Credibility-Aware Multimodal Fusion Using Probabilistic Circuits

Sahil Sidheekh[1]
The University of Texas at Dallas,
Richardson, TX, USA

Pranuthi Tenali[1]
The University of Texas at Dallas,
Richardson, TX, USA

Saurabh Mathur[1]
The University of Texas at Dallas,
Richardson, TX, USA

Erik Blasch
Air Force Research Laboratory,
Rome, NY, USA

Kristian Kersting
{ TU Darmstadt, hessian.AI },
Germany

Sriraam Natarajan
The University of Texas at Dallas,
Richardson, TX, USA

Abstract

We consider the problem of late multimodal 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 multimodal 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 being competitive with the state-of-the-art.

1 INTRODUCTION

Decision-making in many real-world domains, such as healthcare, requires reliable learning and reasoning from diverse modalities of available data sources. Although multimodal data offer rich representations and multiple views of underlying phenomena (e.g., MRIs, EHRs, and, blood tests in clinical settings), they can present significant challenges for learning and inference due to the heterogeneity of the data from these diverse views. Additionally, raw data from different sources is often noisy, incomplete, and inconsistent, thus posing a significant obstacle to effective decision-making.

Multimodal fusion methods have emerged as a promising direction to integrate such complementary information from different modalities to achieve better performance and reliability in discriminative learning tasks (Baltrusaitis et al., 2019; Atrey et al., 2010). However, a crucial aspect that often remains overlooked is the *explicit modeling of the credibility* of the information sources. In many applications, such as medical diagnosis (Kline et al., 2022), sensor fusion (Khaleghi et al., 2013), and financial analysis (Sawhney et al., 2020), the quality and reliability of the information sources vary significantly. Assuming equal credibility for all information can lead to suboptimal or even incorrect conclusions. Hence, distinguishing reliable sources from unreliable ones is vital for accurate and informed decision-making.

Existing works on reliable multimodal fusion predominantly focus on the late (or decision) fusion setting (owing to the difficulty of modeling source-specific credibility in joint data or feature representations) and have employed weighted average(Rogova and Nimier, 2004), discounting factors(Elouedi et al., 2004a), Bayesian networks(Wright and Laskey, 2006) as well as Neural Networks(Subedar et al., 2019) to combine modality-specific predictions. However, these approaches either oversimplify complex dependencies by making strong assumptions (such as linearity) or require approximations to make inference tractable (as in Bayesian and Neural networks).

We thus focus on **multimodal 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 while modeling their credibility. PCs are a class of generative models that are expressive enough to model complex dependencies while remaining *tractable for exact inference.* We use PCs to define a probabilistic measure for efficiently assessing

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

[1] These authors contributed equally.

the credibility of each modality. Additionally, explicit probabilistic modeling allows a principled way of dealing with missing, noisy, and uncertain data. Overall, we make the following key contributions:

1. To our knowledge, we introduce the first theoretically grounded multimodal fusion with strong probabilistic semantics based on PCs
2. We present two 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 an overview of the relevant background and related work, followed by the formulation of our problem at hand and the PC-based fusion method, including credibility assessment. We then experimentally evaluate the effectiveness of our method and finally conclude with a summary of our findings and future directions.

2 BACKGROUND

2.1 Multimodal Fusion

Multimodal fusion methods(Baltrusaitis et al., 2019) aim to integrate information from diverse sources and modalities, such as images, text, and audio. They exploit complementarity between different information sources to improve decision-making performance. There are three broad approaches to multimodal fusion: 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. (Baltrusaitis et al., 2019). 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.

Intermediate fusion involves processing features extracted from each modality to create a unified, higher-level representation(Joze et al., 2020; Zhang et al., 2019; Pérez-Rúa et al., 2019). This approach offers more flexibility than early fusion since it can account for the unique characteristics of each modality to a greater extent. This can improve representation learning, enabling fusion even with missing modality information (Zhang et al., 2019). However, assessing the reliability of individual input modalities remains challenging due to the combined nature of the classifier's representation.

Late fusion approaches combine information from multiple sources by independently making predictions on each source and fusing the predictions. Combining rules(Natarajan et al., 2005; Manhaeve et al., 2018) such as weighted mean(Shutova et al., 2016) and Noisy-OR(Tian et al., 2020) are commonly used for late fusion. While these rules allow explicit modeling of the importance of each source, they assume independence of the influence of each source on the target(Heckerman and Breese, 1994). Late fusion in deep learning models is implemented via additional feedforward layers (Simonyan and Zisserman, 2014; Wu et al., 2016). This allows them to model complex interactions between the sources. 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 by explicitly modeling source-specific reliability. These reliability models are either defined using domain knowledge (Nimier, 1998; Fabre et al., 2001) or are learned from training data(Rogova and Kasturi, 2001; Elouedi et al., 2004b; Benediktsson et al., 1990).

2.3 Probabilistic Circuits (PCs)

Probabilistic circuits (PCs, Choi et al. (2020)) are a class of generative models that use computational graphs to represent joint probability distributions over a set of random variables (say $\mathbf{X}$). These graphs consist of three node types: *Sum nodes* representing a weighted sum (*i.e.,* mixture) of the distributions represented by their child nodes, *Product nodes* representing a product (*i.e.,* factorization) of the distributions represented by their child nodes, and *Leaf nodes* representing simple tractable distributions, such as categorical or Gaussian distributions. Formally, a PC is defined as the tuple $\langle G, \theta \rangle$ where the rooted Directed Acyclic Graph G represents the computational graph structure and θ is the set of learnable parameters. The output of the root of G gives the joint distribution modeled by the PC,

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

where $\mathbf{ch}(n)$ gives the children of node n, $\mathbf{sc}(n)$ gives the scope of node n and ψ_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 infer marginal and conditional distributions from the PC tractably. However, imposing such structural constraints often limits their expressivity compared to unconstrained neural models.

