EXPLAINING MULTIMODAL CLASSIFICATION VIA INSTANCE-WISE FEATURE SELECTION

Li Tan[1] *Hong Zhou*[2] *Fei Ye*[3,*]

[1]Adobe [2]Peking University [3]University of Electronic Science and Technology of China

ABSTRACT

Multimodal classification is increasingly important in machine learning and signal processing, yet the decision processes of black-box deep multimodal models remain difficult to interpret. In this paper, we introduce a novel framework for instance-wise interpretability that integrates feature selection with causal reasoning. For each modality, our method identifies the most informative input components or elements that exert the strongest causal influence on model predictions. The causal influence is quantified through conditional mutual information (CMI) and estimated using a Cauchy-Schwarz divergence–based CMI estimator. To enable efficient optimization, we adopt continuous subset sampling under a sufficiency assumption on the encoder. Experiments on two multimodal benchmarks show that our approach accurately recovers ground-truth causal features within each modality and provides faithful explanations of black-box decisions, revealing their intrinsic strengths and limitations.

Index Terms— Multi-omics classification, Cauchy-Schwarz divergence, Gács-Körner common information

1. INTRODUCTION

Deep learning has achieved remarkable success in many domains, especially medical diagnosis. Yet, its "black-box" nature limits clinical adoption, as medical experts demand interpretability to ensure safety and trust. Explainable AI (XAI) addresses this gap by highlighting important features or regions, with methods such as Grad-CAM [1] and LIME [2] widely applied to medical images [3] and single-modality clinical data [4].

Medical diagnosis, however, often requires integrating heterogeneous sources such as genomics, clinical data, and imaging. Multimodal learning frameworks can jointly leverage these complementary modalities, leading to improved diagnostic performance [5].

Despite these advances, explaining multimodal models remains challenging, as the added complexity and heterogeneity obscure how each modality contributes to predictions.

Motivated by this challenge, we propose a novel explanation framework for multimodal classification models. Our approach identifies the most informative features in each modality that contribute to diagnostic decisions, thereby improving reliability and trustworthiness. To achieve this, we design modality-specific explainers and introduce a conditional mutual information regularization term to ensure that the extracted explanations are causally related to the model's predictions. For tractable estimation, we assume sufficient modality-specific encoders and employ Cauchy–Schwarz divergence–based information-theoretic measures [6].

Our paper mainly introduces two technical contributions:

- We propose an instance-wise feature selection approach grounded in information theory and causal inference. Using conditional mutual information as a measure of causal strength, our method highlights features with genuine causal impact on model predictions.
- We derive a practical training objective for our framework by implementing differentiable subset selection with the Gumbel–Softmax trick [7]. To enable efficient estimation of conditional mutual information, we assume encoder sufficiency [8] and leverage a recently proposed Cauchy–Schwarz divergence [6].

2. RELATED WORK

Various techniques have been developed to enhance the transparency of AI-based diagnostic models, with two of the most popular being saliency-based and perturbation-based approaches. The former is widely applied in medical imaging data to visualize important areas of interest, such as tumors or lesions, using methods like Grad-CAM [1] and SmoothGrad [9]. The latter assesses the model's response by altering data points [10] or features [11], allowing for an understanding of feature importance. A recent trend is to involve medical professionals or guidelines directly in the explanation process [12, 13]. Many of these techniques are also tailored to specific data modalities, with notable examples including Dynamask in clinical time series [14] and GNNExplainer [15] for brain network data [3].

On the other hand, with the growing availability of extensive training data and easy access to data from diverse modalities, the explainability of multimodal diagnostic models is becoming increasingly important. Most existing explanation

Corresponding author: feiye@uestc.edu.cn

methods for multimodal medical AI are simple extensions of single-modality approaches, which limits their applicability in complex multimodal settings. For example, in [4], image data and metadata are combined for skin lesion diagnosis, where Grad-CAM is applied to interpret image features, and kernel SHAP [11] is used to explain metadata.

