Measuring Cross-Modal Interactions in Multimodal Models

Laura Wenderoth[1] , Konstantin Hemker[1], Nikola Simidjievski[2,1], Mateja Jamnik[1]
[1]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
[2]PBCI, Department of Oncology, University of Cambridge, Cambridge, United Kingdom
{lw457, kh701, ns779, mj201}@cam.ac.uk

AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

Code

Paper

Problem

While numerous explainability (XAI) methods exist for unimodal models, existing XAI methods fail in multimodal settings.

- Need for scalable solutions to explain multimodal model behavior, focusing on cross-modal interactions and modality contributions.
- Existing methods are limited to only two modalities, require labeled data, and are not performance-agnostic:

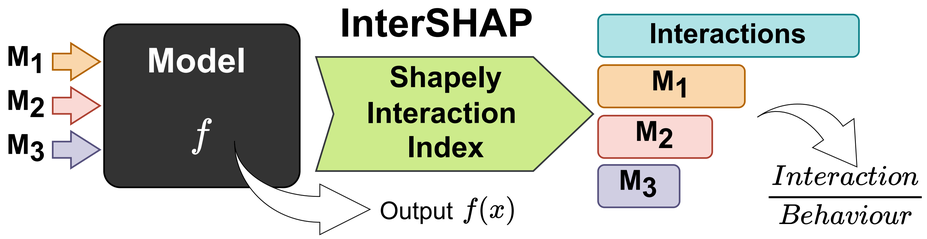
Score	Modalities > 2	Local	Unsupervised	Performance Agnostic
PID	✗	✗	✓	✓
EMAP	✗	✗	✗	✗
SHAPE	✓	✗	✗	✗
InterSHAP	✓	✓	✓	✓

Solution

Novel cross-modal interaction score, InterSHAP, based on the Shapley interaction index. It handles more then two modalities, works with unlabelled data, and provides both local and global explanations, offering a performance-agnostic approach to understanding cross-modal interactions and modality contributions.

Contributions

1. Validated on synthetic datasets
2. Works > 2 modalities
3. Seamlessly integrates into SHAP visualization
4. Application to multimodal healthcare datasets



$$\Phi_{ij} = \left| \frac{1}{N} \sum_{a=1}^{N} \phi_{ij}(m_1^a, \ldots, m_M^a, f) \right| \tag{1}$$

$$\Phi = \begin{bmatrix} \Phi_{11} & \Phi_{12} & \ldots & \Phi_{1M} \\ \Phi_{21} & \Phi_{22} & \ldots & \Phi_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{M1} & \Phi_{M2} & \ldots & \Phi_{MM} \end{bmatrix} \tag{2}$$

$$Interactions = \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \Phi_{ij}, \qquad Behavior = \sum_{i,j=1}^{M} \Phi_{ij} \tag{3}$$

$$InterSHAP = \frac{Interactions}{Behaviour} \tag{4}$$

1 Validated on synthetic datasets

	Uniqueness		Synergy		Redundancy	Random
	XOR	FCNN	XOR	FCNN	FCNN	FCNN
InterSHAP	0.0	0.2 ± 0.1	99.7	98.0 ± 0.5	38.6 ± 0.5	57.8 ± 1.1
InterSHAP$_{local}$	1.8	3.4 ± 5.2	96.9	85.8 ± 12.1	37.3 ± 25.0	40.0 ± 13.3
PID	0.01	0.01 ± 0.01	0.39	0.39 ± 0.01	0.14 ± 0.02	0.48 ± 0.01
EMAP$_{gap}$	0	0 ± 0	49.1	43.5 ± 1.0	1.8 ± 0.4	0.6 ± 0.4
SHAPE	1.6	16.5 ± 0.6	33.1	27.6 ± 1.5	-47.1 ± 0.5	15.1 ± 1.9

Table 2: InterSHAP values as percentages for both the XOR function and FCNN with early fusion on synthetic HD-XOR datasets. Results align with expectations, confirming the effectiveness of InterSHAP.

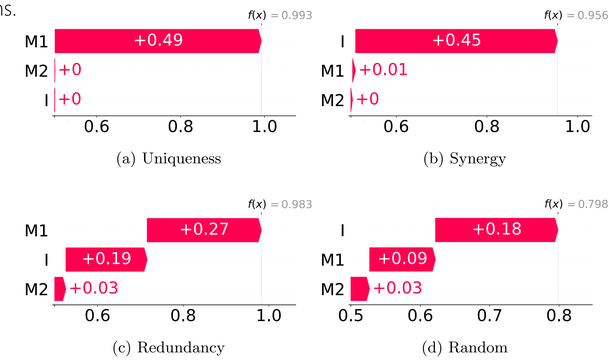
2 Works > 2 modalities

	Uniqueness	Synergy	Redundancy
2 Modalities	0.2 ± 0.1	98.0 ± 0.5	38.6 ± 0.5
3 Modalities	0.6 ± 0.2	88.8 ± 0.5	51.9 ± 0.3
4 Modalities	1.2 ± 0.1	64.1 ± 0.8	40.2 ± 0.2

Table 3:InterSHAP values, expressed as percentages, for FCNN with early fusion on HD-XOR datasets with two, three, and four modalities. The results indicate, that InterSHAP works for more than two modalities.

3 Seamlessly integrates into SHAP visualization

Figure 2: Visualization of Table 2 results. M1 represents modality 1, M2 modality 2, and I interactions.



(a) Uniqueness (b) Synergy (c) Redundancy (d) Random

4 Application to multimodal healthcare datasets

MIMIC III

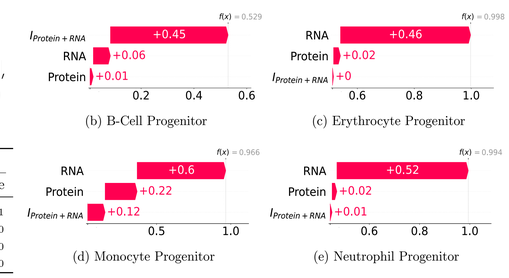
- Modalities: 12 physiological measurements (e.g. heart rate, 24h), static information on patients
- Tasks: ICD and Mortality Classification

	ICD-9		Mortality	
	baseline	MVAE	baseline	MVAE
InterSHAP	1.2 ± 0.2	6.8 ± 1.3	11.0 ± 0.5	12.3 ± 2.8
PID	0.06 ± 0.01	0.09 ± 0.01	0.10 ± 0.01	0.11 ± 0.01
EMAP$_{gap}$	0 ± 0	1.2 ± 0.0	-0.8 ± 0.1	0.9 ± 0.1
SHAPE	0.2 ± 0	0.6 ± 0	0.2 ± 0.2	0.7 ± 0.2

Multimodal Single Cell Dataset

- Modalities: RNA, Protein
- Task: Cell Class Classification (B-Cell, Erythrocyte, Monocyte, Neutrophil)

	Single-Cell		
	early	intermediate	late
InterSHAP	1.9 ± 0.4	1.5 ± 0.4	0.4 ± 0.1
PID	0.08 ± 0.01	0.08 ± 0.01	0.06 ± 0.0
EMAP$_{gab}$	0 ± 0	0 ± 0	0 ± 0
SHAPE	1.0 ± 0.2	0.7 ± 0.2	0 ± 0



(b) B-Cell Progenitor (c) Erythrocyte Progenitor (d) Monocyte Progenitor (e) Neutrophil Progenitor

Figure 3: FCNN early fusion model