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). These models have been shown to be highly expressive and capable of modeling complex distributions while retaining tractability for exact probabilistic inference. We refer the reader to(Sidheekh and Natarajan, 2024) for a detailed review.

3 MULTIMODAL FUSION via PCs

We begin by formalizing the *late multimodal fusion setting for discriminative learning* that we focus on in this work. 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 $\mathcal{D} = \{(\mathbf{x}_1^i, \mathbf{x}_2^i \ldots \mathbf{x}_M^i, y^i)\}_{i=1}^N$ 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 is its target.

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[1] as

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

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.

Real-world applications often involve noisy data, which can affect the reliability of different modalities. Although multiple modalities provide complementary insights into the target Y, noise can introduce conflicting information (e.g., predictions using only an MRI scan may contradict that looking only at a blood test). An ideal fusion function ($\mathcal{M}_\theta$) must not only combine the information from each modality effectively but also assess the credibility of each modality-specific prediction. 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 multimodal 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.

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

where δ is a divergence measure. We use KL-Divergence for the δ in our theoretical analysis and experiments. 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

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

[1] We use uppercase to denote random variables and lowercase to denote their corresponding values.

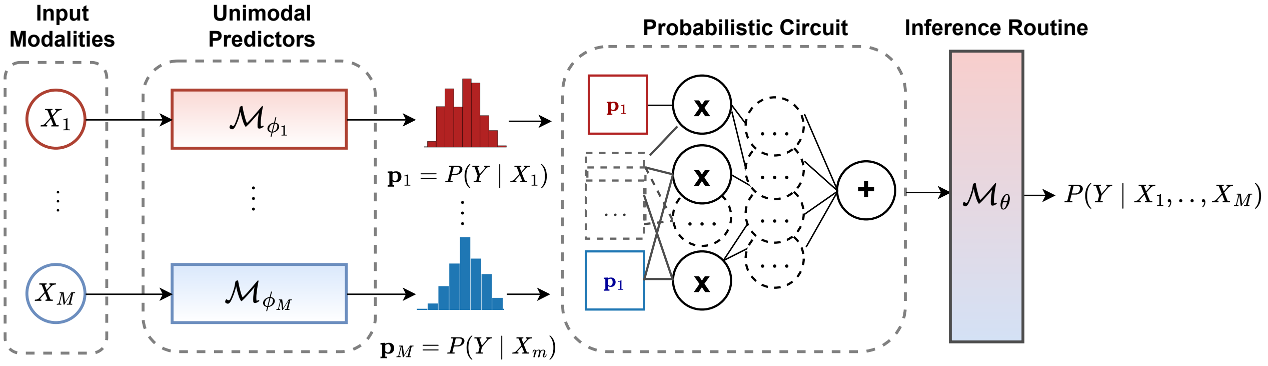

Figure 1: **Model Diagram** for our proposed fusion method. Each input modality $\mathbf{X}_i$ is processed by a unimodal predictor $\mathcal{M}_{\phi_i}$ to get the corresponding predictive distribution $\mathbf{p}_i$ over the target Y. A probabilistic circuit θ is used to model the joint distribution 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 ($\mathcal{M}_\theta$).

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 distributions. A well-established method for quantifying the uncertainty and information content within a random variable is through the concept of entropy. The following theorem establishes a direct connection between the credibility of a modality and the entropy of its predictive distribution.

Theorem 1. *The expected credibility* $(\mathbb{E}[\mathcal{C}^j])$ *of a modality* j *in predicting the target variable* Y *equals the reduction in entropy* $(\mathbb{H})$ *over the joint predictive distribution due to the inclusion of modality* j*.*

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

Proof. We provide a short sketch of the proof and defer the detailed version to the supplementary. For ease, let us use the notation $\mathbf{F} = \{\mathcal{F}_{\phi_i}\}_{i=1}^M$ and $\mathbf{F}^{-j} = \{\mathcal{F}_{\phi_i}\}_{i=1}^M \setminus \{\mathcal{F}_{\phi_j}\}$. The credibility $\mathcal{C}^j$ is defined as the KL divergence between the predictive distributions $P(Y|\mathbf{F})$ and $P(Y|\mathbf{F}^{-j})$, which can be written as:

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

Now, taking the expectation with respect to the joint distribution $P(\mathbf{F})$ over $\mathbf{F}$, we get:

$$\begin{aligned}\mathbb{E}[\mathcal{C}^j] = &\int_{\mathbf{F}} P(\mathbf{F}) \sum_y P(y|\mathbf{F}) \log P(y|\mathbf{F}) d\mathbf{F} \\ &- \int_{\mathbf{F}} P(\mathbf{F}) \sum_y P(y|\mathbf{F}) \log P(y|\mathbf{F}^{-j}) d\mathbf{F}\end{aligned}$$