Parallel to our work, recent studies have approached the problem of multimodal explainability at the modality level [16, 17],aiming to identify which modalities or modality interactions contribute most to a decision. In contrast, our paper focuses on feature-level explainability.

3. PROPOSED METHOD

3.1. Problem Formulation

We consider a well-trained, possibly black-box multimodal classification model $f^\star$. Given N i.i.d. multimodal observations from K modalities, we denote the dataset as $\{\{\mathbf{x}_i^k\}_{k=1}^K, y_i\}_{i=1}^N$, where $y_i = f^\star(\{\mathbf{x}_i^k\}_{k=1}^K)$ represents the model prediction from $f^\star : \{\mathcal{X}^k\}_{k=1}^K \to \mathcal{Y}$, rather than the ground-truth labels. Here, $\mathbf{x}_i^k$ denote the i-th sample in the k-th modality, and $y_i \in \mathbb{R}^C$, where C is the number of classes. Each $\mathbf{x}_i^k$ may take different forms: a vector representation (e.g., genomic data), or a 3D volume (e.g., chest CT scans).

Our objective is to interpret the model's decisions by identifying, for each instance in each modality, the subset of input elements whose presence most significantly influences the prediction. We thereby introduce a set of K feature selection networks that deterministically map $X^k \mapsto E^k$, where E^k has the same dimensionality as X^k and serves as its explanation. Each entry of E^k is a binary variable, with value 1 indicating an informative feature and 0 indicating a non-informative one. In general, E^k is both sample-specific and modality-specific.

In this paper, we focus on a broad class of multimodal classification models where each modality is equipped with its own modality-specific encoder, and fusion is performed either in the latent representation space or at the final decision level. Such architectures have become prevalent in multimodal learning [18, 19, 5], in contrast to early fusion.

3.2. Objective and Overall Framework

Our first objective is to ensure that, after feature selection within each modality, the decision of $f^\star$ remains as close as possible to its original decision without feature selection. To quantify this difference, we employ the Kullback-Leibler (KL) divergence, which for a random variable $\mathbf{x}$ with densities p and q is defined as:

$$D_{\mathrm{KL}}(p\|q) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)\right]. \tag{1}$$

We define the following loss to regularize the closeness between the predictions of the original input $\mathbf{x}$ and the masked input $\tilde{\mathbf{x}}$ under $f^\star$, serving as our first objective (see also Fig. 1 for an illustrative example). Here, $\tilde{\mathbf{x}}$ denotes the multimodal input retaining only the informative features in each modality, i.e., $\tilde{\mathbf{x}}^k = \mathbf{x}^k \odot e^k$, where $\odot$ denotes the element-wise product. Formally, the loss is given by:

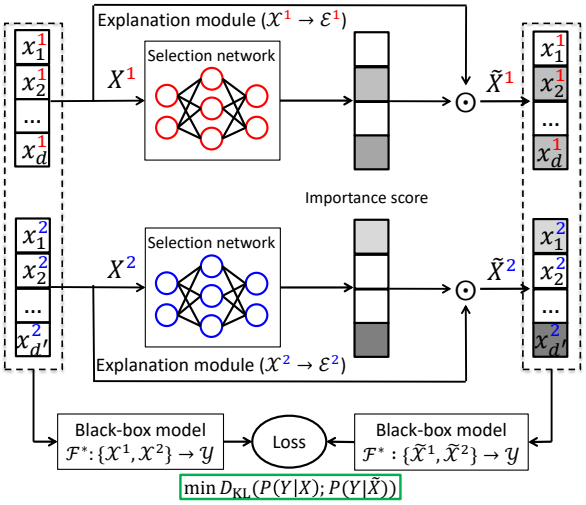


Fig. 1: Overview of our multimodal feature selection framework. We design a modality-specific explainer $\mathcal{X} \to \mathcal{E}$ to identify the top-l informative features within each modality that drive diagnostic decisions. The explainer is trained by minimizing the expected KL divergence between predictions of the black-box model $f^\star$ using the full feature set and those using only the selected subset. This plot illustrates this process for two modalities, with X^1 shown in red and X^2 in blue.

