

Jun 12, 2025

Version 1

Self-Explainable AI and Attention for Interpretable Cancer Analysis with Image and Omics Data (Multi-Modal): A Systematic Review V.1

DOI

<https://dx.doi.org/10.17504/protocols.io.e6nvwqwpdvmk/v1>

Muruganantham Jaisankar^{1,2}, Begoña arcía-Zapirain Soto^{3,4}, Armantas Ostreika^{5,6}

¹Kaunas University of Technology; ²University of Deusto; ³eVida Research group;

⁴University of Deusto, Bilbao, Spain; ⁵Department of Multimedia Engineering;

⁶Faculty of Informatics, Kaunas University of Technology, Kaunas, Lithuania



Muruganantham Jaisankar

Kaunas University of Technology

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.e6nvwqwpdvmk/v1>

Protocol Citation: Muruganantham Jaisankar, Begoña arcía-Zapirain Soto, Armantas Ostreika 2025. Self-Explainable AI and Attention for Interpretable Cancer Analysis with Image and Omics Data (Multi-Modal): A Systematic Review. **protocols.io** <https://dx.doi.org/10.17504/protocols.io.e6nvwqwpdvmk/v1>

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: In development

We are still developing and optimizing this protocol

Created: June 12, 2025

Last Modified: June 12, 2025

Protocol Integer ID: 220038

Keywords: attention for interpretable cancer analysis, ai models for multimodal cancer data analysis, interpretable cancer analysis, multimodal cancer data analysis, explainable ai, context of explainable ai, potential utility for interpretability, deep learning, attention, many deep learning model, ai model, interpretability, multimodal data, including medical image, improving cancer, interpretable for trustworthiness, medical image, attention mechanism, systematic review cancer, explainability, trust, model, artificial intelligence, use of attention mechanism, insights into the model, prognosis, complex dataset, cancer

Abstract

Cancer is a complex and heterogeneous disease that causes death, requiring an integrative multi-approach for effective diagnosis, prognosis, and treatment. Multimodal data, including medical images and omics data, provide complementary information about the disease. Artificial intelligence, specifically deep learning, shows promising results in the analysis of complex datasets and improving cancer-related tasks. However, many deep learning models are 'black boxes', limiting clinical adoption due to lack of transparency and trust. Explainable AI (XAI) helps to address this limitation by developing techniques to understand how AI models arrive at their predictions. This review mainly focuses on Self-Explainable AI (S-XAI), which refers to models designed to be inherently interpretable for trustworthiness. Attention mechanisms, inspired by human visual systems, allow AI models to focus on the most relevant parts of the input data. Although not specifically designed for explainability, attention mechanisms naturally highlight key features and provide insights into the model's decision-making process. This review focuses on the use of attention mechanisms in AI models for multimodal cancer data analysis, with a particular emphasis on their potential for Self-explainability (S-XAI). Considering the innovative approach adopted in the review, it facilitates a thorough exploration of the current applications of attention mechanisms and their potential utility for interpretability. The review includes a wide array of studies that employ attention mechanisms, regardless of whether these studies explicitly frame themselves within the context of Explainable AI (XAI).

Guidelines

This systematic review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

Materials

The following electronic databases will be searched: PubMed/MEDLINE, Scopus, Web of Science Core Collection, IEEE Xplore, The Cancer Genome Atlas. Additional sources if necessary: arXiv, Grey literature, Conference Proceedings, Google Scholar.

Objectives

1 Primary objectives:

To systematically review and synthesize the literature on the application of attention mechanisms in self-explainable AI (SXAI) for interpretable cancer analysis using multimodal integration (or individual) of medical imaging and omics data.

2 Secondary objectives:

- To analyse how attention mechanisms are used to integrate and provide interpretability for multi-modal data (both combined and separately with image and omics).
- To categorise the types of attention mechanism used in this context.
- To identify the specific image and omics modalities used in the context.
- To examine the metrics and methods used to evaluate the interpretability of AI models with attention.
- To identify the challenges and limitations in attention based S-XAI in cancer data analysis.
- To investigate the clinical applicability and potential impact of interpretable AI with attention mechanism in cancer research and practice

Methods

3 Study design

This systematic review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

Eligibility Criteria

Types of studies

• Inclusion:

Original research papers (e.g., retrospective studies), and conference proceedings with full papers.

• Exclusion:

Review articles and systematic reviews without future direction in this field, meta-analysis, editorials, letters to editor, case reports, and theoretical papers without empirical validation on multi-modal (images, omics) cancer data.

Types of participants/Problem

Studies that use AI to analyze image and omics data for cancer research, diagnosis, prognosis, or treatment prediction.

Types of interventions/Exposure

Studies that use AI models, including attention mechanisms, for omics and image data.