The first term reduces to the expected conditional entropy of Y given the full set of unimodal predictive distributions $\mathbf{F}$ and the second term simplifies by integrating over $\mathcal{F}_{\phi_j}$ and results in the conditional entropy of Y given the full set excluding j, giving $\mathbb{E}[\mathcal{C}^j] = \mathbb{H}(Y|\mathbf{F}^{-j}) - \mathbb{H}(Y|\mathbf{F})$ □

The expected credibility score of a modality thus quantifies the reduction in uncertainty about the target variable that results from incorporating modality j. If modality j becomes corrupted or noisy, its inclusion would increase the overall uncertainty (hence the entropy $\mathbb{H}(Y|\{\mathcal{F}_{\phi_i}\}_{i=1}^M)$ increases), leading to a corresponding decrease in its credibility. Thus, the proposed measure of credibility is theoretically grounded and reflects the reliability of each modality. This is particularly valuable in high-stakes domains such as healthcare, where the consequences of decision-making are significant. In such scenarios, credibility assessments can inform the degree of reliance on specific expert systems or allow for the exclusion of unreliable modalities.

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

$\{\mathcal{F}_{\phi_j}\}_{j=1}^M$. 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 capable of representing complex distributions while supporting tractable and linear-time inference of conditional and marginal distributions.

Thus, we define the fusion function using a PC (parameterized by θ) that models the joint distribution of the unimodal predictors and the target Y. Specifically, 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 $P_\theta(Y, \mathbf{p}_1, \ldots, \mathbf{p}_M)$. We use categorical and Dirichlet leaf distributions to represent the target and the unimodal predictive distributions respectively.

The PC can be used to define the fusion function $\mathcal{M}_\theta$ in different ways. One straightforward way would be to use exact conditional density evaluation as

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

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.

Given a multimodal example, $(\mathbf{x}_1, \ldots, \mathbf{x}_M)$, 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.
2. Infer the multimodal predictive distribution over Y given the unimodal distributions $\mathbf{p}_1, \ldots, \mathbf{p}_M$ by performing conditional inference: $P_\theta(Y \mid \mathbf{p}_1, \ldots, \mathbf{p}_M) = \frac{P_\theta(Y, \mathbf{p}_1, \ldots, \mathbf{p}_M)}{P_\theta(\mathbf{p}_1, \ldots, \mathbf{p}_M)}$

The credibility of a modality j can then be estimated using the PC θ as $\mathcal{C}_j^\theta = \delta(P_\theta(Y|\mathbf{p}_1, \ldots \mathbf{p}_M)||P_\theta(Y|\mathbf{p}_1, \ldots \mathbf{p}_{j-1}, \mathbf{p}_{j+1} \cdots \mathbf{p}_M))$

An alternative to the Direct-PC combination function, that 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:

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

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.

