

# Credibility-Aware Multi-Modal Fusion Using Probabilistic Circuits

Sahil Sidheekh<sup>\*1</sup> Pranuthi Tenali<sup>\*1</sup> Saurabh Mathur<sup>\*1</sup> Erik Blasch<sup>2</sup> Kristian Kersting<sup>3</sup> Sriraam Natarajan<sup>1</sup>

<sup>1</sup>Erik Jonsson School of Engineering & Computer Science, The University of Texas at Dallas

<sup>2</sup>Air Force Research Laboratory, Rome, NY, USA

<sup>3</sup>Department of Computer Science, TU Darmstadt

## Abstract

We consider the problem of late multi-modal fusion for discriminative learning. Motivated by noisy, multi-source domains that require understanding the reliability of each data source, we explore the notion of *credibility* in the context of multi-modal fusion. We propose a combination function that uses probabilistic circuits (PCs) to combine predictive distributions over individual modalities. We also define a probabilistic measure to evaluate the credibility of each modality via inference queries over the PC. Our experimental evaluation demonstrates that our fusion method can reliably infer credibility while maintaining competitive performance with the state-of-the-art.

## 1 INTRODUCTION

Real-world decision-making in high-profile tasks such as healthcare requires learning and reasoning reliably by utilizing the diverse modalities of available data sources. While such multi-modal data offer rich representations and potentially multiple views of the underlying phenomena (for example, images vs blood tests in a clinical setting), they also make learning and inference quite challenging. Raw data from different sources is often noisy, incomplete, and inconsistent. This heterogeneity poses a significant obstacle to effective data fusion and analysis.

Multi-modal fusion techniques [Baltrušaitis et al., 2018] have emerged as a promising approach to combine information from multiple sources to enhance performance on discriminative learning tasks. These techniques aim to extract and integrate complementary information from different modalities, leading to more robust and reliable results. However, a crucial aspect that often remains overlooked

in multimodal fusion is the *explicit modeling of the credibility* of the information sources. In many applications, such as sensor fusion Khaleghi et al. [2013], medical diagnosis Kline et al. [2022], and financial analysis Sawhney et al. [2020], the quality and reliability of the information sources vary significantly. Distinguishing reliable sources from non-reliable sources is essential for making accurate and informed decisions. Multimodal fusion methods often assume that all sources are equally credible, which can lead to suboptimal performance or even erroneous conclusions.

Credibility-aware methods in the context of late multimodal fusion have previously used weighted average Rogova and Nimier [2004], discounting factors Elouedi et al. [2004b] and Bayesian networks Wright and Laskey [2006]. This results in models of credibility that are either too simple (as in the case of weighted averages and discounting factors) to model complex dependencies or too complex to perform tractable inference/reasoning (as in the case of general Bayesian networks or more recent deep models). We focus on **multi-modal discriminative learning and propose a late fusion method that uses Probabilistic Circuits (PCs)** Choi et al. [2020], to effectively combine the predictive distributions over individual modalities. PCs are a class of generative models that are expressive enough to model complex distributions while tractable for exact inference. Using the tractability of PCs, we define a probabilistic measure for assessing the credibility. Some salient features of our approach are that the use of PCs (1) allows for modeling uncertainty over unimodal predictive distributions effectively; (2) makes the model robust to noise and outliers; (3) enables effective handling of missing data; (4) is grounded in a robust theoretical framework; and (4) finally, makes it possible to obtain faithful estimates of credibility.

Our paper makes the following key contributions: (1) To our knowledge, we introduce the first theoretically grounded multimodal fusion with strong probabilistic semantics based on PCs; Specifically, we identify the class of PCs that are amenable to this task of credibility-aware multi-modal fusion and define their characteristics; (2) We present two

<sup>\*</sup>Equal contribution

versions of our late fusion algorithm with different characteristics; (3) We derive a theoretically grounded measure of credibility and illustrate its connection to the conditional entropy over unimodal predictive distributions, allowing for reliable late fusion; (4) Finally, we experimentally validate the efficacy of PCs in modeling complex interactions between modalities and faithfully estimating their credibility.

The rest of the paper is organized as follows: we begin with a concise overview of essential background and relevant work. Following this, we formulate the problem at hand and our PC-based fusion method, along with the architectural details and methodology for assessing credibility. We then experimentally evaluate the effectiveness of our method and finally conclude by summarizing our findings, contributions, and future work.

## 2 BACKGROUND

### 2.1 MULTI-MODAL FUSION

Multi-modal fusion [Baltrušaitis et al., 2018] involves the integration of information from diverse sources or modalities. This field harnesses the potential of combining data of various types, such as text, images, and audio, to improve decision-making, pattern recognition, and predictive modeling. There are three broad approaches to multi-modal fusion in the discriminative learning setting, namely, early fusion, intermediate fusion, and late fusion.

Early fusion approaches fuse information from multiple sources at the input level, typically ahead of feature extraction. A simple way to achieve this would be to combine raw modality features via concatenation or pooling via operations such as average, min, max, etc. [Baltrušaitis et al., 2018]. In more complex deep learning models, early fusion is typically achieved by learning joint feature spaces [Gadzicki et al., 2020]. Apart from the curse of dimensionality, feature aggregation results in the loss of information about source-specific distributions [Schulte and Routley, 2014]. This makes it difficult to infer the credibility of input sources.

In intermediate fusion, features extracted from each modality undergo further processing and transformation into a combined, higher-level representation [Jozé et al., 2019, Zhang et al., 2019, Pérez-Rúa et al., 2019]. This approach offers more flexibility compared to early fusion, as the fusion process can take into account the characteristics of each modality individually. This can benefit learning representations, which can be used for fusion even when there's information missing from certain modalities [Zhang et al., 2019]. However, inferring the credibility of individual input modality remains difficult due to the combined nature of representation used by the classifier.

On the other hand, late fusion approaches combine the in-

formation from multiple sources by making predictions on each source independently and then combining the predictions. Combining rules [Natarajan et al., 2005, Manhaeve et al., 2018] like weighted mean [Shutova et al., 2016] and Noisy-OR [Tian et al., 2020] are commonly used for late fusion. While these combining rules allow explicit modeling of the importance of each source, they assume independence of the influence of each source on the target. Late fusion in deep learning models is implemented via additional feedforward layers [Glodek et al., 2011, Ramirez et al., 2011]. This allows them to model complex correlations and influences of the sources on the target. However, this also makes it difficult to model the credibility of each source since neural network layers are opaque.

### 2.2 CREDIBILITY

Combining information from multiple, heterogeneous sources requires information fusion systems to account for the credibility of each modality's contribution [De Villiers et al., 2018]. Credibility, as distinct from reliability, focuses on the information's truthfulness, while reliability relates to the source's consistency [Blasch et al., 2013]. While human experts might estimate their information's credibility (self-confidence), automated sources require external evaluation [Blasch et al., 2014].

We follow prior works that approach the problem of accounting for source reliability in multimodal fusion from the perspective of the credibility of the information provided by the source. These works perform multimodal fusion using source-reliability coefficients learned using domain and contextual information [Nimier, 1998, Fabre et al., 2001]. In the absence of such information, an alternate approach involves learning these coefficients from data. This is achieved by minimizing the distance between a vector of beliefs resulting from fusion and a target vector from the training set [Rogova and Kasturi, 2001, Elouedi et al., 2004a]. Another data-driven method for establishing reliability is based on *separability*, wherein the average statistical separability of information classes in each source is considered [Benediktsson et al., 1990]. This category of methods i.e. learning coefficients from training data, proves useful in establishing the relative credibility of the predictions of classifiers.

