On the Robustness and Reliability of Late Multi-Modal Fusion using Probabilistic Circuits

Sahil Sidheekh†, Pranuthi Tenali†*, Saurabh Mathur†*, Erik Blasch‡, Sriraam Natarajan†
†The University of Texas at Dallas
{Sahil.Sidheekh, Pranuthi.Tenali, Saurabhsanjay.Mathur, Sriraam.Natarajan}@utdallas.edu
‡Air Force Research Lab
erik.blasch.1@us.af.mil

***Abstract*—Multimodal fusion is important for building intelligent systems that exploit patterns across diverse data sources for improved decision-making. However, the reliability and robustness of these systems in safety-critical domains are often compromised by the inherent noise and incompleteness of data. Probabilistic Circuits (PCs) have recently emerged as a promising approach for late (or decision) fusion. Their strength lies in being both expressive and capable of inferring source credibility due to their ability to tractably perform exact probabilistic inference. However, their ability to handle missing data and their reliability in practical scenarios remains underexplored. This work investigates the robustness of PCs as fusion functions in scenarios with missing and noisy data; particularly by examining their impact on the calibration and reliability of the resulting classifiers. Our findings show that PCs not only enable the modeling of complex correlations across modalities but also lead to calibrated and reliable classifiers, highlighting their potential as a robust fusion mechanism in multimodal systems.**

***Index Terms*—Multi-modal fusion, reliability, robustness, probabilistic circuits**

I. Introduction

Humans effectively reason about their surroundings by utilizing complementary information from various sensory inputs. Integrating such a *multimodal* reasoning capability into intelligent systems has become increasingly important for enhancing data-driven decision-making, as many domains naturally offer data encoded as different modalities. For example, in the development of autonomous vehicles [1], the fusion of visual data from cameras placed at different angles with distance measurements from LiDAR sensors can provide a more comprehensive representation of the environment, enabling the system to make better-informed decisions. This has led to the rise of multimodal fusion as a significant subfield within artificial intelligence [2].

When deploying multimodal systems, ensuring their reliability and trustworthiness is crucial, especially in safety-critical domains. However, real-world data is often noisy and incomplete, and different modalities can vary significantly in their quality and reliability. These challenges are prevalent in many real-world domains such as sensor fusion [3, 4], medical diagnosis [5], and financial analysis [6]. Consequently, several studies have explored concepts such as reliability and credibility [7, 8] within the context of late multimodal fusion, where predictions from individual modalities are combined using functions like weighted averages [8], discounting factors [9, 10], and Bayesian Networks [11]. However, these models often face challenges in adequately balancing the simplicity required to integrate notions of credibility and the complexity needed to intricate relationships between modalities.

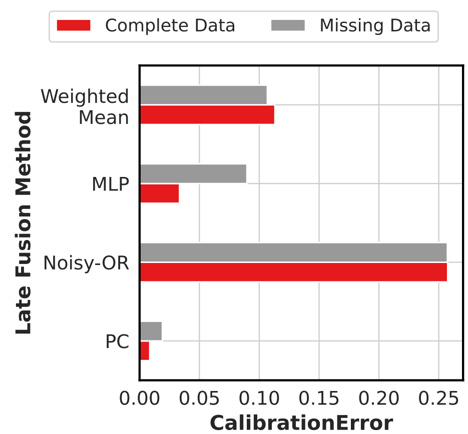


Fig. 1: Mean Expected Calibration Error achieved by late fusion methods on the test split of the Audiovisual MNIST (AVMNIST) dataset [12], when presented when complete data (in red) and when 50% of the modalities are missing (in grey). Using a Probabilistic Circuit (PC) as the fusion function helps achieve better calibrated and hence more reliable classifiers.

Recent work [13] proposes using tractable probabilistic models like Probabilistic Circuits (PCs [14]) as fusion functions, offering a promising approach for integrating unimodal predictions. PCs can capture complex correlations while enabling the inference of theoretically grounded notions of credibility. Although these models provide a probabilistic framework for fusion and can handle missing data through marginalization, there is a lack of comprehensive analysis regarding their robustness and reliability.

This study addresses this gap by examining the performance of late/decision fusion approaches in scenarios with missing and noisy data, reflecting real-world conditions. We assess the reliability of these approaches by evaluating the calibration of the resulting classifiers and their robustness to missing and noisy data. Figure 1 shows a comparison of the mean

* - Equal contribution

calibration error achieved by common late/decision fusion approaches and that using a PC in the presence of both complete and missing data. Our findings indicate that using tractable probabilistic models as fusion functions not only facilitates modeling complex correlations and inferring the credibility of source domains but also leads to calibrated and reliable classifiers that are robust to missing and noisy data.

The rest of this paper is structured as follows. First, we provide background on multimodal fusion in safety-critical domains, including the challenges of credibility and data missingness. We then formally introduce our research questions. Following this, we detail the experimental setup employed to answer these questions and present the results. Finally, we conclude by summarizing our key findings and opportunities for future work.

