Set-based v.s. Distribution-based Representations of Epistemic Uncertainty: A Comparative Study

Kaizheng Wang[*1,2] **Yunjia Wang**[2,5] **Fabio Cuzzolin**[3] **David Moens**[4,5] **Hans Hallez**[2] **Siu Lun Chau**[1]

[1]College of Computing and Data Science, Nanyang Technological University, Singapore
[2]Department of Computer Science, KU Leuven, Belgium
[3]School of Engineering, Computing, and Mathematics, Oxford Brookes University, U.K.
[4]Department of Mechanical Engineering, KU Leuven, Belgium
[5]Flanders Make@KU Leuven, Belgium

Abstract

Epistemic uncertainty in neural networks is commonly modeled using two second-order paradigms: distribution-based representations, which rely on posterior parameter distributions, and set-based representations based on credal sets. These frameworks are often regarded as fundamentally non-comparable due to differing semantics, assumptions, and evaluation practices, leaving their relative merits unclear. Empirical comparisons are further confounded by variations in the underlying predictive models. To clarify this issue, we present a controlled comparative study enabling principled, like-for-like evaluation of the two paradigms. Both representations are constructed from the same finite collection of predictive distributions generated by a shared neural network, isolating representational effects from predictive accuracy. Our study evaluates each representation through the lens of 3 uncertainty measures across 14 benchmarks, including selective prediction and out-of-distribution detection, spanning 6 underlying predictive models and 10 independent runs per configuration. Our results show that meaningful comparison between these seemingly non-comparable frameworks is both feasible and informative, providing insights into how second-order representation choices impact practical uncertainty-aware performance.

1 INTRODUCTION

Recent research has increasingly emphasized the representation and quantification of epistemic uncertainty (EU) in neural networks (NNs) to improve the robustness and reliability, particularly in safety-critical settings [Zhou et al., 2012, Tuo and Wang, 2022, Mukhoti et al., 2023, Mehrtens et al., 2023, Chau et al., 2025a]. EU captures a models incomplete knowledge of the true inputoutput relationship and reflects uncertainty that is, in principle, reducible with additional information. Modeling EU often requires a second-order formalism capable of expressing uncertainty over the models own probabilistic predictions [Hüllermeier and Waegeman, 2021, Wang et al., 2026b].

Two dominant paradigms have emerged for representing such second-order uncertainty. The first adopts a distribution-based representation, where uncertainty is modeled via probability distributions over model parameters or predictions. This perspective underlies Bayesian neural networks (BNNs) as well as practical approximations such as deep ensembles (DE) [Blundell et al., 2015, Krueger et al., 2017, Lakshminarayanan et al., 2017]. In practice, it is often simply a uniform distribution placed over a finite collection of (sampled) predictive distributions. The second paradigm employs a set-based representation, in which uncertainty is encoded by sets of plausible predictive distributions, most commonly credal sets—convex sets of probability distributions—as used in recent credal classification frameworks [Levi, 1980, Wang et al., 2024a, 2026a, 2025a, Löhr et al., 2025].

Despite their shared objective of capturing epistemic uncertainty, these paradigms are frequently regarded as fundamentally non-comparable. The two frameworks differ in semantics, mathematical structure, modeling assumptions, and evaluation methodology. Distribution-based approaches are often interpreted through Bayesian lenses, whereas credal approaches adopt imprecision-aware or set-valued perspectives. Consequently, existing discussions of their relative merits remain largely conceptual, and empirical comparisons are difficult to interpret due to confounding factors. In particular, comparisons typically involve predictors derived from different learning algorithms, architectures, or training procedures, making it unclear whether ob-

[*]This work was initiated at KU Leuven and primarily completed at Nanyang Technological University. Corresponding author: Kaizheng Wang (kaizheng.wang@ntu.edu.sg).

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7072–7102.

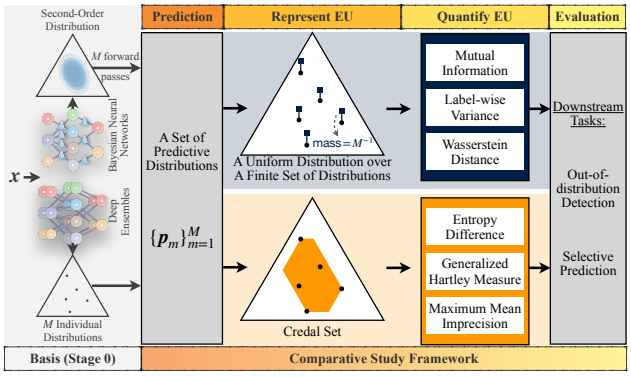

Figure 1: Illustration of our comparative study framework.

served differences arise from predictive accuracy, optimization effects, or the uncertainty representation itself.

This ambiguity raises a central yet unresolved question: how do second-order representation choices themselves influence practical uncertainty-aware behavior? Addressing this question requires isolating representational effects from predictive ones—a requirement rarely satisfied in existing studies. Prior empirical work predominantly focuses on within-paradigm comparisons [Mehrtens et al., 2023, Löhr et al., 2025] or evaluates representations using a single uncertainty metric or downstream task [Sale et al., 2024a, Chau et al., 2025a], limiting the ability to draw more general conclusions. As a result, the practical implications of choosing between distributional and set-based second-order representations remain poorly understood.