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
    $\{(\mathbf{x}_j^i, y^i)_{j=1}^M\}_{i=1}^B \sim \mathcal{D}$      ▷ 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 $\tilde{\theta}, \{\tilde{\phi_j}\}_{j=1}^M$
```

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 predictions as well as the joint multimodal prediction. Further, we optimize the PC parameters to maximize the joint likelihood $P_\theta(Y, \mathbf{p}_1, \ldots, \mathbf{p}_M)$ 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**

Fusion Model	Accuracy	Precision	Recall	F1Score	AUROC
MLP	$\mathbf{72.43 \pm 0.15}$	$\mathbf{72.20 \pm 0.31}$	$\mathbf{71.97 \pm 0.18}$	$\mathbf{71.93 \pm 0.23}$	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	$\mathbf{96.48 \pm 0.07}$

Table 1: Mean test performance 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	$\mathbf{99.47 \pm 0.27}$
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)	$\mathbf{92.49 \pm 1.41}$	$\mathbf{94.03 \pm 1.57}$	$\mathbf{92.50 \pm 1.42}$	$\mathbf{92.49 \pm 1.02}$	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 on the **CUB** dataset, ± standard deviation across 3 trials.

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

4 EMPIRICAL EVALUATION

Our key hypothesis is that PC-based combination functions can help bridge the gap between capturing complex dependencies between modalities while allowing tractable credibility inference. *The focus of our work is not necessarily to surpass all existing methods but to demonstrate that PCs can offer comparable performance while introducing a new capability: credibility assessment.* This allows our method to be more interpretable and robust to noise and missing data, aligning with the goals of reliable machine learning as emphasized in (Rudin et al., 2024). Concretely, we aim to answer the following research questions empirically:

(Q1) Can a PC-based combining rule efficiently capture intricate dependencies between modalities to achieve performance at par with or better than 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 the presence of noise?

Datasets We used **four** multimodal, multi-class classification datasets for our experiments: Caltech UCSD Birds (CUB), NYU Depth (NYUD), SUN RGB-D, and AV-MNIST. We defer the details of the data sets and the preprocessing to the appendix.

Methods We compared our proposed methods with the following 4 fundamental late-fusion approaches

1. *Weighted Mean* combination function that defines the multimodal predictive distribution as:

$$P(Y|\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M) = \textstyle\sum_{i=1}^{M} w_i P(Y|\mathbf{X}_i)$$

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 output is a valid distribution.

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

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

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

4. *TMC* (Han et al., 2021) that uses *Dempster's* combination function which combines evidence from different sources by fusing *belief masses* and *uncertainty masses* which are obtained using an evidence-theory frame-

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

Table 3: Mean test performance on the **NYUD** dataset, ± standard deviation across 3 trials.

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

Table 4: Mean test performance on the **SUNRGBD** dataset, ± standard deviation across 3 trials.

work (Sensoy et al., 2018). It ensures high confidence in the final prediction when input modalities are less uncertain, and lowers confidence when modalities are highly uncertain. In cases of conflicting beliefs, only the shared, confident parts are fused, making the prediction dependent on the most reliable modalities.

More recent methods have introduced regularization schemes and specialized training algorithms to better handle modality conflicts and improve fusion performance (Liu et al., 2022; Xu et al., 2024). However, these schemes can be applied to all the core baselines we have considered, as they extend rather than replace the fundamental fusion mechanisms. We thus do not compare against these methods here, and we leave the integration of such advanced training schemes aimed at enhancing performance for future work.

Setup For each of the fusion methods considered, we used the same backbone architecture to obtain the unimodal predictions. We implemented the PC-based combination functions using Einsum Networks(Peharz et al., 2020a). We trained all models end-to-end using backpropagation to minimize the cross-entropy loss between the targets and predictions. We used an Adam optimizer with a learning rate of 0.001 and a batch size of 128. We defer additional implementation details to the supplementary, and our code is publicly available.[2].

(Q1: Performance) Tables 1, 2, 3, and 4 summarize the test-set performance of the baseline methods and our PC-based combination functions on the AV-MNIST, CUB, NYUD, and SUN RGB-D datasets, respectively. On the large AV-MNIST data set, we observe that **Direct-PC not only outperforms simple probabilistic baselines such as Weighted Mean, Noisy-Or, and TMC on all performance metrics but also achieves performance similar to that of an MLP-based fusion** method. On smaller datasets (CUB, NYUD, and SUN RGB-D), 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. On these data sets, the **Credibility-Weighted Mean combination function achieves better performance than other models on average.** Overall, the results suggest that the PC-based methods are expressive enough to capture intricate de-