II. Background and Related Work

A. Multimodal fusion in safety-critical domains

Safety-critical systems, such as patient monitoring systems, need to combine information from multiple heterogeneous sources. Multimodal fusion [2, 15] offers a promising approach for handling such data from diverse sources. For effective deployment, these multimodal fusion systems need to account for the *credibility* of information from each source [13, 16] while being robust to missing data [17, 18]. For multimodal discriminative learning three three categories of fusion exist: early (signal), intermediate (feature), and late (decision) fusion.

a) Early/Signal Fusion: Early fusion approaches integrate raw data from various sources at the input level, often through aggregation operations, such as pointwise minimum, maximum, and average [2]. Deep learning-based approaches perform early fusion by learning joint feature representations [19]. However, these approaches are unable to reason about the information from each source separately [20], potentially hindering the model's ability to assess source-specific credibility.

b) Intermediate/Feature Fusion: Intermediate fusion approaches first extract features from each data source's raw data. These features are then combined to create a higher-level representation [21]–[23] that can be fed into a classifier. Unlike early fusion, intermediate fusion offers more flexibility for considering each modality's unique characteristics. While this allows intermediate fusion approaches to robustly deal with missing data [22], the combined nature of the intermediate representation still makes it difficult to infer credibility.

c) Late/Decision Fusion: Late fusion, on the other hand, operates by merging the independent predictions from unimodal classifiers at a later stage in the process. This integration is done through combination functions [24, 25]. Common strategies include weighted mean [26] and noisy-OR-based combination functions [27]. Recently, tractable probabilistic models have also been employed as efficient combination functions for late fusion. The strength of late fusion lies in its explainability and its capacity to preserve the autonomy of each data source, thereby facilitating a more granular assessment of source credibility. Thus, we will focus on late fusion in this work and elaborate on the common approaches in detail below.

B. Late/decision fusion and Credibility

In this section, we describe four late fusion approaches and their ability to represent source-specific credibility. We use X to denote a variable, x to denote a value, $\boldsymbol{X}$ to denote a set of variables, and $\boldsymbol{x}$ to denote a set of values corresponding to the set of variables. So, we use $\boldsymbol{X_i}$ to denote a modality, which is a set of variables, and $\boldsymbol{x_i}$ to denote a value of that modality. We consider late fusion given m modalities and a discrete target variable Y. Each modality $i = 1, \ldots, m$ is encoded using a unimodal model representing the conditional probability over the target variable given that modality, $P_i(Y = y \mid \boldsymbol{X_i} = \boldsymbol{x_i})$.

1) Weighted Mean: Weighted Mean combination rule models the fused predictive probability as an explicitly weighted combination of the predictions of unimodal models. This representation allows the modality-specific credibility to be inferred by inspecting the weights.

$$\begin{aligned} &P(Y = y \mid \boldsymbol{X_1} = \boldsymbol{x_1}, \ldots, \boldsymbol{X_m} = \boldsymbol{x_m}) \\ &= \sum_{i=1}^{m} w_i P_i(Y = y \mid \boldsymbol{X_i} = \boldsymbol{x_i}) \end{aligned} \tag{1}$$

where each $w_i \in [0, 1]$ is the weight for modality i such that $\sum_{i=1}^{m} w_i = 1$.

2) Noisy-OR: The Noisy-OR combination function combines multiple unimodal predictions by assuming causal independence of the influence of each modality on a boolean target variable. It models the fused predictive probability of the target being active as the complement of the product of the unimodal probabilities of the target being inactive.

$$\begin{aligned} &P(Y = 1 \mid \boldsymbol{X_1} = \boldsymbol{x_1}, \ldots, \boldsymbol{X_m} = \boldsymbol{x_m}) \\ &= 1 - \prod_{i=1}^{m} (1 - P_i(Y = 1 \mid \boldsymbol{X_i} = \boldsymbol{x_i})) \end{aligned} \tag{2}$$

3) Multilayered Perceptrons: Weighted mean and Noisy-OR combination functions make restrictive assumptions about the relationship between the predictions of the unimodal models and the fused predictive distribution over the target, namely, linear dependence and independence of causal influence respectively. In cases where the validity of such assumptions is not clear apriori, more expressive combination functions like Multilayer perceptions (MLPs) might be used [28, 29]. Formally, given an MLP f_{MLP}, the fused predictive probability over the target is given by the following expression:

$$\begin{aligned} &P(Y = y \mid \boldsymbol{X_1} = \boldsymbol{x_1}, \ldots, \boldsymbol{X_m} = \boldsymbol{x_m}) \\ &= f_{\text{MLP}}(p_1, \ldots, p_m) \end{aligned} \tag{3}$$

where $p_i = P_i(Y = y \mid \boldsymbol{X_i} = \boldsymbol{x_i})$ for each $i = 1, \ldots, m$. While MLPs can approximate arbitrarily complex functions [30, 31], MLP-based combination functions lack a reliable way to quantify modality-specific credibilities.

