Classification with Uncertainty-Aware Multimodal Deep Learning: A Survey

Grigor Bezirganyan
Aix Marseille Univ
CNRS, LIS
Marseille, France
gbezirganyan@gmail.com

Sana Sellami
Aix Marseille Univ
CNRS, LIS
Marseille, France
sana.sellami@univ-amu.fr

Laure Berti-Équille
IRD, ESPACE-DEV
Montpellier, France
laure.berti@ird.fr

Sébastien Fournier
Aix Marseille Univ
CNRS, LIS
Marseille, France
sebastien.fournier@univ-amu.fr

ABSTRACT

Multimodal deep learning has achieved remarkable progress by leveraging complementary information across heterogeneous data sources such as texts, images, audios, and structured signals. While increasingly powerful encoders and fusion mechanisms have improved predictive performance, the reliability of multimodal systems remains a critical challenge. In particular, modality disagreement, distribution shifts, and noisy inputs can lead to overconfident yet incorrect predictions.

Distinct from existing surveys on uncertainty in deep learning [45] or on multimodal learning [8; 135], this survey jointly covers three aspects: (i) the structural foundations of multimodal classification examined through the lens of the uncertainty challenges each design choice introduces; (ii) uncertainty quantification in both unimodal and multimodal settings; and (iii) set-valued classification as a decision-level strategy for cautious multimodal prediction. We first review foundational aspects of multimodal representation learning and fusion strategies, highlighting their structural limitations in modeling inter-modal dependence and the uncertainty challenges each stage introduces. We then examine uncertainty quantification methods in deep learning, including both probabilistic and evidence-theoretic approaches, and analyze how these techniques extend to multimodal settings. Special attention is given to conflict-aware fusion mechanisms and to decision-level strategies such as set-valued classification, which enable more cautious and informative predictions.

Beyond reviewing existing methods, we identify key open challenges, including the modeling of partial dependence between modalities, the need for systematic benchmarking of multimodal uncertainty, and the integration of uncertainty into decision-making pipelines. Finally, we discuss how these reliability challenges extend to emerging multimodal agentic systems. By synthesizing advances across multimodal learning and uncertainty modeling, this survey aims to provide a unified perspective and to outline recent research directions toward more reliable multimodal AI systems.

1. Introduction

In recent years, multimodal deep learning (MDL) has seen increasing adoption in various domains, where fusing information from different data sources, such as images, texts, and audios, can improve predictive performance [129; 34; 115; 74]. Combining information from diverse data sources requires different design choices, including the methods used to extract and encode relevant information from each modality, the choice of when and how to fuse the information, and how to make the final prediction. Consequently, there has been a growing interest in developing multimodal learning architectures that do not only improve prediction accuracy but also reflect the reliability of the fused decision and effectively estimate the uncertainty.

Indeed, these heterogeneous data streams inherently possess fluctuating degrees of noise, occlusion, and missing information. By explicitly modeling both aleatoric uncertainty (data-inherent noise) and epistemic uncertainty (model ignorance), multimodal systems can dynamically calibrate their fusion mechanisms. This allows the model to intelligently down-weight corrupted or ambiguous modalities while actively leaning on more reliable signals, thereby preventing the propagation of silent errors during representation alignment. Furthermore, as these complex models are increasingly deployed in high-stakes environments like clinical diagnostics and autonomous navigation, robust uncertainty estimates are paramount. They provide the vital mechanisms needed for safe out-of-distribution (OOD) detection, mitigating catastrophic multimodal wrong prediction, and ultimately bridging the gap between raw predictive performance and trustworthy AI.

This paper discusses the evolution of multimodal deep learning and multimodal uncertainty quantification methods, highlighting both foundational work and state-of-the-art approaches. It also outlines the key challenges in this domain, including the complexity of existing approaches, the difficulty of han-

dling conflicting or uninformative modalities, and the need for efficient, actionable, and trustworthy predictions from multimodal inputs.

Several surveys address related but distinct problems. [45] provides a comprehensive treatment of uncertainty quantification (UQ) in deep learning, focusing primarily on unimodal architectures. Multimodal learning surveys [8; 135] cover fusion and representation challenges but do not address uncertainty modeling in depth. To our knowledge, no existing survey jointly covers (i) the structural foundations of multimodal classification with an explicit focus on the uncertainty challenges introduced at each stage, (ii) UQ methods in both unimodal and multimodal settings, and (iii) decision-level strategies such as set-valued classification for cautious multimodal prediction. This survey fills this gap by providing a unified perspective connecting all three dimensions, and by showing how limitations in fusion architecture directly motivate uncertainty-aware design.

The remainder of this paper is structured as follows: Section 2 reviews the recent key developments and the current trends in multimodal deep learning. Sections 3 and 4 present uncertainty quantification approaches in unimodal and multimodal settings, respectively. Section 5 discusses the challenges of evaluating multimodal UQ methods. Finally, Section 6 summarizes the main insights and outlines future research directions.

2. Multimodal Deep Learning Foundations

Multimodal (or multi-view) deep learning (MDL) aims to leverage and combine information from diverse data types, such as images, audios, and texts, and other types of signals, that could not previously be jointly integrated, in order to improve the performance of machine learning models. Although there is no universally accepted distinction between the terms multimodal and multi-view, they are often used interchangeably in the literature, with broadly similar definitions [140; 171; 135]. In the same line, we will use the terms multimodal and multi-view interchangeably throughout this paper. We define multimodal data as data that is represented in different forms or collected from different sources, but that describes the same underlying entity or concept [171].

Different modalities often contain complementary information, and combining them can lead to a more comprehensive representation of the underlying real-world entities or phenomena [93]. [8] identify five key challenges in MDL: data representation, translation, modality alignment, fusion, and co-learning described as follows:

- *Representation*: How to represent different modalities in a way that captures both their complementarity and redundancy.
- *Translation*: How to translate information between modalities (e.g., from image to text, or from text to audio). Since we focus on multimodal classification, inter-modal translation will not be discussed further in this article.
- *Alignment*: Ensuring that data from different modalities, such as text, images, and audio, refer to the same underlying concept or instance, whether at the data, feature, or semantic level. For example, in video captioning, alignment ensures that the text accurately describes the visual content. At the semantic level, models may associate concepts across modalities (e.g., linking an image of a cat with the spoken word "cat") despite differences in representation. While some methods enforce alignment explicitly, many models learn it implicitly through supervision on the main task (e.g., classification or captioning) [92].
- *Fusion*: Combining information from multiple modalities to perform a prediction. Different modalities can provide varying levels of information, and may contain noise, inaccuracies, or disagreements.
- *Co-Learning*: Transferring knowledge between modalities. Co-learning mainly explores how models can be benefited from other models trained on different modalities.

Beyond these challenges, we also highlight an additional issue that is highly relevant for multimodal classification: the modality imbalance, closely related to representation learning, when learning rates are different depending on the modality or when one modality dominates the training process, often leading to suboptimal joint representations [162].

In this paper, our primary focus is on multimodal classification, and in this section we review approaches and challenges in representation learning, multimodal fusion, and learning objectives, and modality imbalance. These three aspects are deeply interconnected: the quality of representation learning determines the informativeness of unimodal features; fusion governs how these features are combined into a joint decision; and learning objectives influence both representation and fusion by shaping how modalities are weighted during training. A recurring challenge across all three is modality imbalance, where some modalities dominate while others are underutilized, leading to suboptimal joint models. Addressing these issues is central to designing efficient and reliable multimodal classifiers.

These challenges have direct implications for uncertainty modeling that motivate the content of later sections. The quality of representation learning determines whether per-modality uncertainty can be estimated reliably: noisy or poorly calibrated representations propagate errors into any downstream uncertainty estimate. Fusion design governs how uncertainty is combined across modalities: strategies that assume equal modality reliability cannot detect nor propagate uncertainty arising from inter-modal conflict. Modality imbalance introduces systematic overconfidence when dominant modalities suppress the uncertainty signals of weaker sources. We highlight these implications throughout this section to build a principled motivation for the uncertainty quantification frameworks reviewed in Sections 3 and 4.

Other classical challenges of multimodal learning are less relevant in our scope. The challenge of translation pertains to tasks involving inter-modal generation, which falls outside the focus of this work. Similarly, co-learning typically addresses transfer learning across modalities in different settings. While explicit alignment can sometimes improve classification, it is often not essential, since semantic alignment is usually learned implicitly through supervision.

Scope note. The subsections below review representation learning, fusion strategies, and learning objectives as they

bear on uncertainty in multimodal classification. Rather than providing a comprehensive taxonomy of all MDL architectures, this section deliberately emphasizes the design choices that introduce calibration or conflict challenges, providing principled motivation for the UQ frameworks reviewed in Sections 3 and 4.

2.1 Multimodal Classification Steps

There are various taxonomies for multimodal classification, with one of the most widespread being based on the stage of the pipeline at which fusion occurs. Based on this, multimodal networks are categorized into three types: *early fusion*, *intermediate fusion*, *late fusion* [121]. Some architectures that use a combination of these taxonomies are also often referred as *hybrid fusion*. However, [135] argue that classifying architectures solely by fusion stage is not sufficiently specific to capture the diversity of current multimodal pipelines. Hence, they propose a more detailed taxonomy based on five stages: 1) Preprocessing, 2) Feature extraction, 3) Data fusion, 4) Primary learning, and 5) Final classification. In this paper, we follow the taxonomy provided by [135], while merging the *feature extraction* and *primary learning* stages into a single category referred to as *representation and primary learning*. We thus categorize the stages of multimodal classification pipelines into: 1) Preprocessing, 2) Representation / primary learning, 3) Fusion, and 4) Final classifier (Figure 1).

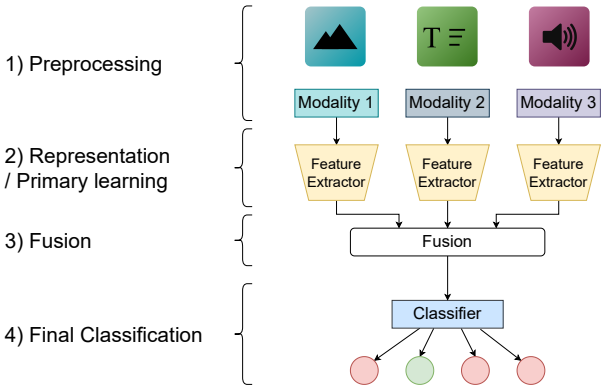

Figure 1: An example of a multimodal classifier divided into the proposed stages.

In many architectures, several of these stages can be shared and performed by the same neural network.

In the following subsections, we will review the main architectures in multimodal deep learning and their advances, focusing on representation learning and data fusion stages.

2.1.1 Preprocessing & Multimodal Representation Learning

This section presents the first two stages of the pipeline, with a particular focus on approaches including multimodal representation learning.

The preprocessing stage can include data cleaning, normalization, addressing missing values, and typically varies depending on the modality. For instance, text preprocessing may include normalization and tokenization; audio preprocessing often involves converting waveforms to spectrograms; and image preprocessing may include cropping, resizing, or normalization. Often, deep learning architectures may omit the preprocessing steps, and perform the learning from raw data.

An important consideration in multimodal classification is how each modality is represented before fusion, since the quality of learned features sets the foundation for cross-modal integration. We therefore review approaches for representation learning, tracing their evolution from handcrafted features to deep encoders and foundation models. A brief overview of representation learning approaches is illustrated in Figure 2.

In the early days of multimodal learning, feature extraction relied primarily on hand-crafted techniques, such as SIFT [98] for images and bag-of-words [53] for text. Fusion was often performed using linear methods such as Canonical Correlation Analysis (CCA) [57], which maximized cross-modal correlations. Extensions such as discriminative CCA [28; 76; 180] incorporated label information, yet these approaches remained limited to modeling linear relationships, motivating the development of more expressive non-linear models.

Neural network-based approaches began addressing these limitations even prior to the deep learning era [32]. With the rise of deep learning, joint training of modality-specific encoders became standard. Early works such as multimodal autoencoders [112] and deep Boltzmann machines [137] demonstrated the benefit of learning shared representations end-to-end, while Deep CCA [4] introduced non-linear cross-modal alignment.