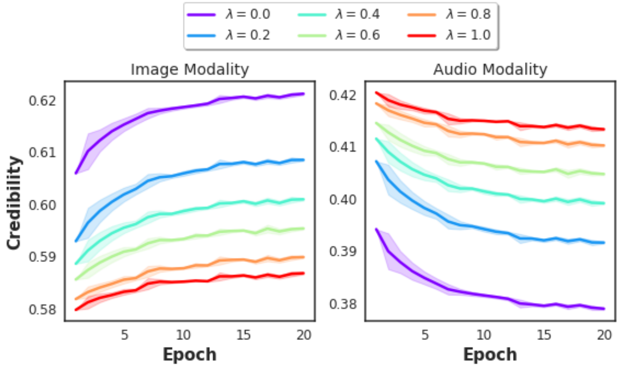

Figure 2: **Mean Validation Relative Credibility** obtained using a PC for the two modalities of the AV-MNIST dataset across training epochs. Varying degrees of noise (controlled by λ) are introduced into the audio modality.

[2] https://github.com/Pranuthi23/Credibility_MultimodalData

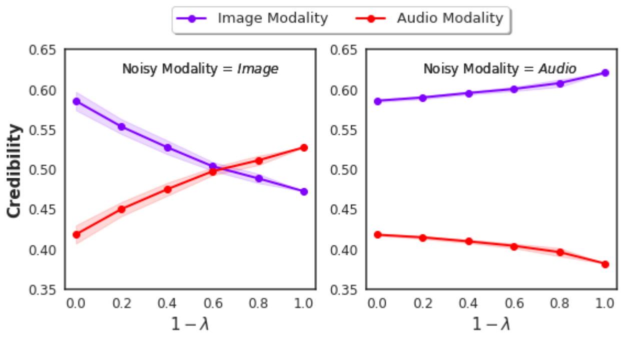

Figure 3: **Mean Test Relative Credibility** outputted by a PC for the two modalities of the AV-MNIST dataset across varying degrees of noise (controlled by λ) introduced into each modality.

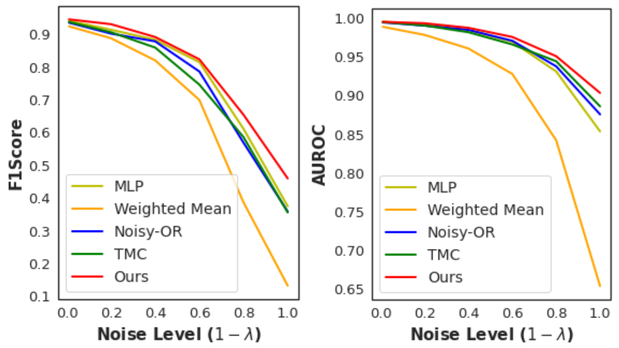

Figure 4: **Robustness to Noise.** Mean test performance of late fusion methods across varying degrees of noise.

pendencies between unimodal predictive distributions and achieve performance at par and at times even better than more complex fusion approaches.

(Q2: Credibility Evaluation) To evaluate whether our PC-based late fusion method can reliably compute the credibility of each modality, we constructed noisy versions of the AV-MNIST dataset by introducing varying degrees of noise into one of the modalities (say i), keeping the others fixed. Since the unimodal predictors are identical for each compared method, we introduce noise directly into their predictive distributions. 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 α, 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. λ controls the amount of information retained in $\tilde{P}$ from the unimodal predictive distribution. Note that as $\lambda \to 0$, $\tilde{P}(Y|\mathbf{X}_i) \to N$, 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 λ. 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 $\lambda \to 0$. 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 $\lambda \to 0$, 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*. However, the credibility of each modality may differ locally for individual data points, which can also be evaluated efficiently using the PC. For instance, the image modality in AV-MNIST seems to have higher global credibility than audio (see $\lambda = 1$) *on average* but the credibilities for each data point can vary.

(Q3: Robustness to noise) To establish the robustness of our approach to noise, we used the realistic CUB data set and constructed a noisy setup similar to the one described previously. Figure 4 illustrates the decline in test performance for the different fusion methods over the CUB dataset when varying degrees of noise λ are introduced in one of the unimodal predictive distributions. We can observe that our method (CWM), which performed best on the CUB dataset, also exhibits the smallest decline in both F1 score and AUROC, validating the robustness of our approach.

5 CONCLUSION

We considered the problem of late multimodal fusion in the noisy discriminative learning setting. We introduced a theoretically grounded measure of credibility and proposed probabilistic circuit (PC) based combination functions capable of modeling complex interactions, handling missing modalities, and making reliable, credibility-aware predictions. Our experimental results demonstrated that our methods are competitive with leading approaches, while offering a principled framework for evaluating the credibility of modalities. Our framework can also be easily adapted to other parallel paradigms like ensemble learning, multi-view learning, and federated learning, that involve learning from multiple sources, to enhance their reliability. One of the inherent limitations with the current setup is that late fusion can be less expressive than intermediate or early fusion. One potential solution to achieve intermediate fusion while supporting credibility evaluations could be introducing an additional flow of information from the intermediate features to the fusion function using conditional PCs. Further exploration into these directions, scaling the approach to domains with more modalities, and extending the framework to allow subgroup-specific credibilities will be a focus for future research.

Acknowledgements

The authors gratefully acknowledge the generous support by AFOSR award FA9550-23-1-0239, the ARO award W911NF2010224 and the DARPA Assured Neuro Symbolic Learning and Reasoning (ANSR) award HR001122S0039.

References

Adel, T., Balduzzi, D., and Ghodsi, A. (2015). Learning the structure of sum-product networks via an svd-based algorithm. In *UAI*.

Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16:345–379.

Baltrusaitis, T., Ahuja, C., and Morency, L. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443.

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

Blasch, E., Jøsang, A., Dezert, J., Costa, P. C. G., and Jousselme, A. (2014). URREF self-confidence in information fusion trust. In *FUSION*, pages 1–8. IEEE.

Blasch, E., Laskey, K. B., Jousselme, A., Dragos, V., da Costa, P. C. G., and Dezert, J. (2013). URREF reliability versus credibility in information fusion (STANAG 2511). In *FUSION*, pages 1600–1607. IEEE.

Choi, Y., Vergari, A., and Van den Broeck, G. (2020). Lecture notes: Probabilistic circuits: Representation and inference.

Dang, M., Vergari, A., and den Broeck, G. V. (2020). Strudel: Learning structured-decomposable probabilistic circuits. In *PGM*, volume 138 of *Proceedings of Machine Learning Research*, pages 137–148. PMLR.

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