$$\mathcal{L}_1 = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[D_{\mathrm{KL}}\big(Y \,|\, X = \mathbf{x} \,\|\, Y \,|\, X = \tilde{\mathbf{x}}\big)\right]. \tag{2}$$

The KL divergence in Eq. (2) can be rewritten as:

$$\begin{aligned} &D_{\mathrm{KL}}\big(Y \,|\, X = \mathbf{x} \,\|\, Y \,|\, X = \tilde{\mathbf{x}}\big) \\ &= \mathbb{E}_{y\sim p_{Y|X=\mathbf{x}}}\left[\log\left(\frac{p_Y(y \mid \mathbf{x})}{p_Y(y \mid \tilde{\mathbf{x}})}\right)\right] \\ &= \mathbb{E}_{y\sim Y|X=\mathbf{x}}[\log p_Y(y \mid \mathbf{x}) - \log p_Y(y \mid \tilde{\mathbf{x}})] \\ &= \int_{\mathcal{Y}} p_Y(y \mid \mathbf{x})\Big[\log p_Y(y \mid \mathbf{x}) - \log p_Y\big(y \mid \tilde{\mathbf{x}}\big)\Big]\, dy, \end{aligned} \tag{3}$$

where $p_Y(\cdot|\cdot)$ denotes the appropriate conditional densities of Y (e.g., the softmax output). $\tilde{\mathbf{x}} = \{\tilde{\mathbf{x}}^k\}_{k=1}^K$ and $\tilde{\mathbf{x}}^k = \mathbf{x}^k \odot e^k$.

We write

$$\begin{aligned} &\int_{\mathcal{Y}} p_Y(y \mid \mathbf{x})\Big[\log p_Y(y \mid \mathbf{x}) - \log p_Y\big(y \mid \tilde{\mathbf{x}}\big)\Big] dy \\ &= \int_{\mathcal{Y}} p_Y(y \mid \mathbf{x})\, \log p_Y(y \mid \mathbf{x}) dy \\ &\quad - \int_{\mathcal{Y}} p_Y(y \mid \mathbf{x})\, \log p_Y\big(y \mid \tilde{\mathbf{x}}\big) dy. \end{aligned} \tag{4}$$

The first term on the RHS of Eq. (4) depends only on the pretrained black-box classifier $f^\star$, and is independent of the

masked input $\widetilde{\mathbf{x}}$. It can therefore be treated as a constant during optimization, and the final loss reduces to:

$$\mathcal{L}_1 = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[-\int_{\mathcal{Y}} p_Y(y \mid \mathbf{x}) \log p_Y\big(y \mid \widetilde{\mathbf{x}}\big) dy\right], \quad (5)$$

which is essentially the cross-entropy loss that can be rewritten as:

$$\mathcal{L}_1 = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}), e\sim\pi_\theta(\mathbf{x})}\left[-\sum_{i=1}^{C} y_i \log\big(f_i^\star(\widetilde{\mathbf{x}}, e)\big)\right], \quad (6)$$

where y_i is the i-th entry of the original prediction of $f^\star(\mathbf{x})$, and π_θ is the distribution induced by our K feature selector networks, which will be defined in the following section.

Eq. (6) alone does not guarantee that the selected features in each modality are causally related to the black-box prediction. To address this, we further introduce an additional causal regularization term.

Theorem 1. *[20] The causal strength of a subset of links in s going from input $\mathbf{x}$ to the response variable of the model y, denoted as CS_s, is given as follows:*

$$CS_s = I(\tilde{\mathbf{x}}; y \mid \bar{\mathbf{x}}). \quad (7)$$

Here, $\tilde{\mathbf{x}}$ denotes the features in set s, $\bar{\mathbf{x}}$ denotes the features not in set s, and $\mathbf{x} = \widetilde{\mathbf{x}} \cup \bar{\mathbf{x}}$.