4) Probabilistic Circuits: Probabilistic circuits (PCs [14]) are a class of probabilistic models that represent the joint distribution over variables using a computational graph. This directed acyclic graph, composed of three types of nodes, sum, product, and leaf nodes, defines the multivariate joint distribution in terms of compositions of functions of simpler distributions. The internal nodes, sum and product, represent the mixture and factorization, respectively, of their input distributions. The leaf nodes represent simple univariate distributions over input variables. Formally, a PC $\mathcal{M}$ is defined as the tuple $\langle \mathcal{G}, \theta \rangle$ where $\mathcal{G}$ is the computational graph and θ is the set of parameters of the sum and leaf nodes. The joint probability distribution represented by the PC is given by the following expression:

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

where $\mathbf{ch}(n)$ is the set of child nodes of a node n, w_c is edge weight corresponding to the child node c of a sum node, $\mathbf{sc}(n)$ is the scope of a node n (*i.e.,* the set of variables over which it is defined) and ϕ_n is the univariate probability distribution function corresponding to a leaf node n.

A key advantage of PCs is their ability to perform exact probabilistic inference in time *polynomial in the computational graph size*. Additionally, the computational graph structure allows PCs to exploit the efficiency of deep learning while maintaining their probabilistic semantics [32].

PC-based late-fusion approaches [33] use a PC to model the joint distribution over the target variable and the predictions of the unimodal models. The fused predictive distribution is defined as the conditional distribution over the target variable given the unimodal predictions. This predictive probability in PC-based combination functions can be computed efficiently and exactly using conditional probability inference on the PC.

Tractable computation of conditional probability queries requires the PC to satisfy two properties – *smoothness* and *decomposability*. A PC is said to be smooth if, for each sum node, all children are defined over the same set of variables. It is said to be decomposable if, for each product node, all children are defined over disjoint sets of variables. Smooth and decomposable PCs are also called sum-product networks (SPNs [34]). The predictive probability in a late multimodal fusion model with the combination function modeled by an SPN $\mathcal{M}$ is given by the following expression:

$$\begin{aligned} &P(Y=y \mid \boldsymbol{X_1}=\boldsymbol{x_1}, \ldots, \boldsymbol{X_m}=\boldsymbol{x_m}) \\ &\quad = \frac{P_{\mathcal{M}}(Y=y, \boldsymbol{P_1}=\boldsymbol{p_1}, \ldots, \boldsymbol{P_m}=p_m)}{P_{\mathcal{M}}(\boldsymbol{P_1}=p_1, \ldots, \boldsymbol{P_m}=p_m)} \end{aligned} \tag{4}$$

where $p_i = P_i(Y=y \mid \boldsymbol{X_i}=\boldsymbol{x_i})$ for each $i=1,\ldots,m$.

C. Data missingness in credibility-aware multimodal fusion

Missing data poses a significant challenge across various fields, including safety-critical domains such as healthcare. Factors contributing to missingness can vary. In healthcare, demographic information might be readily available, but obtaining patient test results can be hindered by test invasiveness, privacy concerns, and cost. Similarly, in robot navigation and autonomous vehicles, certain sensors might be faulty or non-functional.

Three primary strategies to address missing data include – listwise deletion, imputation, and marginalization. *Listwise deletion* [35], a common approach during training, removes any data points containing missing values. While popular, this method can lead to substantial data loss and introduce bias if missingness is not random [36]. Additionally, listwise deletion is impractical for inference, as it leaves the system unable to make predictions on a potentially large number of data points with missing values.

In contrast, *imputation* replaces missing values with estimates [37]–[39]. While effective when the missingness process is understood, naive imputation [40] can introduce significant bias into the training data. Furthermore, imputing a single most likely value for uncertain data points ignores information about other possible values. Some intermediate fusion approaches such as Cross Partial Multi-View Networks (CPM-Nets) [22] avoid explicit data imputation by imposing structure on the latent representation to allow inference without complete data.

Marginalization, on the other hand, addresses missingness through a joint probabilistic model. This method involves aggregating the predictions of a model across all possible values of the missing feature, weighted by the probability of each value. This method respects the inherent uncertainty of missing data and is a more grounded way of dealing with missing data.

III. Experimental Investigation

This study empirically investigates the performance of various late/decision fusion methodologies, particularly their robustness and reliability in real-world scenarios often characterized by noisy and incomplete data. In multimodal fusion, these issues manifest as noise within individual modalities and the absence of certain modalities. We train late fusion approaches on complete data and study their robustness and reliability when presented with missing, incomplete, or absent data. Overall, we aim to answer the following research questions experimentally

- **(Q1)** How *robust* is the performance of late fusion approaches when faced with noisy and incomplete data?
- **(Q2)** How *reliable* are the predictions made by late fusion methods? Specifically, do they yield well-calibrated predictions under missing data conditions?

We first elaborate on the methodology adopted for evaluating the above questions in this section and discuss the results in the next section.

