Credal Ensemble Distillation for Uncertainty Quantification

Kaizheng Wang,[1,2*] Fabio Cuzzolin,[3] David Moens,[2,4] Hans Hallez[1]

[1]Department of Computer Science, KU Leuven, Belgium
[2]Flanders Make@KU Leuven, Belgium
[3]School of Engineering, Computing and Mathematics, Oxford Brookes University, U.K.
[4]Department of Mechanical Engineering, KU Leuven, Belgium
{kaizheng.wang, david.moens, hans.hallez}@kuleuven.be fabio.cuzzolin@brookes.ac.uk

Abstract

Deep ensembles (DE) have emerged as a powerful approach for quantifying predictive uncertainty and distinguishing its aleatoric and epistemic components, thereby enhancing model robustness and reliability. However, their high computational and memory costs during inference pose significant challenges for wide practical deployment. To overcome this issue, we propose *credal ensemble distillation* (CED), a novel framework that compresses a DE into a single model, *CREDIT*, for classification tasks. Instead of a single softmax probability distribution, CREDIT predicts class-wise probability intervals that define a credal set, a convex set of probability distributions, for uncertainty quantification. Empirical results on out-of-distribution detection benchmarks demonstrate that CED achieves superior or comparable uncertainty estimation compared to several existing baselines, while substantially reducing inference overhead compared to DE.

Code — https://gitlab.kuleuven.be/m-group-campus-brugge/distrinet_public/credal-ensemble-distillation
Extended version — https://arxiv.org/abs/2511.13766

1 Introduction

Uncertainty quantification (UQ) in neural networks (NNs) has gained increasing attention, with two primary types of uncertainty distinguished: aleatoric uncertainty (AU), which stems from the inherent randomness in the data generation process and models the stochasticity in the output given an input (i.e., via a conditional distribution $p(\text{output}|\text{input})$); and epistemic uncertainty (EU), which is caused by a lack of evidence and reflects the model's imprecise knowledge of the true conditional distribution (Hüllermeier and Waegeman 2021; Hüllermeier, Destercke, and Shaker 2022; Wang et al. 2024, 2025b). The effective estimation and differentiation of AU and EU can improve a model's trustworthiness and robustness (Senge et al. 2014; Kendall and Gal 2017; Sale, Caprio, and Hüllermeier 2023; Manchingal et al. 2025). For example, proper EU estimates can help avoid misclassifying ambiguous in-distribution (ID) samples as out-of-distribution (OOD), since their ambiguity does not necessarily correspond to regions of high EU within the ID distribution (Mukhoti et al. 2023; Wang et al. 2025a).

To quantify both AU and EU, recent studies propose training NNs to predict a *second-order representation*, capable of expressing the uncertainty about a prediction's uncertainty itself (Malinin, Mlodozeniec, and Gales 2019; Hüllermeier and Waegeman 2021; Caprio et al. 2024; Wang et al. 2024). Bayesian neural networks (BNNs) (Blundell et al. 2015; Gal and Ghahramani 2016; Krueger et al. 2017; Mobiny et al. 2021), in particular, learn posterior distributions over their weights and enable predictions in the form of *second-order distributions* (Caprio et al. 2024). However, BNNs generally face significant challenges in scalability to large datasets and complex architectures due to high computational demands (Mukhoti et al. 2023). Their performance is also sensitive to the choice of prior, likelihood, and training objectives (Henning, D'Angelo, and Grewe 2021; Knoblauch, Jewson, and Damoulas 2022).

Alternative to BNNs, deep ensembles (DE), which combine multiple standard neural networks (SNNs) to predict *a finite set of distributions* (Lakshminarayanan, Pritzel, and Blundell 2017), have been treated as a strong UQ baseline (Ovadia et al. 2019; Gustafsson, Danelljan, and Schon 2020; Abe et al. 2022; Mucsányi, Kirchhof, and Oh 2024). Nevertheless, a key limitation of DEs is their substantial demand for memory and computational resources. To this end, *ensemble distillation* (ED) has become a popular way of significantly reducing inference costs (Hinton, Vinyals, and Dean 2015; Lin et al. 2020), by distilling a DE into an SNN that approximates the mean of DE's predictive distributions. However, one drawback of ED is that the distilled SNN only generates a single predictive distribution, limiting its ability to quantify AU. This is because this single distribution captures randomness in the mapping between the input and output while assuming precise knowledge of this dependency (Hüllermeier, Destercke, and Shaker 2022).

To address this problem, *ensemble distribution distillation* (EDD) (Malinin, Mlodozeniec, and Gales 2019) has been proposed to distill a DE into a single model outputting a *Dirichlet distribution* as the second-order prediction. Yet, one practical challenge in EDD and other Dirichlet-based methods (DBMs) (Malinin and Gales 2018, 2019; Charpentier, Zügner, and Günnemann 2020) is the absence of ground-truth Dirichlet labels for training. In addition, DBMs

*Corresponding author.

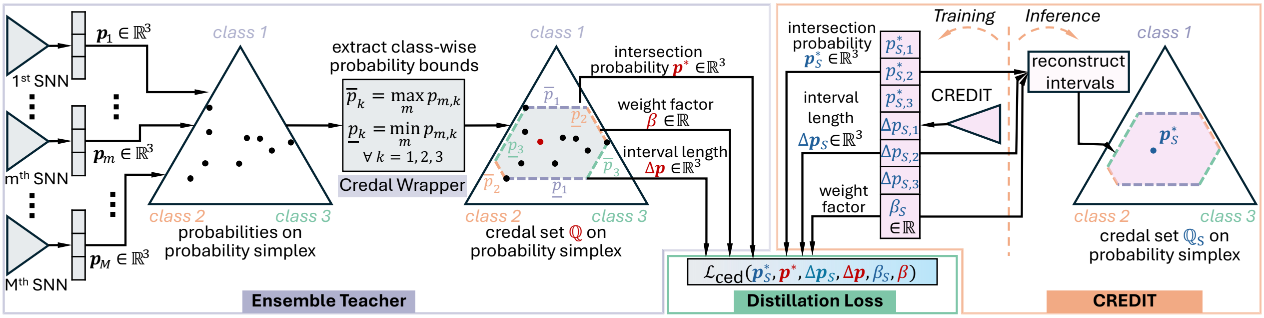

Figure 1: CED framework for three-class classification ($C=3$). Given an ensemble teacher composed of M SNNs, the predicted probabilities can generate class-wise probability bounds via a credal wrapper (see Sec. 3.1). These intervals form a credal set for UQ, from which a unique intersection probability is extracted for class prediction. As described in Sec. 3.2, the proposed credal student is designed to output a vector $\boldsymbol{v} := (\boldsymbol{p}_S^* \in \mathbb{R}^C, \Delta\boldsymbol{p}_S \in \mathbb{R}^C, \beta_S \in \mathbb{R})$, each component representing the intersection probability, the interval length vector, and the weight factor, respectively. The student is trained using a novel distillation loss introduced in Sec. 3.3. At inference time, $\boldsymbol{p}_S^*$ is employed for class prediction, while $\boldsymbol{v}$ can recover a credal set $\mathbb{Q}_S$ for UQ.

have recently faced criticism for departing from the theoretical tenets of epistemic uncertainty (Ulmer, Hardmeier, and Frellsen 2023), and failing to provide a meaningful quantitative interpretation of EU (Juergens et al. 2024).

