantify the information exchanged between cardiovascular signals over time and applied to ultra-short-term analyses (i.e. on time series shorter than 300 beats) for distinguishing rest, postural, and cognitive stress conditions (Raimondi, 2025). Furthermore, an information-theoretic framework based on Partial Information Decomposition was developed to dissect unique, redundant and synergistic contributions of systolic and diastolic arterial pressure to mean arterial pressure (Sparacino, 2025), complemented by linear and nonlinear beat-to-beat prediction models exploring the same relationships (Saputo, 2025). In addition, a thorough comparison of entropy rate measures for the evaluation of time series complexity provided further insights into how and in what extent different linear model-based and nonlinear model-free estimators can capture the multiscale dynamics of complex physiological systems (Barà, 2024). Together, these developments established a solid foundation for extracting physiologically meaningful and information-based features used in subsequent analyses.

The feature extraction stage focused on deriving quantitative indices from cardiovascular and respiratory time series, computed in time and nonlinear domains, to capture complementary aspects of autonomic and cardiovascular regulation. This methodological framework was applied in (Iovino, 2024), combining feature extraction, feature selection, and classification to discriminate among different types of physiological stress using cardiovascular indices. The subsequent feature selection and feature importance stages aimed to improve the efficiency, robustness, and

interpretability of the models. Feature selection methods based on information theory and dependency analysis were implemented to maximize the relevance of features to the target variable, while minimizing redundancy among them. In parallel, feature importance analyses were designed to quantify how each variable contributes individually and in combination with others, revealing cooperative and high-order effects among physiological indices (i.e., effects involving interactions among three or more variables, beyond simple pairwise relationships). These approaches allowed the identification of compact sets of features with maximal predictive power, provided further insights into the underlying mechanisms of physiological control.

Finally, the selected and weighted features were used as input to machine learning models for the classification of physiological and pathological states. In addition to standard classification methods, used for example in the assessment of driving stress using multimodal physiological data (Fruet, 2025), a novel information-theoretic model called the Local Information Classifier (LIC) was developed and applied to perform a multi-feature classification of physiological stress in cardiovascular and cardiorespiratory interactions (Saputo, 2024b). The LIC provides an information-theoretic approach to classification based on local estimates of probability density and information content. This approach allows the model to adapt to nonlinear structures in physiological data and offers interpretable insights into the information patterns driving classification decisions.

Barà C., et al. (2024). Comparison of entropy rate measures for the evaluation of time series complexity: Simulations and application to heart rate and respiratory variability. DOI: 10.1016/j.bbe.2024.04.004.

Fruet D., et al. (2025). A Signal Normalization Approach for Robust Driving Stress Assessment Using Multi-Domain Physiological Data. DOI: 10.3390/eng6110288.

Iovino M., et al. (2024). Comparison of automatic and physiologically-based feature selection methods for classifying physiological stress using heart rate and pulse rate variability indices. DOI: 10.1088/1361-6579/ad9234.

Raimondi A., et al. (2025). Ultra-Short-Term Analysis of Cardiovascular Interactions at Rest and During Stress Conditions Using Mutual Information Rate. Proceedings of the National Congress of Bioengineering, Palermo, Italy, 2025.

Saputo R., et al. (2024a). Assessment of Cerebrovascular Interactions and Control in Coronary Artery Disease Patients Undergoing Anaesthesia through Bivariate Predictability Measures", 2024 Medical & Biological Engineering & Computing – accepted

Saputo R., et al. (2024b). Multi-Feature Classification of Physiological Stress in Cardiovascular and Cardiorespiratory Interactions. DOI: 10.1109/ESGCO63003.2024.10767062.

Saputo R., et al. (2025). The Effect of Systolic and Diastolic Arterial Pressures on Mean Arterial Pressure: Linear and Non-Linear Prediction Approaches”, Proceedings of the National Congress of Bioengineering, Palermo, Italy, 2025

Sparacino L. et al. (2025). Investigating the Effect of Systolic and Diastolic Arterial Pressures on Mean Pressure through Partial Information Decomposition. IEEE Engineering in Medicine and Biology Conference, July 2025.

2.27. XAI for Histopathological Images

The timely and accurate analysis of histopathological samples is a foundation for modern clinical practice, particularly in oncology, where it provides the gold standard for diagnosis and prognosis. This diagnostic precision is fundamental to preventive medicine and early intervention strategies.

A substantial research effort was directed toward addressing the "black box" problem in deep learning, which remains a critical barrier to the clinical adoption of advanced AI as a trusted decision support tool. In high-stakes fields like histopathology, a simple classification output (e.g., "malignant") is insufficient without a transparent, verifiable motivation.

This research in Explainable AI (XAI) directly confronts this challenge, by developing an ensemble of three interconnected sub-models whose reasoning can be inspected and understood by a human expert.

2.27.1. Technical Insight

The three sub-models have been developed and published in almost a sequential manner. They will be introduced here following such order.

