(Implicit) Ensembles of Ensembles: Epistemic Uncertainty Collapse in Large Models

Anonymous authors
Paper under double-blind review

Abstract

Epistemic uncertainty is crucial for safety-critical applications and out-of-distribution detection tasks. Yet, we uncover a paradoxical phenomenon in deep learning models: an *epistemic uncertainty collapse* as model complexity increases, challenging the assumption that larger models invariably offer better uncertainty quantification. We propose that this stems from *implicit ensembling* within large models. To support this hypothesis, we demonstrate epistemic uncertainty collapse empirically across various architectures, from explicit *ensembles of ensembles* and simple MLPs to state-of-the-art vision models, including ResNets and Vision Transformers—for the latter, we examine *implicit ensemble extraction* and decompose larger models into diverse sub-models, recovering epistemic uncertainty. We provide theoretical justification for these phenomena and explore their implications for uncertainty estimation.

1 Introduction

Bayesian deep learning provides us with a principled framework for quantifying uncertainty in complex machine learning models (MacKay, 1992; Neal, 1994). A key concept in this framework is *epistemic uncertainty*, which represents a model's uncertainty about its predictions due to limited knowledge or data (Smith & Gal, 2018; Der Kiureghian & Ditlevsen, 2009). This form of uncertainty is distinct from *aleatoric uncertainty*, which captures inherent noise or randomness in the data (Kendall & Gal, 2017). Epistemic uncertainty is particularly useful in safety-critical applications and decision-making scenarios, where understanding the limits of a model's knowledge can prevent costly errors.

Epistemic uncertainty is usually quantified using the mutual information between the model's predictions and its parameters (Houlsby et al., 2011; Gal et al., 2017). For deep ensembles (Lakshminarayanan et al., 2017), which consist of multiple independently trained neural networks, this amounts to the mutual information between the ensemble member index and the models' predictions. This is equivalent to the disagreement among ensemble members on a particular input (Smith & Gal, 2018): if the models in the ensemble produce widely varying predictions for a given sample, this is indicative of high epistemic uncertainty.

Intuitively, one might expect that as deep learning models grow in size and complexity, their capacity for epistemic uncertainty would increase. As Fellaji & Pennerath (2024) argue, "the more parameters a model has, the more likely it is to fit the data in multiple ways. Put another way, the posterior and thus the posterior predictive will tend to be flatter, making the epistemic uncertainty grow," which is aligned with the conventional understanding of model complexity and uncertainty.

However, our work provides evidence for a simple yet paradoxical phenomenon: when constructing higher-order ensembles, *ensembles of ensembles*, we observe an *epistemic uncertainty collapse*. This collapse occurs because individual ensembles, given sufficient size and training, converge to similar predictive distributions, causing inter-ensemble disagreement to vanish as the ensemble size grows.

We hypothesize that, similar to ensembles of ensembles, *implicit ensembling* might occur within the layers of large over-parameterized neural networks, potentially leading to significant underestimation of epistemic uncertainty for traditional uncertainty estimators that rely on final logits. Hence, similar to deep ensembles that have been found to offer better calibration (Ovadia et al., 2019), implicit ensembling may explain why larger

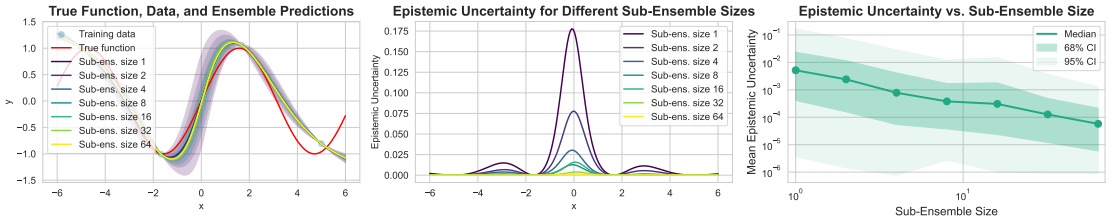


Figure 1: **Epistemic Uncertainty Collapse in a Toy Regression Problem.** As the sub-ensemble size increases, epistemic uncertainty vanishes. Ensembles of 10 sub-ensembles with different sub-ensemble sizes. *Left:* True function, data, and ensemble predictions. *Middle:* Epistemic uncertainty across input space. *Right:* Mean epistemic uncertainty vs. sub-ensemble size.

models also appear more calibrated (Tran et al., 2022). Recent work by Fellaji & Pennerath (2024) provides additional evidence that the epistemic uncertainty collapse can occur even in simple over-parameterized MLPs trained on standard benchmark datasets.

A wide range of applications relies on accurate uncertainty estimates. These include active learning, where uncertainty guides data acquisition; anomaly detection, where uncertainty can signal out-of-distribution inputs; and safety-critical systems, where understanding model confidence is crucial for responsible deployment. The potential underestimation of epistemic uncertainty in these contexts could lead to overconfident predictions and compromised decision-making processes, particularly in large over-parameterized neural networks.

In the remainder of this paper, we provide both theoretical justification and empirical evidence for the epistemic uncertainty collapse phenomenon using toy examples, and MLPs trained on MNIST. Vice-versa, we show that *implicit ensemble extraction,* which decomposes larger models into sub-models, effectively recovers hidden ensemble structure and epistemic uncertainty on simple MLPs and state-of-the-art vision models based on ResNets (He et al., 2015) and Vision Transformers (Dosovitskiy et al., 2021).

2 Theoretical Framework

The quantification of uncertainty is crucial for robust and reliable predictions. We begin by distinguishing between two important types of uncertainty (Der Kiureghian & Ditlevsen, 2009):

1. **Aleatoric uncertainty**: Often referred to as statistical uncertainty, it represents the inherent noise in the data. This type of uncertainty is irreducible given the current set of features and cannot be reduced by collecting more data.
2. **Epistemic uncertainty**: Also known as model uncertainty, it stems from our lack of knowledge about the true data-generating process. This uncertainty can, in principle, be reduced by gathering more data or by using more expressive models.

Traditional Neural Networks vs. Bayesian Neural Networks. Traditional Neural Networks (NNs), while powerful function approximators, typically provide point estimates as outputs and inherently do not capture model uncertainty. This limitation means that regular NNs cannot easily distinguish between aleatoric and epistemic uncertainty.

In contrast, Bayesian Neural Networks (BNNs) infer a distribution over model parameters instead of single-point estimates. This probabilistic approach allows BNNs to capture epistemic uncertainty. When a BNN encounters data similar to its training distribution, the predictive distributions across the parameter distribution tend to be more concentrated, indicating lower epistemic uncertainty. For out-of-distribution data, on the other hand,these distributions become more diffuse, indicating greater epistemic uncertainty.

Bayesian Model Average. The *Bayesian Model Average (BMA)* lies at the core of Bayesian deep learning. Let $\mathrm{p}(\theta \mid \mathcal{D})$ be the distribution over model parameters θ, given observed data $\mathcal{D}$. Then, the predictive

distribution for a new input $\mathbf{x}^*$ is:

$$\mathrm{p}(y^* \mid \mathbf{x}^*, \mathcal{D}) = \int \mathrm{p}(y^* \mid \mathbf{x}^*, \theta)\, \mathrm{p}(\theta \mid \mathcal{D}) d\theta. \tag{1}$$

Information-Theoretic Quantities. To formally quantify and differentiate between aleatoric and epistemic uncertainty, we can use an information-theoretic decomposition (Kendall & Gal, 2017). Let Y be the predicted output, and θ be the model parameters. We define:

1. **Total Uncertainty** as the *entropy* of the predictive distribution of the BMA:

$$\mathrm{H}[Y \mid \mathbf{x}, \mathcal{D}] = -\int \mathrm{p}(Y \mid \mathbf{x}, \mathcal{D}) \log \mathrm{p}(Y \mid \mathbf{x}, \mathcal{D}) dY. \tag{2}$$

2. **Epistemic Uncertainty** ($\mathrm{I}[Y; \Theta \,|\, \mathbf{x}, \mathcal{D}]$) as the *mutual information* which estimates the expected reduction in uncertainty about the prediction Y that would be obtained if we knew the model parameters θ:

$$\mathrm{I}[Y; \Theta \mid \mathbf{x}, \mathcal{D}] = \mathrm{H}[Y \mid \mathbf{x}, \mathcal{D}] - \mathbb{E}_{\mathrm{p}(\theta \mid \mathcal{D})}[\mathrm{H}[Y \mid \mathbf{x}, \theta]]. \tag{3}$$

   It has been shown to be an effective measure of epistemic uncertainty (Houlsby et al., 2011; Gal et al., 2017; Smith & Gal, 2018).