Novelty and contributions. To address this issue, we conduct a controlled comparative study in classification settings, providing a unified framework for evaluating different second-order uncertainty representations. The study, as summarized in Figure 1, is designed to ensure rigorous and like-for-like comparison by explicitly controlling for confounding factors.

(1) To eliminate effects arising from differences in model and learning assumptions, both representations are constructed from a finite collection of predictive distributions generated by the same neural network (either a Bayesian neural network or an ensemble). Within this setting, the distribution-based representation is defined as a uniform distribution over the predictive set, while the credal representation is derived from the identical predictions via class-wise probability interval construction [De Campos et al., 1994]. **(2)** To reduce dependence on any single uncertainty metric, we evaluate each representation using multiple uncertainty measures and perform both intra- and inter-representation comparisons. **(3)** Furthermore, to broaden the evaluation scenarios, experiments are conducted across multiple neural network architectures and widely used benchmarks, including selective prediction—where an EU-aware model abstains on samples with high EU estimates to reduce misclassification risk—and out-of-distribution (OOD) detection.

Three key findings emerge from our study. First, no representation exhibits uniform superiority independent of the associated uncertainty measure; conclusions depend critically on the interaction between representation and metric. Second, out-of-distribution detection more clearly exposes representational differences than selective prediction. Third, reliable uncertainty quantification depends jointly on representation and uncertainty measure, with different metrics yielding substantially different behavior even within the same representation. See Section 4 for how our results support these analyses.

Paper outline. The remainder of this paper is organized as follows. Section 2 introduces the preliminaries about uncertainty representation, quantification, and evaluation, respectively. Section 3 and Section 4 present the experimental setups and the comparative analysis in full detail, respectively. Section 5 concludes this work with discussions.

Further related work. Alternative second-order representations in classification include Dirichlet-based models [Malinin and Gales, 2018, Charpentier et al., 2020] and random sets [Manchingal et al., 2025b]. However, these frameworks do not provide a principled mechanism for construction from general EU-aware predictors or systematic translation across representations, making controlled, like-for-like comparisons difficult. We therefore restrict our study to representations that can be consistently derived from a shared set of predictive distributions. Prior efforts toward cross-representation evaluation remain limited. Manchingal et al. [2025a] converted Bayesian neural network and deep ensemble predictions into credal sets, focusing on the accuracyprecision trade-off for model selection. Mucsányi et al. [2024] examined uncertainty disentanglement but did not consider credal representations. Recent uncertainty measures [Löhr et al., 2025, Chau et al., 2026] have similarly concentrated on credal frameworks, leaving representation-level comparisons largely underexplored.

2 PRELIMINARIES

2.1 PROBLEM SETTINGS

In a supervised K-class classification problem, an NN with learnable parameters θ, denoted by $f_\theta(\cdot)$, is typically trained on i.i.d. samples $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N \subset \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y} = \{1, ..., k, ..., K\}$ is the output space. Given a test input $\boldsymbol{x} \in \mathcal{X}$, the NN outputs a softmax probability vector $\boldsymbol{p} := \big(p(y = 1|\boldsymbol{x}), ..., p(y=K|\boldsymbol{x})\big)$. However, this single conditional distribution captures only aleatoric predictive uncertainty, as it assumes precise knowledge of the underlying inputoutput mapping [Hüllermeier and Waegeman, 2021, Wang

et al., 2026b]. In this setting, representing EU generally requires a second-order formalism that expresses uncertainty about the models probabilistic prediction itself. The notable formalisms include distribution- and set-based representations, which are the main focus of this work.

2.2 UNCERTAINTY REPRESENTATIONS

Distribution-based representations. Bayesian neural networks (BNNs) and deep ensembles (DE) are well-known distribution-based approaches. A BNN [Blundell et al., 2015, Gal and Ghahramani, 2016, Krueger et al., 2017, Mobiny et al., 2021] learns a posterior distribution over parameters, $p(\theta|\mathcal{D})$, obtained by applying Bayes' rule:

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

where $p(\theta)$, $p(\mathcal{D})$, and $p(\mathcal{D}|\theta)$ denote the prior over parameters, the evidence, and the likelihood, respectively. Given a test input $\boldsymbol{x}$, a BNN theoretically marginalizes over this posterior to produce a prediction:

$$p\big(\boldsymbol{p}|\boldsymbol{x},\mathcal{D}\big) = \int_{\theta} p\big(\boldsymbol{p}|\boldsymbol{x},\theta\big)\, p(\theta|\mathcal{D})\, d\theta. \tag{2}$$

Since the network is deterministic given θ, each parameter sample yields exactly one predictive distribution, $\boldsymbol{p} = f_\theta(\boldsymbol{x})$; that is, $p(\boldsymbol{p}|\boldsymbol{x},\theta)$ places all its probability mass on this single $\boldsymbol{p}$. The posterior $p(\theta|\mathcal{D})$ therefore induces a probability density over the predictive distributions $\boldsymbol{p}$ themselves, rather than over the model output for a fixed θ. Eq. (2) can thus be interpreted as a *second-order distribution*—a probability distribution over probability distributions [Shaker and Hüllermeier, 2021, Wang et al., 2026b]—representing epistemic predictive uncertainty.