The first study and resulting model systematically explored metric learning techniques as a method to improve model transparency. This approach moves beyond simple classification to provide interpretable insights based on the similarity between tissue samples. The proposed model achieved a Patient Level Accuracy (PLA) of 88.90% in Magnification Independent Binary (MIB) classification on the BreakHis dataset, outperforming state-of-the-art methods and demonstrating that, in this case, interpretability does not require a trade-off in performance (Amato, 2023).

A second model has been built exploiting the previous one, where a subsequent study introduced a novel framework using Self-Organising Maps (SOMs) from a granular computing perspective. Although the classification results don't exceed the current state-of-the-art, the proposed granular method allows for the precise identification of malignant patches within the tissue samples. This capability significantly increases the explanation capacity of the system, establishing a valuable trade-off between the quality of the explanation, providing context-aware insights, and the model's performance (Amato, 2024).

The third model applied the results above through the implementation of Siamese and Triplet networks and the use of a standard classifier, e.g. SVM. This work demonstrates how specialised network embeddings can be engineered to optimise classification based on learned similarity metrics. The proposed Siamese Network coupled with an SVM classifier achieved remarkable results across multiple tasks, reaching for example an accuracy of about 94% on the multiclass Kather dataset and about 95% on the PatholDCG grading dataset. Furthermore, the introduction

of a Confidence Score allowed the system to filter uncertain predictions, significantly increasing reliability for clinical decision support (Amato, 2025).

| Method | Accuracy | Method | Accuracy |
|-----------------------------------|---------------------|-----------------------------------|---------------------|
| Kather et al. [21] | 87.4 | Calderaro et al. [35] | 96.67 ± 0.98 |
| Yazdi et al. [26] | 93.38 | Yan, R., Yang, Z. et al. [38] | 91.6 |
| Rizalputri et al. [27] | 82.2 | Yan, R., Ren, F. et al. [36] | 93.4 |
| Zeid et al. [30] | 94.73 | ResNet152 + SVM (ablat.) | 53.43 ± 3.97 |
| Ohata et al. [29] | 92.8 | ResNet152 + <i>k</i> -NN (ablat.) | 63.52 ± 0.74 |
| Cascianelli et al. [28] | 84.00 | SNN + SVM | 95.17 ± 0.71 |
| ResNet152 + SVM (ablat.) | 79.46 ± 0.77 | TNN + SVM | 91.21 ± 0.40 |
| ResNet152 + <i>k</i> -NN (ablat.) | 56.54 ± 1.92 | SNN + <i>k</i>-NN | 95.19 ± 0.64 |
| SNN + SVM | 94.73 ± 0.85 | TNN + <i>k</i>-NN | 91.24 ± 0.51 |
| TNN + SVM | 94.37 ± 0.98 | | |
| SNN + <i>k</i>-NN | 84.48 ± 1.05 | | |
| TNN + <i>k</i>-NN | 94.12 ± 1.13 | | |

Figure 26: Comparisons over the Kather Dataset (left) and the PathoIDCG dataset (right). Picture taken from (Amato, 2025).

Amato D., et al. (2023). Metric Learning in Histopathological Image Classification: Opening the Black Box. <https://doi.org/10.3390/s23136003>

Amato D., et al. (2024). Explainable Histopathology Image Classification with Self-organizing Maps: A Granular Computing Perspective. <https://doi.org/10.1007/s12559-024-10312-1>

Amato D., et al. (2025). Classification of Histopathology Images by Siamese and Triplet Network Embeddings. <https://doi.org/10.1109/ACCESS.2025.3613448>

2.28. Semantic Segmentation of gliomas on brain MRIs

This model focuses on the application of advanced computational techniques to neuroimaging, focusing on glioma and glioblastoma. In particular, the goal of the model is the automatic semantic segmentation of gliomas on MRI brain images

2.28.1. Technical Insight

The model, published in the study (Amato, 2024) details the development of Graph Convolutional Neural Networks (GCNNs) for the complex task of semantic segmentation of gliomas on brain MRIs. By representing the MRI scans as graphs of superpixels to capture spatial relationships, this method achieved a Dice Coefficient of 0.67 and a Pixel Error of 0.01. These results significantly outperformed traditional U-Net-based approaches (which achieved a Dice score of about 0.53), demonstrating that graph-based segmentation offers superior precision in delineating tumor boundaries.