2.1 Deep Ensembles

Deep ensembles, introduced by Lakshminarayanan et al. (2017), have emerged as a simple yet effective method for approximating Bayesian inference in deep learning (Wilson & Izmailov, 2020). They have been shown to offer superior performance compared to more complex Bayesian approximations on many tasks (Ovadia et al., 2019), including calibration measurements (Guo et al., 2017), out-of-distribution detection, and classification with rejection (Ashukha et al., 2020). The core idea is to train multiple neural networks independently using different random initializations.

The BMA of a deep ensemble of M models is:

$$\mathrm{p}(Y \mid \mathbf{x}, \mathcal{D}) \approx \frac{1}{M} \sum_{m=1}^{M} \mathrm{p}(Y \mid \mathbf{x}, \theta_m), \tag{4}$$

where $\theta_m \sim \mathrm{p}(\theta_m \mid \mathcal{D})$ represents the parameters of the m-th model in the ensemble.

Given m ensemble members, we can estimate the epistemic uncertainty via the mutual information between the predicted class and the model index (Beluch et al., 2018):

$$\mathrm{I}[Y; M \mid \mathbf{x}, \mathcal{D}] \approx \mathrm{H}(\mathbb{E}_m[\mathrm{p}(Y \mid \mathbf{x}, \theta_m)]) - \mathbb{E}_m[\mathrm{H}[Y \mid \mathbf{x}, \theta_m]]. \tag{5}$$

3 Epistemic Uncertainty Collapse for Ensembles of Ensembles

Let us formally investigate a paradoxical phenomenon: the collapse of epistemic uncertainty when constructing ensembles of ensembles. Consider a scenario where we construct a higher-order ensemble by creating multiple deep ensembles, each comprised of M models. Intuitively, one might expect this approach to yield more robust uncertainty estimates, as it effectively increases the diversity of our model space. However, our analysis has a surprising outcome: as we increase the size of each individual ensemble, the disagreement between these ensembles—and consequently, our estimate of epistemic uncertainty—vanishes.

Let $\mathcal{E}_{1:\mathcal{K}} \coloneqq \{\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_{\mathcal{K}}\}$ be a set of $\mathcal{K}$ deep ensembles, each containing $\mathcal{M}$ models:

$$\mathcal{E}_k \coloneqq \{\theta_1^k, \theta_2^k, ..., \theta_{\mathcal{M}}^k\}, \tag{6}$$

where θ_m^k represents the parameters of the m-th model in the k-th ensemble. For a given input $\mathbf{x}$, the predictive distribution of the k-th ensemble is:

$$\mathrm{p}(y \mid \mathbf{x}, \mathcal{E}_k) = \frac{1}{M} \sum_{m=1}^{M} \mathrm{p}(y \mid \mathbf{x}, \theta_m^k), \tag{7}$$

where we have dropped conditioning on the data $\mathcal{D}$ for brevity. Consequently, the predictive distribution of the ensemble of ensembles is equal to the BMA over all individual members:

$$\mathrm{p}(y \mid \mathbf{x}, \mathcal{E}_{1:\mathcal{K}}) = \frac{1}{\mathcal{K}} \sum_{k=1}^{\mathcal{K}} \mathrm{p}(y \mid \mathbf{x}, \mathcal{E}_k) = \frac{1}{\mathcal{K}} \frac{1}{\mathcal{M}} \sum_{k=1}^{\mathcal{K}} \sum_{m=1}^{\mathcal{M}} \mathrm{p}(y \mid \mathbf{x}, \theta_m^k). \tag{8}$$

By more canonically defining Θ to depend on K and M as categorical random variables with uniform distribution, we can rephrase the epistemic uncertainty as:

$$\mathrm{I}[Y; (K, M) \mid \mathbf{x}] = \mathrm{I}[Y; \theta_M^K \mid \mathbf{x}]. \tag{9}$$

Infinite Sub-Ensemble Size. How does the epistemic uncertainty change with increasing size of the sub-ensembles? For this, we note that if we let $\mathcal{M} \to \infty$,

$$\mathrm{p}(y \mid \mathbf{x}, \mathcal{E}_k) = \frac{1}{\mathcal{M}} \sum_{m=1}^{\mathcal{M}} \mathrm{p}(y \mid \mathbf{x}, \theta_m^k) \to \mathbb{E}_{\mathrm{p}(\theta)}[\mathrm{p}(y \mid \mathbf{x}, \theta)] = \mathrm{p}(y \mid \mathbf{x}) \tag{10}$$

independent of k. Hence, thanks to the central limit theorem, we have:

$$\begin{aligned}
\mathrm{I}[Y; \mathcal{E}_K \mid \mathbf{x}, \mathcal{D}] &= \mathrm{I}[Y; K \mid \mathbf{x}, \mathcal{D}] && (11)\\
&= \mathrm{H}[\mathrm{p}(Y \mid \mathbf{x})] - \mathbb{E}_{\mathrm{p}(k)}[\mathrm{H}[\mathrm{p}(Y \mid \mathbf{x}, k)]] && (12)\\
&\to \mathrm{H}[\mathrm{p}(Y \mid \mathbf{x})] - \mathrm{H}[\mathrm{p}(Y \mid \mathbf{x})] && (13)\\
&= 0. && (14)
\end{aligned}$$

Epistemic Uncertainty Collapse

As the size of the sub-ensemble in an ensemble of ensembles increases, the epistemic uncertainty of the overall ensemble approaches zero, and we observe an *epistemic uncertainty collapse*. This collapse occurs because the individual ensembles converge to similar predictive distributions, leading to a reduction in the mutual information between the predicted class and the ensemble index.

Epistemic Uncertainty Decomposition. To better understand the relationship between the epistemic uncertainty of a single ensemble and that of an ensemble of ensembles, we can leverage the chain rule of the mutual information. Denoting the epistemic uncertainty of a single ensemble consisting of all models as $\mathrm{I}[Y; \theta_M^K \mid \mathbf{x}, \mathcal{D}]$ and the epistemic uncertainty of the ensemble of ensembles as $\mathrm{I}[Y; \mathcal{E}_K \mid \mathbf{x}, \mathcal{D}]$, we have:

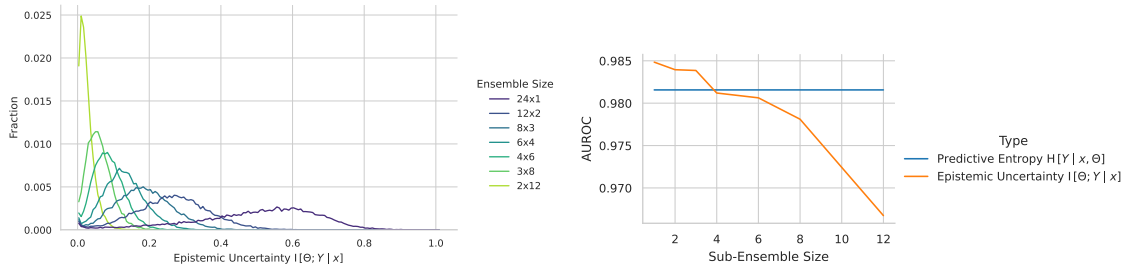
$$\begin{aligned}
\mathrm{I}[Y; \theta_M^K \mid \mathbf{x}, \mathcal{D}] &= \mathrm{I}[Y; K, M \mid \mathbf{x}, \mathcal{D}] && (15)\\
&= \mathrm{I}[Y; K \mid \mathbf{x}, \mathcal{D}] + \mathrm{I}[Y; M \mid K, \mathbf{x}, \mathcal{D}] && (16)\\
&= \mathrm{I}[Y; \mathcal{E}_K \mid \mathbf{x}, \mathcal{D}] + \mathrm{I}[Y; \theta_M^K \mid K, \mathbf{x}, \mathcal{D}]. && (17)
\end{aligned}$$

This decomposition shows that the epistemic uncertainty of a single ensemble can be expressed as the sum of the epistemic uncertainty of an ensemble of ensembles and the expected epistemic uncertainty within each ensemble (conditioned on the ensemble index K). That is, the more epistemic uncertainty is captured by the sub-ensembles, the less epistemic uncertainty remains for the ensemble of ensembles.