### 2.3 PROBABILISTIC CIRCUITS (PCs)

Probabilistic circuits [Choi et al., 2020] are a class of generative models that represent the joint distribution over a set of random variables (say  $X$ ) using computational graphs that comprise three types of nodes - sum and product nodes as internal nodes, and simple tractable distributions at the leaves. Formally, a PC is defined as the tuple  $(G = (V, E), \theta)$  where the Directed Acyclic Graph  $G$  represents the computational graph structure and  $\theta$  is the set of learnable parameters.

ters. The output of the root node gives the joint distribution modeled by the PC, which can be recursively obtained as:

$$P_n(\mathbf{X} = \mathbf{x}) = \begin{cases} \sum_{c \in \mathbf{ch}(n)} w_c P_c(\mathbf{X} = \mathbf{x}) & n \in \text{Sum} \\ \prod_{c \in \mathbf{ch}(n)} P_c(\mathbf{X}^{\mathbf{sc}(c)} = \mathbf{x}^{\mathbf{sc}(c)}) & n \in \text{Product} \\ \psi_n(\mathbf{X} = \mathbf{x}) & n \in \text{Leaf} \end{cases}$$

where  $\mathbf{ch}(n)$  gives the children of node  $n$ ,  $\mathbf{sc}(n)$  gives the scope of node  $n$  and  $\psi_n$  is the probability density (or mass) function associated with the leaf node  $n$ .

The key advantage of PCs is that they admit tractable and often linear time inference for a variety of probabilistic queries under mild assumptions about the structure of  $G$ . In this work, we consider a subclass of PCs that are *smooth* and *decomposable* (typically called sum-product networks Poon and Domingos [2011]). A PC satisfies smoothness if the scope of each sum node is identical to the scope of each of its children. It satisfies decomposability if, for each product node, all the children have disjoint scopes. Smoothness and decomposability allow us to tractably infer marginal and conditional distributions from the learned joint.

The structure of PCs can be learned recursively via greedy heuristics [Gens and Pedro, 2013, Rooshenas and Lowd, 2014, Dang et al., 2020], or by latent-space decomposition [Adel et al., 2015]. However, structure learning can be costly for large-scale data, and recent approaches rely on random and tensorized structures that resemble deep neural models [Mauro et al., 2017, Peharz et al., 2020a,b, Sidheekh et al., 2023] to achieve state-of-the-art performance.

### 3 MULTIMODAL FUSION VIA PCs

We begin by formalizing the *noisy late multi-modal fusion setting for discriminative learning* that we focus on. Given a dataset in which features predictive of a target concept are obtained from multiple different modalities, the late fusion setting involves training an expert over each modality to estimate the unimodal predictive distribution over the target and then combining them using a fusion function (probabilistic combination function in our case) to obtain the final output. More formally,

**Given:** A dataset  $\mathcal{D} = \{(\mathbf{x}_1^i, \mathbf{x}_2^i \dots \mathbf{x}_M^i, y^i)\}_{i=1}^N$  with  $N$  data points, each with information from  $M$  different modalities, i.e. each  $\mathbf{x}_j^i \in \mathbb{R}^{d_j}$  where  $d_j$  denotes the feature dimension corresponding to modality  $j$  for the  $i^{th}$  example, and  $y^i$  denotes its target class.

**To do:** Learn a discriminative model  $\mathcal{M}$  parameterized by  $\{\theta, \phi = \{\phi_i\}_{i=1}^m\}$  that approximates the multimodal

predictive distribution over  $Y$ <sup>1</sup> as

$$\begin{aligned} P(Y|\mathbf{X}_1, \dots, \mathbf{X}_M) &\approx \mathcal{M}_{\theta, \phi}(\mathbf{X}_1, \dots, \mathbf{X}_M) \\ &= \mathcal{M}_\theta(\mathcal{M}_{\phi_1}(\mathbf{X}_1), \dots, \mathcal{M}_{\phi_M}(\mathbf{X}_M)) \end{aligned}$$

where  $\mathcal{M}_\theta$  is the fusion function, and  $\mathcal{M}_{\phi_i}$  (or  $\mathcal{M}_i$ ) is the unimodal predictor corresponding to modality  $i$ .

In several applications, data inherently comes with a degree of noise which can affect the reliability of the information provided by each modality. Different modalities often offer complementary insights into the target  $Y$ ; however in the presence of noise, they can potentially present conflicting information (an image might potentially present a conflicting finding to that of a blood test). This necessitates a fusion method that not only leverages the unique information within each modality to make accurate predictions but is also capable of evaluating the reliability of these predictions, providing a measure of each modality's credibility.

Thus, as a key contribution, we develop *a principled notion of credibility by taking a probabilistic view of the late multimodal fusion setting*. Let us denote by  $\mathcal{F}_{\phi_j}$  the true predictive distribution over target  $Y$  given modality  $j$ , i.e  $\mathcal{F}_{\phi_j} = P(Y|\mathbf{X}_j)$ . We consider the joint distribution over the unimodal predictors and the target  $Y$  and define credibility as the relative amount of information contributed by a modality to the multi-modal predictive distribution over the target  $Y$ , as follows:

**Definition 1.** The **credibility** of a modality  $j$  in predicting the target  $Y$  is defined as the divergence between the conditional distributions over  $Y$  given all unimodal predictive distributions  $\{\mathcal{F}_{\phi_i}\}_{i=1}^M$  including and excluding  $\mathcal{F}_{\phi_j}$ , i.e.

$$\mathcal{C}_j = \delta(P(Y | \{\mathcal{F}_{\phi_i}\}_{i=1}^M) || P(Y | \{\mathcal{F}_{\phi_i}\}_{i=1}^M \setminus \{\mathcal{F}_{\phi_j}\}))$$

where  $\delta$  is a divergence measure, such as the KL-Divergence. It follows that  $\mathcal{C}_j \geq 0 \forall j$ , but can be unbounded. Thus, to facilitate easy comparison across modalities, we define the **relative credibility** score  $\tilde{\mathcal{C}}$  as

$$\tilde{\mathcal{C}}_j = \frac{\mathcal{C}_j}{\sum_j \mathcal{C}_j}.$$

Note that  $0 \leq \tilde{\mathcal{C}}_j \leq 1 \forall j$  and  $\sum_j \tilde{\mathcal{C}}_j = 1$ , and is therefore a normalized and probabilistic measure for assessing the credibility of modality  $j$ .

We now outline more formally how the defined notion of credibility is related to the uncertainty over the unimodal predictive distribution. A well-established method for quantifying the uncertainty and information content within a random variable is through the concept of entropy. We present a theorem that correlates the credibility of a modality with the entropy of its predictive distribution, under mild assumptions, as defined below.

<sup>1</sup>We use uppercase to denote random variables and lowercase to denote their corresponding values.

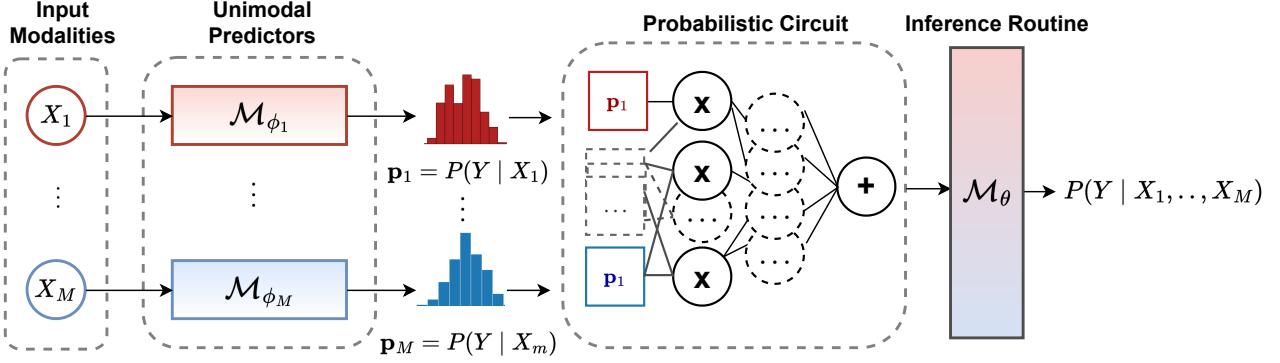


Figure 1: **Model Diagram** for our proposed PC-based fusion method. Each input modality  $X_i$  is processed by a unimodal predictor  $\mathcal{M}_{\phi_i}$  to get the corresponding predictive distribution  $\mathbf{p}_i$  over the target  $Y$ . A probabilistic circuit  $\theta$  is used to model the joint distribution over the unimodal predictive distributions and  $Y$ , and the final prediction is obtained by running an inference routine over it, governed by the form of fusion function employed ( $\mathcal{M}_\theta$ ).