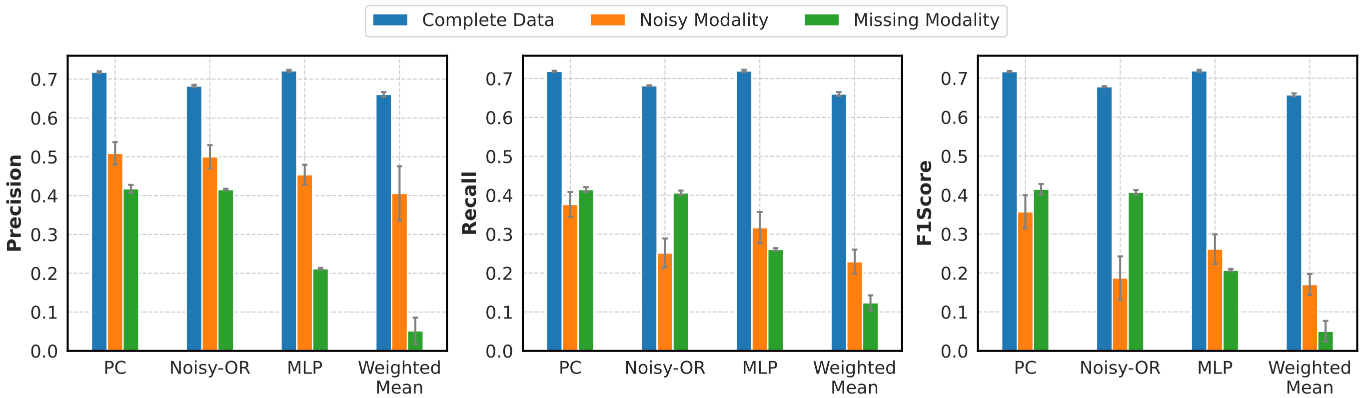


Fig. 2: **Robustness Analysis**: Mean test performance of late fusion approaches on the AVMNIST dataset when presented with complete data (blue), noisy data (orange), and with 50% modalities missing (green), in terms of Precision (left), Recall (middle) and F1 Score (right). Error bars denote standard deviation across 3 independent trials.

A. *Setup*

For our experimental evaluation, we utilize the Audiovisual-MNIST (avMNIST) [12] benchmark dataset, which comprises two modalities: visual and auditory. The visual modality consists of 28×28 pixel images depicting handwritten digits from 0 to 9, while the auditory modality is represented by 112×112 spectrograms corresponding to each digit's sound. The dataset is divided into 55,000 training examples, 5,000 validation examples, and 10,000 test examples.

We implement and compare four late (decision) fusion approaches: a Multilayer Perceptron (MLP), Weighted Mean, Noisy-Or, and a Probabilistic Circuit (PC), following the architecture and hyperparameter settings detailed in [13]. We investigate the robustness of the above late fusion approaches during the test phase under two primary conditions: missing data and noisy input data, the details of which we elaborate on below:

B. *Evaluating Robustness*

1) Missing Data: To evaluate the resilience of the fusion methods to incomplete data, we mask out the information from one of the modalities by multiplying it with a zero vector. Since the AVMNIST dataset has only two modalities, the resulting test data distribution has lost information from 50% of the input modalities. For the Probabilistic Circuit (PC) fusion function $\mathcal{M}$, we handle missing data through **marginalization**, utilizing its tractability. Let k denote the index of the missing modality. The final prediction in the presence of missing data is obtained as follows:

$$\begin{aligned} P(Y = y \mid \mathbf{X}_{-k}) &= P_{\mathcal{M}}(Y = y \mid \boldsymbol{P}_{-k}) \\ &= \frac{\int_{\mathbf{p}_k} P_{\mathcal{M}}(Y = y, \boldsymbol{P}_{-k} = \mathbf{p}_{-k}, \boldsymbol{P}_k = \mathbf{p}_k)}{\sum_y \int_{\mathbf{p}_k} P_{\mathcal{M}}(Y = y, \boldsymbol{P}_{-k} = \mathbf{p}_{-k}, \boldsymbol{P}_k = \mathbf{p}_k)} \end{aligned}$$

where $\mathbf{X}_{-k}$ represents the observed modalities and $\boldsymbol{P}_{-k}$ denotes the predictions made by the unimodal models on each of the observed modalities. Marginalization over the missing modality $\mathbf{X}_k$ can be efficiently performed in linear time for a smooth and decomposable PC by setting the corresponding leaf variables to 1 and conducting a bottom-up evaluation of the PC [14].

2) Noisy Data: To assess the robustness of the fusion methods in the presence of noise, we generate a noisy version of the test dataset by introducing noise into both modalities.

We create a noisy data set by transforming each data point $((\boldsymbol{x_1}, \ldots, \boldsymbol{x_m}), y) \in \mathcal{D}$, to $((\tilde{\boldsymbol{x}}_{\boldsymbol{i}}, \ldots, \tilde{\boldsymbol{x}}_{\boldsymbol{m}}), y)$. We do so by adding noise to each $\boldsymbol{x_i}$ using the following equation:

$$\tilde{\boldsymbol{x}}_i = \alpha \boldsymbol{x}_i + (1 - \alpha)\boldsymbol{n}_i,$$

where $\boldsymbol{n}_i \sim U(\boldsymbol{X}_i^{\min}, \boldsymbol{X}_i^{\max})$ is a noise vector sampled from a uniform distribution over the range of the random variable $\boldsymbol{X_i}$, and $\alpha \in [0, 1]$ is a parameter that controls the noise level. By varying α, we can simulate different levels of noise in the data and evaluate the impact on the fusion method's performance.