In an alternative approach, *credal sets*, i.e., convex sets of probability distributions (Levi 1980), have been employed for UQ in a broader machine learning context (Zaffalon 2002; Corani and Zaffalon 2008; Corani, Antonucci, and Zaffalon 2012; Mauá et al. 2017). This method has recently garnered renewed attention in deep learning. Recent advancements include modeling both NN weights and outputs as credal sets (Caprio et al. 2024), deriving credal set predictions from outputted probability intervals (Wang et al. 2024, 2025c), and wrapping the predictive probabilities of BNNs and DE as a credal set (Wang et al. 2025a), to name a few. Although these credal predictors offer improved UQ compared to BNN and DE baselines, they generally demand even greater computational resources for inference.

In this context, a research question arises: *Can a single NN predicting a credal set as a second-order representation be distilled from a DE, which is capable of improving the UQ performance of existing distillation frameworks?*

Novelty and Contributions In response, this paper proposes an innovative distillation framework, termed as *credal ensemble distillation* (CED), to distill a DE teacher into a single model, called *CREDIT*. The distilled CREDIT can predict class-wise probability intervals (De Campos, Huete, and Moral 1994) that can define a credal set.

As illustrated in Figure 1, given multiple predictive softmax outputs from a DE teacher, the proposed CED framework first applies the credal wrapper method in (Wang et al. 2025a) to construct class-wise probability intervals. From these intervals, an *intersection probability* (Cuzzolin 2022) is computed—a single probability vector used for class prediction—based on the interval lengths and a scale weight factor. For a classification task involving C classes, the proposed CREDIT modifies the final classification layer of a standard neural network architecture to output a vector in $\mathbb{R}^{2C+1}$. This vector consists of the predicted intersection probability in $\mathbb{R}^C$, the corresponding probability interval lengths in $\mathbb{R}^C$, and a scalar scale weight. CREDIT is trained using label information distilled from the DE teacher via a novel distillation loss. At inference time, the intersection probability is used for class prediction, while the full output vector enables the reconstruction of a credal set for uncertainty quantification.

While our work employs a recent credal wrapper (Wang et al. 2025a) to extract credal information from ensembles, we emphasize that *credal ensemble distillation* is a *novel task* that, to our knowledge, has not been previously explored. Our objective is to design a novel and simple single model capable of learning and retaining the credal information from an ensemble, while improving the UQ performance of the ensemble and existing distillation approaches.

Empirical experiments on several OOD detection benchmarks, including different dataset pairs and various network architectures, demonstrate that CED achieves superior or comparable uncertainty estimation compared to several existing ED, EDD, and DE baselines, while substantially reducing inference overhead compared to DE.

The remainder of this paper is organized as follows. Sec. 2 extends the background discussion. Sec. 3 introduces the full details of our CED. Sec. 4 describes the experimental validations. Sec. 5 summarizes our conclusion and future work.

2 Background

This section discusses UQ in deep ensembles and different distillation frameworks.

Deep Ensembles (DE) DE generates an averaged prediction from a set of M individually-trained standard neural networks (SNNs), as follows:

$$\tilde{\boldsymbol{p}} = \tfrac{1}{M}\textstyle\sum_{m=1}^{M}\mathrm{SNN}_m(\boldsymbol{x}) = \tfrac{1}{M}\sum_{m=1}^{M}\boldsymbol{p}_m, \tag{1}$$

where $\boldsymbol{p}_m$ represents the single predictive probability vector from the m-th SNN. In this context, the total uncertainty (TU) and the aleatoric uncertainty (AU) in DE can be approximately quantified by the Shannon entropy of the averaged prediction (denoted as $H(\tilde{\boldsymbol{p}})$) and by averaging the Shannon entropy of each sampled prediction ($\tilde{H}(\boldsymbol{p})$) (Hüllermeier and Waegeman 2021), respectively, as follows:

$$\begin{aligned} H(\tilde{\boldsymbol{p}}) &= -\textstyle\sum_{k=1}^{C}\tilde{p}_k \log \tilde{p}_k \\ \tilde{H}(\boldsymbol{p}) &= -\frac{1}{M}\textstyle\sum_{m=1}^{M}\sum_{k=1}^{C} p_{m,k} \log p_{m,k} \end{aligned}. \tag{2}$$

Here, $\tilde{p}_k$ and $p_{m,k}$ are the k-th elements of the probability vectors $\tilde{\boldsymbol{p}}$ and $\boldsymbol{p}_m$, respectively. The level of epistemic uncertainty (EU) can be estimated as $H(\tilde{\boldsymbol{p}})$–$\tilde{H}(\boldsymbol{p})$ (Depeweg et al. 2018), which can be interpreted as an approximation of mutual information (MI) (Hüllermeier, Destercke, and Shaker 2022; Hüllermeier and Waegeman 2021).

Ensemble Distillation (ED) Given a DE of M trained SNNs, a single model, ED-Net, can be trained within the ED framework by minimizing the cross-entropy between its predictive categorical distribution ($\boldsymbol{p}$) and the soft label ($\tilde{\boldsymbol{p}}$) obtained by averaging the ensemble's output distributions in eq. (1), as follows (Hinton, Vinyals, and Dean 2015):

$$N^{-1}\textstyle\sum_{n=1}^{N}\big(-\sum_{k=1}^{C}\tilde{p}_k^n \log p_k^n\big). \tag{3}$$

Here, the superscript n is the index of the N training samples. ED-Net can merely measure AU by calculating the classical Shannon entropy as $H(\boldsymbol{p}) = -\sum_{k=1}^{C} p_k \log p_k$.