**Definition 2.** A model  $\mathcal{M}$  representing a probability distribution ( $P_{\mathcal{M}}$ ) over  $n$  random variables  $\mathbf{X}$  is said to be **Marginal Dominant** if its marginals are lower bounded by the joint everywhere. *i.e.*,

$$P_{\mathcal{M}}(\mathbf{X}^{-j} = \mathbf{x}^{-j}) \geq P_{\mathcal{M}}(\mathbf{X}^{-j} = \mathbf{x}^{-j}, \mathbf{X}^j = \mathbf{x}^j) \quad \forall (\mathbf{x}^j, \mathbf{x}^{-j}) \in \text{Dom.}(\mathbf{X}^j, \mathbf{X}^{-j})$$

where  $j \subseteq \{1, \dots, n\}$  and we use the notation  $\mathbf{X}^{-j}$  to denote  $\{X_i\}_{i=1}^n \setminus \{X_k\}_{k \in j}$  and  $\text{Dom.}(\mathbf{X})$  to denote the domain set of the variables in  $\mathbf{X}$ .

**Theorem 3.1.** *The expected credibility  $\mathcal{C}^j$  of a modality  $j$  in predicting the target  $Y$ , under a Marginal Dominant distribution is lower bounded by the negative of the conditional entropy ( $\mathbb{H}$ ) of the unimodal predictive distribution of modality  $j$  over  $Y$ , given the predictive distributions of all other modalities, *i.e.**

$$\mathbb{E}[\mathcal{C}^j] \geq -\mathbb{H}(\mathcal{F}_{\phi_j} | \{\mathcal{F}_{\phi_i}\}_{i=1}^M \setminus \{\mathcal{F}_{\phi_j}\})$$

*Proof.* Deferred to the appendix.  $\square$

Intuitively, a modality less corrupted by noise and more informative of the target than others can be expected to have a lower predictive entropy. Thus, by the above theorem, we can conclude that such a modality would always have a higher assigned credibility than others. Conversely, when a modality gets corrupted by noise, its credibility score decreases. Thus the defined measure of credibility is *theoretically grounded*. Its utility becomes evident in critical domains such as healthcare, where the stakes of decision-making are high. In such contexts, credibility assessments can guide the reliance on specific expert systems or enable the discounting of modalities that are deemed unreliable.

### 3.1 PCs AS COMBINATION FUNCTIONS

We now present the details of late fusion models  $\mathcal{M}$  capable of incorporating the above-defined notion of credibility. It is clear that estimating credibility requires access to a generative model that estimates the joint distribution over  $Y$  and the unimodal predictors  $\{\mathcal{F}_j\}_{j=1}^M$ . Additionally, the generative model should support efficient and exact evaluation of both joint and conditional probability densities. Probabilistic Circuits (PCs) are one such class of generative models that can model complex distributions while supporting tractable and linear time inference of conditional and marginal distributions. Further, as we show below, the distribution modeled by a PC is Marginal Dominant under certain structural properties, making it well-suited for credibility-aware fusion.

**Theorem 3.2.** *A Probabilistic Circuit is Marginal Dominant if it is smooth, decomposable, and has leaf distributions with unimodal densities upper-bounded by unity.*

*Proof.* Deferred to the appendix.  $\square$

Thus, we define the fusion function using a PC  $\theta$  that models the joint distribution over the unimodal predictors and the target  $Y$ . More formally, given unimodal experts  $\{\mathbf{p}_j = \mathcal{M}_{\phi_j}(\mathbf{X}_j)\}_{j=1}^M$  typically parameterized as deep neural networks, the PC models the distribution  $P_\theta(Y, \mathbf{p}_1, \dots, \mathbf{p}_M)$ . The PC can be viewed as a computational graph that recursively builds a complex joint distribution by taking sums and products of simpler distributions. We use categorical leaf distributions to model the target  $Y$  and Dirichlet leaf distributions to model the unimodal predictive distributions  $\mathbf{p}_1, \dots, \mathbf{p}_M$ .

The PC  $\theta$  can be used to define the fusion function  $\mathcal{M}_\theta$  in different ways. Since a PC supports exact conditional den-

sity evaluation, a straightforward way would be to define:

$$\mathcal{M}_\theta(\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M) = P_\theta(Y|\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M)$$

We will refer to this as the **Direct-PC (DPC)** combination function. It can explicitly model complex correlations between the influence of each source on the target while still being able to reason about their credibility. The resulting late fusion method allows both predictive inference and credibility assessment as elaborated below.

**Predictive Inference.** Given a multi-modal example,  $(\mathbf{x}_1, \dots, \mathbf{x}_M)$ , we can perform predictive inference over target  $Y$  as follows:

(1) compute  $\mathbf{p}_j = \mathcal{M}_{\phi_j}(\mathbf{x}_j)$  for each modality  $j$  by evaluating the unimodal predictors  $\mathcal{M}_1, \dots, \mathcal{M}_m$ . (2) Infer the multimodal predictive distribution over  $Y$  given the unimodal distributions  $\mathbf{p}_1, \dots, \mathbf{p}_M$  by performing conditional inference:

$$\begin{aligned} P_\theta(Y | \mathbf{p}_1, \dots, \mathbf{p}_M) &= \frac{P_\theta(Y, \mathbf{p}_1, \dots, \mathbf{p}_M)}{P_\theta(\mathbf{p}_1, \dots, \mathbf{p}_M)} \\ &= \frac{P_\theta(Y, \mathbf{p}_1, \dots, \mathbf{p}_M)}{\sum_y P_\theta(Y=y, \mathbf{p}_1, \dots, \mathbf{p}_M)} \end{aligned}$$

**Credibility Assessment.** The credibility of a modality  $j$  can then be estimated using the PC  $\theta$  as

$$\mathcal{C}_j^\theta = \delta(P_\theta(Y|\mathbf{p}_1, \dots, \mathbf{p}_M)) || P_\theta(Y|\mathbf{p}_1, \dots, \mathbf{p}_{j-1}, \mathbf{p}_{j+1}, \dots, \mathbf{p}_M))$$

As smooth and decomposable PCs support linear time evaluation of joint, marginal, and conditional distributions, both predictive and credibility inference can thus be achieved in linear time.

An alternative to the Direct-PC combination function, which explicitly utilizes the credibility scores would be to define the final predictive distribution as a convex sum of credibility-weighted unimodal predictive distributions. i.e:

$$\mathcal{M}_\theta(\mathbf{p}_1, \dots, \mathbf{p}_M) = \sum_{j=1}^M \left( \frac{\mathcal{C}_j^\theta}{\sum_{i=1}^M \mathcal{C}_i^\theta} \right) \mathbf{p}_j$$

We refer to this combination function as the **Credibility-Weighted Mean (CWM)**. This approach allows us to weigh the predictive distributions according to the trustworthiness of the source, and is useful in ensuring that the final prediction reflects the most reliable and pertinent information available. Figure 1 illustrates the overall architecture of our credibility-aware late-fusion approach.

Since PCs are differentiable computational graphs, they can be easily integrated with neural unimodal predictors and learned in an end-to-end manner using backpropagation and gradient descent. We optimize the unimodal predictors to minimize the classification loss over both the unimodal