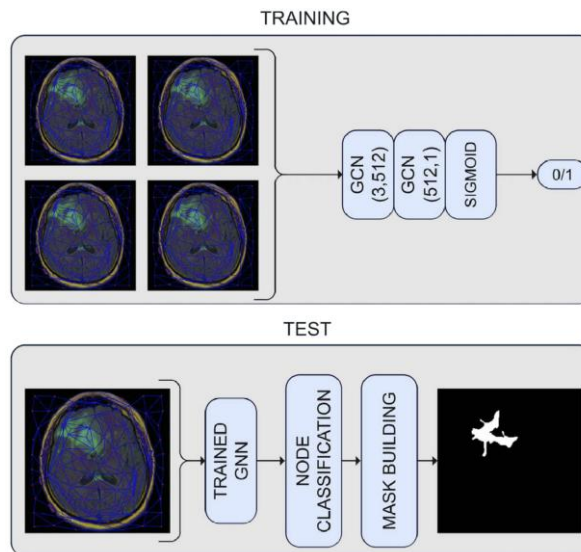


Figure 27: Architecture proposed in this approach

Figure 27 shows the architecture of the Graph Convolutional Neural Networks adopted in this approach, while in **Figure 28** three examples of segmentations obtained by the GCNNs are shown. Finally, **Table 6** shows the results w.r.t. state-of-the-art approaches.

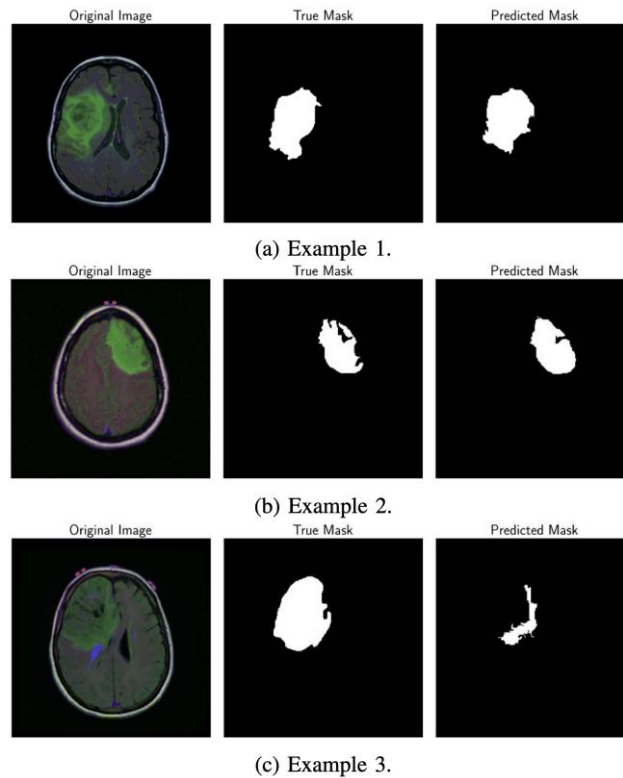


Figure 28: Examples of segmentation. The left column shows the MRI image, the center column shows the mask available in the annotated dataset, and the right column shows the mask predicted by the GCNN-base approach.

Table 6: The obtained results. \uparrow signifies higher is better, vice versa \downarrow lower is better.

| Method/Metric | DC \uparrow | PE \downarrow | RE \downarrow |
|--------------------------|---------------------|---------------------|---------------------|
| U-Net Pred. | 0.5279 \pm 0.0258 | 0.0126 \pm 0.0017 | 0.0057 \pm 0.0008 |
| U-Net FT | 0.4154 \pm 0.0917 | 0.0188 \pm 0.0027 | 0.0165 \pm 0.0035 |
| U-Net Train. | 0.3093 \pm 0.1231 | 0.0406 \pm 0.0215 | 0.0559 \pm 0.0358 |
| This Computational Model | 0.6743 \pm 0.0329 | 0.0107 \pm 0.0029 | 0.0200 \pm 0.0050 |

Amato D., et al. (2024). Semantic Segmentation of Gliomas on Brain MRIs by Graph Convolutional Neural Networks. <https://doi.org/10.1109/AIxDKE63520.2024.00036>

2.29. Glioblastoma Treatment Response Classification

This model, currently under development, focuses on therapeutic outcomes classification according to RANO criteria. The model, currently under review in (Amato, 2025) exploits a sophisticated hybrid model that integrates deep learning architectures with handcrafted radiomic features.

2.29.1. Technical Insight

Through a rigorous grid search methodology, the study (Amato, 2025) investigates the optimal input configuration for the model, providing precise indications to domain experts regarding the most effective combinations of MRI sequences and feature aggregation strategies. The model demonstrated robust performance on the LUMIERE dataset:

- In the binary classification task (Progressive vs. Non-Progressive Disease), the proposed method achieved a Balanced Accuracy of 0.71, surpassing the baseline established by Matoso et al. (0.49) and highlighting the model's ability to handle class imbalance effectively.
- In the challenging multiclass scenario (Complete Response, Partial Response, Stable Disease, Progressive Disease), the approach reached an F1-Score of 0.616 ± 0.048 , drastically improving upon the baseline (0.091).

This indicates that integrating radiomic features provides critical complementary information, making the model a valuable tool for prognostic and clinical decision support.

Amato D., et al. (2025). Integrating Deep Learning and Radiomic Features for Glioblastoma Treatment Response Classification. Under Review.