As unimodal deep architectures matured, multimodal systems increasingly leveraged strong modality-specific encoders [38; 63; 6; 182]. Convolutional networks such as AlexNet [81], VGG [134], and ResNet [54] replaced handcrafted visual features, while distributed word embeddings like Word2Vec [108] and GloVe [118] became standard in text. These advances enabled richer latent representations suitable for cross-modal alignment and fusion.

A major paradigm shift occurred with the introduction of transformers [153], whose attention mechanisms allowed scalable modeling across modalities. Originally proposed for NLP, transformers were extended to vision [30] and audio [47], and became central to multimodal architectures. Models such as ViLBERT [99], LXMERT [141], and VisualBERT [91] demonstrated both dual-encoder and shared-encoder strategies for vision–language integration. While powerful, transformers incur quadratic complexity in sequence length. Alternatives such as MLP-Mixer [146] explored attention-free architectures with improved computational efficiency.

More recently, large-scale pre-trained multimodal models have further advanced representation learning. CLIP [120] introduced contrastive pre-training of image and text encoders in a shared embedding space, achieving strong zero-shot performance. Subsequent models such as ALIGN [66] and Flamingo [3] expanded this paradigm. The emergence of *foundation models* extended multimodal learning to broader modality sets; for example, ImageBind [46] aligned images, text, audio, depth, and other signals into a unified embedding space without requiring fully paired data across all modalities. In general, the representation learning stage in multimodal deep learning is an active area of research, since having a good representation of the data is crucial for the performance of the model. The trade-off between the complexity of the model and the quality of the representa-

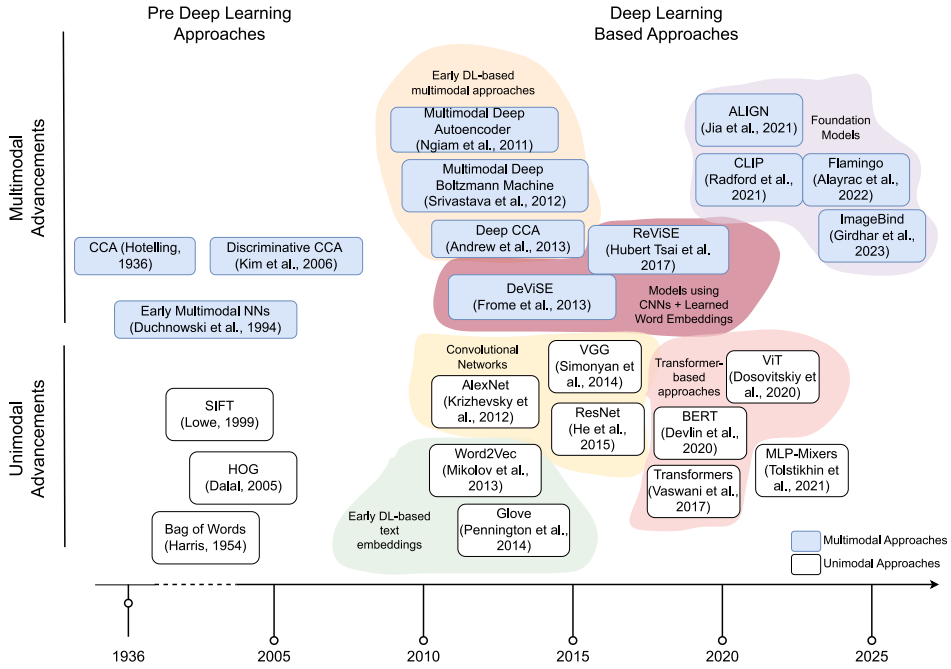

Figure 2: A brief overview of key approaches in unimodal and multimodal representation learning.

tion is an important consideration, as while more complex models can learn better representations, they also require more data and computational resources to train. The learning stage corresponds to the primary learning phase of the pipeline. Depending on the architecture, this learning may occur at a single point, as in early fusion, or at multiple points, as in cross-modality and late fusion strategies.

Uncertainty implication. Representation quality directly bounds the reliability of per-modality uncertainty estimates: a poorly calibrated encoder produces feature distributions whose uncertainty cannot be trusted by downstream fusion mechanisms. Large pre-trained models and foundation models offer better-calibrated representations but introduce epistemic uncertainty about feature-space transferability to the target domain [45].

Having reviewed representation learning, we next discuss multimodal fusion and final classification. Fusion is essential both for integrating features during representation learning and for aggregating predictions in the final decision stage.

2.1.2 Multimodal Fusion & Final Classification

Once suitable modality-specific representations are available, the next design choice concerns how to integrate them. Fusion is the core mechanism by which multimodal systems combine complementary or redundant information, and different fusion stages lead to different trade-offs in flexibility, computational cost, and robustness. Finally, the combined representation or aggregated predictions can be passed through a final classifier, which produces the system's output.

In this section, we will review the main approaches for fusion and final classification in multimodal deep learning. Fusion is usually categorized into *early, intermediate* and *late* fusion strategies. In early fusion, modalities are combined at the input data level and passed through a shared representation learner. In intermediate fusion, each modality is first processed by a separate feature extractor, and their features are then fused for downstream tasks. In late fusion, modality-specific classifiers make independent predictions, which are then aggregated at the decision level. In practice, these strategies are not mutually exclusive, many architectures combine them at different points in the pipeline, which is often referred to as hybrid fusion. Examples of early, intermediate and late fusion strategies are illustrated in Figure 3.

Early (raw data) fusion combines the raw inputs from multiple modalities and processes them with a unified encoder for feature learning and prediction. Because fusion occurs at such an early stage, it is challenging to directly integrate heterogeneous modalities (e.g., text with images, or audio with text). However, when modalities share similar raw representations (e.g., multiple image modalities, audio spectrograms with images, or images with depth maps), early fusion becomes straightforward to implement [138]. In such cases, it can also be computationally efficient, as only a single network is required to process the fused input.

The intermediate (feature) fusion combines information at the intermediate representation (feature) level, rather than at the raw data or classifier output level. This allows for interaction between modalities at a higher level, which is not possible in the early fusion stage. Compared to late fusion strategies, which mostly model modality information independently from one another, intermediate fusion also better models cross-modal correlations.

In the multimodal architectures utilizing intermediate fusion, fusion operation can happen at various points in the architecture. [49] categorized the approaches of intermediate fusion strategies based on when they are fused into *sudden, gradual,* and *multi-flow* fusion. Examples of these strategies can be found in Figure 4. In the *sudden* fusion architectures, all modalities are fused together at the same time in one fu-

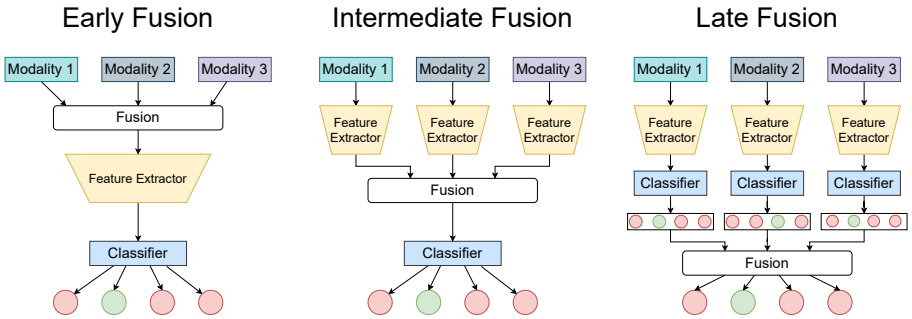

Figure 3: Examples of early, intermediate and late fusion strategies.

sion function. In the *gradual* fusion, a subset of modalities can be fused initially, and then additional modalities can be fused progressively. The gradual fusion approach allows for a hierarchical processing of the modalities. Finally, the *multi-flow* fusion, fuses modalities with different independent fusion functions, which then are fused with each other into a single representation by another function.

The late (decision) fusion combines modality-specific predictions at the decision level. Compared to intermediate fusion, late fusion sacrifices some capacity to model fine-grained cross-modal interactions, but offers greater modularity, robustness to missing modalities, and a natural interface for uncertainty quantification. We organize the existing late fusion approaches into four categories: *simple*, *meta-learning-based*, *optimization-based*, and *uncertainty-aware*, acknowledging that these categories are not exhaustive and that some methods may span multiple groups. A high level overview of the approaches is given in Figure 5.

a) Simple approaches: Simple approaches, as discussed by [78], combine classifier outputs using aggregation functions such as product, averaging, maximum, minimum, and majority voting. These methods typically operate on class-conditional probabilities and include a normalization step to preserve a valid probability distribution. In product fusion, probabilities from each classifier are multiplied and normalized. Averaging computes the mean probability across classifiers, with a common extension being weighted averaging, where modality weights are assigned or learned from a validation set [5; 139; 94]. Maximum and minimum fusion select the highest or lowest probability for each class, respectively, while majority voting assigns the class with the most votes as the final prediction. These simple approaches remain popular because they are easy to implement, but they treat all modalities equally or rely on fixed weights, which may not be optimal for every instance and can perform poorly when modalities are missing or degraded.

b) Meta-learning: To address the limitations of simple approaches, meta-learning-based approaches train a separate model on unimodal score functions to predict instance-specific fusion weights or learn more complex combinations. In the deep learning context, [133] combined two convolutional networks for action recognition via averaging and a multi-class SVM, finding the SVM fusion superior. Other works use logistic regression [152], decision trees, random forests [65], neural networks [62; 122], or adaptive weighting and gating mechanisms [170] to dynamically select or weight modalities. While meta-learning fusion can capture cross-modal dependencies and adapt to varying modality relevance, it requires labeled training data and may not generalize well under distribution shifts.

c) Optimization-based: Optimization-based fusion approaches remove the need for supervised fusion training by formulating fusion as an unsupervised problem, seeking an optimal representation that satisfies predefined structural or statistical constraints. These methods aim to recover a consensus prediction that agrees with all modalities while accounting for noise and outliers. Examples include low-rank matrix recovery [175; 116] and hard-rank-constrained matrix factorization with consistency preservation [29], which leverage structural assumptions in the prediction space. Other strategies estimate modality reliabilities and latent labels jointly using spectral formulations [117], or determine instance-specific fusion weights by optimizing unsupervised criteria such as clarity-index maximization [82]. While these methods avoid the need for supervised fusion training, their performance depends on the validity of their assumptions, and they can be computationally expensive.

d) Uncertainty-aware: Finally, uncertainty-aware approaches explicitly quantify the confidence of each modality and use this information in the fusion function to prioritize more trustworthy and informative sources. For example, [160] estimated uncertainty using deep ensembles and fused predictions with weights derived from uncertainty estimates and modality correlations. A prominent line of work [113; 148; 90; 59] applies the Dempster–Shafer (DS) theory of belief functions [24; 127], a well-established framework for decision-making under uncertainty in which evidence is represented as mass functions and combined using Dempster's rule of combination or its variants [119; 107; 59]. A related family of methods leverages subjective logic [69], which extends DS theory by introducing prior beliefs, subjective opinions, and additional fusion operators [70]. Several recent works [52; 130; 173; 96; 168; 12] have successfully applied subjective logic for multimodal uncertainty quantification, as discussed in Section 4. The uncertainty-aware late fusion category thereby forms the conceptual bridge between fusion architecture design and the principled evidence-theoretic UQ methods reviewed in later sections.

In summary, as outlined in Table 1, late fusion offers flexibility in using modality-specific architectures, handles missing modalities well, and allows straightforward integration of new ones. Its main drawback is weaker modeling of cross-modal interactions and the common assumption that all modalities are equally reliable, which can lead to overcon-

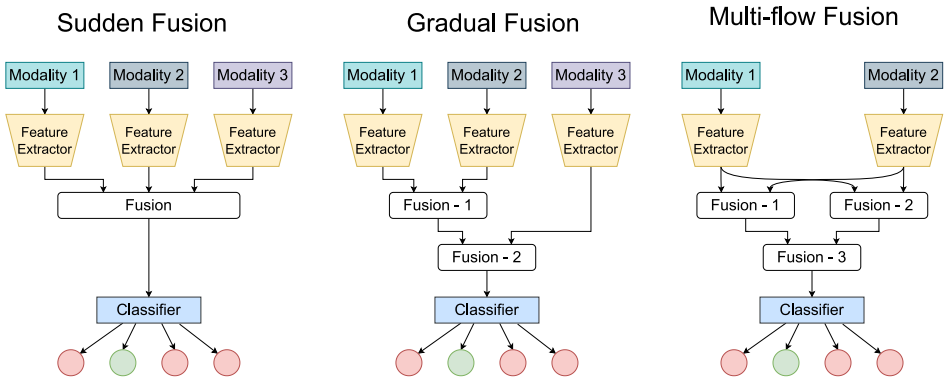

Figure 4: Examples of *sudden, gradual* and *multi-flow* intermediate fusion architectures.