---

**Algorithm 1:** Credibility Aware Late Fusion - Learning

---

**input** : Multimodal Dataset  $\mathcal{D} = \{(\mathbf{x}_j^i, y^i)\}_{j=1}^M\}_{i=1}^N$ ,  
 Unimodal Predictors  $\{\mathcal{M}_{\phi_i}\}_{i=1}^M$   
 Probabilistic Circuit  $\theta$ ,  
 Loss function  $l$ , Divergence Measure  $\delta$   
 Learning rates  $\eta_1, \eta_2$ , #Iterations  $t_{max}$   
**output** : Optimal parameters:  $\tilde{\theta}, \{\tilde{\phi}_j\}_{j=1}^M$   
**initialize**:  $\tilde{\theta} = \theta, \{\tilde{\phi}_j = \phi_j\}_{j=1}^M, t = 1$   
**while**  $t \leq t_{max}$  **do**  
 $\{(\mathbf{x}_j^i, y^i)\}_{j=1}^B\}_{i=1}^B \sim \mathcal{D}$  ▷ Sample a mini-batch  
 For each modality  $j$  and data point  $i$   
 ▷ Compute unimodal predictive distributions  $\mathbf{p}_j^i$   
 $\mathbf{p}_j^i \leftarrow \mathcal{M}_{\tilde{\phi}_j}(\mathbf{x}_j^i)$   
 ▷ Obtain credibility scores  
 $\mathcal{C}_j^i \leftarrow \delta(P_{\tilde{\theta}}(Y|\{\mathbf{p}_k^i\}_{k=1}^M) || P_{\tilde{\theta}}(Y|\{\mathbf{p}_k^i\}_{k=1}^M \setminus \mathbf{p}_j^i))$   
 $\tilde{\mathcal{C}}_j^i \leftarrow \mathcal{C}_j^i / (\sum_{j=1}^M \mathcal{C}_j^i)$   
 ▷ Compute the final predictive distribution  
 $\mathbf{p}^i \leftarrow \sum_{j=1}^M \tilde{\mathcal{C}}_j^i \mathbf{p}_j^i$  if CWM else  $P_{\tilde{\theta}}(Y|\{\mathbf{p}_k^i\}_{k=1}^M)$   
 ▷ Compute the empirical loss  
 $L_j \leftarrow \frac{1}{B} \sum_{i=1}^B l(\mathbf{p}_j^i, y^i)$   
 $L \leftarrow \frac{1}{B} \sum_{i=1}^B l(\mathbf{p}^i, y^i) + \sum_{j=1}^M L_j$   
 ▷ Update the unimodal predictors and PC  
 $\{\tilde{\phi}_j\}_{j=1}^M \leftarrow \{\tilde{\phi}_j\}_{j=1}^M - \eta_1 \nabla_{\{\tilde{\phi}_j\}_{j=1}^M} L$   
 $\tilde{\theta} \leftarrow \tilde{\theta} - \eta_2 \nabla_{\tilde{\theta}} L + \eta_2 \nabla_{\tilde{\theta}} \sum_{i=1}^B P_{\tilde{\theta}}(y^i, \{\mathbf{p}_j^i\}_{j=1}^M)$   
 $t = t + 1$   
**end**  
**return**  $\tilde{\theta}, \{\tilde{\phi}_j\}_{j=1}^M$

---

predictions as well as the joint multimodal prediction. Further, we optimize the PC parameters to maximize the joint likelihood  $P_\theta(Y, \mathbf{p}_1, \dots, \mathbf{p}_M)$  as well as the classification loss over the joint multimodal prediction. Algorithm 1 summarizes the overall training methodology for our proposed credibility-aware late multimodal fusion using PCs.

The adoption of PCs in our approach is primarily **motivated by their tractability for probabilistic inference, which is instrumental in computing the probabilistic measures essential for assessing the credibility of each modality**. This tractability contrasts with the capabilities of more complex combination functions, such as neural networks, which, despite their potential for higher expressiveness and the ability to learn more intricate functions, do not inherently support the derivation of credibility measures. PCs on the other hand offer a balance between expressiveness and tractability. Moreover, through the process of marginalization, PCs can naturally accommodate and adjust to the absence of data from one or more modalities, preserving the integrity of the inference process without requiring imputation or other pre-processing steps. This also enhances the robustness of the fusion method, ensuring reliable performance even when faced with incomplete data.

Fusion Model	Accuracy	Precision	Recall	F1Score	AUROC
MLP	<b>72.43 ± 0.15</b>	<b>72.20 ± 0.31</b>	<b>71.97 ± 0.18</b>	<b>71.93 ± 0.23</b>	96.29 ± 0.07
Weighted Mean	66.00 ± 1.03	65.45 ± 1.28	65.48 ± 1.12	65.23 ± 0.98	95.25 ± 0.05
Noisy-OR	68.62 ± 0.17	68.06 ± 0.46	68.08 ± 0.18	67.76 ± 0.21	94.50 ± 0.16
TMC	69.95 ± 0.11	69.70 ± 0.21	69.45 ± 0.15	69.18 ± 0.14	94.99 ± 0.11
Credibility-Weighted Mean (Ours)	70.41 ± 0.15	70.32 ± 0.31	69.46 ± 0.27	68.09 ± 0.21	94.82 ± 0.16
Direct-PC (Ours)	72.18 ± 0.43	71.70 ± 0.35	71.76 ± 0.40	71.63 ± 0.36	<b>96.48 ± 0.07</b>

Table 1: Mean test performance of late fusion methods on the **AV-MNIST** dataset, ± standard deviation across 3 trials.

Fusion Model	Accuracy	Precision	Recall	F1Score	AUROC
MLP	89.66 ± 1.39	90.38 ± 1.32	89.66 ± 1.39	89.56 ± 1.38	<b>99.47 ± 0.27</b>
Weighted Mean	91.33 ± 2.25	91.97 ± 1.73	91.33 ± 2.25	91.38 ± 2.12	99.39 ± 0.33
Noisy-OR	90.83 ± 2.63	91.39 ± 2.39	90.83 ± 2.63	90.86 ± 2.56	99.41 ± 0.28
TMC	91.50 ± 3.24	92.14 ± 3.03	91.50 ± 3.24	91.47 ± 3.12	99.45 ± 0.29
Credibility-Weighted Mean (Ours)	<b>92.49 ± 1.41</b>	<b>94.03 ± 1.57</b>	<b>92.50 ± 1.42</b>	<b>92.49 ± 1.02</b>	99.42 ± 0.29
Direct-PC (Ours)	91.67 ± 1.02	92.42 ± 1.15	91.67 ± 1.02	91.58 ± 0.94	99.28 ± 0.40

Table 2: Mean test performance of late fusion methods on the **CUB** dataset, ± standard deviation across 3 trials.

## 4 EMPIRICAL EVALUATION

To experimentally validate the utility of the proposed approach, we conducted experiments on **four different** multimodal datasets: Caltech UCSD Birds (CUB), NYU Depth (NYUD), SUN RGB-D, and AV-MNIST, focusing on the task of multi-class classification. Overall, we designed experiments to answer the following research questions:

- (Q1) Can a PC-based combining rule efficiently capture intricate dependencies between modalities to achieve performance at par with existing methods?
- (Q2) Can the tractability of PCs be used to reliably infer credibility scores for each source modality?
- (Q3) Is the proposed credibility-aware fusion robust to noise?

**Baselines** We implemented 4 baseline fusion functions as elaborated below for comparison:

1. **Weighted Mean** combination function that defines the multimodal predictive distribution as:  $P(Y|\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M) = \sum_{i=1}^M w_i P(Y|\mathbf{X}_i)$  where  $w_i$  are learnable weights such that  $0 \leq w_i \leq 1$  and  $\sum_{i=1}^M w_i = 1$ . The constraints on the weights ensure that the combination function outputs a valid distribution.

2. **Noisy-Or** combination function that defines the multimodal predictive distribution as:

$$P(Y|\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M) = 1 - \prod_{i=1}^M (1 - P(Y|\mathbf{X}_i))$$

3. **Multi Layer Perceptron (MLP)** combination function that maps the vector of unimodal predictions  $[P(Y|\mathbf{X}_i)]_{i=1}^M$  to the multimodal predictive distribution

$P(Y|\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M)$  using a feedforward neural network having 2 hidden layers with 64 neurons.

4. **Dempster's** combination function, used in **TMC** (Han et al. [2021]) allows evidence from different sources to be combined by fusing *belief masses* and *uncertainty masses*. This rule ensures that the confidence of the final prediction is high when the input modalities are less uncertain and low when the input modalities are highly uncertain. When faced with different modalities that has conflicting beliefs, this combination rule only fuses the shared parts, making the final prediction dependent only on the confident modalities when some of the modalities are more uncertain.

For each of these fusion methods, we use the same backbone architecture to obtain the unimodal predictions. We train all models end to end via gradient descent and backpropagation to minimize the cross-entropy loss between the targets and predictions, using an Adam optimizer with a learning rate of 0.001 and batch size of 128.

**Datasets.** The CUB (Wah et al. [2011]) dataset comprises of 11,788 images of birds, each annotated with attribute descriptions across 200 bird categories. Following Han et al. (2021), we used a subset of the original dataset consisting of the first 10 bird categories and 336 train images, 144 validation, and 120 test images for our experiments. Deep visual features obtained from using GoogLeNet on images, and the text features extracted using doc2vec are used as two modalities.