2.30. In Silico Trials to reduce the risk of hip fracture in fragile elders

Over the past few years, the group led by Prof. Viceconti at UNIBO has developed and validated a digital twin in healthcare called BBCT-Hip. Starting from a CT scan of the subject, BBCT can predict the risk of hip fracture upon falling with a stratification accuracy that is 10-15% better than the current standard of care, Areal Bone Mineral Density (aBMD) obtained from Dual X-ray Absorptiometry (DXA) [Aldieri 2022]. The CT data are converted into a biophysical finite element model that can predict the impact force required to fracture the subject's femur.

As we utilised BBCT across various clinical cohorts in the UK and Italy, we accumulated sufficient data to develop a statistical population model of the inputs. In particular, from the morpho-

densitometric information derived from the CT scans, we developed a statistical atlas [La Mattina, 2023]. Using this statistical model, we can generate thousands of input sets for the BBCT digital twin, each representing a virtual patient. These virtual cohorts, which can be designed to present statistical distributions of age, weight, height, gender, femoral shape and bone densitometry, according to predefined targets, are then analysed with BBCT to establish their baseline risk of hip fracture.

Using longitudinal observational data on the progression of aBMD in untreated subjects, we then developed a phenomenological population model of disease progression, which enabled the prediction of how the risk of hip fracture evolved over time (typically five to ten years) [Savelli 2024].

Lastly, we enhanced the BoneStrength In Silico Trial by incorporating the ability to simulate the effects of some of the most common interventions aimed at reducing the risk of hip fracture, including alendronates [Oliviero 2025] and hip protectors [Oliviero 2024].

The resulting stochastic model is sampled with multiple realisations (10-20) using a stochastic Markov chain process, to obtain convergence of predicted hip fracture incidence. To simulate the equivalent of a Phase III clinical trial with approximately 1,000 virtual patients per cohort, each realisation requires 2,000-7,000 Finite Element Analysis simulations, depending on the follow-up duration. In the frame of the PNRR CN1 National Centre for Supercomputing, we successfully ported and optimised the execution of BoneStrength on the Leonardo HPC system available at the CINECA supercomputing centre.

In the DARE project, the BoneStrength in silico trial was further developed, particularly with regard to fall modelling and its description in terms of the subject's frailty. This produced one publication:

Savelli G., Oliviero S., Viceconti M., La Mattina A.A. In silico prediction of hip fractures: improved fall modeling and expanded validation across cohorts with diverse risk profiles. (2025) Journal of the Mechanical Behavior of Biomedical Materials, 172, art. no. 107182. DOI: 10.1016/j.jmbbm.2025.107182.

Furthermore, we used this improved model to explore how the combination of pharmacological treatment and rehabilitation programs aimed to reduce the risk of falling could reduce the risk of hip fracture in a cohort of frail women. This second result was presented at a conference this summer:

Savelli G., Oliviero S., Viceconti M. In silico trial to assess the efficacy of intervention strategies for the prevention of hip fractures. Presented at the 30th Congress of the European Society of Biomechanics, 6 – 9 July 2025, Zürich, Switzerland.

A manuscript, “*In Silico Clinical Trial* assessing interventions for hip-fracture prevention”, is currently being written and should be submitted for publication in 2026.

2.30.1. Bibliography

Aldieri A, Terzini M, Audenino AL, Bignardi C, Paggiosi M, Eastell R, Viceconti M, Bhattacharya P. Personalised 3D Assessment of Trochanteric Soft Tissues Improves HIP Fracture Classification Accuracy. *Ann Biomed Eng.* 2022 Mar;50(3):303-313. doi: 10.1007/s10439-022-02924-1.

La Mattina AA, Baruffaldi F, Taylor M, and Viceconti M, 'Statistical Properties of a Virtual Cohort for In Silico Trials Generated with a Statistical Anatomy Atlas', *Ann Biomed Eng.*, vol. 51, no. 1, pp. 117–124, Jan. 2023, doi: 10.1007/s10439-022-03050-8.

Oliviero S, La Mattina AA, Savelli G, and Viceconti M, 'In Silico clinical trial to predict the efficacy of hip protectors for preventing hip fractures', *Journal of Biomechanics*, vol. 176, p. 112335, Nov. 2024, doi: 10.1016/j.jbiomech.2024.112335.

Oliviero S, Savelli G, Viceconti M, La Mattina AA, 'In silico clinical trial to predict the efficacy of alendronate for preventing hip fractures', *Clinical Biomechanics*, Oct. 2025:106689. doi: 10.1016/j.clinbiomech.2025.106689.

Savelli G, Oliviero S, La Mattina AA, and Viceconti M, 'In Silico Clinical Trial for Osteoporosis Treatments to Prevent Hip Fractures: Simulation of the Placebo Arm', *Ann Biomed Eng.*, Nov. 2024, doi: 10.1007/s10439-024-03636-4.