Implications for Large Models and Their Ensembles. The epistemic uncertainty collapse we have derived for ensembles of ensembles could have significant implications for epistemic uncertainty quantification in large neural networks, particularly for foundation models: as models grow in size and complexity, they may exhibit implicit ensembling effects within their layers, leading to a reduction in epistemic uncertainty estimates when ensembling them or even when using individual models.

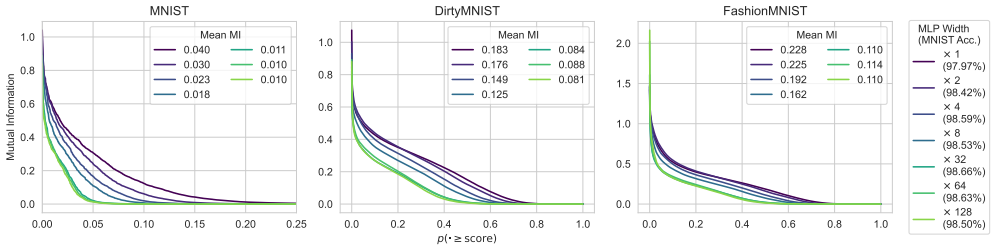
More concretely, large neural networks can be viewed as a hierarchical composition of implicit ensembles because each layer can be thought of as an ensemble of neurons, and successive layers of the network as a whole can then be considered an ensemble of ensembles. As the depth and width increase, we thus might observe the same epistemic uncertainty collapse as we have derived in ensembles of ensembles. Thus, larger models might not necessarily provide better uncertainty quantification.

This hypothesis about implicit ensembling in large neural networks set the stage for our empirical investigation.

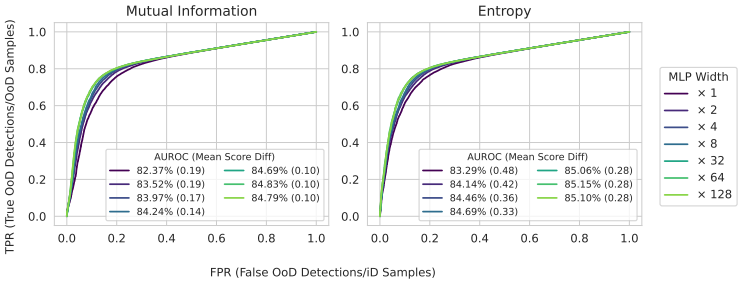


(a) Mutual Information Histogram on SVHN (OoD) (b) OoD Detection for CIFAR10 vs. SVHN (OoD)

Figure 2: **Ensemble of Ensemble Results for CIFAR10 (iD) vs. SVHN (OoD).** Different configurations of 24 ResNet-50 models trained on CIFAR-10. *(a)* As the sub-ensemble size increases, the epistemic uncertainty on SVHN as OoD dataset collapses. *(b)* The area under the receiver-operating characteristic (AUROC ↑) for OoD detection using mutual information slowly deteriorates as the sub-ensemble size increases.



(a) Mutual Information ECDF for Different MLP Widths



(b) MNIST vs. FashionMNIST OoD Detection AUROC

Figure 3: **Epistemic Uncertainty Collapse on MNIST via Implicit Ensembling.** *(a) Mutual Information Empirical Cumulative Distribution Function (ECDF) for Different MLP Widths.* As MLP size increases, mutual information decreases while accuracy remains stable. This trend persists across training and other distributions. *(b) MNIST vs. Fashion-MNIST OoD Detection AUROC Curves.* The mean difference in uncertainty scores between in-distribution and out-of-distribution samples (in parentheses) also decreases with width, further evidencing epistemic uncertainty collapse, while the AUROC for OoD detection slightly improves across both uncertainty metrics.

4 Empirical Results

In the following section, we present a series of experiments that not only demonstrate the epistemic uncertainty collapse in explicit ensembles of ensembles but also find parallels in the behavior of wide neural networks, providing evidence for the hypothesized effects of implicit ensembling. Details on the models, training setup, datasets, and evaluation are provided in §A and in §B in the appendix.

Toy Example. To illustrate the epistemic uncertainty collapse in ensembles of ensembles, we present a one-dimensional regression task with the ground-truth function $f(x) = \sin(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1)$. We

generate a small training dataset of four points over $[-5, 5]$ and fit this data using ensembles of neural networks with one hidden layer of 64 units and tanh activations. The output layer also uses a tanh activation, scaled by a factor of 2 to allow for a output range that surely covers any noised sample.

We create ensembles of 10 sub-ensembles each, with sub-ensemble sizes of 1, 2, 4, 8, 16, 32, and 64 members. Figure 1 presents the results across three panels. The left panel shows ensemble predictions converging to the true function as sub-ensemble size increases, with narrowing uncertainty bands. The middle panel illustrates decreasing epistemic uncertainty across the input space for larger sub-ensembles between training points. The right panel quantifies the inverse relationship between sub-ensemble size and mean epistemic uncertainty.

These results demonstrate the epistemic uncertainty collapse phenomenon: the systematic reduction of epistemic uncertainty, visible as reduced prediction variance and thus higher overall confidence, even in regions far from the training data. This toy example highlights how implicit ensembling within wide neural networks could lead to underestimating epistemic uncertainty, particularly in data-sparse regions.

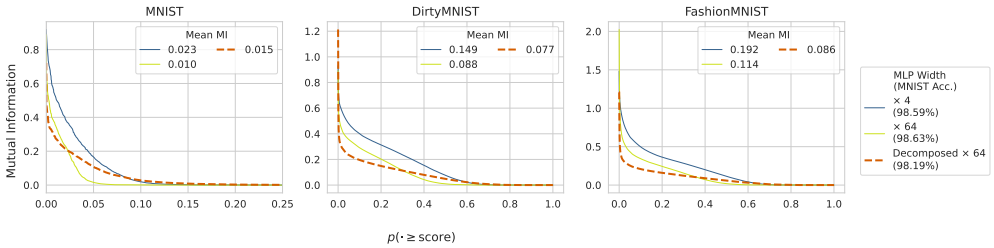
Explicit Ensemble of Ensemble. In Figure 2, we construct a deep ensemble comprising of 24 Wide-ResNet-28-1 models (Zagoruyko & Komodakis, 2016; He et al., 2015) trained on CIFAR-10 (Krizhevsky et al., 2009), which we then partition into ensembles of ensembles with varying sub-ensemble sizes ($24 \times 1, 12 \times 2, 8 \times 3, 6 \times 4, 4 \times 6, 3 \times 8, 2 \times 12$). Figure 2(a) illustrates the epistemic uncertainty of these ensemble configurations. As the sub-ensemble size increases, we observe a clear epistemic uncertainty collapse, manifested by the mutual information concentrating on smaller values. Figure 2(b) presents the AUROC for out-of-distribution (OoD) detection for different sub-ensemble sizes, comparing CIFAR-10 (in-distribution) with SVHN (OoD) (Netzer et al., 2011) using epistemic uncertainty as the detection metric. The AUROC shows a deterioration as the sub-ensemble size increases, directly resulting from the epistemic uncertainty collapse. Crucially, while the decrease in AUROC may appear modest, it is large enough to make the difference between state-of-the-art performance and baseline methods, such as confidence or entropy-based approaches (Hendrycks & Gimpel, 2017), e.g., compare with the results in Mukhoti et al. (2023).