By Theorem 1, we obtain our refined objective by incorporating a regularization term that maximizes the sum of causal strength in each modality:

$$\begin{aligned}\mathcal{L}_2 =& \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}), e\sim\pi_\theta(\mathbf{x})}\left[-\sum_{i=1}^{C} y_i \log\big(f_i^\star(\widetilde{\mathbf{x}}, e)\big) + \lambda\|e\|\right] \\ &- \beta\sum_{k=1}^{K} I(\widetilde{\mathbf{x}}^k; y|\bar{\mathbf{x}}^k),\end{aligned} \quad (8)$$

where λ, β are the regularization coefficients. Note that we further include a sparsity regularization term $\lambda\|e\|$ to encourage a minimal number of selected features in each modality, where $\|\cdot\|$ denotes the ℓ_0-norm. In practice, we approximate the ℓ_0-norm with the ℓ_1-norm to enable stable optimization.

However, directly estimating $I(\widetilde{\mathbf{x}}; y \mid \bar{\mathbf{x}})$ is generally infeasible, since $\mathbf{x}$ may be a high-dimensional vector or even complex structured data (e.g., graphs or images). In practice, we therefore optimize the following final objective:

$$\begin{aligned}\mathcal{L}_2 =& \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}), e\sim\pi_\theta(\mathbf{x})}\left[-\sum_{i=1}^{C} y_i \log\big(f_i^\star(\widetilde{\mathbf{x}}, e)\big) + \lambda\|e\|\right] \\ &- \beta\sum_{k=1}^{K} I(g^k(\widetilde{\mathbf{x}}^k); y|g^k(\bar{\mathbf{x}}^k)),\end{aligned} \quad (9)$$

where g^k denotes the modality-specific encoder of the black-box classifier $f^\star$, which maps the k-th modality input $\mathbf{x}^k$ into a vector representation. Theorem 2 justifies the validity of our approximation.

Theorem 2. *Let $(\tilde{X}, \bar{X}, Y)$ be random variables and let g be a (possibly stochastic) encoder applied to both $\tilde{X}$ and $\bar{X}$. Assume the encoder is sufficient [8], then:*

$$I\big(g(\tilde{X}); Y \mid g(\bar{X})\big) = I\big(\tilde{X}; Y \mid \bar{X}\big). \quad (10)$$

Proof. Proofs can be found in the supplementary material `https://shorturl.at/G2UX3`. □

To optimize Eq. (9), we detail the feature selection network in Section 3.3 and the estimation of conditional mutual information in Section 3.4.

3.3. Feature Selection Network

Here, we introduce the design of our built-in explainer. For an instance from the m-th modality $x_i^m \in \mathbb{R}^d$, given a fixed size l ($l > 0$), we can define $\wp_l = \{s_i^m \subset 2^d, |s_i^m| = l\}$, which represents the collection of all possible subsets of features of size l drawn from a set of d features. Our aim is to find a subset $s_i^m \in \wp_l$ that best explains or predicts the prediction made by the model.

However, a direct estimation requires summing over $\binom{d}{l}$ combinations of feature subsets, which is intractable. Therefore, we employ the Gumbel-Softmax [7, 21] to overcome this problem in a differentiable manner. Suppose we aim to approximate a categorical random variable represented as a one-hot vector in $\mathbb{R}^d$ with category probabilities $p_1, p_2, \ldots, p_d$, we start by adding a random perturbation to the log probability of each category $\log p_i$:

$$\begin{aligned}G_i &= -\log(-\log u_i) \quad \text{where } u_i \sim \text{Uniform(0,1)}, \\ C_i &= \frac{\exp\{(G_i + \log p_i)/\tau\}}{\sum_{i=1}^{d}\exp\{(G_i + \log p_i)/\tau\}},\end{aligned} \quad (11)$$

where τ is a tuning parameter for the temperature of the Gumbel-Softmax distribution. Then, we can define a Concrete random vector $C = [C_1, \cdots, C_d]$, which serves as a continuous, differential approximation of a categorical random variable represented as a one-hot vector in $\mathbb{R}^d$.