3. Bibliography of DARE-related publications

- [1] Comparison of automatic and physiologically-based feature selection methods for classifying physiological stress using heart rate and pulse rate variability indices. Iovino, M.; Lazić, I.; Loncar-Turukalo, T.; Javorka, M.; Pernice, R.; and Faes, L. *Physiological Measurement*, 45(11). 2024.
- [2] Comparison of entropy rate measures for the evaluation of time series complexity: Simulations and application to heart rate and respiratory variability. Barà, C.; Pernice, R.; Catania, C.; Hilal, M.; Porta, A.; Humeau-Heurtier, A.; and Faes, L. *Biocybernetics and Biomedical Engineering*, 44(2): 380–392. 2024. Number: 2 Publisher: Elsevier B.V.
- [3] Metric Learning in Histopathological Image Classification: Opening the Black Box. Amato, D.; Calderaro, S.; Lo Bosco, G.; Rizzo, R.; and Vella, F. *Sensors*, 23(13): 6003. June 2023.
- [4] Explainable Histopathology Image Classification with Self-organizing Maps: A Granular Computing Perspective. Amato, D.; Calderaro, S.; Lo Bosco, G.; Rizzo, R.; and Vella, F. *Cognitive Computation*. June 2024.
- [5] MUGI-MRI: Enhancing Breast Cancer Classification through Multiplex Graph Neural Networks in DCE-MRI. Ceccarelli, F.; Prinzi, F.; Liò, P.; Vitabile, S.; and Holden, S. *Proceedings of the International Joint Conference on Neural Networks*. 2024.
- [6] Breast cancer classification through multivariate radiomic time series analysis in DCE-MRI sequences. Prinzi, F.; Orlando, A.; Gaglio, S.; and Vitabile, S. *Expert Systems with Applications*, 249. 2024. Publisher: Elsevier Ltd
- [7] Shallow and deep learning classifiers in medical image analysis. Prinzi, F.; Currieri, T.; Gaglio, S.; and Vitabile, S. *European Radiology Experimental*, 8(1). 2024. Number: 1 Publisher: Springer Science and Business Media Deutschland GmbH
- [8] Rad4XCNN: A new agnostic method for post-hoc global explanation of CNN-derived features by means of Radiomics. Prinzi, F.; Militello, C.; Zarcaro, C.; Bartolotta, T. V.; Gaglio, S.; and Vitabile, S. *Computer Methods and Programs in Biomedicine*, 260: 108576. March 2025.
- [9] Assessing the Use of AutoML for Data-Driven Software Engineering. Calefato, F.; Quaranta, L.; Lanubile, F.; and Kalinowski, M. In *2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–12, New Orleans, LA, USA, 2023. IEEE
- [10] An MLOps Solution Framework for Transitioning Machine Learning Models into eHealth Systems. Basile, A.; Calefato, F.; Lanubile, F.; Mallardi, G.; and Quaranta, L. In *Napoli, Italia, May 2024*.
- [11] A Lot of Talk and a Badge: An Exploratory Analysis of Personal Achievements in GitHub. Calefato, F.; Quaranta, L.; and Lanubile, F. *Information and Software Technology*, 176: 107561. December 2024

- [12] Professional Insights into Benefits and Limitations of Implementing MLOps Principles. Araujo, G.; Kalinowski, M.; Endler, M.; and Calefato, F. In volume 2, pages 305–312, 2024.
- [13] An MLOps Approach for Deploying Machine Learning Models in Healthcare Systems. Mallardi, G.; Calefato, F.; Quaranta, L.; and Lanubile, F. In 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 6832–6837, Lisbon, Portugal, December 2024.
- [14] Towards Ensuring Responsible AI for Medical Device Certification. Mallardi, G.; Quaranta, L.; Calefato, F.; and Lanubile, F. In 2025 IEEE/ACM International Workshop on Responsible AI Engineering (RAIE), pages 29–32, Ottawa, ON, Canada, April 2025.
- [15] Explainable Gait Analysis for Early Detection of Neurodegenerative Diseases Using Unsupervised Clustering Techniques. Dentamaro, V.; Franchini, F.; Massaro, I.; Musti, L.; Pirlo, G.; and Sblendorio, E. In 2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRINE), pages 861–866, St Albans, United Kingdom, October 2024.
- [16] Enhancing early Parkinson's disease detection through multimodal deep learning and explainable AI: insights from the PPMI database. Dentamaro, V.; Impedovo, D.; Musti, L.; Pirlo, G.; and Taurisano, P. *Scientific Reports*, 14(1): 20941. September 2024. Publisher: Nature Publishing Group
- [17] Impact of data quality for automatic issue classification using pre-trained language models. Colavito, G.; Lanubile, F.; Novielli, N.; and Quaranta, L. *Journal of Systems and Software*, 210. 2024. Publisher: Elsevier Inc.
- [18] Leveraging GPT-like LLMs to Automate Issue Labeling. Colavito, G.; Lanubile, F.; Novielli, N.; and Quaranta, L. In *Proc. - IEEE/ACM Int. Conf. Min. Softw. Repos., MSR*, pages 469–480, 2024. Institute of Electrical and Electronics Engineers Inc. Journal Abbreviation: *Proc. - IEEE/ACM Int. Conf. Min. Softw. Repos., MSR*
- [19] Training Future Machine Learning Engineers: A Project-Based Course on MLOps. Lanubile, F.; Martinez-Fernandez, S.; and Quaranta, L. *IEEE Software*, 41(2): 60–67. 2024. Number: 2 Publisher: IEEE Computer Society
- [20] Position paper: Extending Credibility Assessment of In Silico Medicine Predictors to Machine Learning Predictors. Viceconti, M.; Lanubile, F.; Carbonaro, A.; Mellone, S.; Curreli, C.; Aldieri, A.; Ranciati, S.; and Montanari, A. *IEEE Journal of Biomedical and Health Informatics*, 1–9. 2025.
- [21] Towards Ensuring Responsible AI for Medical Device Certification. Mallardi, G.; Quaranta, L.; Calefato, F.; and Lanubile, F. In 2025 IEEE/ACM International Workshop on Responsible AI Engineering (RAIE), pages 29–32, Ottawa, ON, Canada, April 2025.
- [22] A Cluster-Based Approach for Emotion Recognition in Software Development. Grassi, D.; Lanubile, F.; Motca-Schnabel, A.; and Novielli, N. In 2025 IEEE/ACM 18th International Conference on Cooperative and Human Aspects of Software Engineering (CHASE), pages 239–247, Ottawa, ON, Canada, April 2025. IEEE

