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, United Kingdom
[2]PBCI, Department of Oncology, University of Cambridge, Cambridge, United Kingdom
{lw457, kh701, ns779, mj201}@cam.ac.uk

Abstract

Integrating AI in healthcare can greatly improve patient care and system efficiency. However, the lack of explainability in AI systems (XAI) hinders their clinical adoption, especially in multimodal settings that use increasingly complex model architectures. Most existing XAI methods focus on unimodal models, which fail to capture cross-modal interactions crucial for understanding the combined impact of multiple data sources. Existing methods for quantifying cross-modal interactions are limited to two modalities, rely on labelled data, and depend on model performance. This is problematic in healthcare, where XAI must handle multiple data sources and provide individualised explanations. This paper introduces InterSHAP, a cross-modal interaction score that addresses the limitations of existing approaches. InterSHAP uses the Shapley interaction index to precisely separate and quantify the contributions of the individual modalities and their interactions without approximations. By integrating an open-source implementation with the SHAP package, we enhance reproducibility and ease of use. We show that InterSHAP accurately measures the presence of cross-modal interactions, can handle multiple modalities, and provides detailed explanations at a local level for individual samples. Furthermore, we apply InterSHAP to multimodal medical datasets and demonstrate its applicability for individualised explanations.

Code — https://github.com/LauraWenderoth/InterSHAP
Extended version — https://arxiv.org/abs/2412.15828

1 Introduction

In medical decision-making, there is a growing trend toward utilising multimodal machine learning approaches that integrate multiple data sources to improve predictive and diagnostic tasks. These approaches acknowledge that medical data analysis is inherently multimodal, leveraging methods that are able to fuse heterogeneous data that includes clinical, genomic, and imaging modalities. Recent advancements in large multimodal models in the medical field, such as Google's Med-PaLM (Tu et al. 2023) and Microsoft's BiomedCLIP (Zhang et al. 2023), further highlight the utility of multimodal approaches in healthcare.

However, an inherent limitation of many of the current multimodal approaches pertains to their explainability.

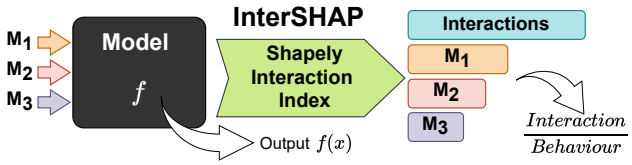


Figure 1: Overview of InterSHAP. The model (black box), takes three different modalities as input and produces an output $f(x)$. Through perturbations of the input modalities and observing the resulting changes in outputs, the Shapley interaction index (Grabisch and Roubens 1999) is used to dissect the model's behaviour into modality contributions and cross-modal interactions. *InterSHAP is defined as the ratio of interactions to model behaviour.*

Thus, healthcare applications of machine learning are often left to rely on high-performance but opaque models that obscure the reasoning behind their predictions. Such models can accurately predict diagnoses, but these predictions may not reflect true causal relationships in the data (Han et al. 2022). The lack of explainability and transparency is widely recognized as a key barrier to the clinical adoption of ML models, as noted by policymakers such as the OECD (Anderson and Sutherland 2024) and emphasized in academic studies (Yang, Ye, and Xia 2022).

In response to the urgent need for XAI, numerous methods such as SHAP (Lundberg and Lee 2017) and LIME (Ribeiro, Singh, and Guestrin 2016) have emerged to explain the results of machine learning systems. While these have shown promise for unimodal models, there are few explanation methods that work effectively across data structures, meaning that cross-modal interactions can often not be quantified with unimodal methods. Multiple modalities allow a machine learning model to utilise and combine information from different sources, enabling cross-modal interactions that often lead to improved performance. However, simply identifying these interactions is insufficient for a meaningful interpretation of the model predictions. In this context, it is crucial to comprehend both the individual impact of each modality and their combined influence. Several methods, such as PID (Liang et al. 2023), EMAP (Hessel and Lee 2020), and SHAPE (Hu, Li, and Zhou 2022), have been developed to assess whether models learn cross-modal

interactions. However, these methods are often (1) limited to only two modalities; (2) applied to the entire dataset, rather than individual samples; and (3) either require labelled data or are not performance-agnostic, which can lead to an incomplete understanding of cross-modal interactions.

In this paper, we introduce **InterSHAP**, a comprehensive, interpretable metric that quantifies the influence of cross-modal interactions on model behaviour. InterSHAP supports any number of modalities, provides local explanations, is agnostic to model performance, and works with unlabelled data. To our knowledge, our proposed variant of the Shapley interaction index is the first approach that is able to effectively separate cross-modal interactions from individual modality contributions. Figure 1 provides a schematic overview of InterSHAP. To validate the effectiveness of InterSHAP, we conduct extensive empirical analyses using synthetic datasets with either no synergy (no cross-modal interactions are needed to solve the task) or exclusively synergy (complete information can only be obtained as a result of cross-modal interaction across all modalities). The results demonstrate that InterSHAP can accurately detect a lack of cross-modal interactions (0%) on datasets with no synergy, and 99.7% on those with exclusive synergy, aligning precisely with the expected behaviour. Furthermore, we demonstrated that InterSHAP is extendable to more than two modalities and achieves favourable results at a local level with individual data points. Additionally, we test InterSHAP on two medical datasets to show its utility in a real-world setting. InterSHAP is not confined to medical domains and can be applied to models trained on any modality.