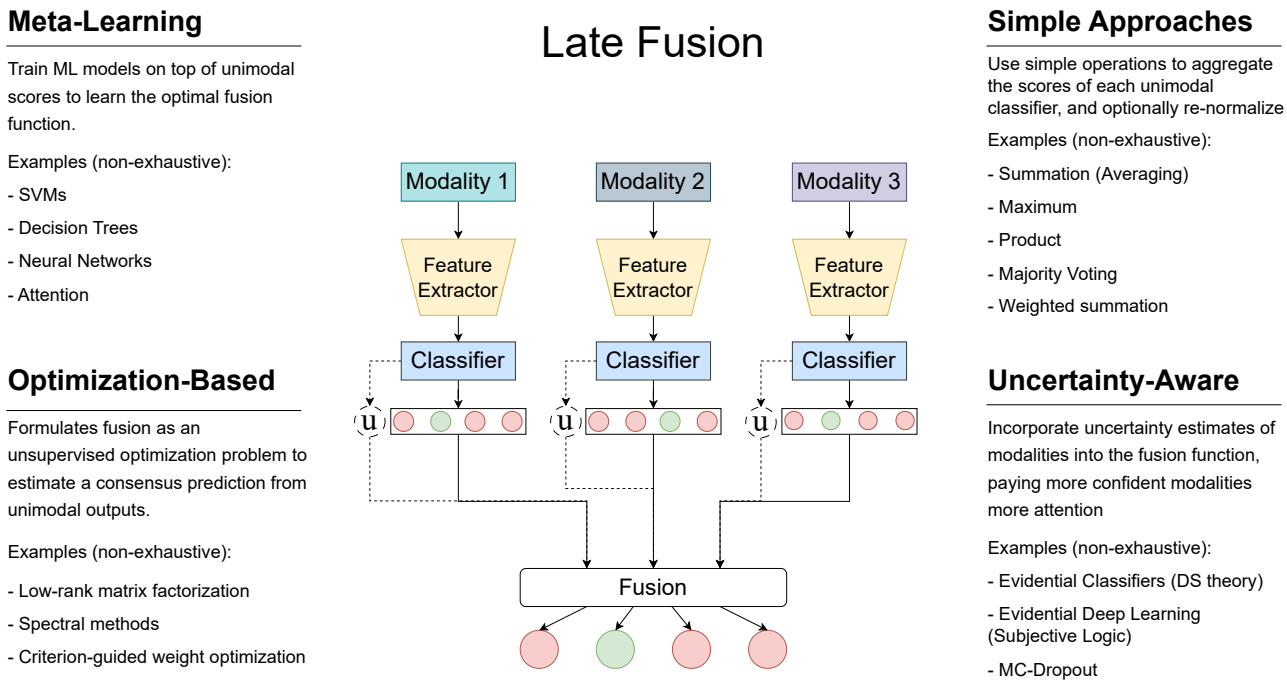

Figure 5: Overview of common multimodal late fusion strategies

fident or inaccurate predictions. Uncertainty-aware methods, such as those based on Dempster–Shafer theory, subjective logic, and other estimation frameworks, address this by weighting modalities according to their estimated reliability and handling potential conflicts.

The final classification stage is the endpoint of the multimodal pipeline, where the fused information is turned into the model's prediction. Depending on the setup, this input may be a joint representation from intermediate fusion or aggregated outputs from late fusion. The final classifier itself can be very simple, such as a linear layer, or more complex, such as a small neural network when richer decision boundaries are needed. The choice of classifier depends mainly on the task, the type of fused input, and the balance between simplicity and accuracy. In late fusion architectures, the fusion function often combines unimodal classifier outputs directly, so the fused result may serve as the final prediction, though some models still add a classifier on top to refine it.

Uncertainty implication. Fusion design governs how per-modality uncertainty propagates to the joint decision. Early and intermediate fusion collapse modality-specific confidence signals into a shared representation, making post-hoc per-modality uncertainty extraction difficult. Late fusion preserves per-modality uncertainty but requires an explicit combination rule; without one, differing modality confidences are silently equalized. Uncertainty-aware fusion strategies that address this limitation are reviewed in Section 4.

While fusion functions govern how information from different modalities is combined, and the final classifier generates

Table 1: Comparison of Fusion Strategies in Multimodal Learning

Fusion Strategy	Advantages	Limitations
Early Fusion	• Simple and effective for homogeneous modalities • Captures low-level cross-modal relationships	• Difficult to apply to heterogeneous modalities • May miss higher-level interactions • Sensitive to alignment and noise • Can struggle with missing modalities
Intermediate Fusion	• Allows modality-specific feature learning • Captures complex high-level interactions • Flexible fusion strategies (e.g., attention, gating)	• More complex design and training • May still struggle with missing modalities
Late Fusion	• Modular and easy to implement • Handles missing or unreliable modalities well • Allows using any architecture per modality	• Weak at modeling cross-modal interactions • Can produce overconfident outputs if modality reliability is not considered

the final prediction, the effectiveness of this combination ultimately depends on how the model is trained. In practice, training dynamics often cause certain modalities to dominate, leading to imbalance problems that undermine the benefits of fusion. We therefore turn next to learning objectives and modality imbalance.

2.2 Learning objectives and modality imbalance

In machine learning, and particularly in deep learning, the loss function plays a central role in shaping the behavior of the model. It determines not only what the model learns, but also how it balances generalization and overfitting, and whether it accounts for uncertainty in its predictions. The loss function encodes the objectives and constraints of the learning process, guiding optimization toward task-aligned solutions.

Beyond influencing output probabilities, the loss function also shapes the internal feature representations learned throughout the network. While in unimodal classification tasks the choice of loss is often straightforward, in multimodal settings it becomes more complex, as it affects both modality-specific learning and cross-modal integration. The optimization strategy has both direct and indirect effects on latent representations before and after fusion. A direct impact arises when loss functions explicitly govern feature representations and their interactions [49]. For example, [67] use contrastive learning with triplet loss to learn modality specific and shared representations. An indirect impact occurs when the loss is applied only at the model's output, while gradients still influence learned features.

Ideally, if all components of the multimodal architecture are trained optimally, the multimodal model should outperform, or at worst behave similarly to, the best unimodal model. However, multimodal models can underperform compared to unimodal models [143; 162]. [162] attribute this partly to overfitting, as multimodal models typically have more parameters and are more susceptible to it. Moreover, different modalities may learn at different rates, allowing faster-learning modalities to dominate training.

To mitigate this issue, [162] proposed optimizing modality-specific and multimodal losses jointly using adaptive weighting based on the overfitting-to-generalization ratio. Although Gradient Blending [162] can improve performance, it is computationally expensive and does not always find optimal weights, potentially leading to suboptimal results [11]. [166] refer to this phenomenon as *greedy learning*, where models rely excessively on easily optimized modalities. They propose estimating each modality's utilization rate from gradient norms and rebalancing training accordingly. Another approach, UMT [31], uses teacher unimodal networks to distill pre-trained unimodal features into the multimodal late fusion architecture. [174] integrate an unsupervised contrastive loss with supervised multimodal classification to address imbalance.

Although effective, these methods increase computational complexity through additional objectives or auxiliary networks. [11] show that such complexity does not necessarily ensure balanced modality learning, and propose simple deterministic weighting as a more efficient alternative. Designing simpler and more efficient modality balancing techniques remains an important research direction.

Uncertainty implication. Taken together, the limitations discussed in this section—unreliable cross-modal representations, fusion strategies that assume equal modality reliability, and training objectives that can amplify modality imbalance—highlight the need for principled uncertainty quantification in multimodal systems. Sections 3 and 4 address this

need, reviewing how uncertainty can be estimated, propagated, and leveraged for more reliable multimodal classification.

3. Uncertainty in Deep Learning

Deep learning models are increasingly being deployed across a wide range of domains, including safety-critical applications such as autonomous driving, medical diagnosis, and financial forecasting. In these contexts, incorrect predictions can lead to serious consequences, such as traffic accidents, misdiagnoses, or substantial financial losses. As a result, understanding the confidence of a model's predictions is essential for ensuring their trustworthiness and reliability. However, modern deep learning models are known to be poorly calibrated and often exhibit overconfidence, even when their predictions are incorrect [50]. To address this issue and improve the trustworthiness of such systems, it is crucial to develop methods that can accurately quantify the uncertainties in the prediction. Based on these uncertainty estimates, a model can either abstain from making a prediction under high uncertainty, or provide several plausible options in the form of a *set-valued classification*, where the true label is expected to be contained within the predicted set.

Although the methods reviewed in this section were primarily developed for unimodal settings, they form the essential building blocks for multimodal uncertainty quantification. In multimodal contexts, per-modality uncertainty estimates serve as inputs to uncertainty-aware fusion mechanisms (Section 4): a modality's reliability score is often derived directly from its predictive uncertainty, and the fusion operator must then combine these estimates in a way that accounts for inter-modal conflict. Understanding the assumptions and limitations of unimodal UQ methods is therefore directly relevant to assessing their suitability for multimodal deployment.

3.1 Types of Uncertainty

In literature, uncertainty is commonly categorized into two main types: *aleatoric* and *epistemic* uncertainty [79]. Figure 6 illustrates the difference between them.

Aleatoric uncertainty arises from inherent randomness in data, such as measurement noise or variability in the underlying process. It is often referred to as *irreducible uncertainty* [1; 55], since it cannot be eliminated by collecting more data or improving the model. However, some works argue that aleatoric uncertainty may be reduced by incorporating additional information, such as more features [64] or more modalities [56].

Epistemic uncertainty, in contrast, stems from a lack of knowledge about the model or data-generating process. It is commonly described as *reducible uncertainty* [1; 55], since it can be decreased with more data or improved modeling. [64] further distinguish between *model uncertainty*, related to the choice of model class, and *approximation uncertainty*, related to training data quality and quantity. [104] additionally introduce *distributional uncertainty*, which arises under distribution shift and is typically high for out-of-distribution (OOD) samples.

Distinguishing between aleatoric and epistemic uncertainty is useful for understanding uncertainty sources and designing appropriate quantification methods. Nevertheless, the distinction remains debated. For example, [110] showed that current methods struggle to disentangle the two types in practice, observing high correlation between them. [165] reported inconsistencies in entropy- and mutual-information-based decompositions of total uncertainty. Alternative definitions have been proposed [48; 124; 123], but no clear consensus has emerged. Therefore, while we acknowledge this distinction and refer to it when relevant, we do not rely heavily on it in the remainder of the paper.

3.2 Uncertainty Quantification Methods

Having defined the aleatoric and epistemic uncertainties in Section 3.1, we can now discuss the methods for quantifying them. Not all methods are able to quantify both types of uncertainty, and some methods are more suitable for certain types than others. In general, uncertainty quantification methods can be broadly categorized into four main categories: (1) Bayesian methods, (2) ensemble methods, (3) single network deterministic methods, and (4) test-time augmentation methods [45]. A high-level summary of these categories is provided in Figure 7.

Bayesian methods [84; 68; 125] apply Bayes' theorem to update prior beliefs $p(h)$ over hypotheses $h \in \mathcal{H}$ given data $\mathcal{D}$:

$$p(h|\mathcal{D}) = \frac{p(\mathcal{D}|h)\,p(h)}{p(\mathcal{D})}, \tag{1}$$

where $p(\mathcal{D}|h)$ is the likelihood and $p(\mathcal{D})$ is the evidence. The posterior $p(h|\mathcal{D})$ reflects the updated belief and captures *epistemic uncertainty*.