However, directly calculating (2) is computationally intractable. In practice, this is approximated using Bayesian model averaging (BMA) [Jospin et al., 2022], i.e., M stochastic forward passes through the BNN are performed:

$$\boldsymbol{p}_m = f_{\theta_m}(\boldsymbol{x}) \text{ for } m = 1, ..., M, \tag{3}$$

where $f_{\theta_m}(\cdot)$ denotes an NN with parameters θ_m sampled from the posterior $p(\theta|\mathcal{D})$ at the m-th pass. The predictive distribution is then approximated by averaging the probability vectors, $\tilde{\boldsymbol{p}} = M^{-1}\sum_{m=1}^{M}\boldsymbol{p}_m$, and the final class prediction is given by $\operatorname{argmax}(\tilde{\boldsymbol{p}})$. Under this approximation, this distribution-based representation can be viewed as *a uniform distribution over a finite set of predictive distributions*, denoted by $\mathcal{B} := \{\boldsymbol{p}_m\}_{m=1}^{M}$, where the predictive distributionsrather than the underlying BNNsare assigned equal weight.[1]

Unlike BNNs, which explicitly infer a distribution over model parameters, DE [Lakshminarayanan et al., 2017] marginalizes over multiple models, $\{f_{\theta_m}(\cdot)\}_{m=1}^{M}$ [Band et al., 2021]. At inference time, DE performs single forward passes across ensemble members to produce *a finite set of predictive distributions*, whose average is used to make the final class prediction. Thus, DE has been viewed as an approximation to BMA by some studies [Wilson and Izmailov, 2020, Abe et al., 2022]. Several variantssuch as batch ensembles [Wen et al., 2020], masked ensembles [Durasov et al., 2021], and packed ensembles [Laurent et al., 2022]have been proposed to predict $\{\boldsymbol{p}_m\}_{m=1}^{M}$ within a single model, achieving comparable uncertainty quantification performance with a lower computational cost.

Following common practice, although these predictions originate from different learning algorithms, we do not distinguish between them in the subsequent evaluation and analysis and denote them as $\mathcal{B} := \{\boldsymbol{p}_m\}_{m=1}^{M}$ in our study.

Set-based representations. Credal sets [Levi, 1980], denoted as convex sets of probability distributions, have been argued to provide a more natural EU representation than single probability distributions [Corani et al., 2012, Hüllermeier and Waegeman, 2021]. For example, sets can better capture ignorance as a lack of knowledge [Dubois et al., 2002], since a single distribution typically requires additional assumptions beyond merely distinguishing plausible from implausible candidates [Löhr et al., 2025].

To eliminate confounding effects arising from differences in model and learning assumptions—performance differences attributable to the base model rather than to the uncertainty representation itselfour comparative study (see Figure 1) focuses on distribution- and set-based representations derived from the same underlying neural network (NN). Although various NN approaches have been proposed for generating credal predictionsincluding methods based on finitely generated credal sets [Caprio et al., 2024, Chau et al., 2025b] and predicted probability intervals [Wang et al., 2024a, 2025b, 2026a]these approaches are typically algorithm-specific. We therefore adopt a common and computationally efficient strategy. Specifically, we transform a finite set of predictive distributions into a credal set via class-wise probability intervals [Wang et al., 2025a, Löhr et al., 2025].

Specifically, for the k-th class, the upper and lower bounds of the probability interval, denoted by p_{U_k} and p_{L_k}, respectively, are obtained from

$$p_{U_k} = \max_{m=1,..,M} p_{k,m} \quad \text{and} \quad p_{L_k} = \min_{m=1,..,M} p_{n,k}, \tag{4}$$

where $p_{n,k}$ denotes the k-th probability element of each $\boldsymbol{p}_n$.

[1]While BMA is theoretically weighted by the posterior over BNN parameters, equal weights are assigned to the finite sampled predictive distributions in practice for approximating, e.g., the final prediction and the mutual information. Here, we emphasize that the uniform distribution here is over these finite predictive samples, not over the BNNs themselves.

Thus, these probability intervals over classes define a credal set $\mathcal{K}$ as follows [De Campos et al., 1994]:

$$\mathcal{K} = \{\boldsymbol{p} \mid p_k \in [p_{L_k}, p_{U_k}]\ \forall k = 1, 2, ..., K\}. \quad (5)$$

where $\boldsymbol{p}$ denotes a probability vector, and each class probability is restricted to the given probability interval.

2.3 EPISTEMIC UNCERTAINTY MEASURES

Quantifying uncertainty requires an appropriate measure that maps a second-order prediction for a given input to a numerical value. We next introduce distinct EU measures for the two representations ($\mathcal{B}$ and $\mathcal{K}$), respectively.

Measures for a distribution-based representation. Using Shannon entropy as a classical measure of uncertainty in classification, the EU for a practical Bayesian representation, $\mathcal{B}=\{\boldsymbol{p}_m\}_{m=1}^{M}$, is computed as follows:

$$\sum_{k=1}^{K} -\tilde{p}_k \log_2 \tilde{p}_k - \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K} -p_{k,m}\log_2 p_{k,m}. \quad (6)$$

Here, $\tilde{p}_k$ and $p_{k,m}$ are the k-th element of the averaged probability vector $\tilde{\boldsymbol{p}}$ and the m-th probability vector $\boldsymbol{p}_m$, respectively. EU is computed using the standard decomposition of total predictive uncertainty into aleatoric and epistemic parts and can be interpreted as an approximate Mutual Information (MI) [Hüllermeier and Waegeman, 2021].