Our main contributions can be summarised as follows:

- **Novel cross-modal interaction score InterSHAP.** We introduce InterSHAP, the first model-agnostic cross-modal interaction score that operates with any number of modalities, offering both local and global explanations, applies to unlabelled datasets and does not rely on the model's performance. We conducted extensive experiments on synthetic datasets and models to validate its functionality.
- **Application to multimodal healthcare datasets.** We demonstrate the benefits of InterSHAP on a cell type classification task based on multimodal protein and RNA data as well as a mortality prediction tasks.
- **Open-source implementation with integration into the SHAP Package.** We have developed an open-source implementation of InterSHAP that seamlessly integrates with the well-known SHAP visualisation package (Lundberg and Lee 2017).

2 Related Work

Existing approaches like MM-SHAP (Parcalabescu and Frank 2023) and Perceptual Score (Gat, Schwartz, and Schwing 2021) focus on understanding the contributions of individual modalities rather than explicitly examining their interactions. They explain modality contributions in isolation without considering or quantifying cross-modal interactions within multimodal networks. In contrast, EMAP (Hessel and Lee 2020) and SHAPE (Hu, Li, and Zhou 2022) detect cross-modal interactions by analysing how perturbations to the input impact the model's output. They systematically vary different modalities or combinations of modalities to assess their importance within the model. Another metric for detecting interactions, Partial Information Decomposition (PID) (Liang et al. 2023), quantifies synergy (interactions between modalities in a dataset with a combined effect surpassing individual contributions, indicating a non-additive relationship) in datasets, which can be interpreted as cross-modal interactions.

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

Table 1: InterSHAP overcomes the limitations of other cross-modal interaction scores: it is unsupervised, performance agnostic, applicable to more than two modalities, and allows for dataset- (global) and sample-level (local) explainability.

However, these existing approaches exhibit various limitations, summarised in Table 1. First, most of them restrict the analysis to just two modalities, which prevents them from being applied in domains that typically have more modalities, like in healthcare. Second, the results cannot be analysed at the level of individual samples, which is particularly critical for medical tasks where explainability for each patient is crucial. Third, some approaches necessitate the availability of labels, rendering them unsuitable for unlabelled datasets employed in self- or unsupervised learning scenarios. Fourth, certain methods assess interactions solely based on performance gain, rendering them highly performance-dependent. However, when assessing a model, it is crucial to understand how each interaction contributed to the present prediction without neglecting interactions only because the prediction was incorrect.

3 Methodology

We introduce InterSHAP to quantify cross-modal interaction learned by multimodal models while overcoming the limitations of SOTA explainability approaches. Following the definition of Liang, Zadeh, and Morency (2022), we define cross-modal interaction as a change in the model response that cannot be attributed to the information gained from a single modality alone, but arises only when both modalities are present. To ensure the accuracy of InterSHAP, we use the Shapley Interaction Index (SII), which has been shown to effectively capture model behaviour (Lundberg et al. 2020) and decompose unimodal feature interactions into its constituent parts (Ittner et al. 2021).

3.1 Preliminaries

Model Fusion. To train multimodal models is model fusion, which comes in three types: (i) early fusion, which merges data features at model input, (ii) intermediate fusion, which combines data within the model, and (iii) late fusion,

which combines the outputs of models trained on different individual data modalities (Stahlschmidt, Ulfenborg, and Synnergren 2022; Azam, Ryabchykov, and Bocklitz 2022).

Shapley Value. The Shapley value ϕ, from cooperative game theory, measures each feature's contribution to overall performance via marginal contributions to feature coalitions (Roth 1988). In machine learning, a model f_M is viewed as a coalition game with M features or modalities as players. To avoid inefficient retraining, masking strategies are applied to features $j \notin S$, where $S \subseteq M$.