The NYUD (Silberman et al. [2012]) is a widely used RGB-D scene recognition benchmark, containing RGB and Depth image pairs. Following previous work by Zhang et al. [2023],

Fusion Model	Accuracy	Precision	Recall	F1Score	AUROC
MLP	$63.55 \pm 0.23$	$64.65 \pm 2.24$	$49.32 \pm 0.95$	$52.35 \pm 0.68$	$86.01 \pm 0.31$
Weighted Mean (WM)	$64.06 \pm 4.30$	$64.70 \pm 1.38$	$57.2 \pm 3.96$	$59.17 \pm 3.22$	$90.99 \pm 0.78$
Noisy-OR	$66.71 \pm 1.42$	$68.85 \pm 1.38$	$59.06 \pm 1.21$	$61.71 \pm 1.31$	$91.23 \pm 0.31$
TMC	$66.97 \pm 0.26$	$68.88 \pm 1.98$	$56.89 \pm 1.09$	$59.94 \pm 0.42$	$91.47 \pm 0.39$
Credibility-Weighted Mean (Ours)	<b><math>68.50 \pm 0.72</math></b>	$67.25 \pm 1.11$	<b><math>60.17 \pm 0.85</math></b>	<b><math>62.03 \pm 0.91</math></b>	<b><math>91.52 \pm 0.41</math></b>
Direct-PC (Ours)	$57.64 \pm 2.01$	$48.80 \pm 1.12$	$49.84 \pm 1.46$	$47.96 \pm 0.79$	$79.70 \pm 0.62$

Table 3: Mean test performance of late fusion methods on the **NYUD** dataset,  $\pm$  standard deviation across 3 trials.

Fusion Model	Accuracy	Precision	Recall	F1Score	AUROC
MLP	$54.55 \pm 1.04$	$46.40 \pm 0.15$	$45.59 \pm 1.03$	$43.78 \pm 0.87$	$87.19 \pm 0.38$
Weighted Mean	$51.80 \pm 2.29$	$45.72 \pm 1.98$	$42.94 \pm 0.73$	$41.59 \pm 0.31$	$90.21 \pm 0.78$
Noisy-OR	$54.30 \pm 1.55$	$46.76 \pm 1.34$	$44.26 \pm 1.11$	$43.60 \pm 0.95$	$90.57 \pm 0.40$
TMC	$50.92 \pm 1.66$	$45.21 \pm 2.25$	$42.94 \pm 0.57$	$40.84 \pm 0.76$	$89.84 \pm 0.32$
Credibility-Weighted Mean (Ours)	<b><math>57.97 \pm 1.05</math></b>	<b><math>48.88 \pm 0.70</math></b>	<b><math>46.04 \pm 0.67</math></b>	<b><math>45.71 \pm 0.71</math></b>	<b><math>91.25 \pm 0.35</math></b>
Direct-PC (Ours)	$53.46 \pm 1.31$	$41.97 \pm 0.68$	$42.60 \pm 0.83$	$40.73 \pm 0.76$	$84.34 \pm 0.53$

Table 4: Mean test performance of late fusion methods on the **SUNRGBD** dataset,  $\pm$  standard deviation across 3 trials.

we use a reorganized dataset with 1863 image pairs (795 train, 414 validation, and 654 test) corresponding to 10 classes (9 usual scenes and one "others" category). The SUNRGBD (Song et al. [2015]) is a relatively larger scene classification dataset with 10,335 RGB-depth image pairs. Following Zhang et al. [2023], we use a subset of the original dataset which contains the 19 major scene categories and 3876 train, 969 validation, and 4,659 test examples. In both the NYUD and SUNRGBD datasets, we utilized resnet18 He et al. [2015] pre-trained on ImageNet as an encoder for each modality.

AV-MNIST is a benchmark dataset designed for multimodal fusion. With 55,000 training, 5,000 validation, and 10,000 testing examples, it has two modalities: images of dimension  $28 \times 28$  depicting digits from 0 to 9, and their corresponding audio represented as spectrograms of dimension  $112 \times 112$ . Following Vielzeuf et al. [2018], we used deep neural models with the LeNet architecture to encode the input data and make predictions for each modality. Specifically, we processed the image input through a 4-layer convolutional neural network with filter sizes  $[5, 3, 3, 3]$ . Similarly, the audio input was encoded using a 6-layer convolutional neural network with filter sizes  $[5, 3, 3, 3, 3, 3]$ . For all the datasets, the encodings obtained were processed through a feed-forward neural network to obtain the unimodal predictions.

#### 4.1 PERFORMANCE EVALUATION

Table 1 summarizes the test-set performance of the baseline methods and our PC-based combination functions on the AV-MNIST dataset in terms of the classification metrics -

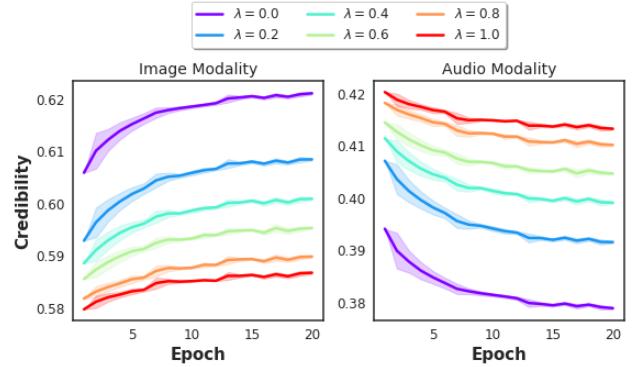


Figure 2: **Mean Validation Relative Credibility** obtained using a PC for the two modalities of the AV-MNIST dataset across training epochs. Varying degrees of noise (controlled by  $\lambda$ ) are introduced into the audio modality.

Accuracy, Precision, Recall, F1-Score, and AUC-ROC, after training for 50 epochs with early stopping. We observe that our PC-based combination functions **not only outperform simple probabilistic baselines such as Weighted Mean, Noisy-Or, and TMC on all performance metrics but also achieve performance similar to that of an MLP-based fusion** method. Table 2 summarizes the test-set performance of the baseline methods and our PC-based combination functions on the CUB dataset. We observe that our Credibility-Weighted Mean combination function achieves better performance than other models on average. As the CUB dataset is very small, we observed that complex models like MLP tend to overfit, impacting the test performance, while simpler combination functions like weighted mean and TMC achieved relatively better performance. Similar

results obtained for the NYUD and SUNRGBD datasets are summarized in Tables 3 and 4 respectively. Note that the NYUD dataset is also very small compared to the capacity of the resnet18-based models used as a backbone to encode the unimodal inputs. Here, again we can observe clearly that the relatively complex models like MLP and Direct-PC overfit, while simpler ones like Noisy-OR and Credibility-Weighted Mean generalize better. Overall, the results suggest that the PC-based methods are expressive enough to capture intricate dependencies between unimodal predictive distributions and achieve performance at par and at times even better than more complex fusion approaches.

## 4.2 CREDIBILITY EVALUATION

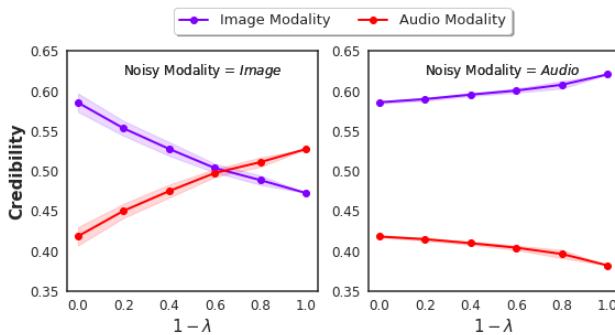


Figure 3: **Mean Test Relative Credibility** outputted by a PC for the two modalities of the AV-MNIST dataset across varying degrees of noise (controlled by  $\lambda$ ) introduced into each modality.