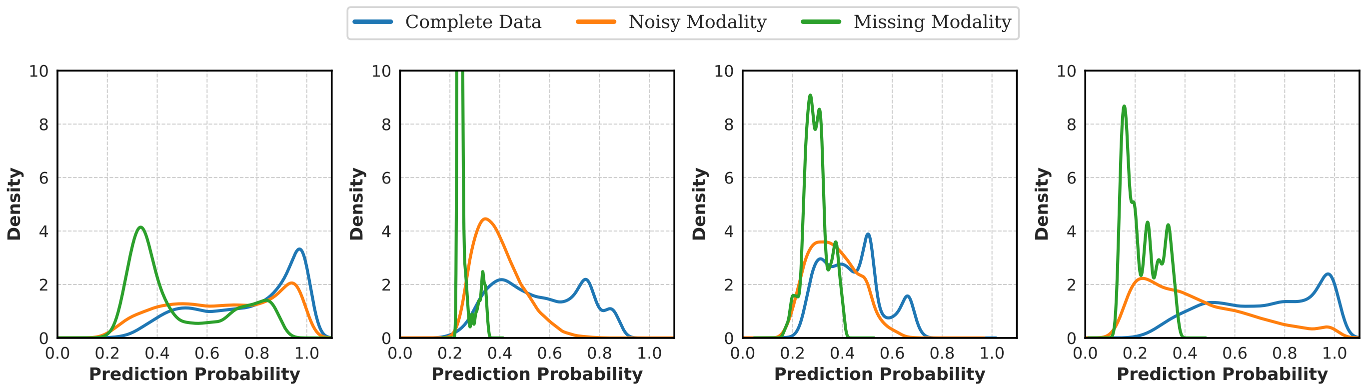
C. *Evaluating Reliability*

Reliability in multimodal fusion systems is closely linked to the *calibration* of predictive outcomes — a concept that ensures that the predicted probabilities of an outcome align closely with its actual occurrence rate [41]. Calibration is crucial not only for the system's efficacy in practical decision-making but also for its interpretability and the trust users place in it [42]. In a scenario where a model predicts a series of events to occur with a confidence level of 0.6, a well-calibrated model would see these events actually happening approximately 60 out of 100 times. Formally, this state of perfect calibration is described as:

$$\mathbb{P}(\hat{y} = y \mid \hat{p} = p) = p \;\forall p \in [0, 1], \tag{5}$$

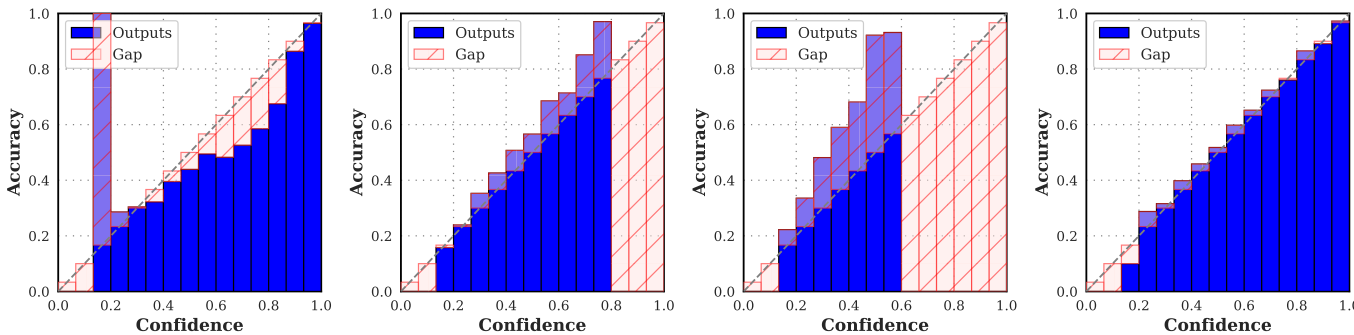
where $\hat{y}$, $\hat{p}$, and y represent the predicted label, the predicted probability, and the actual label, respectively.

Reliability Diagrams are graphical representations that offer an intuitive understanding of a model's calibration [43, 44]. Reliability diagrams plot the model's predicted probabilities against the empirical probability of the predicted outcomes.



(a) MLP (b) Weighted Mean (c) Noisy-Or (d) PC

Fig. 3: **Distribution Shift:** Visualization of distributions of the maximum prediction probabilities outputted by various late fusion approaches when presented with complete data (blue), noisy data (orange), and with 50% modalities missing (green).



(a) MLP (b) Weighted Mean (c) Noisy-Or (d) PC

Fig. 4: **Calibration Analysis**: Reliability Diagrams illustrating the calibration of the four late fusion methods under setting where 50% of modalities are missing. Each subplot displays the alignment of predicted confidence with actual model accuracy. Blue bars represent accuracy, while the red line marks the gap between confidence and observed accuracy within each bin.

A model demonstrating perfect calibration will result in a diagram where the plot lies on the diagonal line, representing a balance between confidence and actual correctness.