The Shapley value ϕ of a feature i is calculated as the weighted average of their marginal contributions to all possible coalitions:

$$\begin{aligned} \phi_i(M, f) &= \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(|M| - |S| - 1)!}{|M|!} \Delta \\ \Delta &= \left[f_{S \cup \{i\}} \left(S \cup \{i\} \right) - f_S \left(S \right) \right]. \end{aligned} \tag{1}$$

Shapley Interaction Index. The Shapley Interaction Index (SII) (Grabisch and Roubens 1999), initially devised to quantify interaction effects between players in cooperative game theory, can be employed to discern interactions between features or modalities. We define its directly adapted version $\phi_{ij}(M, f)$ for M modalities and a model f, where $i, j \in M$ (Lundberg et al. 2020) and $i \neq j$:

$$\phi_{ij}(M, f) = \sum_{S \subseteq M \setminus \{i,j\}} \frac{|S|!(M - |S| - 2)!}{2(M - 1)!} \nabla ij(S, f), \tag{2}$$

$$\begin{aligned} \nabla_{ij}(S, f) = \big[& f_{S \cup \{ij\}} \left(S \cup \{ij\} \right) - f_{S \cup \{i\}} \left(S \cup \{i\} \right) \\ & - f_{S \cup \{j\}} \left(S \cup \{j\} \right) + f_S(S) \big]. \end{aligned} \tag{3}$$

The interaction index for a modality with itself, denoted as ϕ_{ii}, is defined as the difference between the Shapley value ϕ_i and the sum of all interaction values:

$$\phi_{ii}(M, f) = \phi_i(M, f) - \sum_{j \in M} \phi_{ij}(M, f) \quad \forall i \neq j. \tag{4}$$

3.2 Global InterSHAP

We aim to quantify the impact of cross-modal interactions on model behaviour using InterSHAP. Model behaviour is a cumulative process, beginning with the model's base output and progressively adding each modality and their interactions to shape the final prediction. To isolate interactions from the model behaviour, we apply the SII. Let $f \in \mathbb{R}^c$ be a trained model, where c is the number of per-class probabilities, evaluated on a dataset with N samples and M modalities, represented as $\{(m_1^i, \ldots, m_M^i)\}_{i=1}^N$ and corresponding labels $\{y^i\}_{i=1}^N$. The SII value ϕ_{ij} measures the interaction between modalities i and j for sample a. Φ_{ij} is defined as the absolute value of the mean over ϕ_{ij} of all samples in the dataset:

$$\Phi_{ij} = \left| \frac{1}{N} \sum_{a=1}^{N} \phi_{ij}(m_1^a, \ldots, m_M^a, f) \right|, \tag{5}$$

resulting in the matrix Φ containing only positive values:

$$\Phi = \begin{bmatrix} \Phi_{11} & \Phi_{12} & \ldots & \Phi_{1M} \\ \Phi_{21} & \Phi_{22} & \ldots & \Phi_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{M1} & \Phi_{M2} & \ldots & \Phi_{MM} \end{bmatrix}, \tag{6}$$

where interactions appear in green, while modality-specific contributions are shown in black. Note that, our choice of obtaining Φ_{ij} allows for a global interpretability view, accounting for opposing effects that may cancel out across dataset. This prevents score inflation and aligns with our main motivation – a general and robust measure of interaction strength. In turn, these are aggregated to obtain the total interactions contributions as well as the overall model behaviour:

$$Interactions = \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \Phi_{ij}, \quad Behavior = \sum_{i,j=1}^{M} \Phi_{ij}. \tag{7}$$

Finally, these are used to calculate InterSHAP, which is the fraction of model behaviour that can be explained through cross-modal interactions:

$$InterSHAP = \frac{Interactions}{Behavior}. \tag{8}$$

3.3 Modality Contributions

Based on the definition of cross-modal interactions in Equation 7, the contribution of all modalities $\mathcal{M}$ and the contribution of each modality $\mathcal{M}_i$ to the overall result is:

$$\mathcal{M} = 1 - InterSHAP, \quad \mathcal{M}_i = \frac{|\Phi_{ii}|}{Behaviour}. \tag{9}$$

3.4 Local InterSHAP

After determining the amount of interactions that a model f has learned across the entire dataset, we assess the extent to which cross-modal interactions among the M modalities contribute to the model's behaviour for each sample. Thus, we apply the definition of InterSHAP directly to a sample a:

$$I_a = \frac{\sum_{\substack{i,j=1 \\ i \neq j}}^{M} \varphi_{ij}^a}{\sum_{i,j=1}^{M} \varphi_{ij}^a}, \tag{10}$$

where $\varphi_{ij}^a = |\phi_{ij}((m_1^a, \ldots, m_M^a), f)|$, and $i, j \in \{1, \ldots, M\}$. To determine if InterSHAP is effective at the level of all individual data points, we aggregate I_a across the entire dataset to obtain the local interaction score for the dataset as a whole:

$$InterSHAP_{local} = \frac{1}{N} \sum_{a=1}^{N} I_a. \tag{11}$$

3.5 Asymptotic Bound

InterSHAP is characterised by a computational complexity of $O(N^M)$, where N represents the number of samples and M signifies the number of modalities (Grabisch and Roubens 1999; Lundberg et al. 2020). The constraint of

	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: Local and global InterSHAP values are presented in percentages for both the XOR function and the FCNN with early fusion on HD-XOR dataset with two modalities. EMAP$_{gap}$ (EMAP - F1 of the model) and SHAPE indicate F1 score improvement from cross-modal interactions, with negative values signalling performance decline. For PID, the synergy value is provided, though it is not expressed in any standard unit. For all metrics, higher scores indicate greater measured cross-modal interactions. Results for the XOR function align with expectations, confirming the effectiveness of the InterSHAP implementation.

exponential growth emerges with the increasing number of modalities rather than the feature count. InterSHAP's dependency on the number of features within a modality follows a constant pattern, meaning its runtime remains unaffected by the feature count.