Implicit Ensembling on MNIST. Surprisingly, the effect of the epistemic uncertainty collapse is even visible when training relatively small MLP models of varying width on MNIST in a controlled setting. We reproduce the results from Fellaji & Pennerath (2024) in Figure 3. That is, we expand the size of the inner linear of a 2-hidden-layer MLP by a varying factors and train 10-model ensembles for 100 epochs each.

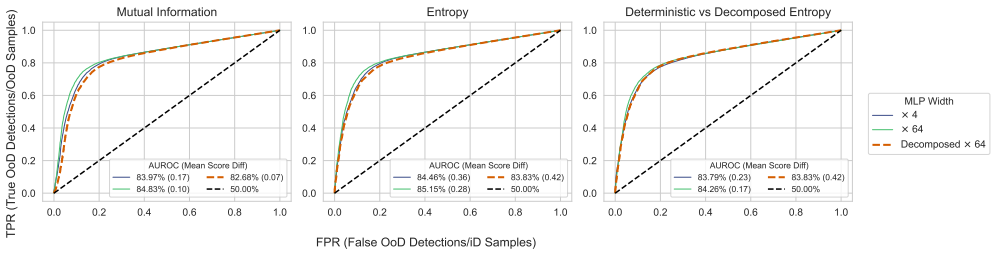
To quantify uncertainty, we employ two metrics that capture different aspects of the model's predictive distribution. Specifically, we utilize: *(1)* The predictive entropy of the ensemble, $\mathrm{H}[Y \mid \mathbf{X}, \mathcal{D}]$, which measures the overall uncertainty in the prediction (Hendrycks & Gimpel, 2017); and *(2)* the mutual information between the predicted class and the ensemble index, as an empirical estimate for $\mathrm{I}[Y; \Theta \mid \mathbf{X}, \mathcal{D}]$, which quantifies the epistemic uncertainty or model uncertainty (Malinin, 2019).

These results provide compelling evidence for the epistemic uncertainty collapse due to implicit ensembling in neural networks. As the width of the MLP increases, we observe a clear trend of decreasing mutual information across all datasets: MNIST (LeCun & Cortes, 1998), Dirty-MNIST (Mukhoti et al., 2023), and Fashion-MNIST (Xiao et al., 2017). In Figure 3(a), the effect is most pronounced for the in-distribution MNIST test data, but the collapse is also clearly visible for out-of-distribution data (Fashion-MNIST & SVHN test sets). The decrease in mutual information indicates a reduction in the model's epistemic uncertainty as it grows larger, despite maintaining similar accuracy. In Figure 3(b), the mean difference in uncertainty scores between in-distribution (MNIST) and out-of-distribution (Fashion-MNIST) samples also decreases with increasing model width, further corroborating the collapse of epistemic uncertainty. However, the AUROC for OoD detection using different uncertainty metrics slightly improves as the model width increases, which is in line with the results by Fellaji & Pennerath (2024), who report a significant deterioration of OoD performance for such MLP ensembles trained on CIFAR-10 but not MNIST.

This shows that epistemic uncertainty collapse due to implicit ensembling might be a pervasive phenomenon in neural networks, even in relatively simple MLP models.



(a) Mutual Information ECDF for Extracted Implicit Ensemble



(b) MNIST vs. FashionMNIST OoD Detection AUROC for Extracted Implicit Ensemble

Figure 4: **Recovering Epistemic Uncertainty through Implicit Ensemble Extraction.** *(a)* The extracted implicit ensemble (dashed line) largely recovers the mutual information scores of a fully trained ensemble of the same width, supporting the hypothesis of latent ensemble structures in large neural networks. The 10-member implicit ensemble is extracted from a single MLP with width factor 64 (§4.1). The regular 10-member ensembles comprise MLPs with width factors 4 and 64 trained on MNIST. Ensembles are evaluated on MNIST, Dirty-MNIST, and Fashion-MNIST test sets. *(b)* The extracted implicit ensemble shows comparable AUROC scores across all metrics relative to a fully trained deep ensemble of the same width. OoD detection is performed using mutual information or entropy scores. The final panel compares the softmax entropy of the original wide MLP with the predictive entropy of its extracted implicit ensemble. The mean entropy difference between iD and OoD samples is larger for the extracted ensemble. At the same time, the OoD performance does *not* match the single wider MLP.

4.1 Implicit Ensemble Extraction

To further substantiate the hypothesis that implicit ensembling is the underlying mechanism driving epistemic uncertainty collapse, we demonstrate that it is possible to mitigate this collapse by decomposing larger models into constituent sub-models. This approach, which we term implicit ensemble extraction, allows us to recover the diversity of an ensemble from a single large model, effectively reversing the collapse of epistemic uncertainty. This not only offers insights into the nature of epistemic uncertainty collapse but also points towards potential practical strategies for improving model robustness and out-of-distribution detection.

Extracted Implicit Ensemble from a Single MNIST MLP. First, we extract implicit ensembles from the MLPs trained on MNIST above. For a given model, an ensemble member with the largest width factor, 64, we train boolean masks on the weights such that we recover 10 individual models with maximally different masks and low individual loss on MNIST's training set. We find that the resulting deep ensemble performs as well as the ensemble of smaller models, even though we started from a single model.

Concretely, we add binary mask to each linear layer of the network which we relax to probabilities of binomial variables by applying sigmoid activations, effectively selecting subsets of the original weights. We optimize separate 1D masks for its rows and columns. The outer product of these 1D masks determines which weights from the original layer are included in each sub-model as a dense sub-matrix. We maximize the diversity among the resulting sub-models by regularizing the mutual information $\mathrm{I}[\text{mask}; M]$ between the masks and sub-model index M while minimizing the loss on the training set. This allows us to extract an implicit ensemble from a single trained network.

In Figure 4, we see the effectiveness of this implicit ensemble extraction technique. The results demonstrate that this decomposition can recover much of the epistemic uncertainty of an ensemble from a single model, providing support for our hypothesis about the mechanism underlying epistemic uncertainty collapse:

- **Effectiveness of Model Decomposition.** The dashed lines in both subfigures of Figure 4 represent the extracted ensemble, as described in the previous section. Remarkably, this decomposed model nearly recovers the mutual information scores of a fully trained ensemble with the same width factor despite being based on the weights of a single model, and its OoD detection performance approaches that of a narrower model. Since we can successfully extract an ensemble, this result supports our hypothesis that implicit ensembling is indeed occurring within larger models and that we can unlock this ensemble structure through decomposition.
- **Predictive Entropy of the Extracted Ensemble vs. the Softmax Entropy of the Original Model.** The last panel of Figure 4(b) compares the softmax entropy of the original large model with the predictive entropy of the decomposed ensemble. While we cannot compare the OoD detection performance of the extracted ensemble to the single wider MLP using mutual information (as the latter does not have a well-defined mutual information), we see that the extracted ensemble does not match the performance of the single wider MLP when using its predictive entropy. We speculate this might be because the masks do not fully cover the weights of the original model, and thus the extracted ensemble does not capture the full predictive entropy of the original ensemble.