In *Bayesian Neural Networks* (BNNs), the hypotheses correspond to weight configurations θ. Instead of a single estimate, BNNs learn a distribution $p(\theta|\mathcal{D})$ and compute predictions via Bayesian model averaging:

$$p(y|x,\mathcal{D}) = \int p(y|x,\theta)\,p(\theta|\mathcal{D})\,d\theta. \tag{2}$$

This is generally intractable for modern networks, so *variational inference* (VI) approximates the posterior by $q_\phi(\theta)$, minimizing the KL-divergence $\mathrm{KL}(q_\phi \,\|\, p(\theta|\mathcal{D}))$. Since $p(\theta|\mathcal{D})$ is unknown, variational inference instead maximizes a loss called evidence lower bound (ELBO), which is equivalent to minimizing the KL-divergence loss up to a constant. *Bayes-by-Backprop* [15] enables training via the reparameterization trick.

Monte Carlo Dropout (MC-Dropout) [39] offers a lightweight approximation by training with dropout and sampling predictions at inference by keeping dropout active. The variance (or entropy) of these predictions estimates uncertainty. MC-Dropout is easy to implement, but studies [154] show its estimates are sensitive to dropout rate, model size, and target magnitude, and may not decrease with more data, limiting its reliability for epistemic UQ.

Gaussian Processes (GPs) [125] are a non-parametric Bayesian approach that models a distribution over functions rather than over finite-dimensional network weights. Formally, a GP is a collection of random variables such that any finite subset has a joint Gaussian distribution:

$$f(x) \sim \mathcal{GP}\big(m(x), k(x,x')\big), \tag{3}$$

where $m(x)$ is the mean function and $k(x,x')$ is a positive-definite kernel encoding correlations between inputs. Starting from a GP prior, conditioning on observed data yields a GP posterior for predictions and uncertainty estimation.

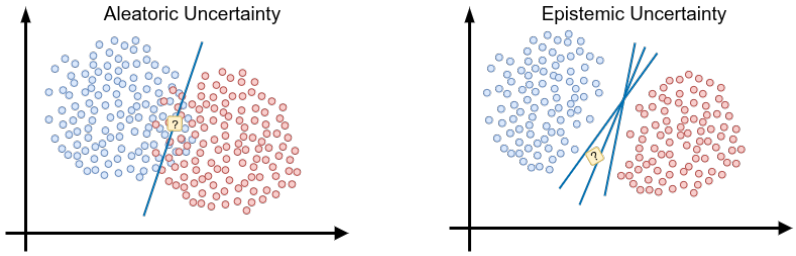

Figure 6: Examples of aleatoric and epistemic uncertainties. In the case of aleatoric uncertainty, even with infinite data and a perfect model, the sample marked with "?" cannot be confidently classified due to inherent overlap between classes. In the case of epistemic uncertainty, the sample cannot be confidently classified because multiple plausible decision boundaries exist, leading to different possible labels. Figure inspired by [64].

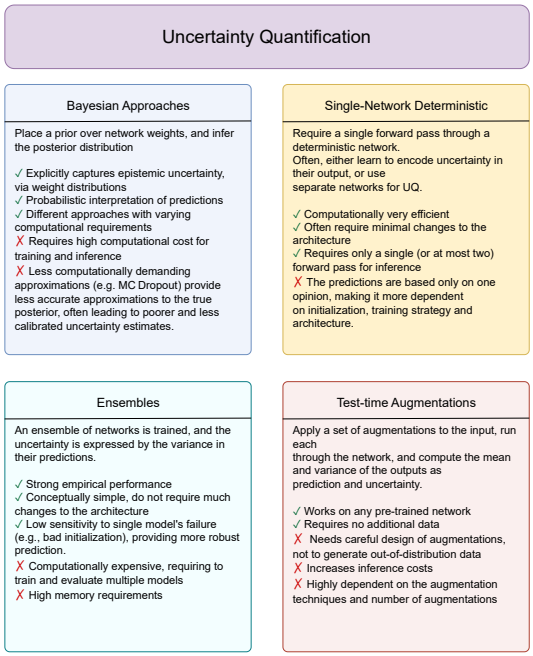

Figure 7: High-level summary of the main uncertainty-quantification paradigms in deep learning. Each quadrant lists general pros (✓) and cons (✗); individual methods may vary in their exact strengths and weaknesses.

While GPs provide well-calibrated uncertainty, their standard form scales as $\mathcal{O}(N^3)$ in the number of training points N. Sparse GPs [136; 145] reduce this to $\mathcal{O}(M^2N)$ using $M \ll N$ inducing points. Performance also depends strongly on kernel choice, which is challenging for structured data such as images. *Deep kernel learning* [164] addresses this by learning feature transformations with deep networks before applying the kernel, and *Deep Gaussian Processes* [22] stack multiple GPs to capture more complex, hierarchical functions.

Neural Processes (NPs) [43; 44] combine neural networks with features of GPs to learn distributions over functions in an end-to-end manner. They divide data into a *context set*, used to condition the model, and a *target set* for prediction. An encoder maps each context pair (x, y) to a latent representation, which is aggregated into a global latent variable. A decoder then combines this variable with target inputs to produce predictions.

Unlike GPs, NPs scale linearly as $\mathcal{O}(N + M)$ for N context and M target samples, making them suitable for large datasets. However, inference quality depends on the chosen context set. Extensions include *Convolutional Neural Processes* [37] for spatial data and *Attentive Neural Processes* [75] for improved context–target interactions.

In the multimodal setting, Bayesian methods and Gaussian Processes have been used to estimate per-modality epistemic uncertainty that is then passed as a reliability weight to uncertainty-aware fusion mechanisms, as discussed in Section 4.

The second category identified by [45] are **Ensembling methods**, which combine predictions from multiple models to improve accuracy and robustness [40]. They are related to Bayesian approaches through the idea of Bayesian model averaging [64]. [83] introduced *deep ensembles* as a practical alternative to Bayesian neural networks: networks trained with different random initializations learn diverse weight configurations, and their predictions are averaged. Uncertainty is estimated from the variance or entropy across ensemble members.

Deep ensembles are easy to implement and often highly performant, but incur high training and inference costs due to multiple full models. To reduce overhead, variants include *Snapshot Ensembles* [58], *Multi-head networks* [87], and *Ensemble Distillation* [106].

Deep ensembles and their lightweight variants have been applied to multimodal UQ by running modality-specific ensemble branches whose cross-member variance serves as a per-modality reliability signal in uncertainty-aware fusion (Section 4).

The third strategy focuses on **single-network deterministic methods**, which estimate uncertainty from a single forward pass of a deterministic network. Many of these methods predict the parameters of a second-order probability distribution over class probabilities, commonly the Dirichlet distribution. For example, *Dirichlet prior networks* [104] parameterize a prior over predictions and are trained to produce sharp Dirichlets for in-distribution (ID) inputs and uniform Dirichlets for out-of-distribution (OOD) inputs, requiring both ID and OOD samples. Similarly, *Evidential Deep Learning* (EDL) [126] parameterizes the *poste-*

rior Dirichlet directly, maximizing evidence for the correct class while encouraging uncertainty (uniform Dirichlet) for others. EDL requires only ID data and will be discussed in more detail later, as it forms a core component of this thesis. Other variants [105; 151; 111; 181] explore both prior and posterior formulations, with differing OOD training needs.

EDL [126] should not be confused with *evidential classification* approaches [26; 102; 147] based on Dempster–Shafer theory (DST) [24; 127]. While subjective logic [69] is conceptually related to DST, the methods differ: EDL predicts Dirichlet parameters to model epistemic uncertainty, whereas DST-based approaches compute mass functions over class hypotheses and combine them via Dempster's rule, deriving uncertainty from residual mass or pignistic probability. In this article, *EDL* refers to the subjective logic formulation unless otherwise specified.

Another group of deterministic approaches are *gradient-based methods*, which infer uncertainty from the magnitude of network gradients at inference. Large gradients indicate greater parameter adjustment would be needed to fit the input, implying higher epistemic uncertainty. [85] used gradients with confounding labels to train an OOD detector, while [61] proposed GradNorm, measuring the gradient norm of the KL divergence between the softmax output and the uniform distribution, requiring no extra classifier. Recent work includes low-rank gradient norms [10] and extensions to segmentation [103]. Gradient-based UQ can be applied post-hoc to trained models without retraining.

Evidential Deep Learning (EDL) via subjective logic is the single-network deterministic method most directly extended to the multimodal setting: per-modality EDL networks generate subjective opinions that are combined by specialized fusion operators, forming the core of TMC, ECML, and related approaches reviewed in Section 4.

Test Time Augmentation methods [158; 7; 101] try to create several augmentations for each test sample, and then pass them through the model to obtain the predictive distribution. Although it is a simple technique, there are many open questions such as what types of (valid) transformations one shall use, how many, and what's the quality of the quantified uncertainty. One solution to the problem of choosing the right transformations was suggested by [158], who proposed a search algorithm to find test-time augmentations policy, based on the predictive performance on validation set. Similar to deep ensembles and Bayesian methods, test time augmentation also requires multiple forward passes through the model, which can be computationally expensive.

Test-time augmentation has been used in multimodal settings as a lightweight alternative to ensemble-based reliability estimation: augmentation variance for each modality provides an uncertainty signal that can be integrated into adaptive fusion strategies (Section 4).

In summary, uncertainty quantification in deep learning can be achieved through a variety of approaches, ranging from Bayesian inference and ensemble strategies to deterministic single-pass and gradient-based methods. Each paradigm offers trade-offs in terms of computational cost, scalability, and the type of uncertainty captured, with no single approach being universally optimal. In the multimodal setting, these challenges are amplified by potential modality conflicts and varying information quality, making reliable UQ essential for robust decision-making. Specifically, Bayesian and ensemble methods are well suited for estimating per-modality epistemic uncertainty, while single-network deterministic methods such as Evidential Deep Learning (EDL) offer the computational efficiency required for real-time fusion. Section 4 shows how these paradigms are extended to multimodal architectures, where per-modality uncertainty estimates must be combined alongside conflict detection.

Table 2: Mapping of unimodal UQ paradigms to their multimodal extensions reviewed in Section 4. Each paradigm contributes a specific type of per-modality signal that uncertainty-aware fusion mechanisms consume.

UQ Paradigm (§3.2)	**Role in multimodal UQ (§4)**
Bayesian / MC-Dropout	Per-modality epistemic uncertainty → reliability weights in ensemble fusion
Deep Ensembles	Cross-member variance → modality confidence in uncertainty-aware late fusion
EDL / Subjective Logic	Subjective opinions → BCF/CBF/ECML multimodal fusion (TMC, ECML, MMLF)
Dempster–Shafer	Belief functions → DS combination rule for multimodal evidence fusion
Test-time Augmentation	Augmentation variance → per-modality reliability in adaptive fusion

One way to utilize uncertainty estimates in safety-critical or high-stakes applications is to allow the model to abstain from overly confident single-label predictions when uncertainty is high. *Set-valued classification* (SVC) formalizes this idea by returning a set of plausible labels whose size reflects the model's uncertainty, thereby reducing the risk of critical misclassifications while still providing actionable information. In the next section, we review the principles and methods of SVC, their advantages and shortcomings.

3.3 Set-valued classification

Most UQ techniques operate in the *precise classification* setting[1], where the model outputs one class or abstains under high uncertainty. However, in assisted decision-making scenarios such as medical diagnosis or risk assessment, suggesting a small set of plausible labels is often more useful than fully rejecting a prediction. *Set-valued classification* (SVC) [19] addresses this by returning a subset of candidate labels in uncertain cases. An example of application of SVC is illustrated in Figure 8. Empirical evidence [21] shows that such set-valued predictions can improve human decision-making compared to fixed top-k suggestions or no assistance.

There is a natural correspondence between epistemic uncertainty and the size of the prediction set [64]. Large sets reflect high uncertainty, while singleton sets correspond to

[1] *Precise classification* refers to returning a single label, in contrast to *set-valued* or *imprecise* classification, which returns multiple labels.