4 Experimental Validation on Synthetic Data

To verify InterSHAP's functionality, we control three factors influencing cross-modal interactions: dataset synergy (non-additive interactions between modalities), model learning capacity, and metric effectiveness.

4.1 Setup

We utilise synthetic high-dimensional tabular datasets, generated using an XOR function (HD-XOR), with varying degrees of synergy to regulate the extent of cross-modal interactions. Each generated dataset contains 20,000 samples, with two to four modalities represented as 1D vectors consisting of around 100 features. The datasets are categorised on their degrees of synergy. *Uniqueness* represents datasets with no synergy, where all information is contained within a single modality. *Synergy* refers to datasets with complete synergy, where information is distributed across a number of modalities and can only be obtained through cross-modal interactions. *Redundancy* denotes datasets where the same information is present across modalities. *Random* applies to datasets with no meaningful information. For each number of modalities (two, three and four), we created a synergy, uniqueness and redundancy dataset as well as a random dataset for the case with two modalities.

To control for model variability in our experiments, we use the XOR function for the uniqueness and synergy datasets. The XOR function cannot be meaningfully applied to random and redundancy datasets, because it is unclear which information from each modality contributes to the solution, given the multiple possibilities. We train fully connected neural networks (FCNNs) on all datasets with three layers: input, hidden (half the input size), and output (half the hidden size). Three fusion strategies are tested: early fusion (data concatenated before input), intermediate fusion (data fused before the hidden layer), and late fusion (data fused before the output layer). Since early fusion captured the most cross-modal interactions, we use it in all subsequent FCNN experiments. Further details on the dataset generation and experimental setup are provided in Appendix A. The appendices are available in our code repository and extended version of the manuscript.

4.2 Results

Verification of InterSHAP. Generally, we anticipate that InterSHAP for the uniqueness dataset will show less cross-modal interaction (expected to be close to 0%) compared to the synergy dataset (expected to be close to 100%). For the random and redundancy datasets, it is unclear how many cross-modal interactions the FCNN should learn. In the redundancy setting, the FCNN might focus on just one modality or use all modalities since information is distributed across them. In the random setting, there is no inherent information, making the FCNN's learning unpredictable.

Table 2 presents the results of the experiments conducted on the HD-XOR datasets. InterSHAP predicts the expected amount of cross-modal interaction in the XOR function with 0% cross-modal interaction for uniqueness and 99.7% for synergy. In contrast, the FCNN model showed an increase in cross-modal interaction for the uniqueness dataset (0.2% $\pm$ 0.1) and decrease in synergy (98.0% $\pm$ 0.5) for the synergy setting. This indicates that the FCNN does not fully capture the systematic structure underlying the dataset's creation with the XOR function. Instead, it sometimes uses both modalities for predictions even when unnecessary, and conversely, may underutilise them when they are required.

InterSHAP indicates that the FCNN attributes approximately 40% of its behaviour to cross-modal interactions on the redundancy dataset. Especially noteworthy is the outcome on the random dataset, which shows that InterSHAP operates performance-agnostic, as only random performance could be achieved on this dataset. Nevertheless, the model appears to have learned something – though not relevant to the F1 Score – evidenced by approximately 60% of cross-modal interactions. In summary, InterSHAP demonstrates the expected behaviour and adequately quantifies cross-modal interaction within a model.

Modality Scalability. To test InterSHAP's scalability to more than two modalities, we create HD-XOR datasets with

	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.

three and four modalities. The InterSHAP values for the FCNN with early fusion across uniqueness, synergy, and redundancy datasets are presented in Table 3.

The expected behaviour is that InterSHAP values for three and four modalities should be the same or similar to those for two modalities, given that the dataset creation process remains unchanged except for the synergy setting. InterSHAP exhibits the expected behaviour on the uniqueness dataset, with values ranging from 0.6 ± 0.2 to 2.6 ± 0.6. Similarly, values on the redundancy setting show consistent behaviour with larger fluctuations, ranging from a minimum of 38.6 ± 0.5 to a maximum of 51.9 ± 0.3. However, for the synergy setting, InterSHAP indicates a decrease from 98% cross-modal interactions to 64% as the number of modalities increases.