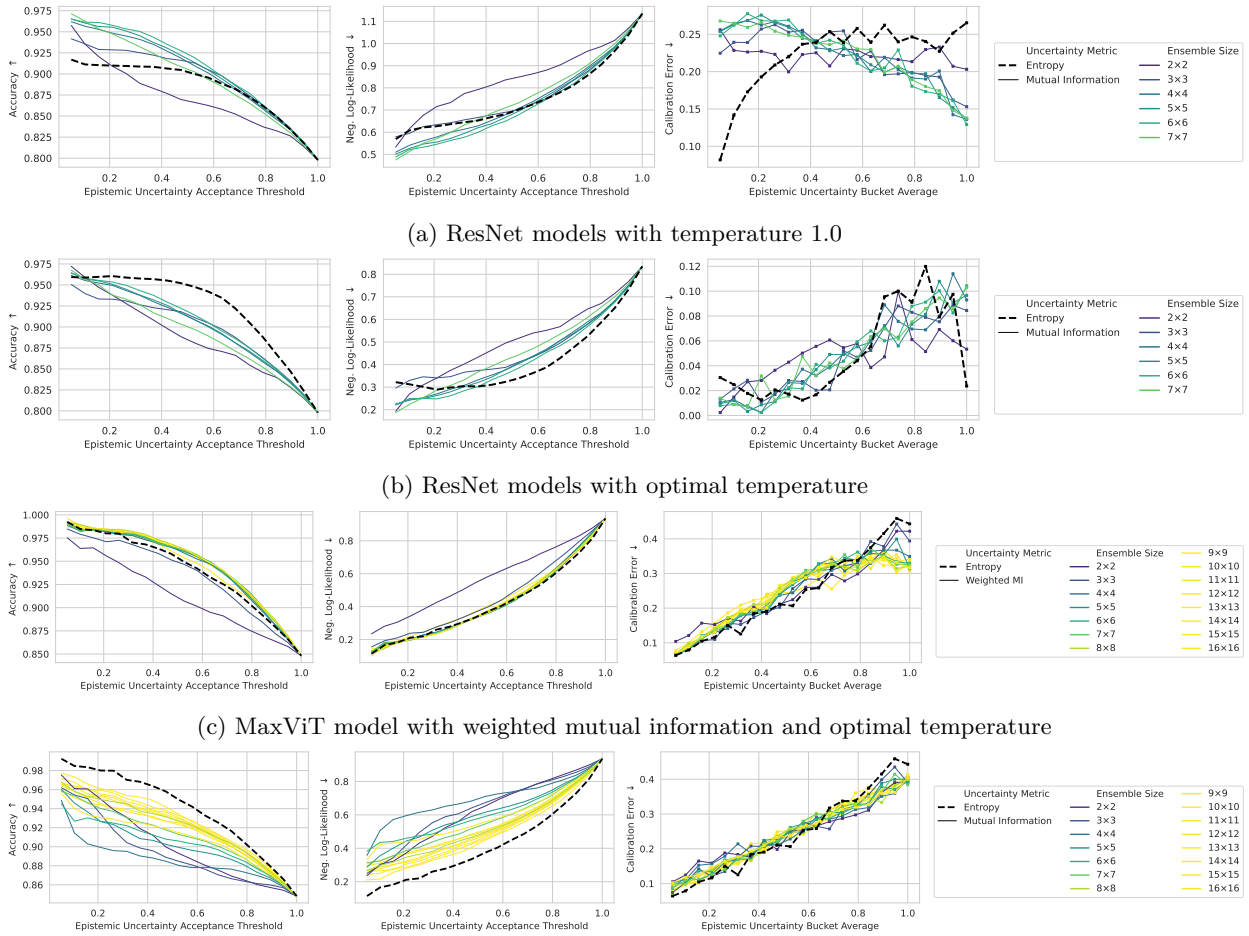
The initial success of this model decomposition technique in recovering epistemic uncertainty suggest a promising direction for improving uncertainty quantification and out-of-distribution detection in individual over-parameterized deep learning models. Collectively, this demonstrates that epistemic uncertainty collapse occurs even in relatively simple MLP models.

Extracting Ensembles from Pre-Trained Vision Models. Second, we explore implicit ensemble extraction from pre-trained vision models based on ResNet (He et al., 2015) and Vision Transformer (Dosovitskiy et al., 2021) model architectures. Leveraging the common use of average pooling in these models to aggregate spatial information, we extract implicit ensembles without optimizing masks. Concretely, we remove the global average pooling layer. This allows us to obtain per-tile class logits, which we average with different target sizes to create differently-sized ensembles.

We evaluate these models in-distribution on the ImageNet-v2 dataset (Recht et al., 2019), which serves as a more challenging test set for ImageNet-trained models (Russakovsky et al., 2015). Specifically, we compare pre-trained ResNet-152 (He et al., 2015), Wide ResNet-101-2 (Zagoruyko & Komodakis, 2016), ResNeXt-101-64x4d (Xie et al., 2017), and MaxViT (Tu et al., 2022) models, with the pre-trained weights retrieved from PyTorch's torchvision (maintainers & contributors, 2016) and timm (Wightman, 2019), respectively. Our evaluation pipeline computes various uncertainty metrics and performance measures for different ensemble sizes, ranging from 2×2 to 7×7 sub-models (respectively, 16×16 for MaxViT) extracted from a single pre-trained network. We compare these extracted ensembles against the original single model performance, using mutual information as the primary uncertainty metric for ensembles and entropy for the single model.

Figure 5 shows three key performance metrics—accuracy, negative log-likelihood (NLL), and calibration error—plotted against epistemic uncertainty quantiles of various extracted ensemble sizes for the original models and for temperature-scaled models (Guo et al., 2017) The solid lines represent extracted ensembles of increasing size, while the dashed black line represents the original single model. For ResNet-based ensembles, we use mutual information between predictions as the measure of epistemic uncertainty, whereas for the single model, we use entropy. For MaxViT, we use a weighted mutual information between predictions and ensemble size as the measure of epistemic uncertainty, which assign a weight to each ensemble member based on the logit sum of the member as it performs better.

For ResNet models, we find that the largest extracted ensembles (7x7) have a better response overall for accuracy and NLL, while the calibration error is worse throughout. For temperature-scaled model, we observe an almost linear response in accuracy and NLL, with better sensitivity in the low-uncertainty regime, but otherwise worse performance, except for a lower calibration error. For MaxViT, we find that the weighted mutual information performs better than the mutual information. The effects are less pronounced for



(a) ResNet models with temperature 1.0

(b) ResNet models with optimal temperature

(c) MaxViT model with weighted mutual information and optimal temperature

(d) MaxViT model with mutual information and optimal temperature

Figure 5: **Classification with Rejection for Implicit Ensemble Extractions from Pre-Trained Models.** Each subfigure shows three performance metrics (Accuracy, Negative Log-Likelihood, and Calibration Error) as a function of epistemic uncertainty quantiles for different ensemble sizes. Solid lines represent extracted ensembles of increasing size (from 2 to 7/16), while the dashed black line represents the original single model. *(a)* The mutual information between predictions is used as the epistemic uncertainty measure for ensembles, while entropy is used for the single model. As the ensemble size increases, we observe improved performance for the area under curve (AUC), which indicates better epistemic uncertainty calibration (with the notable exception of the calibration error). This demonstrates that extracting larger ensembles from a single pre-trained model can enhance performance and uncertainty quantification. *(b)* Temperature scaling improves epistemic uncertainty calibration in general but benefits the original model most. Accuracy and NLL for extracted epistemic uncertainty only benefit in the low-uncertainty regime. *(c)* For VIT models, we find that a mutual information weighted by the logit sum of each ensemble performs better than the mutual information (*(c)* vs *(d)* with mutual information).

temperature-scaled models, likely because these models already are better calibrated. Overall, the results are mixed but promising.

Limitations. These *exploratory* results provide initial evidence that ensemble extraction technique can unlock meaningful epistemic uncertainty, which is useful for downstream tasks. This is aligned with the hypothesis of implicit ensembling. At the same time, we already see that mutual information is not always the best uncertainty metric, the comparative performance changes depending on the model temperature, and the

pool size of the best implicit ensemble is not always the same for different metrics and model architectures. More results that show this are shown in §D.

5 Related Work

Our study of epistemic uncertainty collapse in large neural networks and ensemble extraction intersects with several recent works in adjacent areas.