To empirically validate whether our PC-based late fusion method can reliably compute the credibility of each modality, we designed another experiment. We considered the AV-MNIST dataset and the Direct PC-based fusion model trained over it for 30 epochs. We introduced varying degrees of noise into one of the modalities (say  $i$ ), keeping the others fixed, and trained the PC to maximize the joint predictive likelihood. More specifically, we defined

$$\tilde{P}(Y|\mathbf{X}_i) = \lambda P(Y|\mathbf{X}_i) + (1 - \lambda)N$$

where  $N \sim \text{Dir}(\alpha)$  is a noisy probability vector sampled from a Dirichlet distribution with parameters  $\alpha$ , and  $0 \leq \lambda \leq 1$ .  $\tilde{P}(Y|\mathbf{X}_i)$  is thus a convex combination of two probability distributions and is therefore a valid distribution.  $\lambda$  controls the amount of information retained in  $\tilde{P}$  from the unimodal predictive distribution.

Note that as  $\lambda \rightarrow 0$ ,  $\tilde{P}(Y|\mathbf{X}_i) \rightarrow N$ , and thus has less predictive information about modality  $i$ . Thus, the credibility score should ideally decrease for modality  $i$  and increase for the other modalities. Figure 2 shows how the mean relative credibility outputted by the PC over the validation set varies as it is trained over the noisy unimodal distributions with noise introduced into the audio modality, for varying values of  $\lambda$ . As expected, we can see that the credibility of

the audio modality decreases as training progresses, while that of the image modality increases. Further, we can also observe that the decrease in credibility increases as  $\lambda \rightarrow 0$ . To demonstrate this correlation more evidently, we plot the Mean Relative Credibility outputted by the trained PC for each modality on the test set, for the two settings where noise is introduced into one of image/audio modalities in Figure 3. We can clearly see that in both settings, the credibility score of the noisy modality decreases as  $\lambda \rightarrow 0$ , while that of the non-noisy modality increases. Thus, the credibility score outputted by the PC is a reliable measure that is reflective of the information contributed by each modality to the final predictive distribution.

By averaging the credibility of each modality over all data points, we have so far looked at a *global measure*, and the image modality seems to have higher global credibility than audio for AV-MNIST (see  $\lambda = 1$ ). However, the credibility of each modality may differ locally for individual data points, which can also be evaluated efficiently using the PC.

## 4.3 ROBUSTNESS TO NOISE

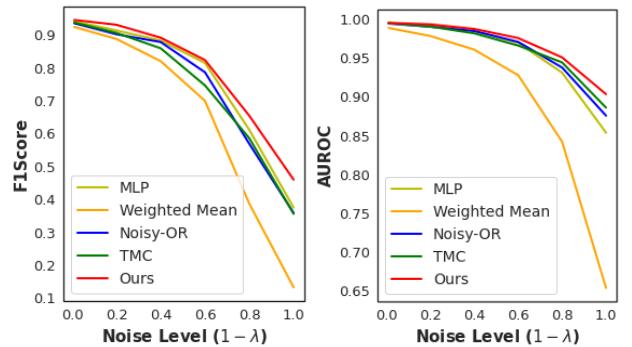


Figure 4: **Robustness to Noise**. Mean test performance of late fusion methods across varying degrees of noise.

We also evaluated the robustness of our proposed credibility-aware late fusion methodology to noisy unimodal predictive distributions. Figure 4 depicts the decline in test performance for the different fusion methods over the CUB dataset when varying degrees of noise  $\lambda$  are introduced in one of the unimodal predictive distributions. We can observe that our approach suffers the lowest decline in terms of both F1 score and AUROC, validating the robustness of our approach.

## 5 CONCLUSION

We considered the problem of late multi-modal fusion in the noisy discriminative learning setting. We derived a theoretically grounded measure of credibility and proposed probabilistic circuit-based combination functions for late-fusion that are expressive enough to model complex interactions, robust to missing modalities, and capable of making reliable

and credibility-aware predictions. Our experiments demonstrated that the proposed approach is competitive with the state-of-the-art while allowing for a principled way to infer the credibility of each modality. Scaling the approach to domains with more sources and extending the framework to allow subgroup-specific credibilities are promising directions for future research.

## Acknowledgements

The authors (SS, PT, SM and SN) gratefully acknowledge the generous support by the AFOSR award FA9550-23-1-0239, the ARO award W911NF2010224 and the DARPA Assured Neuro Symbolic Learning and Reasoning (ANSR) award HR001122S0039. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, by AFOSR, ARO, DARPA or the US government.

## References

Tameem Adel, David Balduzzi, and Ali Ghodsi. Learning the structure of sum-product networks via an svd-based algorithm. In *Conference on Uncertainty in Artificial Intelligence*, 2015.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

J.A. Benediktsson, P.H. Swain, and O.K. Ersoy. Neural network approaches versus statistical methods in classification of multisource remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 28(4):540–552, 1990. doi: 10.1109/TGRS.1990.572944.

Erik Blasch, Kathryn B Laskey, Anne-Laure Jousselme, Valentina Dragos, Paulo CG Costa, and Jean Dezert. Urref reliability versus credibility in information fusion (stanag 2511). In *Proceedings of the 16th International Conference on Information Fusion*, pages 1600–1607. IEEE, 2013.

Erik Blasch, Audun Jøsang, Jean Dezert, Paulo CG Costa, and Anne-Laure Jousselme. Urref self-confidence in information fusion trust. In *17th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2014.

YooJung Choi, Antonio Vergari, and Guy Van den Broeck. Lecture notes: Probabilistic circuits: Representation and inference. February 2020.

Meihua Dang, Antonio Vergari, and Guy Van den Broeck. Strudel: Learning structured-decomposable probabilistic circuits. In *Proceedings of the 10th International Conference on Probabilistic Graphical Models*, volume 138, pages 137–148. PMLR, 23–25 Sep 2020.

JP De Villiers, G Pavlin, AL Jousselme, S Maskell, A de Waal, K Laskey, E Blasch, and P Costa. Uncertainty representation and evaluation for modeling and decision-making in information fusion. *Journal for Advances in Information Fusion*, 13(2):198–215, 2018.

Z. Elouedi, K. Mellouli, and P. Smets. Assessing sensor reliability for multisensor data fusion within the transferable belief model. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):782–787, 2004a. doi: 10.1109/TSMCB.2003.817056.

Zied Elouedi, Khaled Mellouli, and Philippe Smets. Assessing sensor reliability for multisensor data fusion within the transferable belief model. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):782–787, 2004b.

Sophie Fabre, Alain Appriou, and Xavier Briottet. Presentation and description of two classification methods using data fusion based on sensor management. *Inf. Fusion*, 2: 49–71, 2001.

Konrad Gadzicki, Razieh Khamsehshari, and Christoph Zetsche. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd international conference on information fusion (FUSION)*, pages 1–6. IEEE, 2020.

Robert Gens and Domingos Pedro. Learning the structure of sum-product networks. In *International conference on machine learning*, pages 873–880. PMLR, 2013.

Michael Glodek, Stephan Tschechne, Georg Layher, Martin Schels, Tobias Brosch, Stefan Scherer, Markus Kächele, Miriam Schmidt, Heiko Neumann, Günther Palm, et al. Multiple classifier systems for the classification of audio-visual emotional states. In *Affective Computing and Intelligent Interaction*, pages 359–368. Springer, 2011.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296, 2019.

Bahador Khaleghi, Alaa Khamis, Fakhreddine O Karray, and Saeideh N Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information fusion*, 14(1):28–44, 2013.

Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171, 2022.

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Nicola Di Mauro, Antonio Vergari, Teresa Maria Altomare Basile, and Floriana Esposito. Fast and accurate density estimation with extremely randomized cutset networks. In *ECML/PKDD (1)*, volume 10534 of *Lecture Notes in Computer Science*, pages 203–219. Springer, 2017.

Sriram Natarajan, Prasad Tadepalli, Eric Altendorf, Thomas G Dietterich, Alan Fern, and Angelo Restificar. Learning first-order probabilistic models with combining rules. In *Proceedings of the 22nd international conference on Machine learning*, pages 609–616, 2005.

Vincent Nimier. Supervised multisensor tracking algorithm. In *9th European Signal Processing Conference (EUSIPCO 1998)*, pages 1–4, 1998.

Robert Peharz, Steven Lang, Antonio Vergari, Karl Stelzner, Alejandro Molina, Martin Trapp, Guy Van den Broeck, Kristian Kersting, and Zoubin Ghahramani. Einsum networks: Fast and scalable learning of tractable probabilistic circuits. In *ICML*, 2020a.