For a more precise evaluation of calibration, we employ the Expected Calibration Error (ECE) [45], which measures the average calibration gap across the model's predictions:

$$\mathop{\mathbb{E}}_{\hat{P}}\left[\left|\mathbb{P}(\hat{Y}=Y \mid \hat{P}=p)-p\right|\right]. \tag{6}$$

The ECE metric is calculated by dividing the range of predicted probabilities into M distinct bins $B_1, \ldots, B_M$, and it is computed as follows:

$$\mathrm{ECE}=\sum_{i=1}^{M}\frac{|B_i|}{N}\left|\mathrm{acc}(B_i)-\mathrm{conf}(B_i)\right|, \tag{7}$$

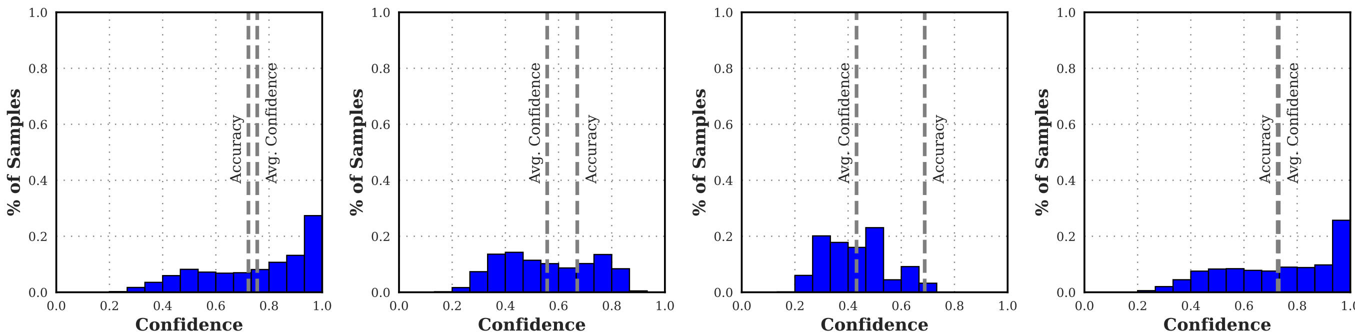
where $|B_i|$ indicates the number of predictions in the i-th bin, $\mathrm{acc}(B_i)$ is the accuracy of predictions within that bin, and $\mathrm{conf}(B_i)$ is the mean predicted confidence of the bin. This measure provides an average sense of how much the model's confidence deviates from ideal calibration.

IV. Results

(Q1) How robust are late fusion approaches when faced with noisy and incomplete data?

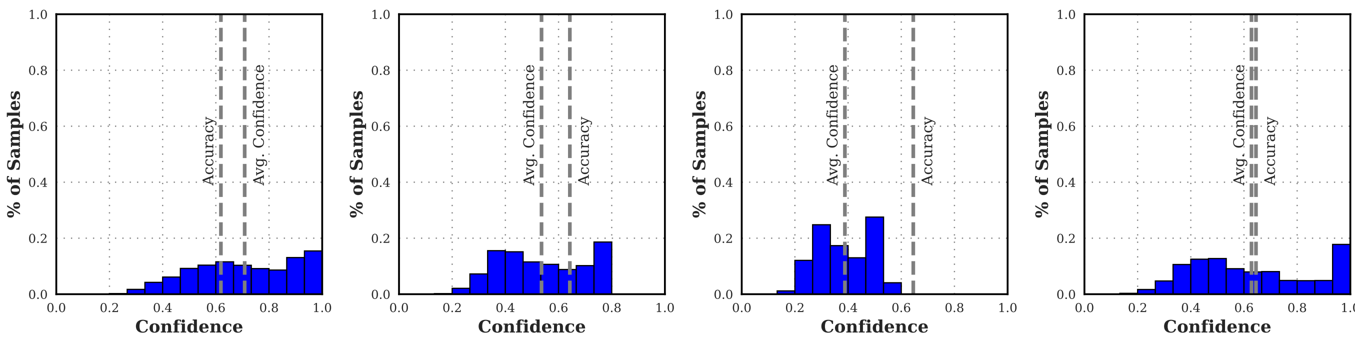
Figure 2 visualizes the mean test performance of each approach when presented with complete data (blue), noisy data (orange), and with 50% modalities missing (green), considering metrics such as Precision (left), Recall (middle), and F1 Score (right). As expected, all methods experience performance degradation with noise or missing data. However, the PC-based fusion exhibits the smallest decrease in performance across both settings, suggesting greater robustness.

A second aspect of robustness is a classifier's ability to appropriately adjust confidence in its predictions in response to noise or missing information. The introduction of noise can lead to an out-of-distribution dataset, where a robust and reliable classifier should exhibit both minimal performance degradation and reduced prediction confidence to reflect increased uncertainty [46]. Similarly, the absence of a modality should naturally increase the model's predictive uncertainty, as the "essence of information is to remove uncertainty" [47].



(a) MLP (b) Weighted Mean (c) Noisy-Or (d) PC

Fig. 5: **Confidence Histograms of Late Fusion Approaches on Complete Data**: Each subplot depicts the distribution of test samples across confidence levels, providing a visual breakdown of how many samples fall into each confidence interval. The grey lines denote the average accuracy and confidence.