Epistemic Uncertainty Collapse. Recent work by Fellaji & Pennerath (2024) observed decreasing epistemic uncertainty as model size and dataset size vary, even in simple MLPs. They termed the decrease in epistemic uncertainty the "epistemic uncertainty hole" but left its explanation to future work. Our study provides an explanation through both a theoretical framework and additional empirical evidence via ensembles of ensembles, implicit ensembling, and ensemble extraction. A more detailed comparison can be found in §C.

Extracting Sub-Models from a Larger Network. The concept of extracting sub-models from a larger network shares similarities with several existing approaches in the literature. The "lottery ticket hypothesis" (Frankle & Carbin, 2018) proposes that dense, randomly-initialized networks contain sparse subnetworks capable of training to similar accuracy. However, our approach differs in that we do not retrain the subnetworks, but rather identify diverse substructures within the pre-trained model. Our method is more closely related to "sub-network ensembles" (Durasov et al., 2021), where multiple subnetworks are extracted from a single trained network to form an ensemble. Unlike previous work that primarily focused on pruning for efficiency, our approach aims to recover epistemic uncertainty. We introduce a novel mutual information-based objective to obtain diverse masks, emphasizing diversity rather than pruning. This allows us to extract an ensemble that better captures the model's internal epistemic uncertainty.

Learning from Underspecified Data. Our work can also related to efforts in learning from underspecified data, such as the approach by Lee et al. (2022) to diversify and disambiguate model predictions. While their focus is on training strategies, our method extracts diverse sub-models from already trained networks, offering a complementary approach.

In contrast to these works, this work addresses fundamentally different research questions. We provide a novel, unified explanation for epistemic uncertainty collapse in large models and hierarchical ensembles, a phenomenon not previously explored in depth. Furthermore, we introduce and examine implicit ensemble extraction to mitigate epistemic uncertainty collapse.

6 Conclusion

This work has uncovered and analyzed the collapse of epistemic uncertainty in large neural networks and hierarchical ensembles. Our theoretical framework and empirical results demonstrate this phenomenon across various architectures and datasets, from simple MLPs to state-of-the-art vision models.

To address this issue, we examined implicit ensemble extraction, a method to decompose larger models into sub-models. This approach effectively recovers hidden ensemble structures and improves epistemic uncertainty estimates, supporting our hypothesis that implicit ensembling contributes to the observed uncertainty collapse.

Our findings challenge the assumption that more complex models invariably offer better uncertainty quantification out of the box. They open new avenues for research into uncertainty estimation in large-scale machine learning models, particularly for safety-critical applications and out-of-distribution detection tasks. Future work should focus on developing robust methods to maintain reliable uncertainty estimates as model complexity increases, ensuring the continued advancement of trustworthy AI systems.

References

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJxI5gHKDr.

William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9368–9377, 2018.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2): 105–112, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Nikita Durasov, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Masksembles for Uncertainty Estimation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13534–13543. IEEE Computer Society, 2021.

Mohammed Fellaji and Frédéric Pennerath. The Epistemic Uncertainty Hole: an issue of Bayesian Neural Networks. *arXiv preprint arXiv:2407.01985*, 2024.

Jonathan Frankle and Michael Carbin. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations*, 2018.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. arxiv e-prints. *arXiv preprint arXiv:1512.03385*, 10, 2015.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Máté Lengyel. Bayesian Active Learning for Classification and Preference Learning. *arXiv preprint*, abs/1112.5745, 2011.

Alex Kendall and Yarin Gal. What Uncertainties Do we Need in Bayesian Deep Learning for Computer Vision? In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Yann LeCun and Corinna Cortes. The mnist database of handwritten digits. 1998. URL https://api.semanticscholar.org/CorpusID:60282629.

Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and disambiguate: Learning from underspecified data. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022.

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3): 448–472, 1992.

TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016.

Andrey Malinin. *Uncertainty estimation in deep learning with application to spoken language assessment*. PhD thesis, 2019.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep Deterministic Uncertainty: A New Simple Baseline for Uncertainty Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24384–24394, 2023.

Radford M Neal. *Bayesian learning for neural networks.* PhD thesis, University of Toronto, 1994.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.

Dustin Tran, Jeremiah Liu, Michael W Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, et al. Plex: Towards reliability using pretrained large model extensions. *arXiv preprint arXiv:2207.07411*, 2022.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MaxViT: Multi-axis vision transformer. In *European conference on computer vision*, pp. 459–479. Springer, 2022.

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 5987–5995. Institute of Electrical and Electronics Engineers Inc., 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016.

A Model and Dataset Details

A.1 MNIST Experiments

For the MNIST experiments, we use a simple Multi-Layer Perceptron (MLP) architecture with two hidden layers . The model structure is as follows:

- Input layer: 784 units (28x28 flattened MNIST images)
- First hidden layer: 64 units multiplied by a width multiplier (ranging from 0.5x to 256x)
- Second hidden layer: 32 units multiplied by the same width multiplier
- Output layer: 10 units (one for each digit class)
- Activation function: ReLU after each hidden layer
- Dropout layers: Applied after each hidden layer with p=0.1

We train these models using the following configuration:

- Optimizer: SGD
- Learning rate: 0.01
- Batch size: 128
- Epochs: 100
- Loss function: Cross-entropy loss

We create ensembles of 10 models for each width configuration. The width multipliers used are 1x, 2x, 4x, 8x, 32x, 64x, and 128x the base width.

Datasets used:

- MNIST (LeCun & Cortes, 1998): Standard handwritten digit dataset (in-distribution)
- Fashion-MNIST (Xiao et al., 2017): Clothing item dataset (out-of-distribution)
- Dirty-MNIST (Mukhoti et al., 2023): MNIST with added noise (high aleatoric uncertainty)

A.2 CIFAR-10 Experiments

For the CIFAR-10 experiments, we use Wide-ResNet-28-1 models (Zagoruyko & Komodakis, 2016). We create an ensemble of 24 independently trained models, which we then partition into sub-ensembles of various sizes, following the training details of (Mukhoti et al., 2023).

Training configuration:

- Optimizer: SGD with momentum (0.9)
- Learning rate: 0.1, decayed by a factor of 10 at epochs 150 and 250
- Weight decay: 5e-4
- Batch size: 128
- Epochs: 350
- Loss function: Cross-entropy loss

Datasets used:

- CIFAR-10 (Krizhevsky et al., 2009): 10-class image classification dataset (in-distribution)
- SVHN (Netzer et al., 2011): Street View House Numbers dataset (out-of-distribution)

A.3 ImageNet Experiments

For the ImageNet experiments, we use pre-trained models from the torchvision and timm libraries:

- ResNet-152 (He et al., 2015)
- Wide ResNet-101-2 (Zagoruyko & Komodakis, 2016)
- ResNeXt-101-64x4d (Xie et al., 2017)
- MaxViT (Tu et al., 2022)

These models are evaluated on the ImageNet-v2 dataset (Recht et al., 2019), which serves as a more challenging test set for ImageNet-trained models.

A.4 Implicit Ensemble Extraction

For the implicit ensemble extraction experiments on MNIST, we use the following approach:

- Starting model: MLP with width factor 64
- Extraction method: Optimizing binary masks for each layer
- Number of extracted sub-models: 10
- Optimization objective: Maximize mask diversity (mutual information between masks and sub-model index) while minimizing the cross-entropy loss on training set
- Mask diversity weight: 2.0

For the pre-trained vision models, we extract implicit ensembles by:

- Removing the global average pooling layer
- Obtaining per-tile class logits
- Averaging these logits with different target sizes (from 2x2 to 7x7 for ResNets, and up to 16x16 for MaxViT)

Table 1: AUROC and Mean Difference for Different Metrics and MLP Widths

	AUROC		Mean MI Difference	
OoD Metric	MI	Entropy	MI	Entropy
MLP Width				
×**1**	0.824	0.833	0.188	0.482
×**2**	0.835	0.841	0.194	0.420
×**4**	0.840	0.845	0.169	0.362
×**8**	0.842	0.847	0.144	0.329
×**32**	0.847	0.851	0.099	0.285
×**64**	0.848	0.851	0.104	0.280
×**1.3e+02**	0.848	0.851	0.100	0.275

B Evaluation Details

This section provides detailed information about our evaluation metrics, with a particular focus on the weighted mutual information and calibration error calculations.