Robert Peharz, Antonio Vergari, Karl Stelzner, Alejandro Molina, Xiaoting Shao, Martin Trapp, Kristian Kersting, and Zoubin Ghahramani. Random sum-product networks: A simple and effective approach to probabilistic deep learning. In *UAI*, 2020b.

Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Patteux, Moez Baccouche, and Frédéric Jurie. Mfas: Multimodal fusion architecture search. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6959–6968, 2019.

Hoifung Poon and Pedro Domingos. Sum-product networks: A new deep architecture. In *UAI*, 2011.

Geovany A Ramirez, Tadas Baltrušaitis, and Louis-Philippe Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II*, pages 396–406. Springer, 2011.

G Rogova and Jyotsna Kasturi. Reinforcement learning neural network for distributed decision making. In *Proc. of the Forth Conf. on Information Fusion*, 2001.

Galina L Rogova and Vincent Nimier. Reliability in information fusion: literature survey. In *Proceedings of the seventh international conference on information fusion*, volume 2, pages 1158–1165, 2004.

Amirmohammad Rooshenas and Daniel Lowd. Learning sum-product networks with direct and indirect variable interactions. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 710–718, Beijing, China, 22–24 Jun 2014. PMLR.

Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. Multimodal multi-task financial risk forecasting. In *Proceedings of the 28th ACM international conference on multimedia*, pages 456–465, 2020.

Oliver Schulte and Kurt Routley. Aggregating predictions vs. aggregating features for relational classification. In *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 121–128. IEEE, 2014.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 160–170, 2016.

Sahil Sidheekh, Kristian Kersting, and Sriram Natarajan. Probabilistic flow circuits: Towards unified deep models for tractable probabilistic inference. In *The 39th Conference on Uncertainty in Artificial Intelligence*, 2023.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, 2012.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 567–576, 2015.

Junjiao Tian, Wesley Cheung, Nathaniel Glaser, Yen-Cheng Liu, and Zsolt Kira. Uno: Uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5716–5723. IEEE, 2020.

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset, 2011.

Edward J Wright and Kathryn Blackmond Laskey. Credibility models for multi-source fusion. In *2006 9th International Conference on Information Fusion*, pages 1–7. IEEE, 2006.

Changqing Zhang, Zongbo Han, yajie cui, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. Cpm-nets: Cross partial multi-view networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *Proceedings of the 40th International Conference on Machine Learning*, ICML’23. JMLR.org, 2023.

## APPENDIX

### A THEOREMS AND PROOFS

**Theorem A.1.** *The expected credibility  $\mathcal{C}^j$  of a modality  $j$  in predicting the target  $Y$ , under a Marginal Dominant distribution, is lower bounded by the negative of the conditional entropy ( $\mathbb{H}$ ) of the unimodal predictive distribution of modality  $j$  over  $Y$ , given the predictive distributions of all other modalities, i.e.*

$$\mathbb{E}[\mathcal{C}^j] \geq -\mathbb{H}(\mathcal{F}_{\phi_j} | \{\mathcal{F}_{\phi_i}\}_{i=1}^M \setminus \{\mathcal{F}_{\phi_j}\})$$

*Proof.* For ease, let us use the notation  $\mathbf{F} = \{\mathcal{F}_{\phi_i}\}_{i=1}^M$  and  $\mathbf{F}^{-j} = \{\mathcal{F}_{\phi_i}\}_{i=1}^M \setminus \{\mathcal{F}_{\phi_j}\}$ . We have from the definition of credibility, using KL divergence as the divergence measure,

$$\begin{aligned} \mathcal{C}^j &= \delta(P(Y|\mathbf{F}) || P(Y|\mathbf{F}^{-j})) \\ &= \sum_y P(y|\mathbf{F}) \log \frac{P(y|\mathbf{F})}{P(y|\mathbf{F}^{-j})} \\ &= \frac{1}{P(\mathbf{F})} \sum_y P(y, \mathbf{F}) \log \frac{P(y, \mathbf{F})P(\mathbf{F}^{-j})}{P(\mathbf{F})P(y, \mathbf{F}^{-j})} \\ &= \frac{1}{P(\mathbf{F})} \sum_y P(y, \mathbf{F}) \log \frac{P(y, \mathbf{F})}{P(y, \mathbf{F}^{-j})} + \log \frac{P(\mathbf{F}^{-j})}{P(\mathbf{F})} \end{aligned}$$

Now, we know that  $\log \frac{P(\mathbf{F}^{-j})}{P(\mathbf{F})} \geq 0$  as  $P$  is assumed to be Marginal Dominant. Thus, we have

$$\mathcal{C}^j \geq \frac{1}{P(\mathbf{F})} \sum_y P(y, \mathbf{F}) \log \frac{P(y, \mathbf{F})}{P(y, \mathbf{F}^{-j})}$$

Now, applying the log sum inequality  $\sum_i a_i \log \frac{a_i}{b_i} \geq \bar{a} \log \frac{\bar{a}}{\bar{b}}$  where,  $\bar{a} = \sum_i a_i$ ,  $\bar{b} = \sum_i b_i$ , and taking expectations, we get

$$\begin{aligned} \mathbb{E}[\mathcal{C}^j] &\geq \mathbb{E}\left[\frac{1}{P(\mathbf{F})} \left( \sum_y P(y, \mathbf{F}) \right) \log \frac{\sum_y P(y, \mathbf{F})}{\sum_y P(y, \mathbf{F}^{-j})} \right] \\ &= \mathbb{E}\left[\log \frac{P(\mathbf{F})}{P(\mathbf{F}^{-j})}\right] = \mathbb{E}\left[\log \frac{P(\mathbf{F}^{-j}, \mathcal{F}_{\phi_j})}{P(\mathbf{F}^{-j})}\right] \end{aligned}$$

Using the definition of conditional entropy  $\mathbb{H}(Y|X) = \mathbb{E}[-\log \frac{P(X, Y)}{P(X)}]$ , the above inequality reduces to

$$\mathcal{C}^j \geq -\mathbb{H}(\mathcal{F}_{\phi_j} | \mathbf{F}^{-j})$$

□

**Theorem A.2.** *A Probabilistic Circuit is Marginal Dominant if it is smooth, decomposable and has leaf distributions with unimodal densities upper-bounded by unity.*

*Proof.* Consider a PC  $\mathcal{M}$  representing the distribution over  $n$  variables  $\mathbf{X}$ . Without loss of generality let  $j$  denote the index of the variable being marginalized. Recall that  $\mathcal{M}$  is said to be Marginal Dominant if  $P_{\mathcal{M}}(\mathbf{X}^{-j} = \mathbf{x}^{-j}) \geq P_{\mathcal{M}}(\mathbf{X}^{-j} = \mathbf{x}^{-j}, \mathbf{X}^j = \mathbf{x}^j) \quad \forall (\mathbf{x}^j, \mathbf{x}^{-j}) \in \text{Dom.}(\mathbf{X}^j, \mathbf{X}^{-j})$ .

As PCs are recursively defined as compositions of three types of nodes - sums, products and univariate leaf distributions in the form of a rooted directed acyclic graph, we can prove by induction on the height of a PC that the introduction of each type of node preserves marginal dominance under the structural properties of smoothness, decomposability and unity bounded leaf densities.

As the base case, consider any univariate leaf node  $l$  in  $\mathcal{M}$ . We have

$$P_l(\mathbf{X} = \mathbf{x}) = \psi_l(\mathbf{X}^{\mathbf{s}e(l)} = \mathbf{x}^{\mathbf{s}e(l)})$$

where  $\psi_l$  denotes the leaf density function and  $\mathbf{sc}(l)$  denotes the scope of  $l$ . Now, if  $\mathbf{sc}(l) = j$ , then

$$P_l(\mathbf{X}^{-j} = \mathbf{x}^{-j}) = \int_{\mathbf{x}^j} P_l(\mathbf{X}^{-j} = \mathbf{x}^{-j}, \mathbf{X}^j = \mathbf{x}^j) d\mathbf{x}^j = 1 \geq P_l(\mathbf{X}^{-j} = \mathbf{x}^{-j}, \mathbf{X}^j = \mathbf{x}^j)$$

since the leaf densities are upper bounded by unity. On the other hand, if  $\mathbf{sc}(l) \neq j$ , then  $P_l(\mathbf{X}^{-j}) = P_l(\mathbf{X}^{-j}, \mathbf{X}^j)$  trivially. Thus under both cases, leaf nodes are Marginal Dominant, hence the base case is satisfied.