Ensemble Distribution Distillation (EDD) In contrast to SNNs predicting a single softmax probability, the EDD-Net model distilled within the EDD framework generates the parameter vector (denoted as $\boldsymbol{\alpha} = \mathbb{R}_+^C$) of a Dirichlet distribution $\boldsymbol{q} := \mathrm{Dir}(\boldsymbol{\alpha})$. Let $\boldsymbol{z}_{\text{edd}} \in \mathbb{R}^C$ be the output logits of EDD-Net; both $\boldsymbol{\alpha}$ and the expected categorical distribution (denoted as $\boldsymbol{\pi}$) under this Dirichlet prior can be calculated as follows (Malinin, Mlodozeniec, and Gales 2019):

$$\begin{aligned} \boldsymbol{\alpha} &= \exp(\boldsymbol{z}_{\text{edd}}) \\ \boldsymbol{\pi} &= \frac{\boldsymbol{\alpha}}{\sum_{k=1}^{C}\alpha_k} = \frac{\exp(\boldsymbol{z}_{\text{edd}})}{\sum_{k=1}^{C}\exp(z_{\text{edd}_k})} = \mathrm{softmax}(\boldsymbol{z}_{\text{edd}}) \end{aligned}. \tag{4}$$

$\boldsymbol{\pi}$ is employed for making class prediction. TU and AU can be measured by the Shannon entropy of $\boldsymbol{\pi}$ and the expected entropy of a categorical distribution sampled from the $\mathrm{Dir}(\boldsymbol{\alpha})$, respectively, as follows:

$$\begin{aligned} H(\boldsymbol{\pi}) &= -\textstyle\sum_{k=1}^{C}\pi_k \log \pi_k \\ \mathbb{E}_{\boldsymbol{p}\sim\mathrm{Dir}(\boldsymbol{\alpha})}(H(\boldsymbol{p})) &= -\textstyle\sum_{k=1}^{C}\frac{\alpha_k}{\alpha_0}\big(\psi(\alpha_k+1)-\psi(\alpha_0+1)\big) \end{aligned}. \tag{5}$$

Here $\alpha_0 = \sum_{k=1}^{C}\alpha_k$, while ψ denotes the digamma function. EU is calculated as $H(\boldsymbol{\pi})$–$\mathbb{E}_{\boldsymbol{p}\sim\mathrm{Dir}(\boldsymbol{\alpha})}(H(\boldsymbol{p}))$ (Malinin, Mlodozeniec, and Gales 2019).

3 Methodology

This section details our *credal ensemble distillation* (CED) framework. As shown in Figure 1, a DE teacher comprising M SNNs generates class-wise probability intervals using a credal wrapper. These intervals define a credal set for UQ, from which a unique intersection probability is derived for class prediction (see Sec. 3.1). To distill a credal predictor from the DE teacher, we introduce an innovative credal student, termed as CREDIT (Sec. 3.2), capable of predicting the intersection probability, the interval length vector, and a weight factor for collectively reconstructing the probability interval systems. Sec. 3.3 proposes our distillation strategy for effectively training the credal student.

3.1 Credal Wrapper for Ensemble Teacher

A credal wrapper method has been recently proposed to enhance the UQ capabilities of DE by generating a credal set from DE's M predicted probabilities (Wang et al. 2025a). Specifically, a set of probability intervals over C classes, denoted as $[\underline{\boldsymbol{p}}, \overline{\boldsymbol{p}}] := \{[\underline{p}_k, \overline{p}_k]\}_{k=1}^{C}$, can be computed from

$$\overline{p}_k = \max_{m=1,..,M} p_{m,k}, \quad \underline{p}_k = \min_{m=1,..,M} p_{m,k}, \tag{6}$$

where $p_{m,k}$ is the k-th class value of the m-th predicted probability $\boldsymbol{p}_m$ from DE. These intervals can determine a valid credal set $\mathbb{Q}$ (De Campos, Huete, and Moral 1994) as

$$\mathbb{Q} = \{\boldsymbol{p} \,|\, p_k \in [\underline{p}_k, \overline{p}_k] \ \forall k\}, \tag{7}$$

while satisfying

$$\textstyle\sum_{k=1}^{C}\underline{p}_k \le 1 \le \sum_{k=1}^{C}\overline{p}_k. \tag{8}$$

As a result, $\mathbb{Q}$ consists of a convex set of valid (normalized) probability vectors $\boldsymbol{p}$, whose any k-th probability value p_k is constrained by the probability interval $[\underline{p}_k, \overline{p}_k]$.

To derive a unique class prediction from the credal set defined in eq. (7), the credal wrapper (Wang et al. 2025a) computes a normalized *intersection probability*, $\boldsymbol{p}^*$, under the assumption of equal trust in the probability intervals across all classes (Cuzzolin 2022). Mathematically, the k-th element of $\boldsymbol{p}^*$ is obtained from

$$p_k^* = \underline{p}_k + \beta(\overline{p}_k - \underline{p}_k), \tag{9}$$

where the weight factor $\beta \in [0, 1]$ can be computed as

$$\beta = \big(1 - \textstyle\sum_{k=1}^{C}\underline{p}_k\big)/\big(\sum_{k=1}^{C}\Delta p_k\big). \tag{10}$$

Here, Δp_k denotes the k-th element of the interval length vector $\Delta\boldsymbol{p} = \overline{\boldsymbol{p}} - \underline{\boldsymbol{p}}$.

3.2 Credal Student Design

Architecture Our proposed CREDIT merely modifies the final classification layer and is compatible with any NN backbone. Specifically, it replaces standard C output neurons with $2C+1$ nodes to predict a vector $\boldsymbol{v} := (\boldsymbol{p}_S^* \in \mathbb{R}^C, \Delta\boldsymbol{p}_S \in \mathbb{R}^C, \beta_S \in \mathbb{R})$, each component corresponding to the intersection probability, the interval length vector, and the weight factor, respectively. Let $\boldsymbol{z}_S \in \mathbb{R}^{2C+1}$ be the final layer logits outputs; then, $\boldsymbol{v}$ is computed as follows:

$$\begin{aligned} &\boldsymbol{p}_S^* = \mathrm{softmax}(\boldsymbol{z}_{S1:C});\ \Delta\boldsymbol{p}_S = \mathrm{sigmoid}(\boldsymbol{z}_{SC+1:2C}); \\ &\beta_S = \mathrm{sigmoid}(z_{S2C+1}) \end{aligned}. \tag{11}$$

This ensures that $\boldsymbol{p}_S^*$ is normalized, that each k-th interval length $\Delta p_{S,k} := \overline{p}_{S,k} - \underline{p}_{S,k} \in [0, 1]$, and that $\beta_S \in [0, 1]$. From eq. (9), it is evident that the probability interval vector $[\underline{\boldsymbol{p}}_S, \overline{\boldsymbol{p}}_S]$ can be reconstructed from the elements of $\boldsymbol{v}$. Namely, its k-th element, $[\underline{p}_{S,k}, \overline{p}_{S,k}]$, can be computed as

$$\underline{p}_{S,k} = p_{S,k}^* - \beta_S\Delta p_{S,k},\ \overline{p}_{S,k} = p_{S,k}^* + (1-\beta_S)\Delta p_{S,k}. \tag{12}$$

