TUAI MSCA DOCTORAL NETWORK · DC3 · UNIVERSITY OF OVIEDO

# Reliable & Explainable AI for Chronic Diseases

Validating that explanations are stable, faithful and trustworthy, not just accurate

**First step towards the TUAI mid-term report**




**Milad Jafari Barani**

Doctoral Candidate (DC3) · Supervisor: Prof. Vicente García Díaz


TUAI Summer School on Sensor Fusion — Gliwice · 22 July 2026


Funded by the European Union

Universidad de Oviedo

# DC3 scope & objectives

**As described in the Grant Agreement (GA No. 101168344) and my Career Development Plan. This defines the scope against which today's progress is measured.**

### Trustworthy Health Informatics
Interpretable & explainable AI: associate cause to effect, justify results.

### Dimensionality reduction & feature selection
The GA's “first and necessary condition” — automatic feature selection and extraction.

### Chronic disease · early intervention
Onset and progression of diabetes, heart disease, Alzheimer's.

### Experiment with different sets of data
GA expected result: test across health-informatics datasets and XAI technologies.

**Summery Of My Project Details**

***“Reliable and Explainable AI solutions for chronic diseases”***

Work package
**WP2 · Advanced Time Series Analysis through Explainable AI**

Applied partners
**BioKeralty · TheNextPangea**

Secondments (GA)
**UPM · NTNU · SUT (4 months each)**

# Why accuracy is not enough

## Accuracy-only model

*A high-scoring “black box”*

- Reports a number, hides the reasoning
- Explanation can flip under tiny input noise
- Clinicians have no rationale to act on

## Trustworthy explanation

*Stable, Faithful, Formally validated*

- Shows which features drive the decision
- Explanation stays consistent under noise
- Explanations are formally tested, not assumed

**Central research question**

**“ Can explanations stay stable and faithful under realistic clinical conditions? ”**

# Domain independent methodology

**A domain-independent methodology**
for building and validating **reliable AI under noise**

## Temporal modelling — T1

Capture disease progression from longitudinal sequences

## Explanation stability — T2

Does the reasoning survive input perturbation?

## Systematic validation — T3

Sanity checks vs faithfulness tested, not assumed

**Main Domain - Clinical (WP2: DC3 core)**
Alzheimer's diagnosis - 3 studies

**Workshop· AGV perception (WP3: network)**
**Shared principles**: reliability & systematic validation - 2 papers

# Modelling disease progression over time

Alzheimer's is progressive: a sequence of visits carries more diagnostic signal than any single cross-sectional assessment. The architecture matured across the two clinical studies.

STUDY 1

## Attention-based LSTM

- Temporal attention over per-subject clinical sequences
- Masking layer; sequences zero-padded to a fixed horizon
- Strict subject-level separation

→

STUDY 2

## Dual-input stacked LSTM

- Mask-aware temporal attention
- Longitudinal sequences AND static biomarkers in parallel
- Three diagnostic classes and  targets the MCI “gray zone”

**The gap this addresses:** Recent reviews find temporal modelling and rigorous XAI validation almost never combined, only a single study explicitly models temporal dynamics with XAI. Nearly 75% of AD-AI studies include no XAI component at all.

# Does the explanation survive perturbation?

**Model robustness**
the prediction stays correct (well studied)

**Explanation stability**
the reasoning stays consistent (minimal attention)

*These are not the same property. Clinical data carries noise from inter-rater variability, instrument calibration and biological fluctuation.*

**Inject noise**
Gaussian · Uniform · Salt-&-Pepper

→

**Re-explain**
recompute attributions

→

**Compare**
cosine similarity · Spearman ρ

→

**Sweep**
across 20 noise levels

**Result** Explanations remain stable across all three noise families and the full noise sweep. This establishes stability as a measurable property in its own right, separate from accuracy.

*Validated in both clinical studies · IEEE FLICS 2026 (Valencia) and the journal manuscript under review*

# Testing the explanations themselves

Most AD-XAI work judges explanations by visual inspection or informal agreement with clinical intuition. Statistical sanity checks are rare in this context, and faithfulness metrics are underused.

## A 14-test validation battery, applied to every explanation method



**9 sanity checks**

*does the explanation depend on what the model learned?*



**5 faithfulness tests**

*do the attributed features actually drive the prediction?*

+ explanation stability under perturbation
+ model robustness under noise corruption
+ 5 complementary XAI methods compared

## Three methodological gaps this closes

### 1 Validation is inconsistent

XAI outputs judged qualitatively, secondary to accuracy reporting

### 2 No standard for stability

explanation consistency under perturbation rarely assessed

### 3 Temporal + validation split

longitudinal modelling and rigorous XAI validation not combined

# Explanation quality is not monolithic (the “IG Paradox”)

## Sanity checks

- Three of four methods pass all checks
- **Integrated Gradients FAILS rank-based sanity checks**

## Faithfulness

- All methods show positive faithfulness
- **But IG achieves the HIGHEST (ρ = 0.721, p = 0.019)**

**The IG Paradox :** An explanation method can fail sanity checks and still be the most faithful. We attribute this to confidence-induced gradient saturation and name it as a phenomenon the field needs to account for.