Alternative to the entropy-based measure in (6), a Label-Wise Variance (LWV) has recently been proposed [Sale et al., 2024b]. Given $\mathcal{B} = \{\boldsymbol{p}_m\}_{m=1}^{M}$, EU is quantified by

$$\sum_{k=1}^{K} \tilde{p}_k(1-\tilde{p}_k) - \sum_{k=1}^{K}\frac{1}{M}\sum_{m=1}^{M} p_{k,m}(1-p_{k,m}). \quad (7)$$

Here, the quantified EU is regarded as an approximation of the expected reduction in squared-error loss, analogous to mutual information, which quantifies the expected reduction in log-loss [Sale et al., 2024b].

In addition, a Wasserstein Distance (WD) measure, inspired by optimal transport theory, has been proposed to quantify EU as follows [Sale et al., 2024a]:

$$\frac{1}{2}\,\text{minimize}_{\boldsymbol{p}^*} \sum_{m=1}^{M} \parallel \boldsymbol{p}_m - \boldsymbol{p}^* \parallel_1, \quad (8)$$

where $\boldsymbol{p}^* \in \Delta^{K-1}$ is the decision probability vector on the full probability simplex Δ^{K-1}. Under this context, the quantified EU corresponds to the minimal Wasserstein distance between the approximated second-order prediction $\mathcal{B}$ and any possible distribution on the probability simplex of K classes. In the binary case, the optimization in (8) simplifies and admits the following closed-form solution:

$$\sum_{m=1}^{M} \mid p_m - \text{median}(p_1, ..., p_M) \mid. \quad (9)$$

Measures for a credal representation. To quantify EU of a credal set $\mathcal{K}$, a widely used measure is the Shannon entropy difference (H_{diff}) [Abellán et al., 2006], defined as the difference between the upper and lower entropy:

$$\text{maximize}_{\boldsymbol{p}\in\mathcal{K}}\, H(\boldsymbol{p}) - \text{minimize}_{\boldsymbol{p}\in\mathcal{K}}\, H(\boldsymbol{p}). \quad (10)$$

Here, $H(\boldsymbol{p})$ is the classical entropy of a single probability vector. Thus, solving the maximization problem in (10) amounts to finding the maximum entropy over $\mathcal{K}$, that is,

$$\begin{aligned} &\text{maximize} \sum\nolimits_{k=1}^{K} -p_k \log_2 p_k \quad \text{s.t.} \\ &p_k \in [p_{L_k}, p_{U_k}] \text{ for } k = 1, ..K \text{ and } \textstyle\sum_{k=1}^{K} p_k = 1. \end{aligned} \quad (11)$$

Similarly, solving the minimization problem in (10) requires replacing maximize by miximize. In the binary case, the credal set reduces to an interval $[p_L, p_U]$, and these optimization problems admit analytical closed-form solutions, which simplify the computation of H_{diff}, as follows:

$$\begin{aligned} &H(0.5) - \min\big(H(p_L), H(p_U)\big) \textit{ if } 0.5 \in [p_L, p_U] \\ &\max\big(H(p_L), H(p_U)\big) - \min\big(H(p_L), H(p_U)\big) \textit{ else} \end{aligned}. \quad (12)$$

An alternative measure to quantify EU of a credal set is the Generalized Hartley (GH) measure [Abellán and Moral, 2000], which corresponds to the expected Hartley measure [Hartley, 1928] taken over all subsets $\mathcal{Q}$ of the output space:

$$\sum_{\mathcal{Q}\subseteq\mathcal{Y}} \text{m}_{\mathcal{K}}(\mathcal{Q}) \log_2(|\mathcal{Q}|), \quad (13)$$

where $\text{m}_{\mathcal{K}}$ denotes the mass function induced by $\mathcal{K}$ and $|\mathcal{Q}|$ is the cardinality of $\mathcal{Q}$. The quantity $\text{m}_{\mathcal{K}}(\mathcal{Q})$ is computed from the Möbius inverse of the capacity function $\nu_{\mathcal{K}}$ [Hüllermeier and Waegeman, 2021]:

$$\text{m}_{\mathcal{K}}(\mathcal{Q}) = \sum_{\mathcal{A}\subseteq\mathcal{Q}} (-1)^{|\mathcal{Q}\setminus\mathcal{A}|} \nu_{\mathcal{K}}(\mathcal{A}), \quad (14)$$

with $\mathcal{Q}\setminus\mathcal{A} = \{k \mid k \in \mathcal{Q},\ k \notin \mathcal{A}\}$, and $\nu_{\mathcal{K}}(\mathcal{A})$ denoting the lower probability of $\mathcal{A} \subseteq \mathcal{Q}$. For a credal set $\mathcal{K}$ defined by probability intervals in (5), $\nu_{\mathcal{K}}(\mathcal{A})$ can be computed directly as follows [De Campos et al., 1994]:

$$\max\Big(\textstyle\sum_{j\in\mathcal{A}} p_{L_k}, 1 - \sum_{j\notin\mathcal{A}} p_{U_k}\Big). \quad (15)$$

The full GH calculation process [Wang et al., 2024a] is presented in Algorithm 1 in the Appendix.