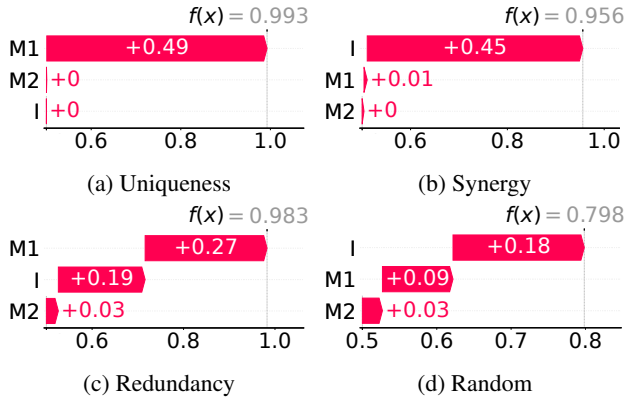
Extension to Local Method. InterSHAP can be applied to individual samples, as outlined in Section 3.4. InterSHAP values are averaged over each data point for three runs and standard deviation is calculated, as shown in Table 2.

On the HD-XOR uniqueness dataset using the XOR baseline function, the local method overestimates the synergy effect, with averages of 1.8%, compared to the global averages of 0%. Conversely, for the synergy setting, the local method underestimates the synergy effect, with averages of 96.9% locally compared to 99.7% globally. This difference is likely attributable to the masking error source. While this error averages out when considering the entire dataset, it is more pronounced at the data point level. The difference becomes more apparent when examining the FCNN results. Particularly noteworthy is the increased standard deviation, which indicates a discrepancy between data points. In contrast to the XOR function, this difference is caused by the neural network and reflects the model's behaviour.

In summary, we have shown that with a controlled model (XOR), cross-modal interactions are slightly overestimated for datasets with near 0% synergy and slightly underestimated for datasets with near 100% synergy. For an overall interaction score across the entire dataset, we recommend using the global InterSHAP. However, if the goal is to understand the interaction score at the level of individual data points, local InterSHAP is more appropriate.

Comparison with SOTA. To compare InterSHAP with SOTA cross-modal interaction scores, we calculated PID, EMAP and SHAPE, as shown in Table 2.

PID. As expected, PID measures fewer cross-modal interactions on the uniqueness compared to the synergy dataset. However, PID is not able to quantify the extent of



(a) Uniqueness (b) Synergy

(c) Redundancy (d) Random

Figure 2: Visualisation of InterSHAP using the SHAP package integration (Lundberg and Lee 2017). The results on the HD-XOR with two modalities for FCNN with early fusion are presented. The x-axes show predicted class probabilities, with baseline of $\sim$0.5 due to the binary classification. M1 denotes modality 1, M2 modality 2, and I interactions.

cross-modal interactions, unlike InterSHAP. In random and redundancy datasets PID also aligns with InterSHAP. Both show higher values for the random dataset than for the redundancy dataset, indicating that the model learned synergistic relationships.

EMAP. For clearer comparison, we present EMAP_{gap}, calculated as EMAP minus the model's F1 score (see Appendix C, Table 11), which reflects the extent of cross-modal interactions. Like other metrics, EMAP also follows the expected pattern, with less synergy observed for the uniqueness dataset than for the synergy dataset. However, EMAP measures no interaction in the model on the random dataset, because it is not performance-agnostic.

SHAPE. SHAPE exhibits the expected behaviour on the baseline XOR function for uniqueness and synergy datasets but not for the FCNN model. There are major deviations between the baseline XOR and FCNN results that are not due to model behaviour. While the uniqueness dataset shows similar values for the baseline XOR, the FCNN results differ significantly, with 16.5 ± 0.06 for uniqueness. Additionally, the redundancy dataset results do not align with the previous three metrics. We attribute these discrepancies to the masking and calculation of base values.

4.3 Visualisation

We designed InterSHAP to be compatible with the visualisation modules of the SHAP implementation (Lundberg and Lee 2017). Figure 2 illustrates the breakdown of model behaviour on the HD-XOR datasets into their components. This representation uniquely shows the relevance of modalities in absolute numbers in addition to cross-modal interactions. For instance, in Figure 2 (d), it is evident that modality 1 is, on average, more important for the prediction than the interactions. Additionally, the prediction on the random dataset is significantly less certain, with an average probability of only 79.8%, compared to 95-99% for the other datasets.

	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

Table 4: Cross-modal interactions scores on the multimodal single-cell dataset for FCNN with early, intermediate and late fusion. InterSHAP aligns with other SOTA methods, capturing the decline in cross-modal information from early to late fusion.

5 Application to Healthcare Domain

The importance of cross-modal interactions in real-world healthcare datasets is examined by analysing two publicly available multimodal datasets with two modalities with distinct input types.