To guarantee that any $\underline{p}_{S,k}$ and $\overline{p}_{S,k}$ fall in $[0, 1]$, our credal ensemble distillation enforces:

$$\underline{p}_{S,k} \leftarrow \max\{\underline{p}_{S,k}, 0\}, \overline{p}_{S,k} \leftarrow \min\{\overline{p}_{S,k}, 1\}. \tag{13}$$

As a result, it can be proved that $\overline{p}_{S,k} - \underline{p}_{S,k} = \Delta p_{S,k} \in [0,1]$ guarantees a valid probability interval for each class and that

$$\begin{aligned}\textstyle\sum_{k=1}^{C} \underline{p}_{S,k} &= \textstyle\sum_{k=1}^{C} p^*_{S,k} - \beta_S \Delta p_{S,k} \leq \sum_{k=1}^{C} p^*_{S,k} = 1 \\ &\leq \textstyle\sum_{k=1}^{C} p^*_{S,k} + (1-\beta_S)\Delta p_{S,k} = \sum_{k=1}^{C} \overline{p}_{S,k}\end{aligned} \tag{14}$$

satisfy the condition in eq. (7) for defining a valid credal set for UQ. $\boldsymbol{p}^*_S$ is then employed to predict a unique class.

Uncertainty Quantification To estimate both AU and EU from a credal set, one can use several measures, including the generalized entropy (Abellán, Klir, and Moral 2006) and the generalized Hartley measure (Abellán and Moral 2000). Due to its broad applicability and computational efficiency (Wang et al. 2024, 2025a), this study adopts the generalized entropy measure. In this framework, an upper and a lower Shannon entropy, $\overline{H}(\mathbb{Q}_S)$ and $\underline{H}(\mathbb{Q}_S)$, are used to quantify total uncertainty (TU) and AU, respectively (Hüllermeier and Waegeman 2021). EU can then be estimated via $\overline{H}(\mathbb{Q}_S) - \underline{H}(\mathbb{Q}_S)$. Calculating $\overline{H}(\mathbb{Q}_S)$ here requires solving the following optimization problem:

$$\begin{aligned}&\overline{H}(\mathbb{Q}_S) = \text{maximize} \textstyle\sum_{k=1}^{C} -p_k \cdot \log p_k \\ &\text{s.t.} \textstyle\sum_{k=1}^{C} p_k = 1 \text{ and } p_k \in [\underline{p}_{S,k}, \overline{p}_{S,k}]\ \forall k = 1, ..., C\end{aligned}, \tag{15}$$

which searches for the maximum entropy value of any possible probability vector $\boldsymbol{p}$ within $\mathbb{Q}_S$. The computation of $\underline{H}(\mathbb{Q}_S)$, for which the maximize is replaced by minimize, returns the minimal entropy. Standard solvers, such as the SciPy optimization package (Virtanen et al. 2020), can be used to efficiently solve these problems. Moreover, empirical evidence (Wang et al. 2024, 2025a) indicates that the computational overhead of eq. (15) is marginal, particularly when $C \leq 10$.

3.3 Distillation Strategy

Generalizing cross-entropy (CE) loss, which corresponds to the Kullback-Leibler divergence, to the task of learning a credal set (defined by lower and upper probabilities) from a credal label remains an open research problem (Soubaras 2011; Lienen, Demir, and Hullermeier 2023; Wang et al. 2024). In this context, we propose minimizing the following loss, $\mathcal{L}_{\text{ced}}$, to distill a CREDIT student from the DE teacher:

$$\begin{aligned}N^{-1}\sum_{n=1}^{N}\Big(&\textstyle\sum_{k=1}^{C} -p_k^{*n} \log p_{S,k}^{*n} + \\ &\textstyle\sum_{k=1}^{C} (\Delta p_k^n - \Delta p_{S,k}^n)^2 + (\beta^n - \beta_S^n)^2\Big)\end{aligned}. \tag{16}$$

Here, the superscript n indicates the index of the N training samples, and $\underline{p}^n_{S,k}$. The first item of $\mathcal{L}_{\text{ced}}$ is the CE loss between the intersection probabilities of the DE ($\boldsymbol{p}^*$) and the CREDIT ($\boldsymbol{p}^*_S$). This enables CREDIT to retain the ensemble's predictive performance—i.e., making a unique class prediction—since the intersection probability serves as the most representative single estimate for approximating probabilistic interval systems (Cuzzolin 2022). The second and third items correspond to classical mean-squared error losses that guide the student in learning the interval length $\Delta \boldsymbol{p}_S$ and the weight factor β_S from the DE teacher, i.e., capturing the imprecision of the probability interval system (credal set) implied by the DE teacher's label. Following training, the probability interval systems in CREDIT are reconstructed via eq. (12). Semantically, these last two items are equivalent to

$$\textstyle\sum_{k=1}^{C} (\overline{p}_k^n - \overline{p}_{S,k}^n)^2 + \sum_{k=1}^{C} (\underline{p}_k^n - \underline{p}_{S,k}^n)^2,$$

where $\overline{p}^n_{S,k}$ and $\underline{p}^n_k$ can be calculated from ${p^*_{S,k}}^n$, $\Delta p^n_{S,k}$, and β^n_S using eq. (12).

Temperature scaling applied to the CE loss has been shown to enhance distillation performance (Hinton, Vinyals, and Dean 2015). This technique is also compatible with our CED framework, and the complete training procedure incorporating temperature scaling is outlined in Algorithm 1.

Algorithm 1: CED with Temperature Scaling

Input: trained DE teacher ($\{\text{SNN}_m\}_{m=1}^{M}$), temperature T
Output: credal student (CREDIT)

while training **do**
  compute scaled logits of each SNN_m given input $\boldsymbol{x}$: $\boldsymbol{z}_m = \text{SNN}_m(\boldsymbol{x})/T$
  calculate the predicted probability of each SNN_m: $\boldsymbol{p}_m = \text{softmax}(\boldsymbol{z}_m)$
  extract $\boldsymbol{p}^*$, $\Delta\boldsymbol{p}$, and β using eqs. (6), (9) and (10)
  compute scaled logits of the student given input $\boldsymbol{x}$: $\boldsymbol{z}_S = (\boldsymbol{z}_{1:C}/T, \boldsymbol{z}_{C+1:2C+1})$ where $\boldsymbol{z} = \text{CREDIT}(\boldsymbol{x})$
  calculate $\boldsymbol{p}^*_S$, $\Delta\boldsymbol{p}_S$, and β_S using eq. (11)
  minimize $T^2 \cdot \mathcal{L}_{\text{ced}}$ in eq. (16)
end while

---

4 Experimental Validation

To assess the UQ performance of our CED framework, empirical validations are conducted on standard OOD detection benchmarks (Mukhoti et al. 2023; Mucsányi, Kirchhof, and Oh 2024). OOD detection is formulated as a binary classification task. ID and OOD samples are labeled as 0s and 1s, and the model's uncertainty estimate for each sample serves as its prediction. In this context, models are expected to assign higher EU/TU scores to OOD samples than to ID inputs. Thus, improved OOD detection performance serves as an indicator of better UQ. To quantify OOD detection performance, we report AUROC (Area Under the Receiver Operating Characteristic curve) and AUPRC (Area Under the Precision-Recall Curve) scores, where higher values reflect enhanced UQ performance.