More recently, an imprecise probability metric framework [Chau et al., 2025a] introduces the maximum mean imprecision (MMI) measure, employing the total variance distance to quantify credal epistemic uncertainty. In classification, the MMI is given as

$$\sup_{\mathcal{A}\subseteq\mathcal{Y}} \overline{P}(\mathcal{A}) - \underline{P}(\mathcal{A}), \quad (16)$$

where $\underline{P}(\mathcal{A})$ and $\overline{P}(\mathcal{A})$ denote the lower and upper probabilities of a subset $\mathcal{A}$, respectively. The lower probability $\underline{P}(\mathcal{A})$ can be computed from (15) for a credal set $\mathcal{K}$ in (5), while $\overline{P}(\mathcal{A})$ is the conjugate of $\overline{P}(\mathcal{A})$, defined as follows:

$$\overline{P}(\mathcal{A}) = 1 - \underline{P}(\mathcal{A}^c), \tag{17}$$

where $\mathcal{A}^c$ is the complement of $\mathcal{A}$ on the output space $\mathcal{Y}$. For the binary case, the MMI in (16) reduces to the interval length $p_U - p_L$, which coincides with the GH measure.

2.4 DOWNSTREAM EVALUATION TASKS

Since ground-truth epistemic uncertainty (EU) is unavailable, the quality of EU quantification is generally assessed through practical downstream tasks. Following common practice, our comparative study considers two widely used benchmarks: *selective prediction* [Hüllermeier et al., 2022, Chau et al., 2025a] and *out-of-distribution (OOD) detection* [Wang et al., 2024a, Löhr et al., 2025].

Selective prediction. The rationale for using selective prediction to assess EU quantification quality is that an EU-aware NN is expected to assign higher EU values to misclassified samples than to correctly classified ones. In practical batch processing, instances with high EU estimates are abstained from and referred to an expert to reduce the risk of misclassification. See Algorithm 2 for details.

Under this setting, an accuracyrejection curve (ARC) is used to characterize the relationship between prediction accuracy on retained samples and the rejection rate [Hühn and Hüllermeier, 2008, Hüllermeier et al., 2022]. Reliable uncertainty estimates produce a monotonically increasing ARC, whereas random rejection results in a flat curve [Hüllermeier et al., 2022]. In addition, the area under the accuracyrejection curve (AUARC) provides a scalar summary metric [Jaeger et al., 2023], where larger AUARC values indicate stronger selective prediction performance.

OOD detection. As a practical benchmark for evaluating EU quantification quality, stronger OOD detection performance suggests that the estimated uncertainty is more informative [Löhr et al., 2025, Wang et al., 2026c]. The intuition is that accurate EU estimation helps avoid misclassifying ambiguous in-distribution (ID) samples as OOD instances. Such ambiguity does not stem from regions of higher EU within the ID distribution, so a valid EU estimate should distinguish these cases [Mukhoti et al., 2023].

In this setting, as summarized in Algorithm 3, OOD detection is formulated as a binary classification problem where ID and OOD samples are assigned labels 0 and 1, respectively. The models EU estimate is used as the prediction score, and performance is evaluated using the area under the receiver operating characteristic curve (AUROC). A higher AUROC value indicates a better performance.

3 EXPERIMENTS[2]

Predictive models. As illustrated in Figure 1, our comparative study constructs a *finite set of predictive distributions* from a *common second-order predictor*. The goal is not to contrast distinct inference paradigms, but to ensure robustness and fairness of the analysis across representative implementations. We consider the following model families: **i)** Stochastic variational inference (SVI) [Blundell et al., 2015, Graves, 2011], a classical BNN method that approximates the parameter posterior with a Gaussian distribution. **ii)** Monte Carlo Dropout (MCDO) [Gal and Ghahramani, 2016], which estimates the posterior via stochastic forward passes with dropout enabled. **iii)** Deep ensembles (DE) [Lakshminarayanan et al., 2017], which approximate Bayesian inference by marginalizing predictions from independently trained models. In addition, we include three recent, computationally efficient DE variants: **iv)** Batch ensembles (BatchEns) [Wen et al., 2020], which factorize each weight matrix into a shared component and a rank-one, member-specific term; **v)** Masked ensembles (MaskEns) [Durasov et al., 2021], which use fixed binary masks to control correlations between ensemble members; and **vi)** Packed ensembles (PackEns) [Laurent et al., 2022], which exploit grouped convolutions to parallelize ensemble members within a shared backbone.

Specifically, for each experiment setting, DE is constructed by training 5 individual neural networks ($M = 5$) with different random seeds, following standard practice. BatchEns, MaskEns, and PackEns use the default ensemble size of $M = 4$. For SVI and MCDO, we perform inference with 10 forward passes ($M = 10$). All underlying models are trained independently for 10 runs per setting.

Datasets and benchmarks. We evaluate our approach on CIFAR10 [Krizhevsky et al., 2009] and two real-world settings: **i)** a medical diagnosis task using whole-slide images from the Camelyon17 dataset [Bandi et al., 2018]; **ii)** a ship classification task reflecting realistic visual variability. The downstream tasks include selective prediction and OOD detection, see Section 2.4.

For **CIFAR10**, all models are trained on the standard training split, with selective prediction evaluated on the test split. For out-of-distribution (OOD) detection, CIFAR10 is treated as the in-distribution (ID) dataset, while SVHN and FMNIST serve as OOD benchmarks. For **medical image classification**, Camelyon17 comprises histopathology images collected from five medical centers in the Netherlands and scanned using three devices, naturally inducing realistic distribution shifts. We treat centers 0, 1, and 3 (3DHistech scanners) as ID data, and centers 2 and 4 (Philips and Hamamatsu scanners) as the distribution-shift set. Dataset

[2]Code is at: `https://github.com/Kaizheng-WANG/set-vs-distribution-epistemic-representation`.