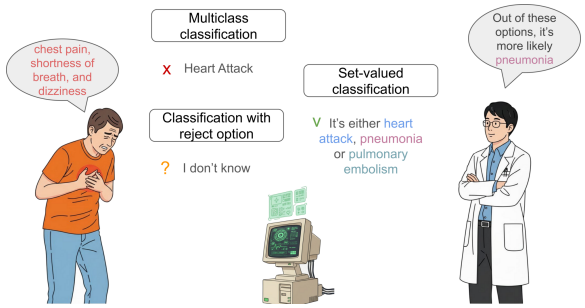

Figure 8: An example of application of set-valued classification in healthcare. Unlike multi-class classification and classification with reject option, set-valued classification proposes several plausible diagnosis options to the doctor, who makes the final decision.

confident predictions. However, under the open-world assumption, high epistemic uncertainty may indicate out-of-distribution samples. In such cases, predicting the full label set is inappropriate, and many approaches combine SVC with rejection or "null set" predictions [64].

Set-valued classification thus acts as a decision-level complement to the UQ methods surveyed in Section 3: where those methods estimate uncertainty, SVC translates it into actionable prediction sets. This relationship is especially important in multimodal settings, where disagreement between modalities often produces structured, irreducible ambiguity that is better expressed by a set of plausible classes than by a single forced prediction. The methods reviewed below are therefore directly relevant to the uncertainty-aware multimodal architectures discussed in Section 4.

Several methodological families have been proposed.

3.3.1 Top-k and Thresholding Approaches

Top-k classification returns the k most probable labels, but uses a fixed set size regardless of confidence. Threshold-based approaches include all labels whose predicted probability exceeds a predefined threshold, allowing adaptive set sizes but being sensitive to calibration errors [109]. Average-k strategies [25; 42] control the expected set size during training. While simple and efficient, these approaches rely directly on predicted probabilities and do not provide formal uncertainty guarantees.

3.3.2 Coverage-based Approaches

Conformal Prediction (CP) [155; 128] constructs prediction sets with distribution-free coverage guarantees. Efficient variants such as split conformal prediction [18] make the method practical for large-scale settings. However, CP guarantees only marginal coverage and may produce overly large sets [109; 163]. Moreover, classical CP does not provide calibrated probabilities for labels within the set, though recent extensions address this limitation [80].

3.3.3 Utility-based Approaches

Utility-based methods [23; 20; 178; 109] formalize SVC as an expected utility maximization problem. Let $u(y, \hat{Y})$ denote the utility of predicting set $\hat{Y}$ when the true label is y:

$$u(y, \hat{Y}) = \begin{cases} 0, & y \notin \hat{Y}, \\ g(|\hat{Y}|), & y \in \hat{Y}, \end{cases} \tag{4}$$

where g controls the trade-off between accuracy and set size. The Bayes-optimal set maximizes the expected utility under $P(c|x)$. In practice, efficient algorithms evaluate only a limited number of candidate sets [109]. These approaches assume well-calibrated probabilities, which motivates integrating uncertainty-aware models [50].

3.3.4 Evidence-Theoretic and Imprecise Probability Approaches

Dempster–Shafer (DS) theory [24; 127] and related evidential neural networks [26] naturally extend to SVC by assigning belief mass to subsets of labels. Extensions to deep learning [102; 147; 27] enable set-valued decisions while preserving uncertainty modeling, though computational complexity can increase with the number of classes. The same DS framework is extended to multimodal fusion in Section 4, where combination rules operate on modality-level evidence sources rather than on class-level belief functions within a single model.

Imprecise probability approaches [177; 172; 156; 150] represent uncertainty via sets of probability distributions (credal sets), supporting cautious decision rules. While theoretically appealing, such methods may be computationally demanding or difficult to scale [114].

Subjective logic–based approaches [69; 126] also provide mechanisms for representing composite hypotheses. Extensions such as HENN [88] model classification ambiguity but are often tailored to multi-label rather than classical SVC settings. The EDL component of these methods is extended to multimodal classification in Section *4.1.2*, where per-modality evidential networks produce subjective opinions that are combined via specialized SL fusion operators such as BCF, CBF, and ECML.

3.3.5 Summary of Limitations

Despite their diversity, existing SVC approaches face common challenges. Thresholding methods lack robustness to calibration errors. Conformal prediction provides coverage guarantees but may produce large sets. Utility-based methods depend on reliable probability estimates. Evidence-theoretic and imprecise probability approaches can become computationally demanding for large label spaces. Overall, effective SVC requires accurate uncertainty quantification, making it closely tied to the quality of the underlying UQ method. In multimodal settings, this dependency becomes even more pronounced: unreliable per-modality uncertainty estimates (e.g., due to modality imbalance or conflict, as identified in Section 2) propagate directly into suboptimal prediction sets. Addressing these limitations motivates the conflict-aware multimodal UQ approaches reviewed in the following section.

SVC in multimodal settings. The SVC methods reviewed above were developed primarily for unimodal models, yet they apply directly to multimodal systems via late fusion: each modality produces an independent evidence mass or probability distribution, which are combined before the SVC decision rule is applied. In this sense, SVC is a natural companion to uncertainty-aware late fusion (Section 2.1): where fusion produces a joint uncertainty estimate over the label space, SVC translates that estimate into a calibrated prediction set. Disagreement between modalities—a form of structured aleatoric uncertainty that cannot be reduced by gathering more data from a single source—is especially well handled by evidence-theoretic SVC methods, since DS theory and subjective logic can represent conflictive opinions as

high-uncertainty mass distributions before any decision rule is applied. The multimodal UQ frameworks reviewed in the following section leverage precisely these properties.

4. Multimodal uncertainty quantification

In the previous sections, we discussed uncertainty quantification (UQ) and set-valued classification (SVC) in the unimodal setting. Multimodal deep learning introduces additional challenges, especially for UQ and SVC. Ideally, incorporating complementary information from multiple modalities should reduce aleatoric uncertainty [56]. In practice, however, uncertainty can also increase when modalities are misaligned or contradictory. For example, in medical diagnosis, an X-ray may suggest one condition while an MRI suggests another. In such cases, it is essential to estimate both per-modality uncertainty and the combined uncertainty of the multimodal system.

While many UQ methods exist for unimodal learning, multimodal UQ remains relatively less explored. A straightforward approach is to treat the multimodal architecture as a single model with one input and one output, then apply unimodal UQ methods directly. However, this ignores modality-specific properties and prevents separate uncertainty estimation for each modality. Hence, various frameworks and approaches have been proposed for quantifying and integrating uncertainty into multimodal deep learning. These methods build directly on the fusion architectures reviewed in Section 2: uncertainty-aware late fusion (Section 2.1) already incorporates elementary modality-reliability weighting. The approaches discussed here extend this idea with principled probabilistic and evidence-theoretic frameworks, enabling systematic conflict detection and uncertainty propagation that simpler fusion strategies cannot provide. Figure 9 provides a brief timeline of some of the key approaches in this domain. For clarity,

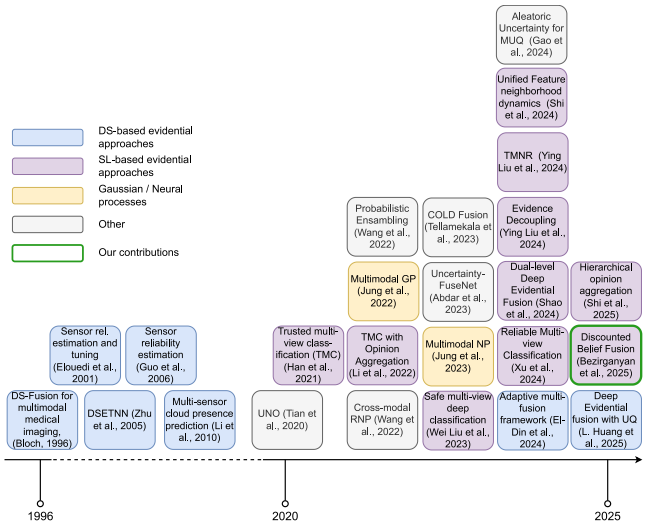

Figure 9: A timeline of non-exhaustive collection of key approaches in multimodal uncertainty quantification.

4.1 Evidence-Theoretic Approaches

One family of methods adopts evidence-theoretic frameworks to explicitly model uncertainty, belief, and inter-modal conflict. These approaches depart from classical probabilistic representations and operate on mass functions or opinion-based representations.

4.1.1 Dempster–Shafer Theory

Among the frameworks for fusing uncertain information, *Dempster–Shafer* (DS) theory [24; 127] is one of the most widely used in multimodal learning. It offers a principled way to combine evidence from multiple sources and to model both uncertainty and imprecision. [14] highlighted features that make DS theory particularly suitable for multimodal classification:

- flexible modeling of uncertainty and imprecision,
- ability to handle variable source reliability, and
- a well-defined combination rule for merging independent evidence.

The *Dempster's rule of combination* fuses multiple beliefs into a single coherent representation. However, in cases of strong conflict between sources, it can yield counter-intuitive results [176]. This limitation has led to numerous alternative combination functions and conflict management strategies [107].

Before the deep learning era, DS theory was already applied to multimodal classification and sensor fusion [35; 9; 51; 89], laying the groundwork for modern architectures. More recently, DS theory has been incorporated into multimodal deep learning [33; 59]. The work of [59] introduced a learnable discounting factor for modality reliability. While effective, their approach assigns a fixed discount per modality and class, limiting adaptability to discrepancies between training and deployment. In contrast, [12] compute sample-specific reliability estimates, enabling dynamic adjustment to modality misalignment, noise, and other real-world inconsistencies.

4.1.2 Subjective Logic

Subjective Logic (SL) [69] extends DS theory, retaining its evidence-combination capabilities while offering a flexible representation of uncertainty and imprecision. [52] introduced one of the first SL-based multimodal deep learning methods, *Trusted Multi-view Classification* (TMC), in which each modality is modeled by an evidential deep learning network [126], and evidences are fused via the *Belief Constrained Fusion* (BCF) rule. Similar BCF-based strategies have been explored by [130] and [173], but since BCF is a SL adaptation of Dempster's rule, it inherits its limitations under high conflict.

[96] employed *Cumulative Belief Fusion* (CBF), an SL operator designed for independent sources contributing new evidence, thereby always reducing uncertainty [69]. However, this behavior can be problematic when strongly conflicting views are fused.

[168] proposed *Evidential Conflictive Multi-view Learning* (ECML), which uses average pooling in belief space and a conflict-penalizing loss. Intended for dependent sources, ECML decreases uncertainty if the new view is less uncertain and increases it if more uncertain. For two views, the combined uncertainty is the harmonic mean of individual uncertainties. Nonetheless, two low-uncertainty but conflicting views can produce undesirably low combined uncertainty, and the non-associative nature of the operator [69] makes it sensitive to fusion order.

To further enhance reliability, [100] introduced trust-based discounting, learning modality- and input-specific trust scores via a separate network. While effective, this adds computational cost and may not generalize well from clean training data to noisy or misaligned test data.

[95] approach the problem from a different perspective, proposing a fusion method that guarantees the fused prediction is at least as accurate as the best unimodal prediction. In other words, their approach ensures that fusion does not deteriorate the prediction quality. [169] addressed the potential label noise in training data, by estimating per-view uncertainty and refining mislabeled samples using pseudo-labels and mixup [179].

In summary, SL-based multimodal UQ methods encompass a range of fusion strategies, including adaptation of Dempster's rule, cumulative and averaging fusion, and more recent approaches that introduce conflict mitigation or trust-based discounting. While these approaches improve robustness and reliability, they still face challenges, especially under strong inter-modal disagreement, and generalization to noisy or misaligned modalities.

4.2 Non-Evidence-Based Multimodal UQ Approaches

In contrast to evidence-theoretic methods, the following approaches estimate uncertainty through probabilistic modeling, representation-level signals, or uncertainty-guided fusion mechanisms, without explicitly modeling inter-modal conflict. For example, [41] proposed *Embracing Aleatoric Uncertainty* (EAU), which models per-modality aleatoric uncertainty via Gaussian embeddings and learns a fusion that is robust to noisy inputs. [16] introduced *HyperDUM*, a deterministic feature-level UQ method based on hyperdimensional computing, quantifying channel- and patch-level epistemic uncertainty before fusion. Other works, such as [17] and [2], leverage ensemble-based techniques or Monte Carlo dropout to estimate predictive uncertainty, without explicitly modeling conflicts between modalities.