During implementation, we begin by approximating the sampling of l distinct features out of d features with the following scheme: sample one feature from d candidates independently l times. This can be implemented by drawing l independent Concrete random vectors.

3.4. Conditional Mutual Information Estimation

Note that the conditional dependence between $\tilde{x}$ and y given $\bar{x}$ can be quantified by measuring the discrepancy between the two conditional distributions $p(y \mid \bar{x})$ and $p(y \mid \bar{x}, \tilde{x})$.

In fact, when the Kullback-Leibler divergence is used as the discrepancy measure, we recover the standard definition of conditional mutual information:

$$I(\tilde{x}; y \mid \bar{x}) = \mathbb{E}\left[D_{\mathrm{KL}}\big(p(y \mid \bar{x}, \tilde{x}) \,\|\, p(y \mid \bar{x})\big)\right]. \tag{12}$$

However, KL divergence is notoriously difficult to estimate. In this work, we adopt the Cauchy-Schwarz (CS) divergence [22] as an alternative measure. The CS divergence admits an elegant non-parametric sample estimator on conditional mutual information, as demonstrated in Proposition 1.

Proposition 1. [6] Given observations $\psi = \{(\bar{\mathbf{x}}_i, \tilde{\mathbf{x}}_i, y_i)\}_{i=1}^{B}$ from a mini-batch of size B. Let $K \in \mathbb{R}^{B\times B}$, $Q \in \mathbb{R}^{B\times B}$ and $L \in \mathbb{R}^{B\times B}$ denote, respectively, the Gram or kernel matrices for the variable $\bar{\mathbf{x}}$, the concatenation of variables $\{\bar{\mathbf{x}}, \tilde{\mathbf{x}}\}$, and the variable y. That is,

$$Q_{ji} = \kappa\left(\begin{bmatrix}\bar{\mathbf{x}}_j \\ \tilde{\mathbf{x}}_j\end{bmatrix} - \begin{bmatrix}\bar{\mathbf{x}}_i \\ \tilde{\mathbf{x}}_i\end{bmatrix}\right) = \kappa(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_i)\,\kappa(\tilde{\mathbf{x}}_j - \tilde{\mathbf{x}}_i),$$

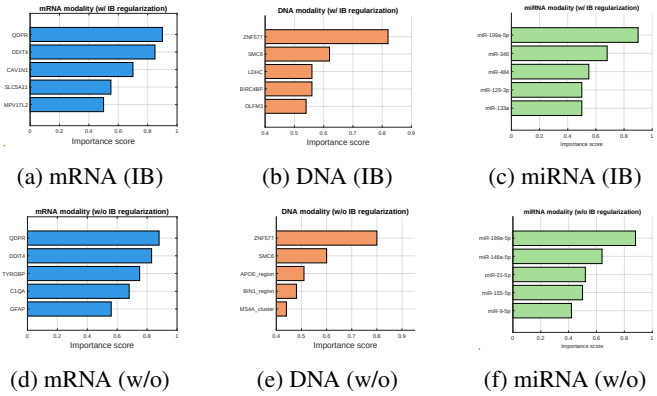
with κ represents a valid kernel function such as Gaussian.