Setups The empirical evaluation employs various dataset pairs (ID vs. OOD data), including CIFAR10 (Krizhevsky, Nair, and Hinton 2009) vs. SVHN (Hendrycks et al. 2021) and CIFAR10 vs. CIFAR10-C (Hendrycks and Dietterich 2019). The CIFAR10-C dataset consists of instances that apply 15 types of corruptions to the CIFAR10 test sets, respectively, with five severity levels per corruption type.

The experiment begins by training 15 SNNs with different random initializations using the ID training sets. Choosing 15 runs aims to improve the statistical significance of the results. DEs are constructed by five SNN models ($M = 5$).

		ID Performance		OOD Detection (CIFAR10 vs. SVHN)				OOD Detection on (CIFAR10 vs. CIFAR10-C)			
		Test Accuracy	ECE	AUROC		AUPRC		AUROC		AUPRC	
				Using EU	Using TU	Using EU	Using TU	Using EU	Using TU	Using EU	Using TU
VGG16	DE	**93.52±0.07**	**1.46±0.13**	89.99±0.79	91.53±0.72	93.78±0.67	95.09±0.49	93.18±1.99	**96.51±1.70**	89.41±4.07	**95.42±2.07**
	SNN	91.79±0.11	6.39±0.15	/	89.44±1.78	/	93.71±1.24	/	93.90±2.41	/	91.68±3.48
	CED	92.23±0.17	6.71±0.18	**93.56±2.17**	**92.51±1.96**	**96.09±1.72**	**95.21±1.52**	**96.51±1.81**	95.56±1.75	**95.09±2.36**	93.58±2.44
	ED	92.18±0.16	6.85±0.16	/	91.07±1.27	/	94.51±0.89	/	94.71±2.20	/	92.72±2.94
	EDD*	91.13±0.18	3.84±0.25	90.94±2.41	90.96±2.66	93.66±1.72	93.78±2.11	93.83±1.88	95.45±2.10	87.91±4.32	92.11±3.65
	MCDO	91.95±0.13	6.15±0.12	51.42±0.46	89.12±1.63	74.72±0.42	93.64±1.17	51.32±0.50	94.74±2.40	56.58 ± 1.92	93.12±3.10
ResNet50	DE	**93.40±0.12**	**1.32±0.16**	89.50±1.05	**94.89±0.50**	92.22±1.19	**97.32±0.33**	87.78±2.28	94.08±3.48	78.92±3.67	93.08±3.92
	SNN	91.60±0.38	5.81±0.27	/	93.55±0.96	/	96.54±0.65	/	93.24±3.46	/	91.64±4.27
	CED	91.77±0.74	6.34±0.59	**96.69±1.14**	94.80±1.07	**98.44±0.64**	97.12±0.67	**96.80±2.81**	**95.23±2.74**	**96.09±4.14**	**93.78±3.72**
	ED	92.02±0.22	6.64±0.24	/	94.02±0.55	/	96.50±0.43	/	94.09±2.81	/	92.22±3.69
	EDD*	80.38±6.10	10.79±8.12	86.80±8.08	88.75±5.93	93.27±4.74	93.61±3.94	89.48±9.52	91.04±6.75	86.30±12.72	86.76±9.45

Table 1: ID performance (test accuracy and ECE) and OOD detection performance comparison across methods on benchmarks (CIFAR10 vs. SVHN/CIFAR10-C) involving VGG16 and ResNet50 (in a pre-trained model setting) as backbones. Results (all in %) are averaged from 15 runs. In terms of OOD detection, ranking legend for consistent use of uncertainty estimates (either EU or TU): **Best**, Second best. The highest scores across both EU and TU are indicated with ☐.

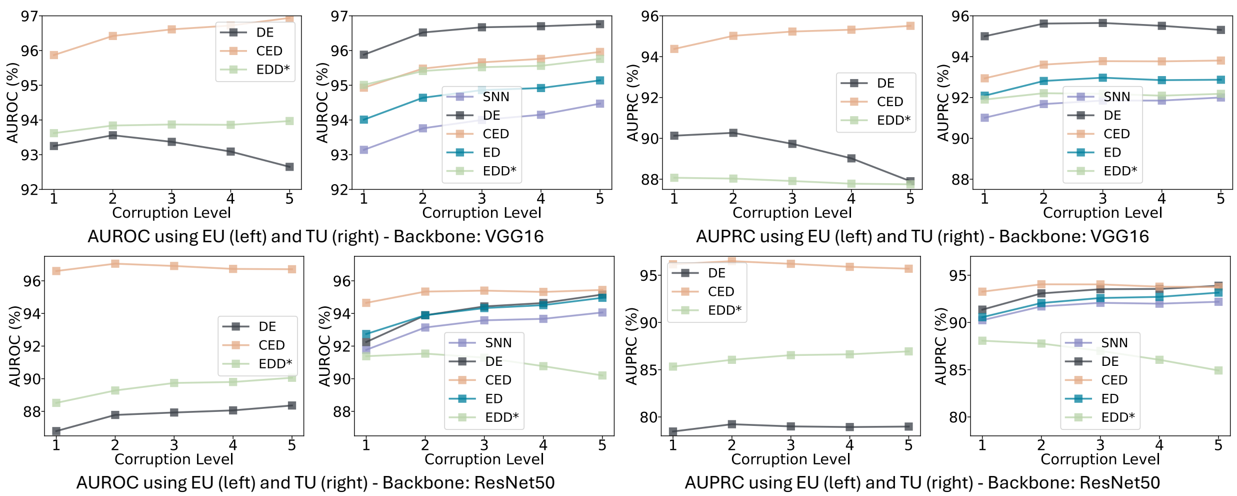

Figure 2: OOD detection (CIFAR10 vs. CIFAR10-C) comparison over increased corruption levels on various backbones.