Cross-modal Random Network Prediction [161] estimates uncertainty by comparing the outputs of a fixed, randomly initialized network with a smaller, trainable predictor over the feature space, leveraging discrepancies to assess uncertainty. These uncertainty estimates then inform a fusion mechanism that adaptively weights modalities during classification or segmentation tasks.

The *uncertainty-aware noisy-or (UNO)* approach [144] combines multiple uncertainty metrics, such as predictive entropy, mutual information, and a novel spatial temperature network, and propose a novel Noisy-Or fusion, which takes into account the uncertainties of the modalities, prioritizing more confident ones. *COLD Fusion* [142] models each modality as a latent Gaussian distribution and interprets the variance of these distributions as a measure of modality's confidence.

[73] and [72] propose alternative multimodal UQ approaches based on Gaussian processes and neural processes, respectively. While these methods perform well, the Multimodal Gaussian Process is computationally expensive due to its non-parametric nature. The Multimodal Neural Process is relatively faster; however, its results are highly dependent on the chosen context set, and there are currently no theoretically guaranteed methods to obtain an optimal context set.

While these approaches often improve robustness to noise or missing modalities, they generally assume either statistical independence or implicit alignment between modalities. As a result, they may produce overconfident predictions when strong inter modal disagreement occurs. To address this issue, several works have proposed to explicitly separate different types of information carried by each modality. For instance, [131] suggested disentangling common and view specific information between modalities. The objective is to isolate features consistently shared across views from those that remain unique to each modality. This separation allows the model to rely on common representations for cross view agreement while preserving view specific cues that may provide complementary discriminative information. Building on this idea, [97] proposed a Dynamic Evidence Decoupling framework that operates in the evidential space. In this formulation, each view's opinion is decomposed into *consistent* evidence, shared across modalities and trained to align with the ground truth, and *complementary* evidence, which captures modality specific signals and is allowed to remain uncertain. However, separating shared and modality specific evidence does not fully address situations where modalities strongly disagree. To explicitly account for such conflicts, [12] introduced a multimodal classification framework based on evidential deep learning and subjective logic. Their method detects conflictive modalities and applies a sample specific discount factor to their evidence, increasing predictive uncertainty when modalities disagree while maintaining low uncertainty for well aligned inputs. Since conflict detection, discounting, and fusion are parameter free, the approach remains efficient and can handle conflicts at test time even when such cases were absent during training.

textbfSummary. Overall, no single uncertainty quantification approach universally dominates in multimodal settings. Probabilistic methods provide strong theoretical grounding but often face scalability challenges, while ensemble-based approaches offer robustness at the cost of computational efficiency. Evidence-theoretic frameworks are particularly well suited to multimodal learning, as they explicitly model inter-modal conflict, but may suffer from instability under high disagreement. These trade-offs highlight the need for hybrid approaches that balance robustness, scalability, and principled uncertainty modeling.

5. Evaluation of Multimodal UQ

Evaluation protocols for multimodal uncertainty quantification remain heterogeneous across studies. Most works report task-specific metrics such as accuracy, F1 score, or correlation to evaluate predictive performance, while uncertainty-related indicators such as predictive entropy are sometimes used to assess the confidence of model predictions. However, there is currently no standardized evaluation protocol specifically designed for multimodal uncertainty quantification.

In practice, evaluating uncertainty-aware models is further complicated by the fact that the true level of uncertainty present in the data is rarely known, and datasets may not accurately reflect the variability encountered in real-world environments. As a result, it becomes difficult to assess how well uncertainty quantification algorithms perform under different uncertainty conditions. Moreover, deep learning models may behave differently depending on the level of uncertainty in the data, for example when inputs are corrupted by noise or contain ambiguous information. For this reason,

several works introduce controlled perturbations in the data, such as additional noise, in order to analyze how uncertainty estimates and predictive performance evolve under varying uncertainty levels. Since different uncertainty quantification approaches target different types of uncertainty, having mechanisms that allow the injection of diverse uncertainty sources can facilitate more systematic evaluation..

Several datasets have been used in multimodal uncertainty quantification settings. A notable line of work [52; 73; 71] has employed datasets such as HandWritten[2], CUB[3], Scene15[4], and Caltech101[5]. These datasets typically extract different features from unimodal sources to create a multi-view setup. While they have been instrumental, they primarily repurpose unimodal data for multimodal tasks, underscoring the need for more comprehensive and inherently multimodal datasets to better evaluate uncertainty in deep learning models.

Furthermore, the current approaches that introduce uncertainty into the data [52; 73; 71] add Gaussian noise to the views or the extracted features. While Gaussian noise does increase uncertainty, it does not accurately reflect the noise that can be found in real-world datasets and this process lacks fine-grained control over the type of uncertainty being injected.

Additionally, how different modalities' uncertainties interact significantly impacts the overall multimodal uncertainty. When both modalities encode redundant information, the total uncertainty might not decrease. Conversely, conflicting information can lead to increased uncertainty, while complementary information can reduce it. A deeper understanding of these phenomena is crucial. Fine-grained control over individual modalities' uncertainties opens the way for more theoretical research based on empirical observations.

To support the analysis of uncertain multimodal data and the evaluation of uncertainty quantification techniques in multimodal learning, [13] introduce a dataset together with an uncertainty generator package. This package provides several mechanisms for injecting uncertainty, including controlling data diversity, adding different types of real-world noise, randomly switching labels to their closest class, and injecting out-of-distribution (OOD) data.

6. Discussion and Research Directions

This survey presented an integrated perspective on multimodal classification under uncertainty, highlighting how design choices in representation learning, fusion strategies, and training objectives shape the emergence of uncertainty. We showed that uncertainty quantification is not merely an auxiliary component, but a central element for building reliable multimodal systems, particularly in the presence of noisy, incomplete, or conflicting modalities. However, current multimodal pipelines still tend to treat fusion, uncertainty estimation, and decision-making as loosely coupled stages. Fusion mechanisms often fail to account for partial dependence between modalities; uncertainty quantification methods frequently rely on restrictive independence assumptions or conflict-prone training settings; and decision layers such as set-valued classification are commonly implemented as post-hoc components rather than being tightly integrated with uncertainty-aware fusion. Crucially, the value of uncertainty quantification lies not only in estimating uncertainty, but also in enabling more informed and reliable decision-making processes.

Table 3: Cross-reference: structural limitations identified in Section 2 and the multimodal UQ methods in Section 4 that directly address them. This table illustrates the coherent progression from fusion architecture to uncertainty-aware design.

§2 Limitation	§4 Addressing Method
Equal-reliability fusion assumption	Uncertainty-aware late fusion (DS, SL); sample-specific discounting [12]
No per-modality confidence signal	EDL-based modality networks (TMC [52], ECML [168])
Modality imbalance / dominant modality	Trust-based discounting [100]; conflict-penalizing loss [168]
No inter-modal conflict detection	DS combination with conflict management [107; 59]; sample-specific discounting [12]
Overconfident joint prediction	SVC decision layer translating uncertainty into prediction sets (§3.3)

Beyond uncertainty estimation, we emphasized the importance of decision-level strategies such as set-valued classification, which translate uncertainty into actionable predictions. This is particularly relevant in multimodal settings, where disagreement between modalities naturally leads to ambiguity that cannot be adequately captured by single-label predictions.

Taken together, these observations highlight the need for a unified view that connects multimodal design, uncertainty estimation, and decision-making, with fusion mechanisms playing a central role in propagating and resolving uncertainty across modalities. Moving forward, key research directions include the development of scalable and conflict-aware uncertainty quantification methods, the establishment of standardized evaluation protocols, and tighter integration between uncertainty modeling and downstream decision-making processes. These advances are essential for deploying trustworthy multimodal AI systems in real-world applications.

6.1 Axis I: Modeling Inter-Modal Dependence for Reliable Fusion

A key open challenge in multimodal uncertainty quantification is moving beyond binary assumptions of full independence or full dependence between modalities. As discussed in Section *4.1.2*, subjective logic provides different fusion operators for independent and dependent opinions. However, real-world multimodal data typically exhibit varying degrees of partial dependence that cannot be adequately captured by these extreme assumptions.

For instance, Cumulative Belief Fusion (CBF) assumes inde-

[2]https://archive.ics.uci.edu/ml/datasets/Multiple+Features
[3]http://www.vision.caltech.edu/visipedia/CUB-200.html
[4]https://serre-lab.clps.brown.edu/resource/hmdb-a-largehuman-motion-database
[5]https://data.caltech.edu/records/mzrjq-6wc02

pendence and aggregates evidences additively, while Averaging Belief Fusion (ABF) assumes dependence and averages them as if redundant. In practice, modalities may share some information while also contributing distinctive signals. A promising direction is therefore to model inter-modal dependence in a continuous manner rather than as a binary choice.

One potential approach is disentangled information modeling [86; 157], where modality-specific opinions are decomposed into shared and modality-unique components. By separating overlapping and distinctive information before fusion, it becomes possible to apply more defensible independence assumptions at the aggregation stage. Such disentangled fusion mechanisms could lead to more accurate uncertainty estimates by explicitly accounting for information redundancy and partial dependence.

6.2 Axis II: Systematic Evaluation of Multimodal Uncertainty

Another major limitation identified in our review is the lack of standardized and controlled benchmarks for multimodal uncertainty quantification. Without systematic evaluation protocols, it remains difficult to compare methods fairly or to assess their robustness under varying levels of conflict, noise, or modality misalignment.

Future efforts should extend datasets such as LUMA by incorporating additional modalities with controlled interdependencies. For example, adding automatically generated textual descriptions of images would introduce causal relationships between modalities and allow controlled study of uncertainty propagation across correlated inputs.

In parallel, developing standalone toolkits capable of injecting controlled perturbations—such as noise, misalignment, modality dropout, and out-of-distribution samples—into existing multimodal datasets would enable reproducible and systematic evaluation of multimodal UQ methods. Such tools would facilitate rigorous comparison across approaches and promote standardized evaluation protocols for conflict-aware learning.

6.3 Axis III: From Multimodal Classification to Agentic Systems

The reliability challenges discussed in this survey naturally extend beyond classification to emerging agentic AI systems, where multimodal models are entrusted with autonomous decision-making [167]. Large language models, often forming the reasoning backbone of agents, are known to produce hallucinations and confidently generate incorrect outputs [60]. In addition, agents integrate information from external sources such as web search results, APIs, or knowledge bases, which may themselves be noisy or unreliable. Multimodal inputs—including text, images, audio, and structured data—can also contain internal conflicts or redundancies.

Ensuring reliability in such systems requires mechanisms for quantifying uncertainty and resolving conflicts both within multimodal inputs and across interacting agents [159; 36]. Uncertainty quantification in large language models [132], as well as conflict resolution in multi-agent systems [149], are active research areas. Recent discussions [77] suggest that new conceptual frameworks may be required to adequately capture the unique uncertainty sources of large-scale agentic systems.

Future research should explore how uncertainty can become an explicit component of agent reasoning rather than a residual by-product. Agents should be able to assess the trustworthiness of external sources [183; 149], reconcile conflicting information streams, and adapt their decisions accordingly. Ultimately, advancing conflict-aware multimodal fusion and uncertainty-driven decision strategies will be crucial for building transparent and reliable multimodal agents capable of reasoning about their own limitations.

Overall, advancing multimodal reliability requires integrating representation learning, uncertainty quantification, and decision strategies within a unified framework that explicitly models dependence, conflict, and uncertainty propagation across heterogeneous information sources.

7. REFERENCES

[1] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

[2] M. Abdar, S. Salari, S. Qahremani, H.-K. Lam, F. Karray, S. Hussain, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi. Uncertainty-fusenet: robust uncertainty-aware hierarchical feature fusion model with ensemble monte carlo dropout for covid-19 detection. *Information Fusion*, 90:364–381, 2023.

[3] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[4] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013.