Types of outcome measures**• Inclusion:**

- Description of the S-XAI method used (general or attention-based).
- Specific attention mechanisms used (for instance, spatial, channel).
- How attention mechanisms are developed (for instance, attention gates, self-attention layers).
- How the S-XAI method, specifically attention mechanisms, contributes to interpretability (for example, feature importance, attention maps).
- Evaluation metrics and methods for model performance (e.g., accuracy, AUC) and interpretability (e.g., fidelity, plausibility), if reported in the studies.
- Types of image and omics data used.
- Clinical applications.
- Discussion of post-hoc explainability methods, and their limitations, if used in the studies.

• Exclusion:

- Studies that fail to provide detailed descriptions of the AI methods employed, making it impossible to determine whether they apply S-XAI or attention mechanisms, or those that neglect to discuss limitations of post-hoc methods, when applicable.

4 Setting

Studies conducted in any settings where multimodal cancer data are analyzed, or any of the domains medical imaging or omics (e.g., hospitals, universities, research institutions).

5 Language

Primary language is English, other languages will be considered if applicable.

6 **Publication Date Range**

No Range.

7 **Information sources**

The following electronic databases will be searched:

- PubMed/MEDLINE
- Scopus
- Web of Science Core Collection
- IEEE Xplore
- The Cancer Genome Atlas

Additional sources if necessary:

- arXiv
- Grey literature, Conference Proceedings
- Google Scholar

8 **Search strategy**

To obtain studies relevant to the review objectives, the following search strategy will be followed: The search will include keywords related to cancer, multimodal data (both combined and separate image and omics data), AI, explainable AI, self-explainable AI, and attention mechanisms. Specific terms will include: (Cancer OR neoplasm OR tumor) AND (Multi-Modal OR Integrative analysis) AND (Image analysis OR medical Imaging OR genomics OR Transcriptomics OR Proteomics OR OMICS) AND (Artificial intelligence OR deep learning OR neural network OR Vision transformer OR Vision-language) AND (Explainable AI OR Intrinsic Interpretable AI OR XAI OR Transparency) AND (Causal Attention mechanism OR Attention mechanism OR Attention Layer OR Self-Attention OR Spatial Attention OR Channel attention OR Transformer OR feature Selection OR Rule based OR prototype based OR post-hoc OR saliency map OR LIME OR SHAP). The search approach and keywords will be adjusted for each database to account for differences in syntax and search functionalities.

9 **Study selection**

Title and abstract screening

The PhD student will screen the titles and abstracts of all records identified through the database searches and other sources of scientific evidence. Before starting the screening process, the PhD student will:

- Comprehensively review and understand the inclusion and exclusion criteria.
- Develop a screening form, either within the chosen software (e.g., Rayyan, Covidence) or as a separate excel sheet, with the columns for:

- Record ID
- Title
- Abstract
- Inclusion/Exclusion decision (Include, Exclude, Unclear)
- Reason for exclusion (if excluded)
- Reviewer initials (Muruganantham Jaisankar)

The PhD student will conduct a pilot test, individually screening a subset of [50–100] records. The results of this pilot test will be reviewed by the supervisors to ensure the consistent use case of criteria. Each record of scientific articles will be classified as “Include”, “Exclude”, or “Unclear” based on the information in the title and abstract. Studies marked as “Exclude” will be excluded from the record. The reason for the exclusion will be recorded in the screening form to understand the details. Studies marked as “Include” or “Unclear” will be retained for full text review. The PhD student will meet the supervisors after the initial screening of a large batch of scientific records, for instance, the first 200, to discuss any uncertainties or difficulties encountered during the screening process. The screening process will be facilitated using chosen software (e.g., Rayyan, Covidence, AS review) Systematic review management software, or manual process.

Disagreement Resolution During the screening process, if the PhD student is not clear about a decision, the student can discuss with one of the supervisors. If consensus is reached in the above discussion, the decision will be recorded. If consensus is not reached in the above discussion, the second supervisor can make the final decision. All disagreements and their resolutions should be documented in a separate log.

Full-Text assessment

Full Text Retrieval

The PhD student will obtain the full text copies of all studies that were included during the title and abstract screening process. This will be done by:

- Accessing library databases.
- Using institutional (KTU, UD) access to journals.
- Contacting authors directly if necessary.

Full text Review

The PhD student will assess the full-text articles based on the inclusion and exclusion criteria. A standard form will be used to document the reason for inclusion or exclusion at the full-text

review stage, this form also includes:

- Record ID
- Full-Text article details
- Inclusion/Exclusion decision (Include, Exclude)
- Detailed reason for exclusion (if excluded)

- Reviewer initials (Muruganantham Jaisankar)

Studies that do not meet the inclusion criteria will be excluded, and the reason for exclusion will be recorded in the full-text assessment form.

Disagreement Resolution

During the full-text process, if the PhD student is not clear about a decision, the student can discuss with one of the supervisors. If consensus is reached in the above discussion, the decision will be recorded. If consensus is not reached in the above discussion, the second supervisor can make the final decision. All disagreements and their resolutions should be documented in a separate log.