B.1 Weighted Mutual Information

For the MaxViT model, we introduce a weighted mutual information metric to measure epistemic uncertainty. This metric assigns a weight to each ensemble member based on the logit sum of that member. The weighted mutual information is calculated as follows:

1. For each input, we compute the logits for all ensemble members.
2. We calculate the sum of logits for each ensemble member.
3. We normalize these sums to create weights for each member.
4. We compute the mutual information between the predictions and the ensemble index, weighting each member's contribution by its normalized logit sum.

Formally, let $l_{i,j}$ be the logit sum for the i-th input and j-th ensemble member. The weight w_j for this member is:

$$w_j = \frac{\sum_i \exp(l_{i,j})}{\sum_{i,k} \exp(l_{i,k})}$$

The weighted mutual information is then computed using these weights in place of the uniform weights used in standard mutual information calculations.

B.2 Calibration Error

We compute the calibration error as the absolute difference between the mean confidence and mean accuracy. This is calculated by first computing the softmax of the logits to get probabilities, then taking the maximum probability as the confidence for each prediction. We then compare the predictions to the true labels to determine accuracy. The calibration error is the absolute difference between the mean confidence and mean accuracy across all samples.

B.3 Metric Computation by Uncertainty Score

To analyze how different metrics vary with uncertainty, we compute metrics for different quantiles of an uncertainty score. This process involves:

1. Sorting the inputs based on the provided uncertainty scores.
2. Dividing the sorted inputs into quantiles.
3. Computing the specified metric for each quantile ("Bucket Average") or up to the given quantile ("Acceptance Threshold").

This approach allows us to observe how metrics like accuracy, negative log-likelihood, or calibration error change as a function of the model's uncertainty.

B.4 Other Evaluation Metrics

In addition to the above, we used several standard evaluation metrics:

1. **Accuracy**: The proportion of correct predictions.
2. **Negative Log-Likelihood (NLL)**: The negative log-likelihood of the true labels under the model's predictions.
3. **Entropy**: The entropy of the model's predictive distribution, used as a baseline uncertainty measure for single models.
4. **Mutual Information**: For ensembles, we use the mutual information between the predicted class and the ensemble index as a measure of epistemic uncertainty.
5. **AUROC**: The Area Under the Receiver Operating Characteristic curve, used for evaluating out-of-distribution detection performance.

These metrics were computed across different uncertainty quantiles to analyze how model performance and uncertainty estimates correlate. For the AUROC calculations, we used the uncertainty scores (entropy for single models, (weighted) mutual information for ensembles) as the ranking criterion to distinguish between in-distribution and out-of-distribution samples.

C Comparison with Fellaji & Pennerath (2024)

Our work explains the main findings of Fellaji & Pennerath (2024), who observe what they term the "epistemic uncertainty hole" in Bayesian neural networks. While their work also provides empirical evidence for this phenomenon, our study offers several key extensions and insights:

- **Theoretical Framework:** We provide a theoretical explanation for the epistemic uncertainty collapse through the lens of ensembles of ensembles and implicit ensembling, offering a mechanistic understanding of why this phenomenon occurs:
  - **Ensembles of Ensembles:** Our work introduces the concept of ensembles of ensembles, showing how the epistemic uncertainty collapse manifests in hierarchical ensemble structures. This provides a novel perspective on the phenomenon not explored in the original work.
  - **Implicit Ensemble Extraction:** We propose and evaluate a novel technique for mitigating the epistemic uncertainty collapse through implicit ensemble extraction. This practical approach to addressing the issue goes beyond the observational nature of Fellaji & Pennerath (2024)'s work.
- **Broader Model Architectures:** While Fellaji & Pennerath (2024) primarily focused on MLPs, we demonstrate that this phenomenon extends to more complex architectures, including state-of-the-art vision models based on ResNets and Vision Transformers.

Thus, while Fellaji & Pennerath (2024) observe an epistemic uncertainty collapse, our work provides a more comprehensive theoretical and empirical investigation of this phenomenon. We not only confirm their findings across a broader range of models and datasets but also offer new insights into the mechanisms behind this effect and potential strategies for mitigation.

Furthermore, our analysis of the implications for out-of-distribution detection and the proposed implicit ensemble extraction technique represent initial steps towards addressing the practical challenges posed by the epistemic uncertainty collapse in real-world applications.

D Additional Results

Table 2: **Mean Mutual Information, Accuracy and NLL for Different MLP Widths and Datasets.**

	Mean MI			Accuracy		NLL	
Dataset	MNIST	Dirty-MNIST	Fashion-MNIST	MNIST	Dirty-MNIST	MNIST	Dirty-MNIST
MLP Width							
$\times$**1**	0.0397	0.183	0.228	98	76.1	0.00216	0.477
$\times$**2**	0.0305	0.176	0.225	98.4	77.5	4.81e-05	0.531
$\times$**4**	0.0233	0.149	0.192	98.6	77.7	1.53e-06	0.471
$\times$**8**	0.0182	0.125	0.162	98.5	77.5	1.73e-06	0.415
$\times$**32**	0.011	0.0838	0.11	98.7	77.2	2.64e-07	0.402
$\times$**64**	0.0104	0.0882	0.114	98.6	77.1	3.73e-08	0.46
$\times$**128**	0.00956	0.0805	0.11	98.5	76.6	7.08e-08	0.351

Table 3: **Covariance between the mutual information as uncertainty metric and performance metrics for different ResNet models and extracted ensemble sizes.** Higher absolute values indicate stronger relationships. The arrows denote which (neg) covariance is to be preferred.

			Neg. Bucket Covariance	Neg. Acceptance Covariance	
		Performance Metric	Calibration Error ↓	Accuracy ↑	Neg. Log-Likelihood ↓
Model	Uncertainty Metric	Ensemble Size			
Resnet152-V2	**Entropy**	**Original**	-0.030	0.046	-0.144
	Mutual Information	**2×2**	-0.016	0.050	-0.179
		3×3	-0.030	0.044	-0.150
		4×4	-0.033	0.053	-0.189
		5×5	-0.035	0.054	-0.192
		6×6	**-0.037**	0.053	-0.185
		7×7	-0.031	**0.054**	**-0.203**
Wide_Resnet101_2-V2	**Entropy**	**Original**	-0.030	0.044	-0.141
	Mutual Information	**2×2**	-0.019	0.049	-0.174
		3×3	-0.031	0.039	-0.139
		4×4	-0.038	0.051	-0.187
		5×5	-0.038	0.052	-0.188
		6×6	**-0.039**	**0.052**	-0.191
		7×7	-0.036	0.050	**-0.194**
Resnext101_64X4D	**Entropy**	**Original**	-0.018	0.046	-0.147
	Mutual Information	**2×2**	-0.011	**0.049**	**-0.162**
		3×3	-0.019	0.042	-0.145
		4×4	-0.022	0.042	-0.156
		5×5	-0.021	0.043	-0.156
		6×6	**-0.024**	0.044	-0.159
		7×7	-0.021	0.043	-0.156

Table 4: **Covariance between the weighted mutual information as uncertainty metric and performance metrics for different ResNet models and extracted ensemble sizes.** Higher absolute values indicate stronger relationships. The arrows denote which (neg) covariance is to be preferred.