(a) MLP (b) Weighted Mean (c) Noisy-Or (d) PC

Fig. 6: **Confidence Histograms of Late Fusion Approaches when** 50% **of modalities are missing**: Each subplot depicts the distribution of test samples across confidence levels, providing a visual breakdown of how many samples fall into each confidence interval. The grey lines denote the average accuracy and confidence.

Therefore, we expect the model's confidence distribution to shift toward the lower end of the spectrum in such situations.

To ascertain whether this theoretical expectation is met in practice, we analyzed the distribution over the maximum prediction probabilities across different late fusion algorithms. As depicted in Figure 3, all approaches tend to produce predictions with diminished confidence when modalities are missing. This effect is more pronounced in the case of the PC, Weighted Mean, and Noisy-Or as compared to MLP. However, with noisy data, the MLP and Noisy-Or methods seem to maintain a higher level of confidence, potentially failing to recognize the out-of-distribution nature of the data. On the other hand, the PC exhibits the lowest peak confidence among all approaches when presented with both noisy and missing modalities. Since it also performs better than the other approaches (Figure 2), we can conclude that it more accurately reflects the input data uncertainty and is, therefore, a more robust approach.

(Q2) How reliable are the predictions made by late fusion methods?

Figure 4 visualizes the reliability diagrams for the different late fusion methods when 50% of the modalities are missing. Each plot shows the model's confidence levels on the x-axis against the actual accuracy achieved at each confidence level on the y-axis. Blue bars represent the actual accuracy achieved within each confidence interval on the test set, and the red line indicates the gap between the confidence and the accuracy for each bin. Ideally, in a perfectly calibrated model, the predictions would be along the diagonal line, implying that the model's confidence matches its accuracy, and it is hence a reliable model. Deviations below the diagonal indicate overconfidence, while deviations above suggest underconfidence. While all models exhibit some miscalibration, as indicated by the gaps between the tops of the blue bars and the diagonal line, they vary in their calibration quality. However, the plot for the PC-based fusion method closely follows the diagonal with smaller gaps, suggesting better calibration and more reliable confidence estimates compared to the other methods.

Figures 5 and 6 visualize the distribution of test samples belonging to different confidence intervals for the late fusion approaches when presented with complete and missing data respectively. In the complete data setting, the PC's accuracy closely follows the average confidence, indicating near-perfect calibration, while other methods show a calibration gap. The calibration gap widens slightly for all the approaches when presented with missing data; however, the PC-based fusion method still maintains the smallest gap. Figure 1 quantitatively confirms this observation in terms of the mean calibration error achieved by each of the approaches. The PC-based fusion method achieves the lowest calibration error across both complete and missing data. This suggests that the use of PCs as tractable probabilistic models as combination functions helps achieve reliable late multimodal fusion.

V. Conclusion

To summarize, this paper provided an in-depth experimental comparison of some popular late multimodal fusion techniques, focusing on their robustness and reliability in scenarios reflective of real-world conditions. The paper specifically examined how these methods perform with noisy data and when confronted with missing modalities. The results revealed that employing a tractable probabilistic generative model like a Probabilistic Circuit (PC) as a combination function yields robust and well-calibrated classifiers. The probabilistic semantics and the inherent capability of PCs to handle missing data through marginalization contribute to their reliability, which is important when deploying multi-modal systems for decision-making. Additionally, the results suggest that non-probabilistic models, such as MLPs and weighted means, might require additional regularization or training modifications to promote calibration for enhanced robustness and reliability in real-world applications.

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

[1] G. Seetharaman, A. Lakhotia, and E. P. Blasch, "Unmanned vehicles come of age: The darpa grand challenge," *Computer*, vol. 39, no. 12, pp. 26–29, 2006.

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[3] E. Blasch, É. Bossé, and D. A. Lambert, *High-level information fusion management and systems design*. Artech House, 2012.

[4] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[5] A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, and Y. Luo, "Multimodal machine learning in precision health: A scoping review," *npj Digital Medicine*, vol. 5, no. 1, p. 171, 2022.

[6] R. Sawhney, P. Mathur, A. Mangal, P. Khanna, R. R. Shah, and R. Zimmermann, "Multimodal multi-task financial risk forecasting," in *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, 2020, pp. 456–465.

[7] E. Blasch, K. B. Laskey, A.-L. Jousselme, V. Dragos, P. C. Costa, and J. Dezert, "Urref reliability versus credibility in information fusion (stanag 2511)," in *Proceedings of the 16th International Conference on Information Fusion*. IEEE, 2013, pp. 1600–1607.

[8] G. L. Rogova and V. Nimier, "Reliability in information fusion: literature survey," in *Proceedings of the seventh international conference on information fusion*, vol. 2, 2004, pp. 1158–1165.

[9] 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)*, vol. 34, no. 1, pp. 782–787, 2004.

[10] M. Somero, L. Snidaro, and G. L. Rogova, "Deep classifiers evidential fusion with reliability," in *2023 26th International Conference on Information Fusion (FUSION)*. IEEE, 2023, pp. 1–7.