Table 1: Average AUARC on the selective prediction task, computed per dataset and averaged across six predictive model architectures. We additionally report net wins, obtained from pairwise comparisons both within the same uncertainty-representation category (intra-representation) and across different categories (inter-representation). For both AUARC and net wins, higher values indicate better performance. In each column, the best-performing method is highlighted in **bold red** and the second-best in **bold blue**. Detailed AUARC score as well as the intra- and inter-representation comparisons for each underlying predictive model are provided in Tables A.3, A.4, and A.5, respectively.

	In-distribution Camelyon17			Distribution-shift Camelyon17			In-distribution SeaShip			In-distribution CIFAR10		
	Average scores	Intra	Inter	Average scores	Intra	Inter	Average scores	Intra	Inter	Average scores	Intra	Inter
MI	0.9602±0.0568	-11	-18	0.9637±0.0176	-12	-19	0.9887±0.0156	-12	-18	0.9807±0.0050	-12	-18
LWV	0.9625±0.0532	**0**	-5	0.9669±0.0131	**0**	-6	0.9899±0.0151	**0**	-4	0.9815±0.0049	**0**	-6
WD	0.9633±0.0530	**11**	**21**	0.9683±0.0125	**12**	**30**	0.9904±0.0144	**12**	**25**	0.9823±0.0046	**12**	**28**
GH	0.9633±0.0523	**6**	**15**	0.9680±0.0126	**10**	**12**	0.9903±0.0144	**12**	**19**	0.9822±0.0046	**12**	**20**
H_{diff}	0.9594±0.0517	-12	-28	0.9613±0.0164	-12	-29	0.9848±0.0237	-12	-30	0.9764±0.0068	-12	-30
MMI	0.9633±0.0523	**6**	**15**	0.9680±0.0126	**2**	**12**	0.9901±0.0149	**0**	8	0.9819±0.0048	**0**	6

statistics are reported in Table A.1. Selective prediction is evaluated on both the ID and distribution-shift test splits. For **ship classification**, models are trained on the SeaShip training set and evaluated on the corresponding test split. For OOD detection, we use SeaShip-C (corruption-based shifts) and SeaShip-O, which contains ship images sourced from external datasets (SMD and SSAVE). These classification datasets are derived from ship detection benchmarks; preprocessing details are provided in Appendix A.3.

Across all datasets, predictive models are trained under a shared protocol. Detailed configurations and experimental settings are deferred to Appendix A.1.

4 COMPARATIVE ANALYSIS

4.1 EVALUATION CRITERIA AND RESULTS

Evaluation criteria. For each downstream task and dataset, we evaluate all second-order predictive models (DE, SVI, MCDO, BatchEns, MaskEns, and PackEns). For distribution-based representations, uncertainty is quantified using Mutual Information (MI) in (6), Label-wise Variance (LWV) in (8), and Wasserstein Distance (WD) in (9) for uncertainty quantification. For credal representations, we instead consider the entropy difference (H_{diff}) in (10), the Generalized Hartley (GH) measure in (13), and the Maximum Mean Imprecision (MMI) in (16). Additional details are provided in Section 2.3.

Beyond reporting average performance scores (AUARC for selective prediction and AUROC for OOD detection, respectively), we perform pairwise one-sided Wilcoxon signed-rank tests at the 5% significance level. For each ordered pair of uncertainty measures (m_i, m_j) with $i \neq j$, the null hypothesis (H_0) assumes no systematic performance difference between m_i and m_j, while the alternative hypothesis (H_1) asserts that m_i yields stochastically larger performance scores than m_j. The tests are conducted over 10 independent runs. Results are considered statistically significant when $p < 0.05$, in which case m_i is deemed to outperform m_j.

Based on the statistical tests, we construct a ranking scheme to provide an interpretable quantitative summary. For each predictive model and uncertainty measure, we record the number of significant wins and losses across pairwise comparisons. Each significant win contributes $+1$, each significant loss contributes -1, and non-significant outcomes contribute 0. The resulting net score, defined as $\text{wins}(m) - \text{losses}(m)$, captures the relative dominance of a measure. Global rankings, including both intra- and inter-representation comparisons, are obtained by aggregating net scores across predictive models.

Results. For the selective prediction task, Table 1 reports the average AUARC across six underlying predictive models and 10 runs. For the OOD task, Table 2 reports the average AUROC under the same setting. Both tables also summarize the net wins across the six underlying predictive models for intra- and inter-representation comparisons, based on one-sided paired Wilcoxon tests at the 5% significance level (partially shown in Figure 2 and fully presented in Figures A.4 and A.3).

The accuracy-rejection curves for selective prediction are presented in Figures A.5, A.6, A.7, and A.8, while Receiver Operating Characteristic (ROC) curves for OOD detection are presented in Figures A.9, A.10, A.12, and A.11.

4.2 SUMMARY AND ANALYSIS

(1) The relative merits of an uncertainty representation cannot be assessed independently of the associated uncertainty measures.

Table 2: Average AUROC on the OOD detection task, computed per dataset pair and averaged across six predictive model architectures. We additionally report net wins, obtained from pairwise comparisons both within the same uncertainty-representation category (intra-representation) and across different categories (inter-representation). For both AUROC and net wins, higher values indicate better performance. In each column, the best-performing method is highlighted in **bold red** and the second-best in **bold blue**. Detailed AUROC score as well as the intra- and inter-representation net-win comparisons for each underlying predictive model are provided in Tables A.3, A.6, and A.7, respectively.