- [23] Self-monitoring of Developers' Emotions: the Case of Agile Retrospective Meetings. Grassi, D.; Lanubile, F.; Novielli, N.; Quaranta, L.; and Serebrenik, A. ACM Transactions on Software Engineering and Methodology. September 2025. Place: New York, NY, USA Publisher: Association for Computing Machinery
- [24] Data Science for Health Image Alignment: A User-Friendly Open-Source ImageJ/Fiji Plugin for Aligning Multimodality/Immunohistochemistry/Immunofluorescence 2D Microscopy Images. Piccinini, F.; Tazzari, M.; Tumedei, M.; Stellato, M.; Remondini, D.; Giampieri, E.; Martinelli, G.; Castellani, G.; and Carbonaro, A. Sensors, 24(2). 2024. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute (MDPI)
- [25] Comparative Evaluation of Commercial, Freely Available, and Open-Source Tools for Single-Cell Analysis Within Freehand-Defined Histological Brightfield Image Regions of Interest. Piccinini, F.; Tazzari, M.; Tumedei, M. M.; Normanno, N.; Castellani, G.; and Carbonaro, A. Technologies, 13(3): 110. March 2025.
- [26] EEG Features Learned by Convolutional Neural Networks Reflect Alterations of Social Stimuli Processing in Autism. Borra, D.; Diciotti, S.; and Magosso, E. In volume 15019 LNCS, pages 124–136, 2024.
- [27] Umbrella Review of Systematic Reviews and Meta-Analyses on Consumption of Different Food Groups and Risk of Type 2 Diabetes Mellitus and Metabolic Syndrome. Banjarnahor, R.; Javadi Arjmand, E.; Onni, A.; Thomassen, L.; Perillo, M.; Balakrishna, R.; Sletten, I.; Lorenzini, A.; Plastina, P.; and Fadnes, L. Journal of Nutrition, 155(5): 1285–1297. 2025.
- [28] Umbrella Review of Systematic Reviews and Meta-Analyses on the Consumption of Different Food Groups and the Risk of Overweight and Obesity. Kristoffersen, E.; Hjort, S. L.; Thomassen, L. M.; Arjmand, E. J.; Perillo, M.; Balakrishna, R.; Onni, A. T.; Sletten, I. S. K.; Lorenzini, A.; and Fadnes, L. T. Nutrients, 17(4): 662. February 2025.
- [29] Umbrella Review of Systematic Reviews and Meta-analyses on Consumption of Different Food Groups and Risk of All-cause Mortality. Onni, A. T.; Balakrishna, R.; Perillo, M.; Amato, M.; Javadi Arjmand, E.; Thomassen, L. M.; Lorenzini, A.; and Fadnes, L. T. Advances in Nutrition, 16(4): 100393. April 2025.
- [30] Efficient text-image semantic search: A multi-modal vision-language approach for fashion retrieval. Moro, G.; Salvatori, S.; and Frisoni, G. Neurocomputing, 538. 2023. Publisher: Elsevier B.V.
- [31] Retrieve-and-Rank End-to-End Summarization of Biomedical Studies. Moro, G.; Ragazzi, L.; Valgimigli, L.; and Molfetta, L. In Pedreira O.; and Estivill-Castro V., editor(s), Lect. Notes Comput. Sci., volume 14289 LNCS, pages 64–78, 2023. Springer Science and Business Media Deutschland GmbH Journal Abbreviation: Lect. Notes Comput. Sci.
- [32] Carburacy: Summarization Models Tuning and Comparison in Eco-Sustainable Regimes with a Novel Carbon-Aware Accuracy. Moro, G.; Ragazzi, L.; and Valgimigli, L. In Williams B.;