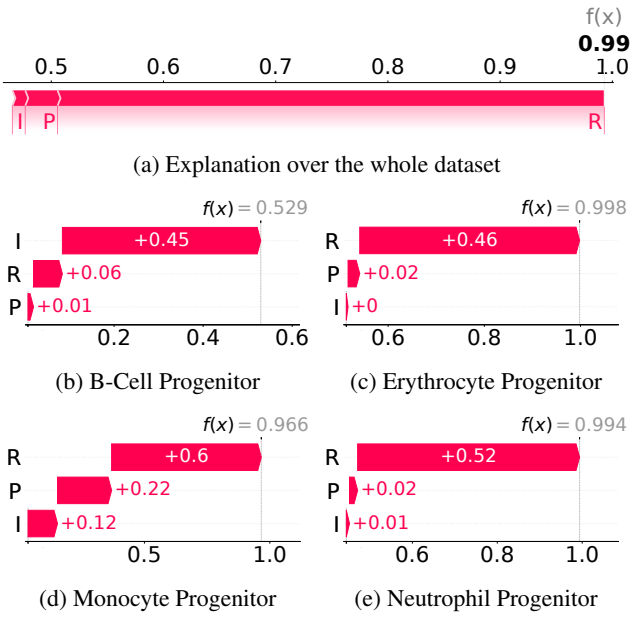
The first dataset, *multimodal single-cell* (Burkhardt et al. 2022), includes RNA and protein data from 7,132 single cells of CD34+ haematopoietic stem and progenitor cells from three donors. Its classification task spans four cell types: neutrophil progenitor (3,121), erythrocyte progenitor (3,551), B-lymphocyte progenitor (97), and monocyte progenitor (363). The data was reduced to 140 proteins and 4,000 genes and split into training (4,993), validation (1,069), and test (1,070) sets (70:15:15 ratio).

The second dataset, *MIMIC-III* (Johnson et al. 2016), contains anonymised clinical data from ICU patients at Beth Israel Deaconess Medical Center (2001–2012). It includes time series data (12 physiological measurements taken hourly over 24 hours, vector size 12×24) and static data (5 variables like age, vector size 5). The dataset has 36,212 data points, and is split into training (28,970), validation (3,621), and test (3,621) sets (80:10:10 ratio) (Liang et al. 2021). The tasks are binary ICD-9 code prediction (group 1) and 6-class mortality prediction. Preprocessing follows Liang et al. (2021).

We use FCNN with all three fusion methods described in Section 4.1 on the multimodal single-cell dataset. For the MIMIC-III dataset, we train two models from MultiBench (Liang et al. 2021): the MultiBench baseline, which combines an MLP for patient information and a GRU encoder, and the MVAE, which uses a product-of-experts mechanism to combine latent representations from an MLP and a recurrent neural network. Table 4 presents the results concerning the amount of cross-modal interactions for each dataset.

5.1 Multimodal Single-Cell

For the multimodal single-cell dataset, we trained three FCNNs using early, intermediate, and late fusion (performance details in Appendix C), and calculated InterSHAP, PID, and EMAP for each model. As shown in Table 4, InterSHAP values range from 0.4% to 1.9%, indicating minimal learned cross-modal interactions. A consistent pattern is observed whereby cross-modal interactions decrease from early to late fusion, which is in accordance with findings from synthetic datasets (see Appendix C.1).



(a) Explanation over the whole dataset

(b) B-Cell Progenitor

(c) Erythrocyte Progenitor

(d) Monocyte Progenitor

(e) Neutrophil Progenitor

Figure 3: SHAP visualisations of InterSHAP computed interactions and modality contributions derived from the FCNN early fusion model trained on the multimodal single-cell dataset. **(a)** Force plot of model's behaviour on the whole dataset, with the x-axis denoting the class probability of the highest probability class. **(b)-(d)** Breakdown by predicted class. R denotes RNA, P Protein, and I interactions.

We employ SHAP visualisations, as shown in Figure 3, to illustrate modality contributions by predicted class. A clear distinction in measured cross-modal interactions is observed between the larger classes (erythrocytes and neutrophils, comprising 93.5% of the training dataset) and the smaller classes (monocytes and B-cells). For smaller classes, the model's predictions depend largely on cross-modal interactions, while for larger classes, RNA is the most significant factor influencing the prediction.

The absence of ground truth in real-world datasets is a significant limitation, as the level of synergy or redundancy is unknown. To address this, additional cross-modal interaction scores, PID and EMAP, were computed for comparison with InterSHAP in Table 4. PID displayed behaviour similar to InterSHAP across the three fusion methods, with higher values for early fusion and lower for late fusion. However, PID indicated much higher synergy within the dataset, which is expected since PID focuses on the dataset rather than directly accounting for model behaviour. The authors also note that PID measures consistently exceed those of the dataset (Liang et al. 2023, 30). Therefore, to quantify the extent of cross-modal interactions utilised by a model in its predictions, InterSHAP is a more effective solution. In contrast, EMAP did not detect cross-modal interactions in any models. This difference likely arises because EMAP targets performance rather than directly measuring model

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

Table 5: Cross-modal interactions scores on the MIMIC III dataset for baseline model and MVAE model from MultiBench implementation. InterSHAP aligns with other SOTA methods, capturing greater cross-modal interaction in the baseline compared to MVAE, while uniquely quantifying the proportional contribution of cross-modal interactions.

behaviour. As depicted in Figure 3, smaller classes benefit significantly from interactions. However, due to their underrepresentation in the dataset, their overall impact on performance remains limited.