	SeaShip v.s. SeaShip-O			SeaShip v.s. SeaShip-C			CIFAR10 v.s. SVHN			CIFAR10 v.s. FMNIST		
	Average scores	Intra	Inter	Average scores	Intra	Inter	Average scores	Intra	Inter	Average scores	Intra	Inter
MI	0.8694±0.0840	**-3**	-6	0.8250±0.0859	**-5**	-11	0.7657±0.0301	-8	-25	0.8639±0.0276	**0**	-11
LWV	0.8655±0.0797	-9	-15	0.8300±0.0776	-7	-13	0.7694±0.0295	**-4**	-21	0.8466±0.0256	-12	-30
WD	0.8822±0.0739	**12**	**19**	0.8512±0.0720	**12**	**18**	0.8174±0.0227	**12**	**18**	0.8731±0.0232	**12**	10
GH	0.8856±0.0720	**12**	**29**	0.8561±0.0703	**12**	**30**	0.8303±0.0197	**12**	**30**	0.8801±0.0226	**9**	**27**
H_{diff}	0.8516±0.0893	-12	-30	0.7929±0.0912	-12	-30	0.7797±0.0343	-12	-8	0.8744±0.0266	**3**	**17**
MMI	0.8768±0.0751	**0**	3	0.8460±0.0729	**0**	6	0.8038±0.0235	**0**	6	0.8630±0.0237	-12	-13

Table 1 and Table 2 show that neither representation consistently dominates across benchmarks. Performance varies systematically with the uncertainty measure. For instance, the credal representation paired with the GH measure attains the strongest results on OOD detection, whereas the distribution-based representation combined with the WD measure ranks highest on selective prediction under the statistical tests.

While our inclusion of multiple second-order predictive models and datasets is intended to improve the robustness of the analysis rather than compare underlying models, Figure 2 shows that uncertainty-aware performance remains sensitive to predictive model choices and dataset. Even when fixing the representation, uncertainty measure, downstream task, and dataset, results vary across different predictive models. Similarly, holding the representation, uncertainty measure, predictive model, and task constant while changing datasets yields different outcomes under the one-sided paired Wilcoxon tests.

Taken together, these findings highlight that empirical comparisons of uncertainty representations must be interpreted conditionally. Claims of effectiveness should therefore be qualified with explicit reference to the uncertainty measure, benchmark, predictive model, and dataset.

(2) OOD detection more readily reveals differences between the two uncertainty representations than selective prediction.

Although one-sided paired Wilcoxon tests indicate measurable differences on selective prediction benchmarks, the gaps in average AUARC remain small (Table 1). This behavior follows naturally from the evaluation protocol. As discussed in Section 2.4, selective prediction combines instance rejection with accuracy on the retained samples. Distinct uncertainty estimates can therefore yield similar rejection sets, leading to nearly identical performance. Even when rejection patterns differ, predictive accuracy often changes only marginally because all measures operate on the same underlying model. Moreover, baseline accuracy without rejection is already high (Figure A.7), which further compresses observable gains.

By contrast, OOD detection constitutes a more sensitive regime for differentiating uncertainty representations, as detection performance directly depends on how uncertainty responds to distributional shifts, despite known limitations of EU-based OOD methods [Li et al., 2025].

These observations suggest that validating new uncertainty representations or measures across multiple downstream tasks is important for establishing robust empirical claims.

(3) Reliable uncertainty quantification depends jointly on the representation and the uncertainty measure.

For distribution-based representations, Table 1 and Table 2 show that WD consistently outperforms MI and LWV across both selective prediction and OOD detection benchmarks. A plausible explanation is that WD more directly captures epistemic predictive uncertainty by quantifying the geometric dispersion of the second-order distribution around its barycenter. This captures variability induced by individual predictive distributions. In contrast, MI relies on entropy-based uncertainty decomposition and is influenced by the global shape of the predictive distribution, while LWV measures label-wise variability without accounting for the geometry of the probability simplex. Moreover, prior analysis suggests that Sale et al. [2024a] WD better satisfies key theoretical desiderata than MI and LWV.

For credal representations, GH consistently achieves the strongest performance. This likely stems from its more expressive characterization of set-valued uncertainty, as GH evaluates the structure of the credal set rather than a single scalar summary. By comparison, the entropy difference H_{diff}, despite its popularity, performs poorly in most set-