The empirical estimation of $D_{\mathrm{CS}}(p(\mathbf{y} \mid \bar{\mathbf{x}})\,;\, p(\mathbf{y} \mid \{\bar{\mathbf{x}}, \tilde{\mathbf{x}}\}))$ is given by:

$$\begin{aligned} D_{\mathrm{CS}}(p(\mathbf{y} \mid \bar{\mathbf{x}})\,;\, p(\mathbf{y} \mid \{\bar{\mathbf{x}}, \tilde{\mathbf{x}}\})) \approx \log\left(\sum_{j=1}^{B}\left(\frac{\sum_{i=1}^{B} K_{ji}L_{ji}}{(\sum_{i=1}^{B} K_{ji})^2}\right)\right) \\ + \log\left(\sum_{j=1}^{B}\left(\frac{\sum_{i=1}^{B} Q_{ji}L_{ji}}{(\sum_{i=1}^{B} Q_{ji})^2}\right)\right) \\ - 2\log\left(\sum_{j=1}^{B}\left(\frac{\sum_{i=1}^{B} K_{ji}L_{ji}}{(\sum_{i=1}^{B} K_{ji})(\sum_{i=1}^{B} Q_{ji})}\right)\right). \end{aligned} \tag{13}$$

4. EXPERIMENTS

We evaluate our multimodal explanation approach on the ROSMAP dataset [5] (351 samples, two classes) for Alzheimer's diagnosis, which includes three modalities: mRNA, DNA methylation, and miRNA expression. We aim to explain the decisions of two multimodal classifiers: the Multimodal Information Bottleneck (MIB) [18] and a baseline model without IB regularization. Both models are trained with Adam and weight decay. On average, MIB achieves 84.9 ± 1.8 accuracy, compared to 79.2 ± 2.4 for the baseline over 30 independent runs. Fig. 2 visualizes the top-5 important features identified in each modality, with the x-axis showing selection probability. In the mRNA modality, both MIB and the baseline model select QDPR and DDIT4, which have been reported as defective in the AD brain [23]. However, the baseline additionally includes immune-heavy redundant genes such as TYROBP and C1QA [24], whereas the IB-selected features are more diverse. A similar pattern is observed in DNA: both models identify ZNF577 [25], which has clinical support, but the baseline also reuses redundant locus features



(a) mRNA (IB) (b) DNA (IB) (c) miRNA (IB)

(d) mRNA (w/o) (e) DNA (w/o) (f) miRNA (w/o)

Fig. 2: Comparison of biomarkers selected with and without IB regularization across modalities (mRNA, DNA, miRNA).

(a) digit 7 (b) digit 6

Fig. 3: Explanation visualizations of MIB (middle) and the baseline without IB (right) on query images (left) of digits 7 and 6 in the MNIST modality. MIB highlights the angular strokes of 7 and the closed loop of 6.

such as APOE and BIN1 [26]. The differences in selected features help explain why MIB outperforms the baseline, as IB regularization reduces redundancy and thereby improves generalization [27].

Next, we evaluate our approach on multimodal image data constructed from paired MNIST ($1 \times 32 \times 32$) and SVHN ($3 \times 32 \times 32$) images, following [28]. Each digit sample from one dataset is paired with 20 random samples of the same digit from the other, yielding 50,000 training and 48,930 validation examples. MIB achieves 0.91 accuracy compared to 0.89 for the baseline. As shown in Fig 3, MIB attends to critical digit structures such as loops and arcs, while the baseline often focuses on background pixels, leading to redundancy. This example further demonstrates that our explanations faithfully reflect model decisions.

5. CONCLUSION

We proposed an instance-wise feature selection framework for multimodal classification that leverages information theory and causal inference to explain black-box decisions. Experiments on medical and image datasets confirmed the faithfulness of the extracted explanations and showed their utility for comparing different classification models. In future work, we plan to extend our framework to larger multimodal datasets and explore additional modalities such as graphs.

6. REFERENCES

[1] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, ""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*, 2016, pp. 1135–1144.

[3] Selene Gallo et al., "Functional connectivity signatures of major depressive disorder: machine learning analysis of two multicenter neuroimaging studies," *Molecular Psychiatry*, vol. 28, no. 7, pp. 3013–3022, 2023.

[4] Sutong Wang, Yunqiang Yin, Dujuan Wang, Yanzhang Wang, and Yaochu Jin, "Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, 2021.