We can conclude that models trained on the multimodal single-cell dataset utilise minimal cross-modal interactions to solve the classification task. Exceptions are the smaller classes, for which integrating both modalities and leveraging cross-modal interactions appears more relevant.

5.2 MIMIC III

For the MIMIC dataset, we trained a baseline model and the MVAE model provided in the MultiBench (Liang et al. 2021) (details in Appendix B).

From the cross-modal interactions presented in Table 5, we observe a consistent pattern: the baseline model shows fewer cross-modal interactions than the more sophisticated MVAE model. InterSHAP values vary between tasks. For deciding if the diagnosis is in ICD-9, lower interactions are measured, ranging between 1.2% and 6.8%, compared to the more challenging task of determining mortality, which shows higher interactions around 12%.

Comparing InterSHAP with PID and EMAP reveals similar behaviour to the multimodal single-cell dataset, discussed above. PID mirrors the behaviour of InterSHAP but with higher synergy values, while EMAP does not indicate performance-relevant cross-modal scores for the ICD-9 task, but does for the mortality task. This behaviour is expected and aligns well with InterSHAP, which, as it measures performance independently, captures even small amounts of cross-modal interaction, whereas EMAP captures larger performance-relevant ones.

6 Discussion and Future Work

InterSHAP quantifies interaction strength as an interpretable percentage and works effectively with unlabelled data. Although the value of a performance-agnostic metric may be debated, InterSHAP is particularly useful for model development and debugging. For example, in our imbalanced single-cell study, smaller classes with minimal impact on the overall performance exhibited a higher reliance on cross-modal interactions, highlighting areas where modality integration could be improved. By capturing all cross-modal interactions – not just those linked to performance – InterSHAP provides deeper insights into low-performing models and can conclusively identify the absence of interactions, avoiding spurious results seen with performance-dependent methods.

Despite these benefits, we note some limitations of InterSHAP. The first is the exponential increase in runtime with the number of modalities, as all possible coalitions must be computed due to the lack of general approximation methods for SII. However, this is less critical in practice, as models are typically trained on fewer than ten modalities (Barua, Ahmed, and Begum 2023; Xu et al. 2024). Future work could explore approximation methods to address this computational challenge.

The second limitation is that we could not clearly demonstrate that InterSHAP functions as intended for synergy datasets with more than two modalities due to two factors: (1) The synergy in the training dataset decreases as the number of modalities increases. This phenomenon is attributed to the definition of XOR for more than two modalities: only one occurrence of 1 will result in the label being true, with the rest being 0. Consequently, if a 1 is present in two modalities, the label (i.e., false) is already known. (2) The F1 score decreases: 94.4 ± 0.3 for two modalities, 79.2 ± 0.2 for three modalities, and 76.4 ± 0.2 for four modalities. This indicates that the synergy present in the dataset is not fully learned, leading to the model likely relying less on cross-modal interactions for prediction. The observed decrease is likely attributable to the two factors, although the potential contribution of InterSHAP cannot be ruled out. Although InterSHAP relies on the Shapley Interaction Index (SII) and captures pairwise modality interactions, it may not fully address complex cross-modality interactions. Future research could explore alternative interaction measures than the SII. Other unimodal indices at the feature level could be also applied to quantify cross-modal interactions in datasets with more than two modalities to overcome this limitation. Potential alternatives for evaluation include Faith-SHAP (Tsai, Yeh, and Ravikumar 2023) and the Shapley-Taylor Interaction Index (Sundararajan, Dhamdhere, and Agarwal 2020).

7 Conclusion

We introduce InterSHAP, a multimodal explainability metric designed to quantify cross-modal interactions and modality contributions. InterSHAP provides a straightforward interpretation of cross-modal interactions, expressed as a percentage of overall model behaviour, which can be applied to multimodal learning tasks with any number of modalities. Besides explainability, it allows for investigating the capabilities of different model architectures in how they exploit multimodal interactions across real-world datasets. As such, InterSHAP overcomes the limitations of existing methods, akin to constraints to only two modalities and the lack of reliable, performance-independent methods. Our experiments on synthetic and real-world data demonstrate the ability of InterSHAP to accurately measure the presence (or absence) of cross-modal interactions, while allowing for global (dataset-level) and local (sample-level) explanations.

Acknowledgments

LW acknowledges support from the Hans-Böckler-Foundation. KH acknowledges support from the Gates Cambridge Trust via the Gates Cambridge Scholarship. NS and MJ acknowledge the support of the U.S. Army Medical Research and Development Command of the Department of Defense; through the FY22 Breast Cancer Research Program of the Congressionally Directed Medical Research Programs, Clinical Research Extension Award GRANT13769713. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense.

References

Anderson, B.; and Sutherland, E. 2024. Collective action for responsible AI in health. Technical Report 10, Organisation for Economic Co-Operation and Development (OECD).

Azam, K. S. F.; Ryabchykov, O.; and Bocklitz, T. 2022. A Review on Data Fusion of Multidimensional Medical and Biomedical Data. *Molecules*, 27(21): 7448.