[11] E. J. Wright and K. B. Laskey, "Credibility models for multi-source fusion," in *2006 9th International Conference on Information Fusion*. IEEE, 2006, pp. 1–7.

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

[13] S. Sidheekh, P. Tenali, S. Mathur, E. Blasch, K. Kersting, and S. Natarajan, "Credibility-aware multi-modal fusion using probabilistic circuits," *ArXiv preprint*, vol. abs/2403.03281, 2024.

[14] Y. Choi, A. Vergari, and G. Van den Broeck, "Lecture notes: Probabilistic circuits: Representation and inference," 2020.

[15] D. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[16] J. De Villiers, G. Pavlin, A. 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*, vol. 13, no. 2, pp. 198–215, 2018.

[17] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, J. Ye, A. D. N. Initiative *et al.*, "Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data," *NeuroImage*, vol. 61, no. 3, pp. 622–632, 2012.

[18] M. Liu, J. Zhang, P.-T. Yap, and D. Shen, "Diagnosis of alzheimer's disease using view-aligned hypergraph learning with incomplete multi-modality data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 308–316.

[19] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *2020 IEEE 23rd international conference on information fusion (FUSION)*. IEEE, 2020, pp. 1–6.

[20] O. Schulte and K. Routley, "Aggregating predictions vs. aggregating features for relational classification," in *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, 2014, pp. 121–128.

[21] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: multimodal transfer module for CNN fusion," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, pp. 13 286–13 296.

[22] C. Zhang, Z. Han, Y. Cui, H. Fu, J. T. Zhou, and Q. Hu, "Cpm-nets: Cross partial multi-view networks," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 557–567.

[23] J. Perez-Rua, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "MFAS: multimodal fusion architecture search," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 6966–6975.

[24] S. Natarajan, P. Tadepalli, E. Altendorf, T. G. Dietterich, A. Fern, and A. C. Restificar, "Learning first-order probabilistic models with combining rules," in *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, ser. ACM International Conference Proceeding Series, L. D. Raedt and S. Wrobel, Eds., vol. 119. ACM, 2005, pp. 609–616.

[25] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, and L. D. Raedt, "Deepproblog: Neural probabilistic logic programming," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle,

K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 3753–3763.

[26] E. Shutova, D. Kiela, and J. 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*. Association for Computational Linguistics, 2016, pp. 160–170.

[27] J. Tian, W. Cheung, N. Glaser, Y.-C. Liu, and Z. Kira, "Uno: Uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 5716–5723.

[28] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm *et al.*, "Multiple classifier systems for the classification of audio-visual emotional states," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 359–368.

[29] G. A. Ramirez, T. Baltrušaitis, and L.-P. 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*. Springer, 2011, pp. 396–406.

[30] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.

[31] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[32] S. Sidheekh and S. Natarajan, "Building expressive and tractable probabilistic generative models: A review," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, 3rd-9th August 2024, Jeju Island, South Korea*, 2024.

[33] S. Sidheekh, K. Kersting, and S. Natarajan, "Probabilistic flow circuits: Towards unified deep models for tractable probabilistic inference," in *The 39th Conference on Uncertainty in Artificial Intelligence*, 2023.

[34] H. Poon and P. M. Domingos, "Sum-product networks: A new deep architecture," in *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14-17, 2011*, F. G. Cozman and A. Pfeffer, Eds. AUAI Press, 2011, pp. 337–346.

[35] E. R. Buhi, P. Goodson, and T. B. Neilands, "Out of sight, not out of mind: strategies for handling missing data." *American journal of health behavior*, vol. 32 1, pp. 83–92, 2008.

[36] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

[37] J. S. Murray, "Multiple imputation: A review of practical and theoretical findings," *Statistical Science*, vol. 33, no. 2, pp. 142–159, 2018.

[38] D. B. Rubin, "Multiple imputation," in *Flexible Imputation of Missing Data, Second Edition*. Chapman and Hall/CRC, 2018, pp. 29–62.

[39] P. Mattei and J. Frellsen, "MIWAE: deep generative modelling and imputation of incomplete data sets," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 4413–4423.

[40] D. B. Rubin, "Multiple imputation after 18+ years," *Journal of the American statistical Association*, vol. 91, no. 434, pp. 473–489, 1996.

[41] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 1321–1330.

[42] E. Blasch, A. Jøsang, J. Dezert, P. C. Costa, and A.-L. Jousselme, "Urref self-confidence in information fusion trust," in *17th International Conference on Information Fusion (FUSION)*. IEEE, 2014, pp. 1–8.

[43] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, ser. ACM International Conference Proceeding Series, L. D. Raedt and S. Wrobel, Eds., vol. 119. ACM, 2005, pp. 625–632.

[44] M. H. DeGroot and S. E. Fienberg, "The comparison and evaluation of forecasters," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 32, no. 1-2, pp. 12–22, 1983.

[45] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, B. Bonet and S. Koenig, Eds. AAAI Press, 2015, pp. 2901–2907.

[46] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[47] H. Ma, Q. Zhang, C. Zhang, B. Wu, H. Fu, J. T. Zhou, and Q. Hu, "Calibrating multimodal learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 23 429–23 450.