Now, let us assume that all nodes at height  $K - 1$  in the PC satisfies marginal dominance. We will show that all nodes at height  $K$  also satisfies marginal dominance. Note that as sums and products constitute the internal nodes in a PC any node at height  $K$  is obtained by introducing either a sum node or a product node over nodes at height  $K - 1$ . Let us consider the two cases separately.

Let  $\times \in \mathcal{M}$  denote a decomposable product node at height  $K$ . We have

$$P_{\times}(\mathbf{X} = \mathbf{x}) = \prod_{c \in \mathbf{ch}(\times)} P_c(\mathbf{X}^{\mathbf{sc}(c)} = \mathbf{x}^{\mathbf{sc}(c)})$$

where  $\mathbf{ch}(\times)$  denotes the children of  $\times$ . Thus, if  $\times$  is decomposable then  $\mathbf{X}^j$  can be present in the scope of only one of its children, say  $\mathcal{N}$ . Thus we have,

$$P_{\times}(\mathbf{X}^{-j} = \mathbf{x}^{-j}, \mathbf{X}^j = \mathbf{x}^j) = \left[ \prod_{c \in \mathbf{ch}(\times), j \notin \mathbf{sc}(c)} P_c(\mathbf{X}^{\mathbf{sc}(c)} = \mathbf{x}^{\mathbf{sc}(c)}) \right] P_{\mathcal{N}}(\mathbf{X}^{\mathbf{sc}(\mathcal{N}) \setminus j} = \mathbf{x}^{\mathbf{sc}(\mathcal{N}) \setminus j}, \mathbf{X}^j = \mathbf{x}^j)$$

Now, since  $\mathcal{N}$  is a node of height atmost  $K - 1$ , by the inductive assumption, it is Marginal Dominant. Hence,  $\forall \mathbf{x}^j \in \text{Dom.}(\mathbf{X}^j)$ , we have

$$\begin{aligned} P_{\times}(\mathbf{X}^{-j} = \mathbf{x}^{-j}) &= \left[ \prod_{c \in \mathbf{ch}(\times), j \notin \mathbf{sc}(c)} P_c(\mathbf{X}^{\mathbf{sc}(c)} = \mathbf{x}^{\mathbf{sc}(c)}) \right] P_{\mathcal{N}}(\mathbf{X}^{\mathbf{sc}(\mathcal{N}) \setminus j} = \mathbf{x}^{\mathbf{sc}(\mathcal{N}) \setminus j}) \\ &\geq \left[ \prod_{c \in \mathbf{ch}(\times), j \notin \mathbf{sc}(c)} P_c(\mathbf{X}^{\mathbf{sc}(c)} = \mathbf{x}^{\mathbf{sc}(c)}) \right] P_{\mathcal{N}}(\mathbf{X}^{\mathbf{sc}(\mathcal{N}) \setminus j} = \mathbf{x}^{\mathbf{sc}(\mathcal{N}) \setminus j}, \mathbf{X}^j = \mathbf{x}^j) \\ &= P_{\times}(\mathbf{X}^{-j} = \mathbf{x}^{-j}, \mathbf{X}^j = \mathbf{x}^j) \end{aligned}$$

Thus, since the product of non-negative terms preserves the direction of the inequality and  $\mathcal{N}$  is Marginal Dominant, the product node  $\times$  is also Marginal Dominant.

Now, let  $+$  denote a smooth sum node at height  $K$ . We have

$$P_+(\mathbf{X}^{-j} = \mathbf{x}^{-j}, \mathbf{X}^j = \mathbf{x}^j) = \sum_{c \in \mathbf{ch}(+)} w_c P_c(\mathbf{X}^{-j} = \mathbf{x}^{-j}, \mathbf{X}^j = \mathbf{x}^j)$$

where  $0 \leq w_c \leq 1 \forall w_c$  and  $\sum_{c \in \mathbf{ch}(+)} w_c = 1$ . Since each  $c \in +$  is a PC node of height atmost  $K - 1$ , it is marginal dominant by the inductive assumption. Thus we have,  $\forall \mathbf{x}^j \in \text{Dom.}(\mathbf{X}^j)$ ,

$$\begin{aligned} P_+(\mathbf{X}^{-j} = \mathbf{x}^{-j}) &= \sum_{c \in \mathbf{ch}(+)} w_c P_c(\mathbf{X}^{-j} = \mathbf{x}^{-j}) \\ &\geq \sum_{c \in \mathbf{ch}(+)} w_c P_c(\mathbf{X}^{-j} = \mathbf{x}^{-j}, \mathbf{X}^j = \mathbf{x}^j) \\ &= P_+(\mathbf{X}^{-j} = \mathbf{x}^{-j}, \mathbf{X}^j = \mathbf{x}^j) \end{aligned}$$

i.e.,  $+$  is Marginal Dominant which follows from the fact that the convex combination preserves the direction of the inequality.

Thus all nodes at height  $K$  are also marginal dominant, and by principle of mathematical induction, we can conclude that the PC is Marginal Dominant.

□

## B IMPLEMENTATION DETAILS

**Datasets.** The CUB (Wah et al. [2011]) dataset comprises of 11,788 images of birds, each annotated with attribute descriptions across 200 bird categories. Following Han et al. (2021), we used a subset of the original dataset consisting of the first 10 bird categories and 336 train images, 144 validation, and 120 test images for our experiments. Deep visual features obtained from using GoogLeNet on images, and the text features extracted using doc2vec are used as two modalities.

The NYUD (Silberman et al. [2012]) is a widely used RGB-D scene recognition benchmark, containing RGB and Depth image pairs. Following previous work by Zhang et al. [2023], we use a reorganized dataset with 1863 image pairs (795 train, 414 validation, and 654 test) corresponding to 10 classes (9 usual scenes and one "others" category). The SUNRGBD (Song et al. [2015]) is a relatively larger scene classification dataset with 10,335 RGB-depth image pairs. Following Zhang et al. [2023], we use a subset of the original dataset which contains the 19 major scene categories and 3876 train, 969 validation, and 4,659 test examples. In both the NYUD and SUNRGBD datasets, we utilized resnet18 He et al. [2015] pre-trained on ImageNet as an encoder for each modality.

AV-MNIST is a benchmark dataset designed for multimodal fusion. With 55,000 training, 5,000 validation, and 10,000 testing examples, it has two modalities: images of dimension  $28 \times 28$  depicting digits from 0 to 9, and their corresponding audio represented as spectrograms of dimension  $112 \times 112$ . Following Vielzeuf et al. [2018], we used deep neural models with the LeNet architecture to encode the input data and make predictions for each modality. Specifically, we processed the image input through a 4-layer convolutional neural network with filter sizes [5, 3, 3, 3]. Similarly, the audio input was encoded using a 6-layer convolutional neural network with filter sizes [5, 3, 3, 3, 3, 3]. For all the datasets, the encodings obtained were processed through a feedforward neural network to obtain the unimodal predictions.

## C EXPERIMENTAL SETUP

For the experiments, we utilized Intel Xeon Platinum 8167M CPU with 24 cores along with NVIDIA Tesla V100 GPUs, each with 16GB memory. Our setup included a total of 2 GPUs, enabling us to distribute the workload efficiently across CUDA cores. However, our experimental results can be reproduced using a single GPU instance of the V100 with the aforementioned configuration.

A total of 8 workers were used to load, preprocess and train the model for each of the datasets. The compute time for the experiment when run on a single GPU instance was approximately an hour for each configuration of the combination functions for the NYUD and AV-MNIST datasets whereas it took only 6 minutes for CUB dataset due to it's compact size. SUN-RGBD, on the other hand, took about 5 hours to run each configuration as it's huge in size, compared to other datasets. Memory utilization was closely monitored, and we observed an approximate average usage of 1, 9, 2 and 9 GB for CUB, NYUD, AVMNIST and SUNRGBD respectively.