Self-Explainable AI and Attention for Interpretable Cancer Analysis with Image and Omics Data (Multi-Modal): A Systematic Review

Authors: Muruganantham Jaisankar[1,2], Armantas Ostreika[1], Begoña García-Zapirain Soto[2]

Affiliation: [1] Kaunas University of Technology, Kaunas, Lithuania.
[2] University of Deusto, Bilbao, Spain.

INTRODUCTION

- Problem: Clinicians need to interpret and trust AI models for cancer diagnosis and prognosis. However, most deep learning models are "black boxes," lacking transparency.

- Objective: To systematically review the literature on attention-based S-XAI models for interpretable cancer analysis using multimodal data.
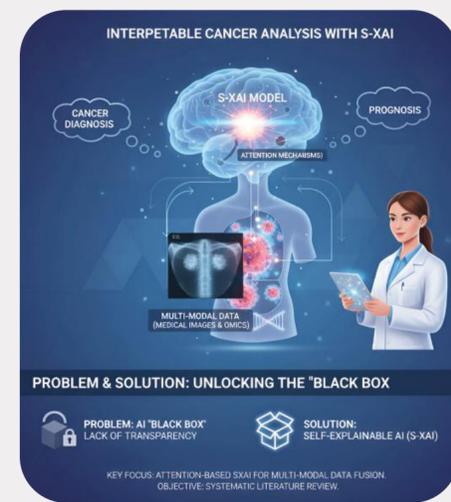
METHODOLOGY

- Study Design: A systematic review following PRISMA guidelines.
- Search Strategy: We searched major databases (e.g., PubMed, IEEE Xplore, arXiv)
- Inclusion Criteria:
  - Studies that use AI/deep learning.
  - Models that incorporate attention mechanisms or are inherently interpretable.
  - Analysis of multimodal data (image + omics) for cancer.
- Exclusion Criteria:
  - Models with purely post-hoc explanations.
  - Studies on non-cancer diseases.
- Data Extraction: We extracted data on model architecture, data types, performance metrics (e.g., C-index, AUC), and interpretability methods.
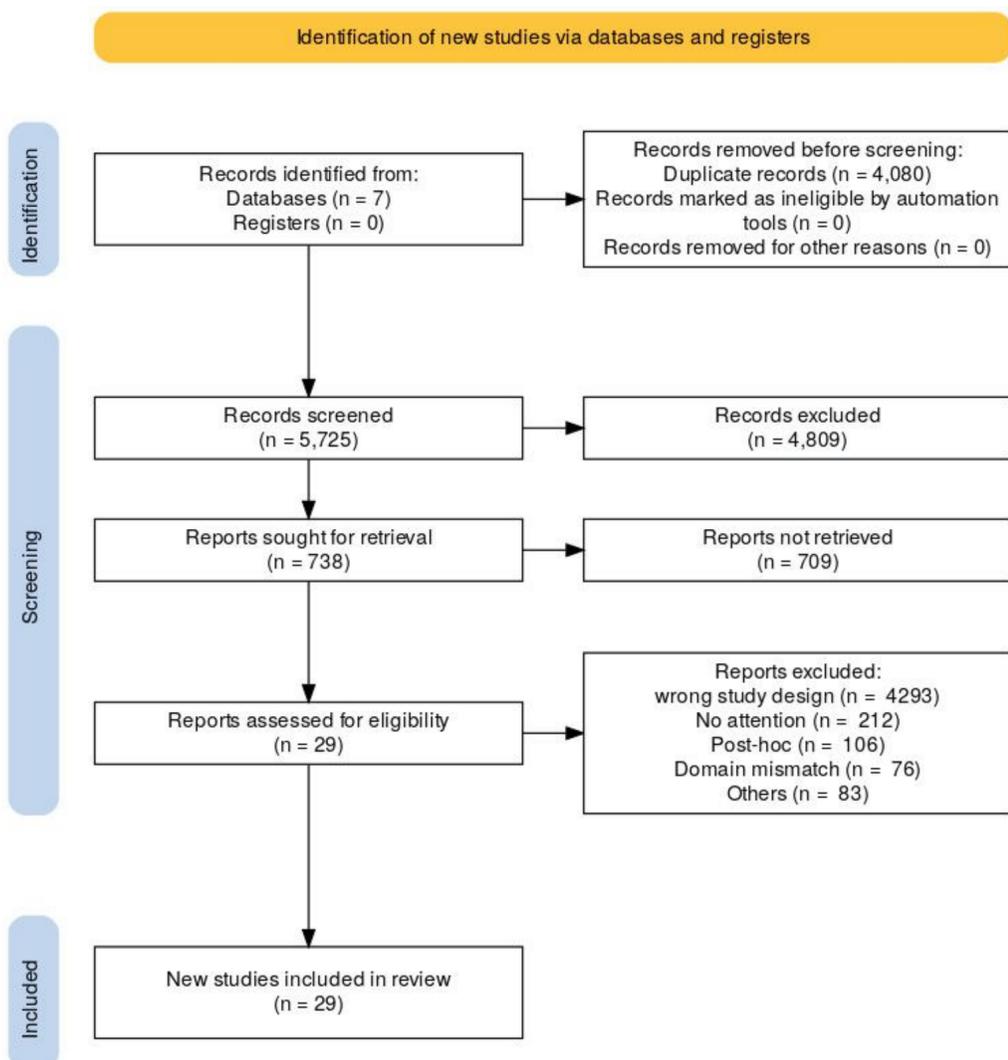

RESULTS

- AI models using attention mechanisms for cancer analysis are highly effective and interpretable.

- These multi-modal models, combining data like images and omics, consistently outperform single-modality approaches.

- For example, one model achieved a C-index of 0.668, surpassing unimodal counterparts.

- Another model achieved an AUC of 0.9704 for predicting a key prognostic marker from MRI scans, highlighting their clinical potential.

- Interpretability is a major strength, with attention heatmaps visually correlating model predictions to features a pathologist would identify on a microscopic level. Although, semantic meaning should be included.

CONCLUSIONS

- The use of attention mechanisms in AI models significantly improves performance for cancer prognosis.

- The resulting models are also more interpretable, fostering trust in clinical applications.

- Semantic meaning can be integrated into the interpretations provided by these models.

PRISMA FLOW DIAGRAM

Identification of new studies via databases and registers

Identification

Records identified from:
Databases (n = 7)
Registers (n = 0)

→ Records removed before screening:
Duplicate records (n = 4,080)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 0)

Screening

Records screened (n = 5,725) → Records excluded (n = 4,809)

Reports sought for retrieval (n = 738) → Reports not retrieved (n = 709)

Reports assessed for eligibility (n = 29) → Reports excluded:
wrong study design (n = 4293)
No attention (n = 212)
Post-hoc (n = 106)
Domain mismatch (n = 76)
Others (n = 83)

Included

New studies included in review (n = 29)

References

1. Muruganantham Jaisankar, Begoña García-Zapirain, Armantas Ostreika. Self-Explainable AI and Attention for Interpretable Cancer Analysis with Image and Omics Data (Multi-Modal): A Systematic Review. PROSPERO 2025 CRD420251032833. Available from https://www.crd.york.ac.uk/PROSPERO/view/CRD420251032833.

2. Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis Campbell Systematic Reviews, 18, e1230.

The review poster has preliminary results.