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

[6] N. Audebert, C. Herold, K. Slimani, and C. Vidal. Multimodal deep networks for text and image-based document classification. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 427–443. Springer, 2019.

[7] M. S. Ayhan and P. Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*, 2018.

[8] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[9] O. Basir, F. Karray, and H. Zhu. Connectionist-based dempster-shafer evidential reasoning for data fusion. *IEEE Transactions on Neural Networks*, 16(6):1513–1530, 2005.

[10] S. Behpour, T. L. Doan, X. Li, W. He, L. Gou, and L. Ren. Gradorth: A simple yet efficient out-of-distribution detection with orthogonal projection of gradients. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[11] G. Bezirganyan, S. Sellami, L. Berti-Équille, and S. Fournier. M2-mixer: A multimodal mixer with multi-head loss for classification from multimodal data. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1052–1058. IEEE, 2023.

[12] G. Bezirganyan, S. Sellami, L. Berti-Équille, and S. Fournier. Multimodal learning with uncertainty quantification based on discounted belief fusion. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pages 3142–3150. PMLR, 2025.

[13] G. Bezirganyan, S. Sellami, L. Berti-Équille, and S. Fournier. Luma: A benchmark dataset for learning from uncertain and multimodal data. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025.

[14] I. Bloch. Some aspects of dempster-shafer evidence theory for classification of multi-modality medical images taking partial volume effect into account. *Pattern Recognition Letters*, 17(8):905–919, 1996.

[15] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.

[16] L. Chen, J. Wang, T. Mortlock, P. Khargonekar, and M. A. Al Faruque. Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22306–22316, 2025.

[17] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, and S. Kong. Multimodal object detection via probabilistic ensembling. In *European Conference on Computer Vision*, pages 139–158. Springer, 2022.

[18] G. Cherubin, K. Chatzikokolakis, and M. Jaggi. Exact optimization of conformal predictors via incremental and decremental learning. In *International Conference on Machine Learning*, pages 1836–1845. PMLR, 2021.

[19] E. Chzhen, C. Denis, M. Hebiri, and T. Lorieul. Set-valued classification–overview via a unified framework. *arXiv preprint arXiv:2102.12318*, 2021.

[20] G. Corani and M. Zaffalon. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *J. Mach. Learn. Res.*, 9:581–621, 2008.

[21] J. C. Cresswell, Y. Sui, B. Kumar, and N. Vouitsis. Conformal prediction sets improve human decision making. In *International Conference on Machine Learning*, pages 9439–9457. PMLR, 2024.

[22] A. Damianou. *Deep Gaussian processes and variational propagation of uncertainty*. PhD thesis, University of Sheffield, 2015.

[23] J. J. del Coz, J. Díez, and A. Bahamonde. Learning nondeterministic classifiers. *J. Mach. Learn. Res.*, 10:2273–2293, 2009.

[24] A. P. Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.

[25] C. Denis and M. Hebiri. Confidence sets with expected sizes for multiclass classification. *Journal of Machine Learning Research*, 18(102):1–28, 2017.

[26] T. Denoeux. A neural network classifier based on dempster-shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(2):131–150, 2000.

[27] L. Deregnaucourt, A. Lechervy, H. Laghmara, and S. Ainouz. An evidential deep network based on dempster-shafer theory for large dataset. *Advances and Applications of DSmT for Information Fusion*, page 907, 2023.

[28] T. Diethe, D. R. Hardoon, and J. Shawe-Taylor. Constructing nonlinear discriminants from multiple data views. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part I 21*, pages 328–343. Springer, 2010.

[29] X. Dong, Y. Yan, M. Tan, Y. Yang, and I. W. Tsang. Late fusion via subspace search with consistency preservation. *IEEE Transactions on Image Processing*, 28(1):518–528, 2018.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[31] C. Du, J. Teng, T. Li, Y. Liu, T. Yuan, Y. Wang, Y. Yuan, and H. Zhao. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning*, pages 8632–8656. PMLR, 2023.

[32] P. Duchnowski, U. Meier, and A. Waibel. See me, hear me: integrating automatic speech recognition and lip-reading. In *ICSLP*, volume 94, pages 547–550. Citeseer, 1994.

[33] D. M. El-Din, A. E. Hassanein, and E. E. Hassanien. An adaptive and late multifusion framework in contextual representation based on evidential deep learning and dempster–shafer theory. *Knowledge and Information Systems*, 66(11):6881–6932, 2024.

[34] B. B. Elallid, N. Benamar, A. S. Hafid, T. Rachidi, and N. Mrani. A comprehensive survey on the application of deep and reinforcement learning approaches in autonomous driving. *Journal of King Saud University-Computer and Information Sciences*, 34(9):7366–7390, 2022.

[35] Z. Elouedi, K. Mellouli, and P. Smets. The evaluation of sensors' reliability and their tuning for multisensor data fusion within the transferable belief model. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 350–361. Springer, 2001.

[36] E. Fadeeva, A. Rubashevskii, R. Vashurin, S. Dhuliawala, A. Shelmanov, T. Baldwin, P. Nakov, M. Sachan, and M. Panov. Faithfulness-aware uncertainty quantification for fact-checking the output of retrieval augmented generation. *arXiv preprint arXiv:2505.21072*, 2025.

[37] A. Foong, W. Bruinsma, J. Gordon, Y. Dubois, J. Requeima, and R. Turner. Meta-learning stationary stochastic process prediction with convolutional neural processes. *Advances in Neural Information Processing Systems*, 33:8284–8295, 2020.

[38] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

[39] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[40] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022.

[41] Z. Gao, X. Jiang, X. Xu, F. Shen, Y. Li, and H. T. Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26876–26885, 2024.

[42] C. Garcin, M. Servajean, A. Joly, and J. Salmon. A two-head loss function for deep average-k classification. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7358–7367. IEEE, 2025.

[43] M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. Rezende, and S. A. Eslami. Conditional neural processes. In *International conference on machine learning*, pages 1704–1713. PMLR, 2018.

[44] M. Garnelo, J. Schwarz, D. Rosenbaum, F. Viola, D. J. Rezende, S. Eslami, and Y. W. Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.

[45] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, pages 1–77, 2023.

[46] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023.

[47] Y. Gong, Y. Chung, and J. R. Glass. AST: audio spectrogram transformer. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 571–575. ISCA, 2021.

[48] C. Gruber, P. O. Schenk, M. Schierholz, F. Kreuter, and G. Kauermann. Sources of uncertainty in machine learning–a statisticians' view. *arXiv preprint arXiv:2305.16703*, 2023.

[49] V. Guarrasi, F. Aksu, C. M. Caruso, F. Di Feola, A. Rofena, F. Ruffini, and P. Soda. A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. *Image and Vision Computing*, page 105509, 2025.

[50] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[51] H. Guo, W. Shi, and Y. Deng. Evaluating sensor reliability in classification problems based on evidence theory. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(5):970–981, 2006.

[52] Z. Han, C. Zhang, H. Fu, and J. T. Zhou. Trusted multi-view classification. *arXiv preprint arXiv:2102.02051*, 2021.

[53] Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[54] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[55] W. He, Z. Jiang, T. Xiao, Z. Xu, and Y. Li. A survey on uncertainty quantification methods for deep learning. *arXiv preprint arXiv:2302.13425*, 2023.

[56] A. Hoarau, B. Quost, S. Destercke, and W. Waegeman. Reducing aleatoric and epistemic uncertainty through multi-modal data acquisition. *arXiv preprint arXiv:2501.18268*, 2025.

[57] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[58] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

[59] L. Huang, S. Ruan, P. Decazes, and T. Denœux. Deep evidential fusion with uncertainty quantification and reliability learning for multimodal medical image segmentation. *Information Fusion*, 113:102648, 2025.

[60] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[61] R. Huang, A. Geng, and Y. Li. On the importance of gradients for detecting distributional shifts in the wild. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 677–689, 2021.

[62] Y. S. Huang, K. Liu, and C. Y. Suen. The combination of multiple classifiers by a neural network approach. *International Journal of Pattern Recognition and Artificial Intelligence*, 9(03):579–597, 1995.

[63] Y.-H. Hubert Tsai, L.-K. Huang, and R. Salakhutdinov. Learning robust visual-semantic embeddings. In *Proceedings of the IEEE International conference on Computer Vision*, pages 3571–3580, 2017.

[64] E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

[65] V. Jayachitra, S. Nivetha, R. Nivetha, and R. Harini. A cognitive iot-based framework for effective diagnosis of covid-19 using multimodal data. *Biomedical Signal Processing and Control*, 70:102960, 2021.

[66] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[67] J. Jiao, H. Sun, Y. Huang, M. Xia, M. Qiao, Y. Ren, Y. Wang, and Y. Guo. Gmrlnet: A graph-based manifold regularization learning framework for placental insufficiency diagnosis on incomplete multimodal ultrasound data. *IEEE Transactions on Medical Imaging*, 42(11):3205–3218, 2023.

[68] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

[69] A. Jøsang. *Subjective logic*, volume 3. Springer, 2016.

[70] A. Jøsang, D. Wang, and J. Zhang. Multi-source fusion in subjective logic. In *2017 20th International Conference on Information Fusion (Fusion)*, pages 1–8, 2017.

[71] M. C. Jung, H. Zhao, J. Dipnall, and L. Du. Beyond unimodal: Generalising neural processes for multimodal uncertainty estimation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 42191–42216. Curran Associates, Inc., 2023.

[72] M. C. Jung, H. Zhao, J. Dipnall, and L. Du. Beyond unimodal: Generalising neural processes for multimodal uncertainty estimation. *Advances in Neural Information Processing Systems*, 36, 2024.

[73] M. C. Jung, H. Zhao, J. Dipnall, B. Gabbe, and L. Du. Uncertainty estimation for multi-view data: The power of seeing the whole picture. *Advances in Neural Information Processing Systems*, 35:6517–6530, 2022.

[74] A. I. Károly, P. Galambos, J. Kuti, and I. J. Rudas. Deep learning in robotics: Survey on model structures and training strategies. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(1):266–279, 2020.

[75] H. Kim, A. Mnih, J. Schwarz, M. Garnelo, S. M. A. Eslami, D. Rosenbaum, O. Vinyals, and Y. W. Teh. Attentive neural processes. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[76] T.-K. Kim, J. Kittler, and R. Cipolla. Learning discriminative canonical correlations for object recognition with image sets. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part III 9*, pages 251–262. Springer, 2006.

[77] M. Kirchhof, G. Kasneci, and E. Kasneci. Position: Uncertainty quantification needs reassessment for large language model agents. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

[78] J. Kittler. Combining classifiers: A theoretical framework. *Pattern analysis and Applications*, 1:18–27, 1998.

[79] A. D. Kiureghian and O. Ditlevsen. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112, March 2009.

[80] N. Kotelevskii, M. Guizani, É. Moulines, and M. Panov. Adaptive temperature scaling with conformal prediction. 2025.

[81] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[82] A. Kumar and B. Raj. Unsupervised fusion weight learning in multiple classifier systems. *arXiv preprint arXiv:1502.01823*, 2015.

[83] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[84] J. Lampinen and A. Vehtari. Bayesian approach for neural networks—review and case studies. *Neural networks*, 14(3):257–274, 2001.

[85] J. Lee and G. AlRegib. Gradients as a measure of uncertainty in neural networks. In *IEEE International Conference on Image Processing, ICIP 2020, Abu Dhabi, United Arab Emirates, October 25-28, 2020*, pages 2416–2420. IEEE, 2020.

[86] M. Lee and V. Pavlovic. Private-shared disentangled multimodal vae for learning of hybrid latent representations. *arXiv preprint arXiv:2012.13024*, 2020.

[87] S. Lee, S. Purushwalkam, M. Cogswell, D. Crandall, and D. Batra. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*, 2015.

[88] C. Li, K. Li, Y. Ou, L. M. Kaplan, A. Jøsang, J.-H. Cho, D. H. JEONG, and F. Chen. Hyper evidential deep learning to quantify composite classification uncertainty. In *The Twelfth International Conference on Learning Representations*, 2024.