To prevent distilled models from relying on one particular DE teacher, we construct 15 DEs by randomly selecting 15 distinct subsets of size 5 from the full SNN index set $\{1, \ldots, 15\}$, ensuring that no ensemble shares an identical composition. Subsequently, three types of student models—ED, EDD, and our CED—are distilled from these DEs, yielding 15 student models per category. All training configurations, including batch size, number of epochs, optimizer, learning rate scheduler, and temperature scaling ($T = 2.5$) as recommended by (Hinton, Vinyals, and Dean 2015), are kept consistent within each student class to ensure fair comparison. Additionally, we include a variant of EDD, denoted as EDD*, which adopts the cyclic learning rate policy, temperature scaling ($T = 10$), and temperature annealing, following the recipe in the original study (Malinin, Mlodozeniec, and Gales 2019). In addition, we include the well-known Monte Carlo Dropout (MCDO) (Gal and Ghahramani 2016) with 10 forward passes during inference as an additional baseline.

As the main experiment, all models are implemented on the established VGG16 architecture (Simonyan and Zisserman 2015), and trained from scratch using the CIFAR10 dataset. As an ablation study, we also evaluate the UQ performance of our CED method in a pre-trained model setting. All models are trained on the CIFAR10 dataset using pre-trained ResNet-50 backbones (He et al. 2016), following the same training pipeline as in the main experiment. To accommodate the pre-trained models, all input images are resized to $(224, 224, 3)$. Because no pretrained models are available, we do not report ResNet50-based MCDO results.

Results Table 1 presents UQ evaluations on EU and TU on the OOD benchmarks (CIFAR10 vs. SVHN/CIFAR10-C). For CIFAR10-C as OOD data, the reported scores are averaged over 15 corruption types and 5 severity levels. The EU and TU estimates of DE are computed from eq. (2) in a standard manner. Figure 2 shows the OOD detection comparison on CIFAR10 vs. CIFAR10-C against increased corruption intensity. Evaluation results for EDD, trained with the same configurations as ED and CED for a fair comparison, are excluded due to its substantially lower prediction accuracy, as shown in Table 2.

	ACC [VGG16]	ECE [VGG16]	ACC [ResNet50]	ECE [ResNet50]
EDD	74.56±2.02	5.51±0.57	61.15±8.47	8.40±2.06

Table 2: Test ACC and ECE of EDD on CIFAR10 test data across distinct architectures. All results are in %.

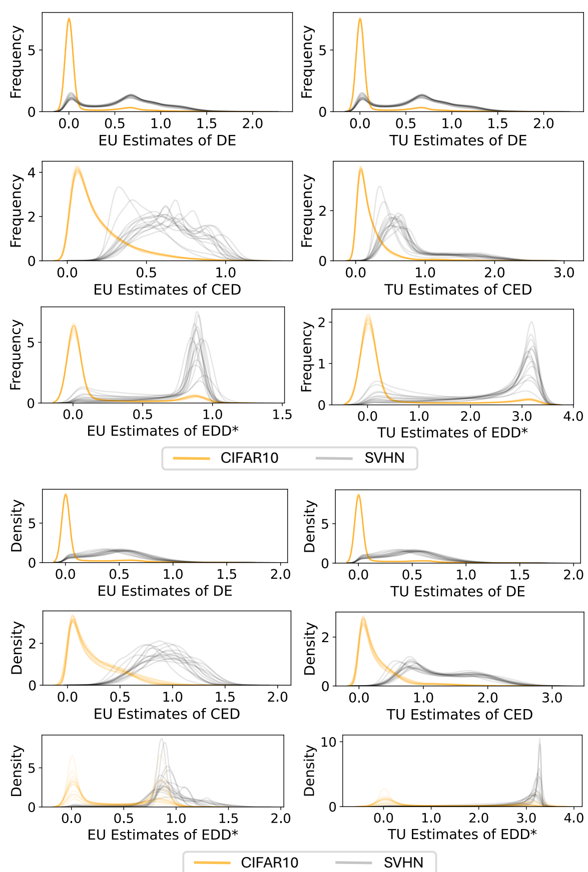

Figure 3: Distributions of EU and TU estimates across models on the VGG16 (top) and ResNet50 (bottom). 15 runs.

These OOD detection results in Table 1 show that our CED framework *significantly* and *consistently* improves EU estimation over baseline methods, as reflected by enhanced OOD detection performance across diverse dataset pairs and backbone architectures. For TU estimation, CED achieves superior or comparable performance over baselines, ranking among the top two in most cases. Furthermore, the comparison of OOD detection scores using both EU and TU in Table 1 and Figure 2 shows that CED's EU estimates *yield the best performance in most cases*. This highlights the importance of improving EU quantification for reliable OOD detection. As a summary, these empirical results establish CED as a principled alternative distillation framework for UQ. In this setting, EU estimation with MCDO becomes unreliable, likely due to limited model diversity.

In addition, Table 1 also presents test accuracy and expected calibration error (ECE) (Guo et al. 2017) on the CIFAR10 test set for various models. A lower ECE value signifies a closer alignment between the model's confidence scores and the true probabilities of the events (Guo et al. 2017; Nixon et al. 2019). The results indicate that distillation enhances the predictive accuracy of individual SNNs, and our CED approach achieves performance comparable to baseline distillation methods. Note that the ECE metric used here is designed for single-probability predictions, and a principled extension of ECE to credal-set predictions is needed for a fair comparison (Wang et al. 2024).

Qualitative Evaluation Figure 3 displays kernel density plots of EU and TU estimates on CIFAR10 (ID) and SVHN (OOD) samples for DE, CED, and EDD*, across different backbone architectures. While the EU and TU estimates are not directly comparable across methods due to differing uncertainty representations, CED consistently exhibits substantially higher EU and TU values for OOD samples relative to ID instances, as qualitatively observed.

It is worth noting that in the VGG16 case in Figure 3, although EDD* yields more distinct density peaks in EU and TU estimates for OOD samples, the uncertainty distributions for ID samples also exhibit considerable density near the OOD peaks. This overlap aligns with the lower OOD detection performance of EDD* compared to DE and CED, as reported in Table 1. The peaks for EDD* on ResNet50 stem from its low test accuracy (Table 1), which results in more misclassified ID samples with high uncertainty.

Ablation Study on Teacher Ensemble Size This experiment assesses the UQ performance of our CED across varying ensemble sizes of the teacher model. Using the same training procedure as in the main experiment, we consider DE teachers with $M \in \{5, 15, 25, 30\}$ for distillation.