- Chen Y.; and Neville J., editor(s), Proc. AAAI Conf. Artif. Intell., AAAI, volume 37, pages 14417–14425, 2023. AAAI Press Journal Abbreviation: Proc. AAAI Conf. Artif. Intell., AAAI
- [33] Graph-Based Abstractive Summarization of Extracted Essential Knowledge for Low-Resource Scenarios. Moro, G.; Ragazzi, L.; and Valgimigli, L. In Gal K.; Gal K.; Nowe A.; Nalepa G.J.; Fairstein R.; and Radulescu R., editor(s), Front. Artif. Intell. Appl., volume 372, pages 1747–1754, 2023. IOS Press BV Journal Abbreviation: Front. Artif. Intell. Appl
- [34] Align-then-abstract representation learning for low-resource summarization. Moro, G.; and Ragazzi, L. Neurocomputing, 548. 2023. Publisher: Elsevier B.V
- [35] Cogito Ergo Summ: Abstractive Summarization of Biomedical Papers via Semantic Parsing Graphs and Consistency Rewards. Frisoni, G.; Italiani, P.; Salvatori, S.; and Moro, G. In Williams B.; Chen Y.; and Neville J., editor(s), Proc. AAAI Conf. Artif. Intell., AAAI, volume 37, pages 12781–12789, 2023. AAAI Press Journal Abbreviation: Proc. AAAI Conf. Artif. Intell., AAAI
- [36] LAWSUIT: a LArge expert-Written SUMmarization dataset of ITalian constitutional court verdicts. Ragazzi, L.; Moro, G.; Guidi, S.; and Frisoni, G. Artificial Intelligence and Law. 2024.
- [37] To Generate or to Retrieve? On the Effectiveness of Artificial Contexts for Medical Open-Domain Question Answering. Frisoni, G.; Cocchieri, A.; Presepi, A.; Moro, G.; and Meng, Z. In volume 1, pages 9878–9919, 2024.
- [38] What Are You Token About? Differentiable Perturbed Top-k Token Selection for Scientific Document Summarization. Ragazzi, L.; Italiani, P.; Moro, G.; and Panni, M. In pages 9427–9440, 2024.
- [39] Evidence, my Dear Watson: Abstractive dialogue summarization on learnable relevant utterances. Italiani, P.; Frisoni, G.; Moro, G.; Carbonaro, A.; and Sartori, C. Neurocomputing, 572. 2024. Publisher: Elsevier B.V.
- [40] Enhancing legal question answering with data generation and knowledge distillation from large language models. Italiani, P.; Moro, G.; and Ragazzi, L. Artificial Intelligence and Law. July 2025.
- [41] A method for the synchronization of inertial sensor signals and local field potentials from deep brain stimulation systems. D'Ascanio, I.; Giannini, G.; Baldelli, L.; Cani, I.; Giannoni, A.; Leogrande, G.; Lopane, G.; Calandra-Buonaura, G.; Cortelli, P.; Chiari, L.; and Palmerini, L. Biomedical Physics and Engineering Express, 10(5). 2024. Number: 5 Publisher: Institute of Physics
- [42] Evaluating gait and postural responses to subthalamic stimulation and levodopa: A prospective study using wearable technology. Cani, I.; D'Ascanio, I.; Baldelli, L.; Lopane, G.; Ranciati, S.; Mantovani, P.; Conti, A.; Cortelli, P.; Calandra-Buonaura, G.; Chiari, L.; Palmerini, L.; and Giannini, G. European Journal of Neurology, 32(1): e16580. January 2025.
- [43] The functional roles of S-adenosyl-methionine and S-adenosyl-homocysteine and their involvement in trisomy 21. Caracausi, M.; Ramacieri, G.; Catapano, F.; Cicilloni, M.; Lajin, B.;