[89] J. Li, S. Luo, and J. S. Jin. Sensor data fusion for accurate cloud presence prediction using dempster-shafer evidence theory. *Sensors*, 10(10):9384–9396, 2010.

[90] L. Li, C. Li, X. Lu, H. Wang, and D. Zhou. Multi-focus image fusion with convolutional neural network based on dempster-shafer theory. *Optik*, 272:170223, 2023.

[91] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[92] S. Li and H. Tang. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*, 2024.

[93] Y. Li, M. Yang, and Z. Zhang. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863–1883, 2018.

[94] T. Liu, J. Huang, T. Liao, R. Pu, S. Liu, and Y. Peng. A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multimodal data. *Irbm*, 43(1):62–74, 2022.

[95] W. Liu, Y. Chen, X. Yue, C. Zhang, and S. Xie. Safe multi-view deep classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8870–8878, 2023.

[96] W. Liu, X. Yue, Y. Chen, and T. Denoeux. Trusted Multi-View Deep Learning with Opinion Aggregation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7585–7593, June 2022.

[97] Y. Liu, L. Liu, C. Xu, X. Song, Z. Guan, and W. Zhao. Dynamic evidence decoupling for trusted multi-view learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7269–7277, 2024.

[98] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[99] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[100] J. Lu, W. Buntine, Y. Qi, J. Dipnall, B. Gabbe, and L. Du. Navigating conflicting views: Harnessing trust for learning. *arXiv preprint arXiv:2406.00958*, 2024.

[101] A. Lyzhov, Y. Molchanova, A. Ashukha, D. Molchanov, and D. Vetrov. Greedy policy search: A simple baseline for learnable test-time augmentation. In *Conference on uncertainty in artificial intelligence*, pages 1308–1317. PMLR, 2020.

[102] L. Ma and T. Denoeux. Partial classification in the belief function framework. *Knowledge-Based Systems*, 214:106742, 2021.

[103] K. Maag and T. Riedlinger. Pixel-wise gradient uncertainty for convolutional neural networks applied to out-of-distribution segmentation. In *VISIGRAPP (2): VISAPP*, pages 112–122, 2024.

[104] A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

[105] A. Malinin and M. J. F. Gales. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. In *Neural Information Processing Systems*, 2019.

[106] A. Malinin, B. Mlodozeniec, and M. Gales. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*, 2019.

[107] A. Martin. Conflict management in information fusion with belief functions. In *Information quality in information fusion and decision making*, pages 79–97. Springer, 2019.

[108] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[109] T. Mortier, M. Wydmuch, K. Dembczyński, E. Hüllermeier, and W. Waegeman. Set-valued prediction in multi-class classification. In *31st Benelux conference on Artificial Intelligence (BNAIC 2019); 28th Belgian Dutch conference on Machine Learning (Benelearn 2019)*, volume 2491. CEUR, 2019.

[110] B. Mucsányi, M. Kirchhof, and S. J. Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *Advances in Neural Information Processing Systems*, 37:50972–51038, 2024.

[111] J. Nandy, W. Hsu, and M. Lee. Towards maximizing the representation gap between in-domain & out-of-distribution examples. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[112] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng, et al. Multimodal deep learning. In *ICML*, volume 11, pages 689–696, 2011.

[113] K. Nguyen, S. Denman, S. Sridharan, and C. Fookes. Score-level multibiometric fusion based on dempster–shafer theory incorporating uncertainty factors. *IEEE Transactions on Human-Machine Systems*, 45(1):132–140, 2014.

[114] V.-L. Nguyen, H. Zhang, and S. Destercke. Learning sets of probabilities through ensemble methods. In *European Conference on Symbolic and Quantitative Approaches with Uncertainty*, pages 270–283. Springer, 2023.

[115] A. M. Ozbayoglu, M. U. Gudelek, and O. B. Sezer. Deep learning for financial applications: A survey. *Applied soft computing*, 93:106384, 2020.

[116] Y. Pan, H. Lai, C. Liu, and S. Yan. A divide-and-conquer method for scalable low-rank latent matrix pursuit. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 524–531, 2013.

[117] F. Parisi, F. Strino, B. Nadler, and Y. Kluger. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 111(4):1253–1258, 2014.

[118] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[119] B. Quost, M.-H. Masson, and T. Denœux. Classifier fusion in the dempster–shafer framework using optimized t-norm based combination rules. *International Journal of Approximate Reasoning*, 52(3):353–374, 2011.

[120] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[121] D. Ramachandram and G. W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

[122] I. Reda, A. Khalil, M. Elmogy, A. Abou El-Fetouh, A. Shalaby, M. Abou El-Ghar, A. Elmaghraby, M. Ghazal, and A. El-Baz. Deep learning role in early diagnosis of prostate cancer. *Technology in cancer research & treatment*, 17:1533034618775530, 2018.

[123] Y. Sale, V. Bengs, M. Caprio, and E. Hüllermeier. Second-order uncertainty quantification: A distance-based approach. In *International Conference on Machine Learning*, pages 43060–43076. PMLR, 2024.

[124] K. Schweighofer, L. Aichberger, M. Ielanskyi, and S. Hochreiter. Introducing an improved information-theoretic measure of predictive uncertainty. In *NeurIPS 2023 workshop: Information-Theoretic Principles in Cognitive Systems*.

[125] M. Seeger. Gaussian processes for machine learning. *International journal of neural systems*, 14(02):69–106, 2004.

[126] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[127] G. Shafer. *A Mathematical Theory of Evidence*. Princeton university press, 1976.

[128] G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[129] S. Shamshirband, M. Fathi, A. Dehzangi, A. T. Chronopoulos, and H. Alinejad-Rokny. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *Journal of Biomedical Informatics*, 113:103627, 2021.

[130] Z. Shao, W. Dou, and Y. Pan. Dual-level deep evidential fusion: Integrating multimodal information for enhanced reliable decision-making in deep learning. *Information Fusion*, 103:102113, 2024.

[131] L. Shi, C. Tang, H. Deng, C. Xu, L. Xing, and B. Chen. Generalized trusted multi-view classification framework with hierarchical opinion aggregation. *arXiv preprint arXiv:2411.03713*, 2024.

[132] O. Shorinwa, Z. Mei, J. Lidard, A. Z. Ren, and A. Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 2025.

[133] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[134] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[135] W. C. Sleeman IV, R. Kapoor, and P. Ghosh. Multimodal classification: Current landscape, taxonomy and future directions. *ACM Computing Surveys*, 55(7):1–31, 2022.

[136] E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18, 2005.

[137] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25, 2012.

[138] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in bioinformatics*, 23(2):bbab569, 2022.

[139] D. Sun, M. Wang, and A. Li. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(3):841–850, 2018.

[140] S. Sun. A survey of multi-view machine learning. *Neural computing and applications*, 23:2031–2038, 2013.

[141] H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019.

[142] M. K. Tellamekala, S. Amiriparian, B. W. Schuller, E. André, T. Giesbrecht, and M. Valstar. Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):805–822, 2023.

[143] J. Thomason, D. Gordon, and Y. Bisk. Shifting the baseline: Single modality performance on visual navigation & qa. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1977–1983, 2019.

[144] 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)*, pages 5716–5723. IEEE, 2020.

[145] M. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR, 2009.

[146] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

[147] Z. Tong, P. Xu, and T. Denoeux. An evidential classifier based on dempster-shafer theory and deep learning. *Neurocomputing*, 450:275–293, 2021.

[148] Z. Tong, P. Xu, and T. Denœux. Fusion of evidential cnn classifiers for image classification. In *International Conference on Belief Functions*, pages 168–176. Springer, 2021.

[149] K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan, and H. D. Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025.

[150] M. C. Troffaes. Decision making under uncertainty using imprecise probabilities. *International journal of approximate reasoning*, 45(1):17–29, 2007.

[151] T. Tsiligkaridis. Information aware max-norm dirichlet networks for predictive uncertainty estimation. *Neural Networks*, 135:105–114, 2021.

[152] J. van Hout, E. Yeh, D. C. Koelma, C. G. Snoek, C. Sun, R. Nevatia, J. Wong, and G. K. Myers. Late fusion and calibration for multimedia event detection using few examples. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4598–4602. IEEE, 2014.

[153] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[154] F. Verdoja and V. Kyrki. Notes on the behavior of mc dropout. In *ICML Workshop on Uncertainty & Robustness in Deep Learning*, 2021.

[155] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[156] P. Walley. Statistical reasoning with imprecise probabilities. 1991.

[157] C. Wang, S. Gupta, X. Zhang, S. Tonekaboni, S. Jegelka, T. S. Jaakkola, and C. Uhler. An information criterion for controlled disentanglement of multimodal data. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[158] G. Wang, W. Li, S. Ourselin, and T. Vercauteren. Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II*, volume 11384 of *Lecture Notes in Computer Science*, pages 61–72. Springer, 2018.

[159] H. Wang, A. Prasad, E. Stengel-Eskin, and M. Bansal. Retrieval-augmented generation with conflicting evidence. *arXiv preprint arXiv:2504.13079*, 2025.

[160] H. Wang, V. Subramanian, and T. Syeda-Mahmood. Modeling uncertainty in multi-modal fusion for lung cancer survival analysis. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1169–1172. IEEE, 2021.

[161] H. Wang, J. Zhang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro. Uncertainty-aware multimodal learning via cross-modal random network prediction. In *European Conference on Computer Vision*, pages 200–217. Springer, 2022.

[162] W. Wang, D. Tran, and M. Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020.

[163] Z. Wang and X. Qiao. Set-valued classification with out-of-distribution detection for many classes. *Journal of Machine Learning Research*, 24(375):1–39, 2023.

[164] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378. PMLR, 2016.

[165] L. Wimmer, Y. Sale, P. Hofman, B. Bischl, and E. Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in artificial intelligence*, pages 2282–2292. PMLR, 2023.

[166] N. Wu, S. Jastrzebski, K. Cho, and K. J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pages 24043–24055. PMLR, 2022.

[167] J. Xie, Z. Chen, R. Zhang, X. Wan, and G. Li. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*, 2024.

[168] C. Xu, J. Si, Z. Guan, W. Zhao, Y. Wu, and X. Gao. Reliable conflictive multi-view learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16129–16137, 2024.

[169] C. Xu, Y. Zhang, Z. Guan, and W. Zhao. Trusted multi-view learning with label noise. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 5263–5271. ijcai.org, 2024.

[170] Z. Xue and R. Marculescu. Dynamic multimodal fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 2575–2584. IEEE, 2023.

[171] X. Yan, S. Hu, Y. Mao, Y. Ye, and H. Yu. Deep multi-view learning methods: A review. *Neurocomputing*, 448:106–129, 2021.

[172] G. Yang, S. Destercke, and M.-H. Masson. Cautious classification with nested dichotomies and imprecise probabilities. *Soft Computing*, 21:7447–7462, 2017.

[173] Q. Yang, Y. Zhao, and H. Cheng. Mmlf: Multi-modal multi-class late fusion for object detection with uncertainty estimation. *arXiv preprint arXiv:2410.08739*, 2024.

[174] Y. Yang, F. Wan, Q.-Y. Jiang, and Y. Xu. Facilitating multimodal classification via dynamically learning modality gap. *Advances in Neural Information Processing Systems*, 37:62108–62122, 2024.

[175] G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang. Robust late fusion with rank minimization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3021–3028. IEEE, 2012.

[176] L. Zadeh. A mathematical theory of evidence (book review). *AI magazine*, 5:81–83, 1984.

[177] M. Zaffalon. The naive credal classifier. *Journal of statistical planning and inference*, 105(1):5–21, 2002.

[178] M. Zaffalon, G. Corani, and D. Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282–1301, 2012. Imprecise Probability: Theories and Applications (ISIPTA'11).

[179] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[180] J. Zhao, X. Xie, X. Xu, and S. Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

[181] X. Zhao, Y. Ou, L. M. Kaplan, F. Chen, and J. Cho. Quantifying classification uncertainty using regularized evidential neural networks. *CoRR*, abs/1910.06864, 2019.

[182] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020.

[183] Y. Zhou, Y. Liu, X. Li, J. Jin, H. Qian, Z. Liu, C. Li, Z. Dou, T.-Y. Ho, and P. S. Yu. Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*, 2024.