Figure 4 presents the OOD detection performance involving the VGG16 backbone for CIFAR10 vs. SVHN and CIFAR10 vs. CIFAR10-C, respectively. The results reveal the following observations: Unlike the DE teacher, where increasing the ensemble size consistently improves UQ performance, no such clear trend is observed for either CED or EDD*. The consistently strong OOD detection scores with EU, along with comparable results using TU, highlight the high potential of our CED approach for UQ, particularly given the significantly lower inference complexity compared to DE with larger ensemble sizes.

Ablation Study on Effect of Temperature Scaling This experiment empirically investigates the effect of various temperature scaling on the UQ performance of our CED. Using the same training procedure as in the main experi-

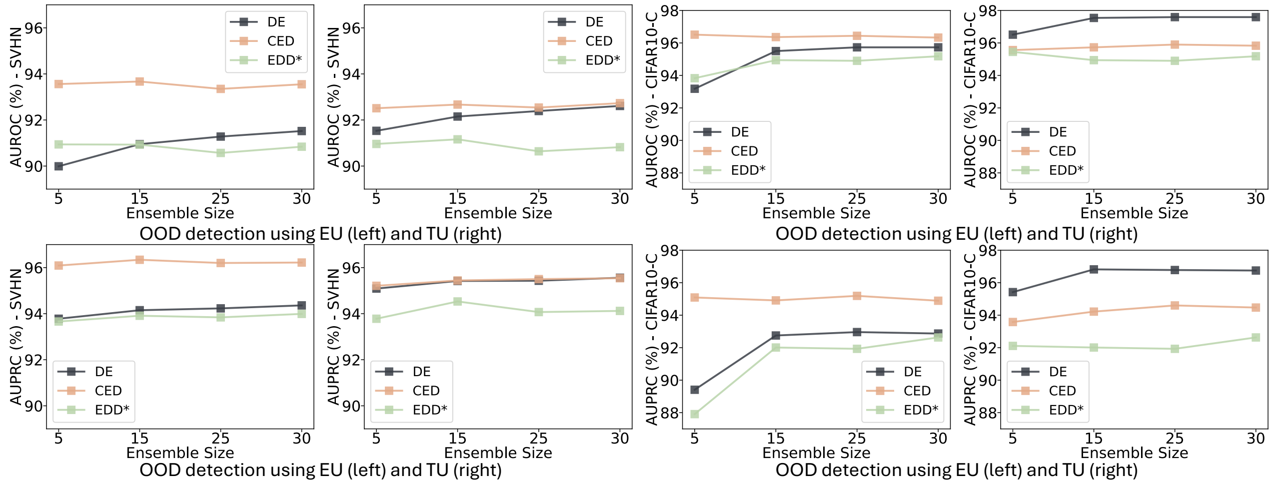

Figure 4: OOD detection performance with increasing ensemble sizes of the DE teacher. Left: CIFAR10 vs. SVHN. Right: CIFAR10 vs. CIFAR10-C. Backbone: VGG16.

ment, we consider $T \in \{1, 2.5, 5, 10\}$ (recommended by (Hinton, Vinyals, and Dean 2015)) for distillation. Figure 5 presents the OOD detection scores on CIFAR10 vs. SVHN/CIFAR10-C using the VGG16 backbone across various temperature scaling values. The results indicate that temperature scaling improves the UQ performance of our CED method, although excessively high values (i.e., $T = 10.0$) degrade performance. Among the tested values, $T = 2.5$ consistently yields the best results, aligning with the findings of Hinton, Vinyals, and Dean (2015).

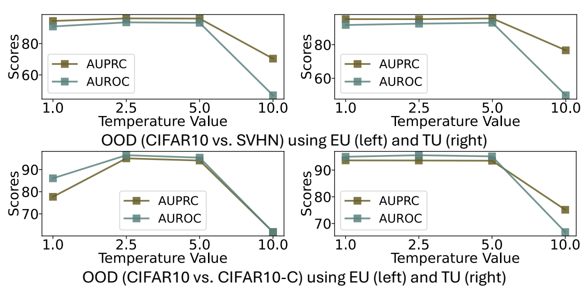

Figure 5: OOD detection performance over increased temperature T values. Backbone: VGG16.

Complexity Table 3 presents the inference times in seconds on the CIFAR10 test set using a single P100 GPU. While CED introduces a slight increase in inference complexity compared to other distillation methods—due to additional output layer nodes—it remains significantly more efficient than DE. In addition, the training time per epoch in seconds using a single P100 GPU in Table 3 shows that training CED is simpler than training EDD*, as CED does not require a sophisticated learning rate scheduler or temperature annealing.

	DE	CED	EDD*
Inference time	5×(2.22±0.20)	2.26±0.23	2.22±0.20
Training time	5×(130.07±0.24)	659.52±11.82	684.54±5.05

Table 3: Inference time (s) comparison on CIFAR10 test set.

Additional experimental details, ablation studies on the ResNet18 backbone, and a case study on real-world medical image classification are provided in the extended version.

5 Conclusion and Future Work

This work presented *credal ensemble distillation* (CED), a novel framework that compresses a deep ensemble (DE) teacher into a single model, *CREDIT*, for classification tasks. Instead of a single softmax probability distribution, CREDIT is capable of predicting class-wise probability intervals that define a credal set for uncertainty quantification. Empirical results on out-of-distribution detection benchmarks demonstrated that CED achieves superior or comparable uncertainty estimation compared to several existing ED, EDD and DE baselines, while substantially reducing inference overhead compared to DE. We believe that these promising results could position our credal ensemble distillation as a principled and scalable alternative for uncertainty quantification in deep neural classifiers.

One direction for future work is to enhance the scalability of CED, enabling its application to classification tasks involving a significantly larger number of classes (e.g., 100 or 1000). A key challenge in the current CED setting is that the softmax activation used in DE teacher produces extremely small probability values (near zero) for most classes, which could potentially destabilize the regression component of the distillation loss during training and undermine the robustness of CED's uncertainty quantification. Another future goal is to integrate calibration considerations into the design of the distillation strategy, with the aim of achieving comparable or better calibration performance than the DE teacher.

Acknowledgments

This work has received funding from the European Horizon 2020 research and innovation program under the FET Open grant agreement No. 964505 (E-pi).

References

Abe, T.; Buchanan, E. K.; Pleiss, G.; Zemel, R.; and Cunningham, J. P. 2022. Deep ensembles work, but are they necessary? *Advances in Neural Information Processing Systems*, 35: 33646–33660.

Abellán, J.; Klir, G. J.; and Moral, S. 2006. Disaggregated total uncertainty measure for credal sets. *International Journal of General Systems*, 35(1): 29–44.

Abellán, J.; and Moral, S. 2000. A non-specificity measure for convex sets of probability distributions. *International journal of uncertainty, fuzziness and knowledge-based systems*, 8(03): 357–367.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight uncertainty in neural network. In *Proceedings of the International Conference on Machine Learning*, 1613–1622. PMLR.