Barua, A.; Ahmed, M. U.; and Begum, S. 2023. A Systematic Literature Review on Multimodal Machine Learning: Applications, Challenges, Gaps and Future Directions. *IEEE Access*, 11: 14804–14831.

Burkhardt, D.; Luecken, M.; Benz, A.; Holderrieth, P.; Bloom, J.; Lance, C.; Chow, A.; and Holbrook, R. 2022. Open Problems - Multimodal Single-Cell Integration.

Gat, I.; Schwartz, I.; and Schwing, A. G. 2021. Perceptual Score: What Data Modalities Does Your Model Perceive? In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 21630–21643.

Grabisch, M.; and Roubens, M. 1999. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4): 547–565.

Han, J.; Xia, T.; Spathis, D.; Bondareva, E.; Brown, C.; Chauhan, J.; Dang, T.; Grammenos, A.; Hasthanasombat, A.; Floto, A.; Cicuta, P.; and Mascolo, C. 2022. Sounds of COVID-19: exploring realistic performance of audio-based digital testing. *npj Digit. Medicine*, 5.

Hessel, J.; and Lee, L. 2020. Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 861–877. Association for Computational Linguistics.

Hu, P.; Li, X.; and Zhou, Y. 2022. SHAPE: An Unified Approach to Evaluate the Contribution and Cooperation of Individual Modalities. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 3064–3070. Vienna, Austria: International Joint Conferences on Artificial Intelligence Organization.

Ittner, J.; Bolikowski, L.; Hemker, K.; and Kennedy, R. 2021. Feature synergy, redundancy, and independence in global model explanations using shap vector decomposition. *arXiv:2107.12436*.

Johnson, A. E. W.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1). Publisher: Nature Publishing Group.

Liang, P. P.; Cheng, Y.; Fan, X.; Ling, C. K.; Nie, S.; Chen, R.; Deng, Z.; Allen, N.; Auerbach, R.; Mahmood, F.; Salakhutdinov, R. R.; and Morency, L.-P. 2023. Quantifying & Modeling Multimodal Interactions: An Information Decomposition Framework. *Advances in Neural Information Processing Systems*, 36: 27351–27393.

Liang, P. P.; Lyu, Y.; Fan, X.; Wu, Z.; Cheng, Y.; Wu, J.; Chen, L.; Wu, P.; Lee, M. A.; Zhu, Y.; Salakhutdinov, R.; and Morency, L.-P. 2021. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. In Vanschoren, J.; and Yeung, S.-K., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2022. Foundations and Recent Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *CoRR*, abs/2209.03430. ArXiv: 2209.03430.

Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; and Lee, S.-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(11): 56–67.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. v.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 4765–4774.

Parcalabescu, L.; and Frank, A. 2023. MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4032–4059.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. San Francisco California USA: ACM.

Roth, A. E. 1988. Introduction to the Shapley value. In Roth, A. E., ed., *The Shapley Value: Essays in Honor of Lloyd S. Shapley*, 1–28. Cambridge: Cambridge University Press. ISBN 978-0-521-36177-4.

Stahlschmidt, S. R.; Ulfenborg, B.; and Synnergren, J. 2022. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23(2).

Sundararajan, M.; Dhamdhere, K.; and Agarwal, A. 2020. The Shapley Taylor Interaction Index. In *Proceedings of*

the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of *Proceedings of Machine Learning Research*, 9259–9268. PMLR.

Tsai, C.-P.; Yeh, C.-K.; and Ravikumar, P. 2023. Faith-Shap: The Faithful Shapley Interaction Index. *J. Mach. Learn. Res.*, 24: 94:1–94:42.

Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.-C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; Mustafa, B.; Chowdhery, A.; Liu, Y.; Kornblith, S.; Fleet, D. J.; Mansfield, P. A.; Prakash, S.; Wong, R.; Virmani, S.; Semturs, C.; Mahdavi, S. S.; Green, B.; Dominowska, E.; Arcas, B. A. y.; Barral, J. K.; Webster, D. R.; Corrado, G. S.; Matias, Y.; Singhal, K.; Florence, P.; Karthikesalingam, A.; and Natarajan, V. 2023. Towards Generalist Biomedical AI. *CoRR*, abs/2307.14334. ArXiv: 2307.14334.

Xu, X.; Li, J.; Zhu, Z.; Zhao, L.; Wang, H.; Song, C.; Chen, Y.; Zhao, Q.; Yang, J.; and Pei, Y. 2024. A Comprehensive Review on Synergy of Multi-Modal Data and AI Technologies in Medical Diagnosis. *Bioengineering*, 11(33): 219. Publisher: Multidisciplinary Digital Publishing Institute.

Yang, G.; Ye, Q.; and Xia, J. 2022. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *An International Journal on Information Fusion*, 77: 29. Publisher: Elsevier.

Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; and others. 2023. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv:2303.00915*.