10 Data Management

Record Keeping

The PhD student will use a (Rayyan, ASReview or Covidence) Systematic review management software, or manual process to manage all records during the study selection process. This includes:

- Importing and eliminating duplicate search results from all databases.
- Tracking the progress of each record through the screening and full-text assessment stages.
- Recording all the decisions of inclusion and exclusion of studies, and the reason for exclusions.
- Storing full-text articles using the software or maintaining a clear system for linking records to the full-text copies.

PRISMA Flow diagram

The PhD student will generate a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) flow diagram to visually represent the study selection process. This diagram should outline the number of records identified, included, and excluded at each stage.

Audit Trail

The PhD student will maintain a detailed audit trail of the study selection process. This should include:

- Date of each stage of screening and full-text assessment.
- A log of all disagreements and how they were resolved.
- Records of any deviations from the protocols related to study selection, and the reason for these deviations.
- Documentation of any consultations with supervisors.

11 Data Extraction

A standard data extraction form will be developed and pilot-tested on a sample of included studies. The form should collect the following information:

- Study characteristics, for instance author, year of publication, journal/conference, study design, sample size.
- Cancer type and stage if applicable.
- Image modalities used, for instance, MRI, CT, Ultrasound.
- OMICS modalities used, for instance, genomics, transcriptomic.
- AI model architecture.
- S-XAI method used, is that using attention based or general.
- Specific attention mechanisms used if applicable, for instance, spatial, channel.
- How attention mechanism are implemented, for instance, attention gates, self-attention layers.
- How the S-XAI method, particularly attention mechanism contributes to the interpretability, for instance, feature importance, attention maps.
- Performance metrics of the AI model, with confidence intervals if reported.
- Methods used to evaluate the interpretability of AI models with attention if reported.
- For instance, qualitative feedback, quantitative metrics like faithfulness.
- Key findings related to the interpretability and explainability.
- Reported limitation of study, including a discussion of the limitations of post-hoc methods if used.

Data extraction will be performed by one reviewer Muruganantham Jaisankar. All extracted data will be

independently verified by a second supervisor (Bego~na Garc'ia-Zapirain Soto) for accuracy and completeness. Any uncertainties or discrepancy will be resolved through discussion between primary and second reviewer. If consensus cannot be reached, the third reviewer (Armantas Ostreika) will be reached for arbitration.

12 **Assessment of Risk of Bias in Individual studies**

The risk of bias in included studies will be assessed using appropriate risk of bias assessment tool(s) based on the study designs, for instance QUADAS-AI for diagnostic accuracy studies with AI. The assessment will cover relevant domains of bias as outlined in the chosen tools, for instance, selection bias, index test bias, reference standard bias, flow and timing. Risk of bias assessment will be performed independently by two reviewers. The results of the risk of bias assessment will be summarized for each included study and presented in the review.

13 **Data synthesis**

The data will be synthesized as follows:

- Overview of S-XAI Methods: Describe general S-XAI methods.
- Frequency of Attention-based S-XAI: Report attention mechanism frequency.

- Characteristics of Attention-based S-XAI:
 - Describe specific attention mechanisms.
 - Analyze interpretability contribution.
- Limitations of Post-hoc Methods: Discuss post-hoc method limitations.
- Evaluation Metrics and Methods for Interpretability: Synthesize interpretability metrics, methods.
- Clinical Applicability and Potential Impact: Investigate clinical impact, applicability.
- Challenges and Limitations: Synthesize challenges and limitations.
- Quantitative Synthesis (Meta-analysis): Perform meta-analysis if homogeneous.
- Assessment of Heterogeneity: Explore and assess heterogeneity.
- Certainty of Evidence: Assess certainty using GRADE.
- Integration of Risk of Bias: Integrate QUADAS-AI results.

14 **Expected outcomes and dissemination**

The findings of this systematic review will be disseminated through:

- Publication in a peer-reviewed academic journal.
- Presentation at relevant conferences or scientific meetings.

15 **Timeline**

- 2025 March - May: Protocol development complete.
- 2025 June: Protocol submission, searching, and screening.
- 2025 July: Full-text, start data extraction.
- 2025 August: Data extraction, risk bias, synthesis.
- 2025 September: Writing, review, submission

16 **Team and Expertise**

- Muruganantham Jaisankar, PhD student, Kaunas University of Technology, University of Deusto.
- Begoña García-Zapirain, Supervisor, eVida Research group, University of Deusto, Bilbao, Spain.
- Armantas Ostreika, Supervisor, Department of Multimedia Engineering, Faculty of Informatics, Kaunas University of Technology, Kaunas, Lithuania

17 **Conflicts of Interest**

No conflicts of Interest.

18 **Funding**

Erasmus+, Lithuanian state fund for research.

19 **Registration**



This protocol will be registered on PROSPERO (International Prospective Register of Systematic Reviews).