Caprio, M.; Dutta, S.; Jang, K. J.; Lin, V.; Ivanov, R.; Sokolsky, O.; and Lee, I. 2024. Credal Bayesian Deep Learning. *Transactions on Machine Learning Research*.

Charpentier, B.; Zügner, D.; and Günnemann, S. 2020. Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts. *Advances in neural information processing systems*, 33: 1356–1367.

Corani, G.; Antonucci, A.; and Zaffalon, M. 2012. Bayesian networks with imprecise probabilities: Theory and application to classification. *Data Mining: Foundations and Intelligent Paradigms: Volume 1: Clustering, Association and Classification*, 49–93.

Corani, G.; and Zaffalon, M. 2008. Learning Reliable Classifiers From Small or Incomplete Data Sets: The Naive Credal Classifier 2. *Journal of Machine Learning Research*, 9(4).

Cuzzolin, F. 2022. The intersection probability: betting with probability intervals. *arXiv preprint arXiv:2201.01729*.

De Campos, L. M.; Huete, J. F.; and Moral, S. 1994. Probability intervals: A tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 02: 167–196.

Depeweg, S.; Hernandez-Lobato, J.-M.; Doshi-Velez, F.; and Udluft, S. 2018. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, 1184–1193. PMLR.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*, 1050–1059. PMLR.

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

Gustafsson, F. K.; Danelljan, M.; and Schon, T. B. 2020. Evaluating scalable Bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 318–319.

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

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *Proceedings of the International Conference on Learning Representations*.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15262–15271.

Henning, C.; D'Angelo, F.; and Grewe, B. F. 2021. Are Bayesian neural networks intrinsically good at out-of-distribution detection? In *Workshop on Uncertainty & Robustness in Deep Learning*, 008. ICML UDL 2021.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hüllermeier, E.; Destercke, S.; and Shaker, M. H. 2022. Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In *Proceedings of the Uncertainty in Artificial Intelligence*, 548–557. PMLR.

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

Juergens, M.; Meinert, N.; Bengs, V.; Hüllermeier, E.; and Waegeman, W. 2024. Is Epistemic Uncertainty Faithfully Represented by Evidential Deep Learning Methods? In *Forty-first International Conference on Machine Learning*.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30.

Knoblauch, J.; Jewson, J.; and Damoulas, T. 2022. An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132): 1–109.

Krizhevsky, A.; Nair, V.; and Hinton, G. 2009. CIFAR-10 (Canadian Institute For Advanced Research). Technical report, CIFAR.

Krueger, D.; Huang, C.-W.; Islam, R.; Turner, R.; Lacoste, A.; and Courville, A. 2017. Bayesian hypernetworks. *ArXiv Preprint ArXiv:1710.04759*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30.

Levi, I. 1980. *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT press.

Lienen, J.; Demir, C.; and Hullermeier, E. 2023. Conformal credal self-supervised learning. In *Conformal and Probabilistic Prediction with Applications*, 214–233. PMLR.

Lin, T.; Kong, L.; Stich, S. U.; and Jaggi, M. 2020. Ensemble distillation for robust model fusion in federated learning. *Advances in neural information processing systems*, 33: 2351–2363.

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

Malinin, A.; and Gales, M. 2019. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. *Advances in Neural Information Processing Systems*, 32.

Malinin, A.; Mlodozeniec, B.; and Gales, M. 2019. Ensemble Distribution Distillation. In *International Conference on Learning Representations*.

Manchingal, S. K.; Mubashar, M.; Wang, K.; Shariatmadar, K.; and Cuzzolin, F. 2025. Random-Set Neural Networks. In *The Thirteenth International Conference on Learning Representations*.

Mauá, D. D.; Cozman, F. G.; Conaty, D.; and Campos, C. P. 2017. Credal sum-product networks. In *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*, 205–216. PMLR.

Mobiny, A.; Yuan, P.; Moulik, S. K.; Garg, N.; Wu, C. C.; and Van Nguyen, H. 2021. Dropconnect is effective in modeling uncertainty of Bayesian deep networks. *Scientific Reports*, 11(1): 1–14.

Mucsányi, B.; Kirchhof, M.; and Oh, S. J. 2024. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *arXiv preprint arXiv:2402.19460*.

Mukhoti, J.; Kirsch, A.; van Amersfoort, J.; Torr, P. H.; and Gal, Y. 2023. Deep Deterministic Uncertainty: A New Simple Baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24384–24394.

Nixon, J.; Dusenberry, M. W.; Zhang, L.; Jerfel, G.; and Tran, D. 2019. Measuring calibration in deep learning. In *CVPR workshops*, volume 2.

Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; and Snoek, J. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

Sale, Y.; Caprio, M.; and Hüllermeier, E. 2023. Is the Volume of a Credal Set a Good Measure for Epistemic Uncertainty? In *Accepted for Conference on Uncertainty in Artificial Intelligence (UAI 2023)*, 0–12.

Senge, R.; Bösner, S.; Dembczyński, K.; Haasenritter, J.; Hirsch, O.; Donner-Banzhoff, N.; and Hüllermeier, E. 2014. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255: 16–29.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society.

Soubaras, H. 2011. Towards an Axiomatization for the Generalization of the Kullback-Leibler Divergence to Belief Functions. In *Proceedings of the 7th conference of the European Society for Fuzzy Logic and Technology*, 1090–1097. Atlantis Press.

Ulmer, D. T.; Hardmeier, C.; and Frellsen, J. 2023. Prior and Posterior Networks: A Survey on Evidential Deep Learning Methods For Uncertainty Estimation. *Transactions on Machine Learning Research*.

Virtanen, P.; Gommers, R.; Oliphant, T. E.; et al. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17: 261–272.

Wang, K.; Cuzzolin, F.; Manchingal, S. K.; Shariatmadar, K.; Moens, D.; and Hallez, H. 2024. Credal Deep Ensembles for Uncertainty Quantification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wang, K.; Cuzzolin, F.; Shariatmadar, K.; Moens, D.; and Hallez, H. 2025a. Credal Wrapper of Model Averaging for Uncertainty Estimation in Classification. In *The Thirteenth International Conference on Learning Representations*.

Wang, K.; Cuzzolin, F.; Shariatmadar, K.; Moens, D.; and Hallez, H. 2025b. A Review of Uncertainty Representation and Quantification in Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20.

Wang, K.; Shariatmadar, K.; Manchingal, S. K.; Cuzzolin, F.; Moens, D.; and Hallez, H. 2025c. CreINNs: Credal-Set Interval Neural Networks for Uncertainty Estimation in Classification Tasks. *Neural Networks*, 185: 107198.

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