Duin, R. (1998). Multiple Features. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5HC70.

Elouedi, Z., Mellouli, K., and Smets, P. (2004a). Assessing sensor reliability for multisensor data fusion within the transferable belief model. *IEEE Trans. Syst. Man Cybern. Part B*, 34(1):782–787.

Elouedi, Z., Mellouli, K., and Smets, P. (2004b). Assessing sensor reliability for multisensor data fusion within the transferable belief model. *IEEE Trans. Syst. Man Cybern. Part B*, 34(1):782–787.

Fabre, S., Appriou, A., and Briottet, X. (2001). Presentation and description of two classification methods using data fusion based on sensor management. *Inf. Fusion*, 2(1):49–71.

Gadzicki, K., Khamsehashari, R., and Zetzsche, C. (2020). Early vs late fusion in multimodal convolutional neural networks. In *FUSION*, pages 1–6. IEEE.

Gens, R. and Pedro, D. (2013). Learning the structure of sum-product networks. In *ICML*, pages 873–880. PMLR.

Han, Z., Zhang, C., Fu, H., and Zhou, J. T. (2021). Trusted multi-view classification. In *ICLR*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society.

Heckerman, D. and Breese, J. S. (1994). A new look at causal independence. In *UAI*, pages 286–292. Morgan Kaufmann.

Joze, H. R. V., Shaban, A., Iuzzolino, M. L., and Koishida, K. (2020). MMTM: multimodal transfer module for CNN fusion. In *CVPR*, pages 13286–13296. Computer Vision Foundation / IEEE.

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

Kline, A. S., Wang, H., Li, Y., Dennis, S. R., Hutch, M., Xu, Z., Wang, F., Cheng, F., and Luo, Y. (2022). Multimodal machine learning in precision health: A scoping review. *npj Digit. Medicine*, 5.

Liu, W., Yue, X., Chen, Y., and Denoeux, T. (2022). Trusted multi-view deep learning with opinion aggregation. In *AAAI*, pages 7585–7593. AAAI Press.

Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., and De Raedt, L. (2018). Deepproblog: Neural probabilistic logic programming. In *NeurIPS*, volume 31.

Mauro, N. D., Vergari, A., Basile, T. M. A., and Esposito, F. (2017). 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.

Natarajan, S., Tadepalli, P., Altendorf, E., Dietterich, T. G., Fern, A., and Restificar, A. (2005). Learning first-order probabilistic models with combining rules. In *ICML*, pages 609–616.

Nimier, V. (1998). Supervised multisensor tracking algorithm. In *EUSIPCO*, pages 1–4. IEEE.

Peharz, R., Lang, S., Vergari, A., Stelzner, K., Molina, A., Trapp, M., den Broeck, G. V., Kersting, K., and Ghahramani, Z. (2020a). Einsum networks: Fast and scalable learning of tractable probabilistic circuits. In *ICML*.

Peharz, R., Vergari, A., Stelzner, K., Molina, A., Shao, X., Trapp, M., Kersting, K., and Ghahramani, Z. (2020b). Random sum-product networks: A simple and effective approach to probabilistic deep learning. In *UAI*.

Pérez-Rúa, J., Vielzeuf, V., Pateux, S., Baccouche, M., and Jurie, F. (2019). MFAS: multimodal fusion architecture search. In *CVPR*, pages 6966–6975. Computer Vision Foundation / IEEE.

Poon, H. and Domingos, P. (2011). Sum-product networks: A new deep architecture. In *UAI*.

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

Rogova, G. L. and Nimier, V. (2004). Reliability in information fusion: literature survey. In *Proc. of the Seventh Conf. on Information Fusion*.

Rooshenas, A. and Lowd, D. (2014). Learning sum-product networks with direct and indirect variable interactions. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 710–718. JMLR.org.

Rudin, C., Zhong, C., Semenova, L., Seltzer, M. I., Parr, R., Liu, J., Katta, S., Donnelly, J., Chen, H., and Boner, Z. (2024). Position: Amazing things come from having many good models. In *ICML*. OpenReview.net.

Sawhney, R., Mathur, P., Mangal, A., Khanna, P., Shah, R. R., and Zimmermann, R. (2020). Multimodal multi-task financial risk forecasting. In *ACM Multimedia*, pages 456–465. ACM.

Schulte, O. and Routley, K. (2014). Aggregating predictions vs. aggregating features for relational classification. In *CIDM*, pages 121–128. IEEE.

Sensoy, M., Kaplan, L. M., and Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In *NeurIPS*, pages 3183–3193.

Shutova, E., Kiela, D., and Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features. In *HLT-NAACL*, pages 160–170. The Association for Computational Linguistics.

Sidheekh, S., Kersting, K., and Natarajan, S. (2023). Probabilistic flow circuits: Towards unified deep models for tractable probabilistic inference. In *UAI*.

Sidheekh, S. and Natarajan, S. (2024). Building expressive and tractable probabilistic generative models: A review. In *IJCAI*, pages 8234–8243. Survey Track.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor segmentation and support inference from RGBD images. In *ECCV (5)*, volume 7576 of *Lecture Notes in Computer Science*, pages 746–760. Springer.

Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576.

Song, S., Lichtenberg, S. P., and Xiao, J. (2015). SUN RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, pages 567–576. IEEE Computer Society.

Subedar, M., Krishnan, R., Lopez-Meyer, P., Tickoo, O., and Huang, J. (2019). Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *ICCV*, pages 6300–6309. IEEE.

Tian, J., Cheung, W., Glaser, N., Liu, Y., and Kira, Z. (2020). UNO: uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation. In *ICRA*, pages 5716–5723. IEEE.

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

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.