- Pelleri, M. C.; Piovesan, A.; Vitale, L.; Locatelli, C.; Pirazzoli, G. L.; Strippoli, P.; Antonaros, F.; and Vione, B. *BioFactors*, 50(4): 709–724. July 2024. Number: 4
- [44] Zinc metabolism and its role in immunity status in subjects with trisomy 21: chromosomal dosage effect. Ramacieri, G.; Locatelli, C.; Semprini, M.; Pelleri, M. C.; Caracausi, M.; Piovesan, A.; Cicilloni, M.; Vigna, M.; Vitale, L.; Sperti, G.; Corvaglia, L. T.; Pirazzoli, G. L.; Strippoli, P.; Catapano, F.; Vione, B.; and Antonaros, F. *Frontiers in Immunology*, 15: 1362501. April 2024.
- [45] Metabolic and genetic imbalance of the homocysteine-methionine cycle in trisomy 21. Vione, B.; Lajin, B.; Antonaros, F.; Cicilloni, M.; Catapano, F.; Locatelli, C.; Pelleri, M. C.; Piovesan, A.; Vitale, L.; Pirazzoli, G. L.; Strippoli, P.; Corvaglia, L. T.; Ramacieri, G.; and Caracausi, M. *Physiological Genomics*, 57(9): 566–574. September 2025.
- [46] Comparison of Machine Learning Algorithms for Heartbeat Detection Based on Accelerometric Signals Produced by a Smart Bed. Hoang, M.; Matrella, G.; and Ciampolini, P. *Sensors*, 24(6). 2024. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute (MDPI)
- [47] Artificial Intelligence Implementation in Internet of Things Embedded System for Real-Time Person Presence in Bed Detection and Sleep Behaviour Monitor. Hoang, M.; Matrella, G.; and Ciampolini, P. *Electronics (Switzerland)*, 13(11). 2024.
- [48] Metrological evaluation of contactless sleep position recognition using an accelerometric smart bed and machine learning. Hoang, M. L.; Matrella, G.; and Ciampolini, P. *Sensors and Actuators A: Physical*, 385: 116309. April 2025.
- [49] Computational Genomics platform: a Cloud-enabled approach. Gasparetto, J.; Magenta, L.; Sinisi, F.; Zotti, S.; Costantini, A.; Martelli, B.; Giangregorio, T.; and Pippucci, T. *Proceedings of Science*, volume 458, 2024.
- [50] A LLMOps-Driven Framework for Clinical Data Harmonization, Marfoggia, Alberto; Robustelli, Antonio; D'Errico, Christian; Mellone, Sabato; Carbonaro, Antonella, in: *Machine Learning Operations 2025*, Aachen, Ceur Workshop proceedings, 2025, pp. 25 - 36
- [51] Comparative Evaluation of Commercial, Freely Available, and Open-Source Tools for Single-Cell Analysis Within Freehand-Defined Histological Brightfield Image Regions of Interest, Piccinini, Filippo; Tazzari, Marcella; Tumedei, M.; Normanno, Nicola; Castellani, Gastone; Carbonaro, Antonella, *TECHNOLOGIES*, 2025, 13, Article number: 110, pp. 1 - 19
- [52] CONNECTED: A Knowledge Graph-Driven Platform for Clinical Data Harmonization and Personalized Digital Twin-Based Healthcare, Marfoggia, Alberto; D'Errico, Christian; Nardini, Filippo; Mellone, Sabato; Carbonaro, Antonella, in: *2025 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events, PerCom Workshops 2025*, Piscataway, IEEE, *PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON PERVASIVE COMPUTING AND COMMUNICATIONS*, 2025, pp. 116 - 121

- [53] Editorial: Implementing digital twins in healthcare: pathways to person-centric solutions, Carbonaro, Antonella; Marfoggia, Alberto; Quaranta, Luigi; Mellone, Sabato; Lanubile, Filippo, FRONTIERS IN DIGITAL HEALTH, 2025, 7, pp. 1 - 3
- [54] Feasibility of MLOps-based healthcare pipelines in ensuring the Cybersecurity Framework, Robustelli, Antonio; Marfoggia, Alberto; D'Errico, Christian; Mellone, Sabato; Carbonaro, Antonella, in: MLOps25: Workshop on Machine Learning Operations 2025, Aachen, Ceur Workshop proceedings, 2025, pp. 1 - 12
- [55] From raw data to research-ready: A FHIR-based transformation pipeline in a real-world oncology setting, Carbonaro, Antonella; Giorgetti, Luca; Ridolfi, Lorenzo; Pasolini, Roberto; Pagliarani, Andrea; Cavallucci, Martina; Andalò, Alice; Del Gaudio, Livia; De Angelis, Paolo; Vespignani, Roberto; Gentili, Nicola, COMPUTERS IN BIOLOGY AND MEDICINE, 2025, 197, Article number: 111051, pp. 1 - 14
- [56] Towards real-world clinical data standardization: A modular FHIR-driven transformation pipeline to enhance semantic interoperability in healthcare, Marfoggia, Alberto; Nardini, Filippo; Arcobelli, VALERIO ANTONIO; Moscato, Serena; Mellone, Sabato; Carbonaro, Antonella, COMPUTERS IN BIOLOGY AND MEDICINE, 2025, 187, Article number: 109745, pp. 1 - 11
- [57] FHIR-standardized data collection on the clinical rehabilitation pathway of trans-femoral amputation patients, Arcobelli, Valerio Antonio; Moscato, Serena; Palumbo, Pierpaolo; Marfoggia, Alberto; Nardini, Filippo; Randi, Pericle; Davalli, Angelo; Carbonaro, Antonella; Chiari, Lorenzo; Mellone, Sabato, SCIENTIFIC DATA, 2024, 11, Article number: 806, pp. 1 - 14
- [58] Representation of Machine Learning Models to Enhance Simulation Capabilities Within Digital Twins in Personalized Healthcare, Alberto Marfoggia, Filippo Nardini, Sabato Mellone, Antonella Carbonaro, in 2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events, PerCom Workshops 2024, IEEE, 2024, pp. 1 - 7