[5] Tongxin Wang et al., "Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification," *Nature communications*, vol. 12, no. 1, pp. 3445, 2021.

[6] Shujian Yu, Hongming Li, Sigurd Løkse, Robert Jenssen, and José C Príncipe, "The conditional cauchy-schwarz divergence with applications to time-series data and sequential decision making," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[7] Eric Jang, Shixiang Gu, and Ben Poole, "Categorical reparameterization with gumbel-softmax," in *International Conference on Learning Representations*, 2017.

[8] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola, "What makes for good views for contrastive learning?," *Advances in neural information processing systems*, vol. 33, pp. 6827–6839, 2020.

[9] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[10] Pang Wei Koh and Percy Liang, "Understanding black-box predictions via influence functions," in *International conference on machine learning*, 2017, pp. 1885–1894.

[11] Scott M. Lundberg and Su-In Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, p. 4768–4777.

[12] Peifei Zhu and Masahiro Ogino, "Guideline-based additive explanation for computer-aided diagnosis of lung nodules," in *International Workshop on Multimodal Learning for Clinical Decision Support*, 2019, pp. 39–47.

[13] Haozhe Luo, Aurélie Pahud de Mortanges, Oana Inel, and Mauricio Reyes, "Dwarf: Disease-weighted network for attention map refinement," *arXiv preprint arXiv:2406.17032*, 2024.

[14] Jonathan Crabbé and Mihaela Van Der Schaar, "Explaining time series predictions with dynamic masks," in *International Conference on Machine Learning*, 2021, pp. 2166–2177.

[15] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," *Advances in neural information processing systems*, vol. 32, 2019.

[16] Paul Pu Liang et al., "Quantifying & modeling multimodal interactions: An information decomposition framework," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[17] Saurabh Varshneya, Antoine Ledent, Philipp Liznerski, Andriy Balinskyy, Purvanshi Mehta, Waleed Mustafa, and Marius Kloft, "Interpretable tensor fusion," *arXiv preprint arXiv:2405.04671*, 2024.

[18] Sijie Mai, Ying Zeng, and Haifeng Hu, "Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations," *IEEE TMM*.

[19] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 2, pp. 2551–2566, 2022.

[20] Pranoy Panda, Sai Srinivas Kancheti, and Vineeth N Balasubramanian, "Instance-wise causal feature selection for model interpretation," in *CVPR*, 2021, pp. 1756–1759.

[21] Seojin Bang, Pengtao Xie, Heewook Lee, Wei Wu, and Eric Xing, "Explaining a black-box by using a deep variational information bottleneck approach," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, vol. 35, pp. 11396–11404.

[22] Robert Jenssen, Jose C Principe, Deniz Erdogmus, and Torbjørn Eltoft, "The cauchy–schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels," *Journal of the Franklin Institute*, vol. 343, no. 6, pp. 614–629, 2006.

[23] Hansruedi Mathys et al., "Single-cell transcriptomic analysis of alzheimer's disease," *Nature*, vol. 570, no. 7761, pp. 332–337, 2019.

[24] Alexander H Stephan, Ben A Barres, and Beth Stevens, "The complement system: an unexpected role in synaptic pruning during development and disease," *Annual review of neuroscience*, vol. 35, pp. 369–389, 2012.

[25] Roy Lardenoije et al., "Alzheimer's disease-associated (hydroxy) methylomic changes in the brain and blood," *Clinical epigenetics*, vol. 11, pp. 1–15, 2019.

[26] Jean-Charles Lambert et al., "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease," *Nature genetics*, vol. 45, no. 12, pp. 1452–1458, 2013.

[27] Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang, "How does information bottleneck help deep learning?," in *International Conference on Machine Learning*. PMLR, 2023, pp. 16049–16096.

[28] Mohammud Junaid Bocus, Xiaoyang Wang, and Robert J Piechocki, "Streamlining multimodal data fusion in wireless communication and sensor networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 10, no. 1, pp. 252–262, 2023.