Wright, E. J. and Laskey, K. B. (2006). Credibility models for multi-source fusion. In *FUSION*, pages 1–7. IEEE.

Wu, D., Pigou, L., Kindermans, P., Le, N. D., Shao, L., Dambre, J., and Odobez, J. (2016). Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(8):1583–1597.

Xu, C., Si, J., Guan, Z., Zhao, W., Wu, Y., and Gao, X. (2024). Reliable conflictive multi-view learning. In *AAAI*, pages 16129–16137. AAAI Press.

Zhang, C., Han, Z., Cui, Y., Fu, H., Zhou, J. T., and Hu, Q. (2019). Cpm-nets: Cross partial multi-view networks. In *NeurIPS*, pages 557–567.

Zhang, Q., Wu, H., Zhang, C., Hu, Q., Fu, H., Zhou, J. T., and Peng, X. (2023). Provable dynamic fusion for low-quality multimodal data. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 41753–41769. PMLR.

SUPPLEMENTARY MATERIAL

In this supplementary material, we present additional implementation details, complete proofs, and extended experimental results that could not be included in the main paper due to space limitations.

Implementation Details

Datasets. We first describe the datasets used and the data preprocessing pipelines employed in depth. The CUB (Wah et al., 2011) dataset comprises $11,788$ 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. We use deep visual features obtained from using GoogLeNet, and text features extracted using doc2vec as the 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 1,863 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 4,845 train and 4,659 test examples. We further divided these 4845 train examples into 3,876 train and 969 validation instances. In both the NYUD and SUNRGBD datasets, we utilized resnet18 (He et al., 2016) 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 × 28 depicting digits from 0 to 9, and their corresponding audio represented as spectrograms of dimension 112 × 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.

We also present results obtained on the Handwritten dataset (Duin, 1998) in this supplementary. It consists of $2,000$ instances of handwritten numerals from 0 to 9 represented in terms of 6 feature sets. Following (Han et al., 2021), we divided these $2,000$ instances into $1,120$ train, 480 validation, and 400 test examples.

Models. We employed the same unimodal feature extraction and prediction backbone architecture across all the combination functions evaluated for each dataset in order to ensure a fair comparison. We used einsum networks (Peharz et al., 2020a), which offer deep tensorized implementation of probabilistic circuits in PyTorch, that can be easily trained on GPUs for the Credibility-Weighted-Mean and Direct-PC combination functions. The exact hyperparameters and configs used for all the experiments can be found here [3].

The approaches TMC (Han et al., 2021) and RCML (Xu et al., 2024) operate at the evidence level, unlike the other combination functions that combine at the prediction level. Hence, we evaluate their performance using their respective frameworks. For the remaining methods, we use the standard cross-entropy loss. QMF typically incorporates a specialized training procedure that leverages historical training trajectories, and hence trained on this regularized loss and evaluated on cross-entropy loss. It is important to note that our methods can be extended to incorporate such specialized training procedures, which could potentially enhance performance. However, such extensions are left for future investigation.

Theorems and Proofs

Theorem 2 (Expected Credibility as Entropy Reduction)**.** *The expected credibility $\mathbb{E}[\mathcal{C}^j]$ of a modality j in predicting the target Y equals the reduction in entropy ($\mathbb{H}$) over the joint predictive distribution due to the inclusion of modality j i.e.*

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

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

[3] `https://github.com/Pranuthi23/Credibility_MultimodalData`

of credibility, using KL divergence as the divergence measure,

$$\mathcal{C}^j = KL(P(Y|\mathbf{F})||P(Y|\mathbf{F}^{-j})) = \sum_y P(y|\mathbf{F}) \log \frac{P(y|\mathbf{F})}{P(y|\mathbf{F}^{-j})} = \sum_y P(y|\mathbf{F}) \log P(y|\mathbf{F}) - \sum_y P(y|\mathbf{F}) \log P(y|\mathbf{F}^{-j})$$

Taking expectations w.r.t $P(\mathbf{F})$, we get

$$\begin{aligned}
\mathbb{E}[\mathcal{C}^j] &= \sum_y \int_{\mathbf{F}} P(\mathbf{F}) P(y|\mathbf{F}) \log P(y|\mathbf{F}) d\mathbf{F} - \sum_y \int_{\mathbf{F}} P(\mathbf{F}) P(y|\mathbf{F}) \log P(y|\mathbf{F}^{-j}) d\mathbf{F} \\
&= \sum_y \int_{\mathbf{F}} P(y, \mathbf{F}) \log P(y|\mathbf{F}) d\mathbf{F} - \sum_y \int_{\mathbf{F}} P(y, \mathbf{F}) \log P(y|\mathbf{F}^{-j}) d\mathbf{F} \\
&= -\mathbb{H}(Y|\mathbf{F}) - \sum_y \int_{\mathbf{F}^{-j}} \log P(y|\mathbf{F}^{-j}) \int_{\mathcal{F}_{\phi_j}} P(y, \mathbf{F}^{-j}, \mathcal{F}_{\phi_j}) d\mathcal{F}_{\phi_j} d\mathbf{F}^{-j} \\
&= -\mathbb{H}(Y|\mathbf{F}) - \sum_y \int_{\mathbf{F}^{-j}} P(y, \mathbf{F}^{-j}) \log P(y|\mathbf{F}^{-j}) d\mathbf{F}^{-j} \\
&= \mathbb{H}(Y|\mathbf{F}^{-j}) - \mathbb{H}(Y|\mathbf{F})
\end{aligned}$$

□

Additional Results

To show that our PC-based combination function can capture intricate dependencies between modalities to achieve performance at par with existing methods, we also include a comparison of the performance of our methods with 2 additional recent late-fusion approaches described below.