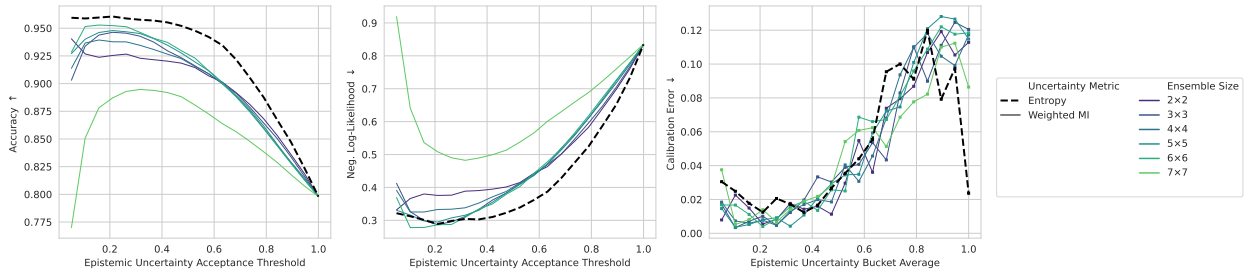
			Neg. Bucket Covariance	Neg. Acceptance Covariance	
		Performance Metric	Calibration Error ↓	Accuracy ↑	Neg. Log-Likelihood ↓
Model	Uncertainty Metric	Ensemble Size			
Resnet152-V2	**Entropy**	**Original**	-0.030	0.046	-0.144
	Weighted MI	**2×2**	-0.032	0.036	-0.134
		3×3	-0.039	0.038	-0.151
		4×4	-0.037	0.038	-0.144
		5×5	**-0.042**	**0.047**	**-0.178**
		6×6	-0.038	0.044	-0.159
		7×7	-0.035	0.020	-0.062
Wide_Resnet101_2-V2	**Entropy**	**Original**	-0.030	0.044	-0.141
	Weighted MI	**2×2**	-0.039	0.040	-0.138
		3×3	-0.039	0.039	-0.148
		4×4	-0.044	0.043	-0.159
		5×5	**-0.047**	0.042	-0.163
		6×6	-0.043	**0.049**	**-0.180**
		7×7	-0.037	0.016	-0.073
Resnext101_64X4D	**Entropy**	**Original**	-0.018	0.046	-0.147
	Weighted MI	**2×2**	**-0.043**	0.036	-0.131
		3×3	-0.042	0.045	-0.160
		4×4	-0.039	0.043	-0.150
		5×5	-0.042	0.045	-0.157
		6×6	-0.040	**0.049**	**-0.171**
		7×7	-0.023	0.009	-0.034

Table 5: **Covariance between the mutual information as uncertainty metric and performance metrics for the MaxVit model and extracted ensemble sizes.** Higher absolute values indicate stronger relationships. The arrows denote which (neg) covariance is to be preferred.

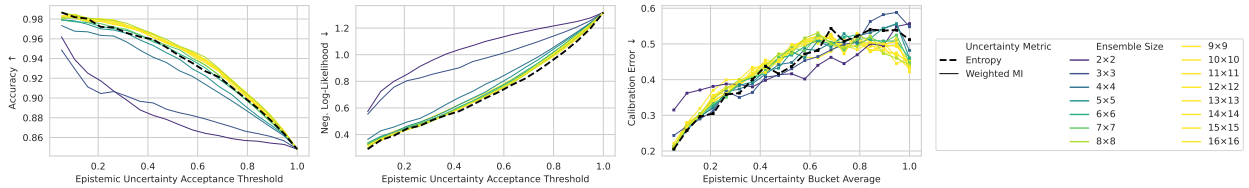
			Neg. Bucket Covariance	Neg. Acceptance Covariance	
		Performance Metric	Calibration Error ↓	Accuracy ↑	Neg. Log-Likelihood ↓
Model	Uncertainty Metric	Ensemble Size			
Timm-Maxvit	**Entropy**	**Original**	**-0.124**	**0.043**	**-0.228**
	Mutual Information	**10×10**	-0.098	0.032	-0.187
		11×11	-0.097	0.031	-0.183
		12×12	-0.101	0.033	-0.197
		13×13	-0.102	0.034	-0.199
		14×14	-0.104	0.035	-0.203
		15×15	-0.105	0.036	-0.207
		16×16	-0.103	0.037	-0.211
		2×2	-0.083	0.040	-0.217
		3×3	-0.103	0.034	-0.200
		4×4	-0.084	0.020	-0.130
		5×5	-0.100	0.029	-0.176
		6×6	-0.094	0.025	-0.155
		7×7	-0.100	0.031	-0.184
		8×8	-0.096	0.033	-0.193
		9×9	-0.091	0.028	-0.161

Table 6: *Covariance between the weighted mutual information as uncertainty metric and performance metrics for the MaxVit model and extracted ensemble sizes.* Higher absolute values indicate stronger relationships. The arrows denote which (neg) covariance is to be preferred.

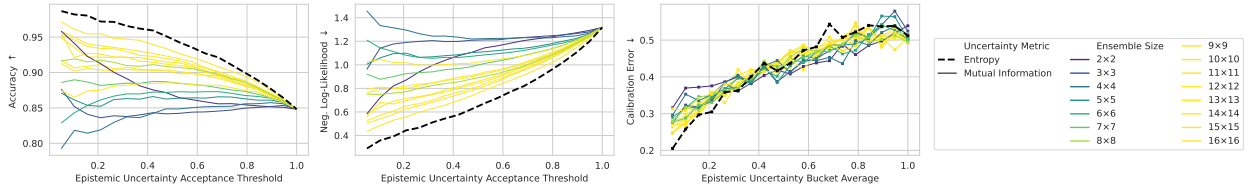
			Neg. Bucket Covariance	Neg. Acceptance Covariance	
		Performance Metric	Calibration Error ↓	Accuracy ↑	Neg. Log-Likelihood ↓
Model	Uncertainty Metric	Ensemble Size			
Timm-Maxvit	**Entropy**	**Original**	**-0.124**	**0.043**	-0.228
	Weighted MI	**10×10**	-0.089	0.041	-0.236
		11×11	-0.089	0.041	-0.235
		12×12	-0.086	0.041	-0.236
		13×13	-0.088	0.041	-0.237
		14×14	-0.086	0.041	-0.236
		15×15	-0.085	0.041	-0.235
		16×16	-0.082	0.043	-0.233
		2×2	-0.093	0.039	-0.213
		3×3	-0.115	0.042	-0.237
		4×4	-0.101	0.041	-0.236
		5×5	-0.100	0.041	-0.236
		6×6	-0.095	0.040	-0.234
		7×7	-0.094	0.040	-0.234
		8×8	-0.085	0.041	**-0.238**
		9×9	-0.091	0.041	-0.237



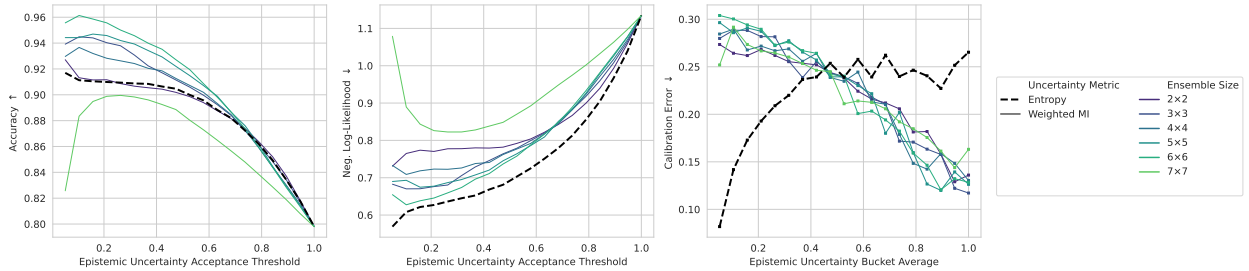
(a) ResNet models with weighted mutual information with optimal temperature



(b) MaxViT model with weighted mutual information and temperature 1.0



(c) MaxViT model with mutual information and temperature 1.0



(d) ResNet models with weighted mutual information and temperature 1.0

Figure 6: **Complementary Plots of Performance metrics for different ensemble sizes extracted from pre-trained models.**