# Validated across cohorts of increasing rigor

The GA expects DC3 to “experiment with different sets of data applied to health informatics”. **The methodology was therefore re-tested on progressively harder cohorts — the same tests, tougher data.**

### Kaggle AD

**150**

subjects · CN 64 / MCI 46 / AD 40

Proof-of-concept: stability under Noise

IEEE FLICS 2026 · published

→

### ADNI *

**1,968**

subjects · CN/MCI/Dementia

Full 14-test battery (The IG Paradox)

Journal · under review

→

### OASIS3 **

**Large**

neuroimaging cohort

Do the Findings Generalize?

In preparation

* Alzheimer's Disease Neuroimaging Initiative
** Open Access Series of Imaging Studies-3

# Main Research output of LM-Rta Workshop

FIRST AUTHOR

## HySDG-ESD

2D-LiDAR obstacle perception for AGVs (Accepted SWE 2026 (IEEE), Rende, Italy)

DBSCAN | Kasa fit | Kalman | Hungarian

Classifies obstacles as static / dynamic / unknown, in real time on AGV-grade hardware

| 0.977 | 0.083 m | >92% | 0 |
|---|---|---|---|
| MOTA (Sc. 1) | RMSE (Sc. 1) | Acc. (Sc. 2–3) | dyn→static err. |

CO-AUTHOR

## Benchmarking Framework

Simulation-driven benchmark for LiDAR collision avoidance · (IEEE FLICS 2026, Valencia)

Detection | Fusion | Avoidance

Each stage isolated and tested separately under controlled sensor / map degradation

| 480 | 11,520 | 8 | 3 |
|---|---|---|---|
| worlds | runs | pipelines | sensor cfg. |

# Progress against the GA objectives

| GA / CDP objective | Delivered so far | Status |
|---|---|---|
| **Interpretable & explainable AI for trustworthy health informatics** | 5 XAI methods + a 14-test validation battery; explanation stability established as a property | On track |
| **Dimensionality reduction ·automatic feature selection** | Boruta as an independent, model-agnostic feature-selection reference | On track |
| **Chronic disease: diabetes · heart disease · Alzheimer's** | Alzheimer's, in depth with three studies (Kaggle → ADNI → OASIS3) | Focused |
| **Early intervention strategies** | Longitudinal progression modelling; targets the MCI “gray zone” | On track |
| **Experiment with different sets of data** | Three cohorts of increasing scale and rigor; plus, a synthetic-simulation domain (WP3) | On track |

## Two deliberate changes

**Alzheimer's only, of the three named diseases**
A validation methodology must go deep in one domain before it can claim anything. Alzheimer's offers the longitudinal data and the hard MCI boundary the method needs.

**A WP3 sensor-fusion line was added**
Network collaboration (LM-RtA). Methodologically linked, and it supplies the second domain that a domain-independence claim requires.

# Training, outputs and mobility

CDP agreed 21 October 2025. Progress against its short-term (1–2 year) objectives:

## Training completed

- **LM-RtA (SUT Doctoral School)**
- **TUAI Spring School (FLICS 2026)**
- **TUAI Summer School, Gliwice**
  July 2026 (this event)

## Research outputs

CDP target: 1–2 conference papers

**3 delivered**
FLICS 2026 (1st author, published)
FLICS 2026 benchmark (co-author, published)
SWE 2026 (1st author, Accepted)

CDP target: 2–3 journal manuscripts

**1 under review:** "Beyond Accuracy" (ADNI)
**1 in preparation:** OASIS3 progression

## Mobility & collaboration

Secondments per GA (4 months each)

- **UPM** Deep visual analytics for time series (D. Camacho)
- **NTNU** Trustworthy AI · real-time anomaly detection (S. Yin)
- **SUT** Perception accuracy & XAI (J. C.-W. Lin)

*Applied partners*
**BioKeralty · TheNextPangea**
clinical data access and validation

LOOKING AHEAD

# Next steps and risks

## Next steps

### 1 Complete the OASIS3 study

### 2 Schedule the secondments

UPM / NTNU / SUT
the main outstanding CDP item.

### 3 Activate applied collaboration

BioKeralty & TheNextPangea for clinical data and real-world validation.

## Risks & mitigation

### Clinical partner data access

*Governance and timelines are outside my control*

Public cohorts (ADNI, OASIS3) already secured as the primary evidence base

### Journal review timeline

*Decision date is not in my hands*

OASIS3 study advanced in parallel rather than waiting

### Secondment scheduling

*12 months across three hosts must fit the remaining period*

Early planning with supervisors, starting now

**All of the above feeds directly into the TUAI mid-term report** *(formal submission — September 2026)*



IN SUMMARY

# Accuracy is the start; trustworthy explanation is the goal.

## Thank you

Milad Jafari Barani (DC3) University of Oviedo
jafaribaranimilad@uniovi.es
Supervisor: Prof. Vicente García Díaz

Funded by the European Union

Universidad de Oviedo