- *RCML* (Xu et al., 2024) uses a conflictive opinion aggregation approach based on the framework presented in (Sensoy et al., 2018). It modifies the late fusion setup by replacing each unimodal classifier's final softmax activation with the softplus activation function to obtain evidence. It fuses these unimodal evidences using the average function.
- *QMF* (Zhang et al., 2023) uses a dynamic weighing mechanism for the combination function. This method captures uncertainty across multiple modalities using an energy score and performs fusion by weighing each modality based on this uncertainty.

As discussed in the main paper, these recent late-fusion methods incorporate *regularization strategies* and *specialized training algorithms* to address modality conflicts and enhance fusion performance. While such techniques could be integrated into our fusion framework to further improve results, we reserve this exploration for future work. Below, we focus on comparing our base models directly with these advanced methods to demonstrate that even without additional enhancements, our base approach achieves performance similar to state-of-the-art techniques.

Tables 5, 6, 7, 8, and 9 present the performance of the compared models (including QMF and RCML) on the AV-MNIST, CUB, NYUD, SUNRGBD, and Handwritten datasets respectively. On the larger AV-MNIST dataset, Direct-PC demonstrates superior performance than other simple probabilistic baselines like Weighted Mean, TMC, RCML, and Noisy-OR. We observe that while QMF marginally outperforms Direct-PC on AVMNIST, the difference in performance is statistically non-significant. However, on smaller datasets, complex models tend to overfit, resulting in Direct-PC underperforming compared to simpler models. In these scenarios, our Credibility-Weighted Mean method proves effective, either surpassing the performance of all the other methods (like in CUB and SUNRGBD) or achieving similar to that of the best-performing approach.

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.

Fusion Model	Accuracy	Precision	Recall	F1Score	AUROC
MLP	**72.43 ± 0.15**	**72.20 ± 0.31**	**71.97 ± 0.18**	**71.93 ± 0.23**	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
RCML	67.56 ± 0.29	67.15 ± 0.67	67.04 ± 0.32	66.93 ± 0.44	91.82 ± 0.15
QMF	72.38 ± 0.33	72.04 ± 0.37	71.94 ± 0.32	71.87 ± 0.40	96.56 ± 0.09
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	**96.48 ± 0.07**

Table 5: Mean test performance 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	99.47 ± 0.27
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
RCML	89.33 ± 5.01	90.04 ± 4.87	89.33 ± 5.01	89.08 ± 5.22	99.34 ± 0.32
QMF	90.50 ± 2.40	90.99 ± 2.42	90.50 ± 2.40	90.35 ± 2.40	**99.53 ± 0.40**
Credibility-Weighted Mean (Ours)	**92.49 ± 1.41**	**94.03 ± 1.57**	**92.50 ± 1.42**	**92.49 ± 1.02**	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 6: Mean test performance on the **CUB** dataset, ± standard deviation across 5 trials.

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 and Handwritten datasets due to their 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, 1, 9, 2, and 9 GB for CUB, Handwritten, NYUD, AVMNIST, and SUNRGBD respectively.

Fusion Model	Accuracy	Precision	Recall	F1Score	AUROC
MLP	63.55 ± 0.23	64.65 ± 2.24	49.32 ± 0.95	52.35 ± 0.68	86.01 ± 0.31
Weighted Mean (WM)	64.06 ± 4.30	64.70 ± 1.38	57.2 ± 3.96	59.17 ± 3.22	90.99 ± 0.78
Noisy-OR	66.71 ± 1.42	68.85 ± 1.38	59.06 ± 1.21	61.71 ± 1.31	91.23 ± 0.31
TMC	66.97 ± 0.26	68.88 ± 1.98	56.89 ± 1.09	59.94 ± 0.42	91.47 ± 0.39
RCML	**68.64 ± 2.34**	**69.46 ± 0.59**	59.84 ± 3.41	62.48 ± 2.88	90.46 ± 0.49
QMF	68.19 ± 1.99	67.39 ± 0.69	**62.20 ± 1.74**	**63.49 ± 1.44**	**92.06 ± 0.54**
Credibility-Weighted Mean (Ours)	68.50 ± 0.72	67.25 ± 1.11	60.17 ± 0.85	62.03 ± 0.91	91.52 ± 0.41
Direct-PC (Ours)	57.64 ± 2.01	48.80 ± 1.12	49.84 ± 1.46	47.96 ± 0.79	79.70 ± 0.62

Table 7: Mean test performance on the **NYUD** dataset, ± standard deviation across 3 trials.

Fusion Model	Accuracy	Precision	Recall	F1Score	AUROC
MLP	54.55 ± 1.04	46.40 ± 0.15	45.59 ± 1.03	43.78 ± 0.87	87.19 ± 0.38
Weighted Mean	51.80 ± 2.29	45.72 ± 1.98	42.94 ± 0.73	41.59 ± 0.31	90.21 ± 0.78
Noisy-OR	54.30 ± 1.55	46.76 ± 1.34	44.26 ± 1.11	43.60 ± 0.95	90.57 ± 0.40
TMC	50.92 ± 1.66	45.21 ± 2.25	42.94 ± 0.57	40.84 ± 0.76	89.84 ± 0.32
RCML	53.44 ± 1.02	44.51 ± 2.00	43.15 ± 0.66	41.77 ± 0.87	80.86 ± 0.35
QMF	57.95 ± 1.38	**51.30 ± 1.17**	**47.98 ± 0.57**	**46.91 ± 0.64**	90.09 ± 0.57
Credibility-Weighted Mean (Ours)	**57.97 ± 1.05**	48.88 ± 0.70	46.04 ± 0.67	45.71 ± 0.71	**91.25 ± 0.35**
Direct-PC (Ours)	53.46 ± 1.31	41.97 ± 0.68	42.60 ± 0.83	40.73 ± 0.76	84.34 ± 0.53

Table 8: Mean test performance on the **SUNRGBD** dataset, ± standard deviation across 3 trials.

Fusion Model	Accuracy	Precision	Recall	F1Score	AUROC
MLP	97.33 ± 0.14	97.38 ± 0.14	97.33 ± 0.14	97.33 ± 0.15	99.91 ± 0.00
Weighted Mean (WM)	97.33 ± 1.13	97.39 ± 1.10	97.33 ± 1.12	97.32 ± 1.13	99.90 ± 0.01
Noisy-OR	97.17 ± 0.95	97.24 ± 0.90	97.17 ± 0.95	97.17 ± 0.95	99.70 ± 0.06
TMC	97.41 ± 1.15	97.46 ± 1.14	97.41 ± 1.15	97.40 ± 1.15	99.92 ± 0.05
RCML	96.41 ± 1.50	96.54 ± 1.40	96.41 ± 1.51	96.42 ± 1.50	99.73 ± 0.33
QMF	**98.08 ± 1.23**	**98.14 ± 1.18**	**98.08 ± 1.23**	**98.08 ± 1.23**	**99.96 ± 0.03**
Credibility-Weighted Mean (Ours)	97.25 ± 1.15	97.32 ± 1.09	97.25 ± 1.15	97.24 ± 1.15	98.87 ± 0.66
Direct-PC (Ours)	96.67 ± 1.52	96.72 ± 1.51	96.67 ± 1.52	96.66 ± 1.53	99.74 ± 0.23

Table 9: Mean test performance on the **Handwritten** dataset, ± standard deviation across 3 trials.