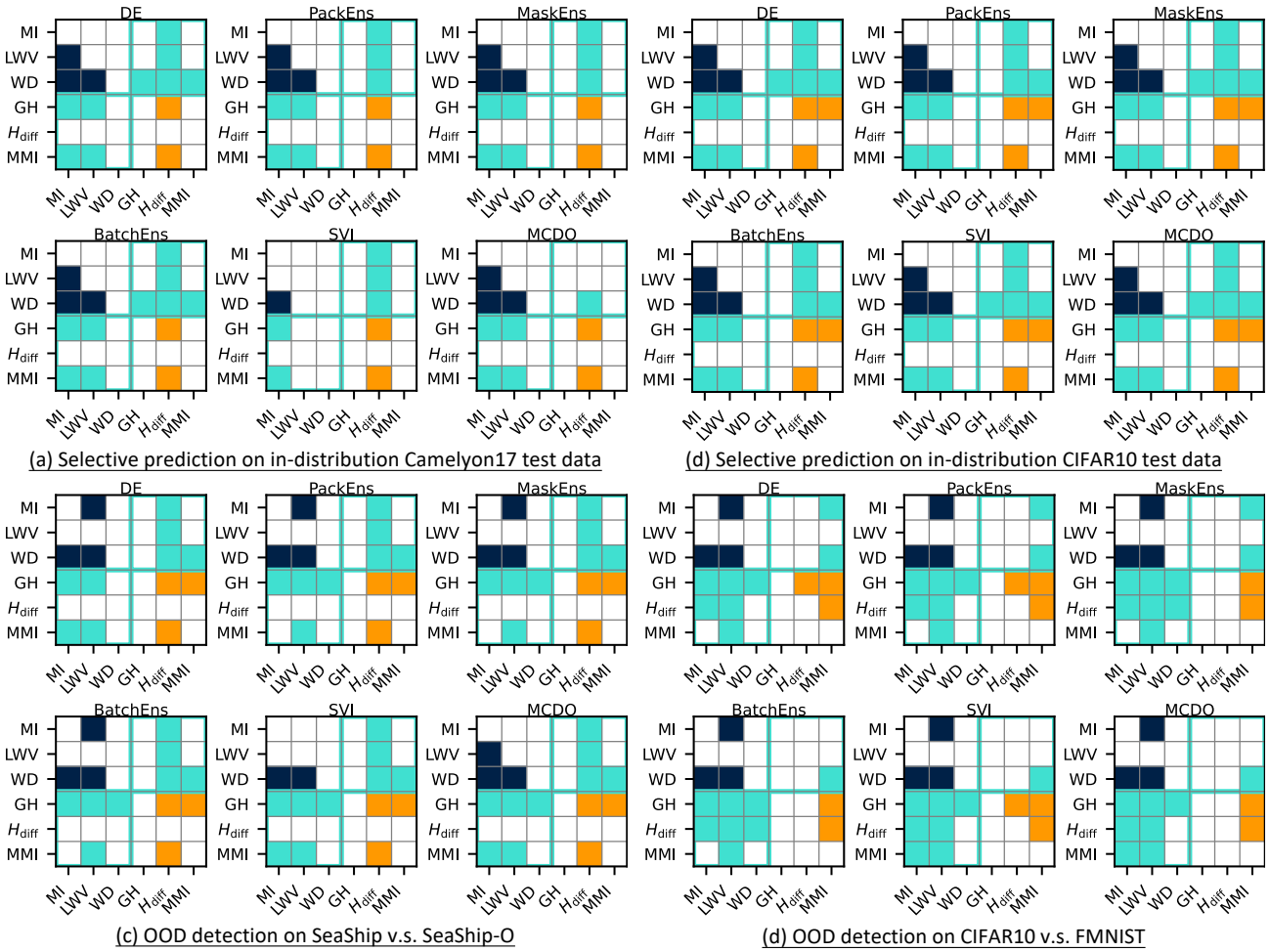

Figure 2: Statistical significance plots on different selective prediction (a, b) and OOD detection (c, d) benchmarks across different underlying predictive models. A cell is shaded if the measure in the i-th row is statistically significantly better than that in the j-th column according to a pairwise one-sided Wilcoxon signed-rank test at the 5% significance level. Intra-representation comparisons are shown in blue (distribution-based measures) and orange (credal-based measures), while inter-representation comparisons are shown in green.

tings. One limitation is that it depends solely on the width of the Shannon entropy interval. For example, in a 2D probability simplex, any credal set that includes the center and one vertex results in the same EU, regardless of the shape of the set itself. This behavior is not attributable to numerical optimization artifacts, as closed-form solutions exist in the binary case. We note, however, that GH becomes computationally impractical for a large label space in K classification, requiring calculations over 2^K subsets. For theoretical discussions on GH and H_{diff}, see Hüllermeier et al. [2022]. This also leads to the recent development of MMI and its linear time upper bound, as an alternative approach to GH [Chau et al., 2025a].

Overall, these findings emphasize that improving uncertainty quantification requires not only principled representations but also carefully designed and computationally efficient uncertainty measures.

Further Experiments. To further examine generalizability, we conduct reciprocal experiments in which SVHN and FMNIST are each treated as the ID dataset in turn, with the remaining dataset and CIFAR10 serving as OOD counterparts. The same six predictive models are trained on FMNIST and SVHN, then evaluated on both selective prediction and OOD detection under the same protocol. Detailed results are reported in Appendix C; they consistently support and further strengthen the generalizability of our three main findings discussed in Section 4.2.

5 CONCLUSION

This paper present a systematic comparison of distribution-based and credal uncertainty representations for classification. To isolate representational effects, both formalisms were derived from an identical finite set of predictive dis-

tributions produced by shared predictive models. The study covers 6 uncertainty measures, 14 selective prediction and OOD detection benchmarks, 6 predictive model families, and 10 independent runs per configuration.

Three main observations emerge from the empirical analysis. First, neither representation demonstrates uniform superiority; performance depends critically on the interaction between representation and uncertainty measure. Second, OOD detection more readily reveals representational differences than selective prediction, highlighting task-dependent sensitivity of uncertainty evaluation. Third, reliable uncertainty quantification depends jointly on representation and metric choice, as different measures induce markedly different behaviors even under fixed predictive models.

These results underscore the importance of conditional interpretation in empirical studies of uncertainty. Claims regarding uncertainty representations or EU-aware predictors should therefore be grounded in clearly specified evaluation settings, including the uncertainty measure, benchmark, and dataset. Robust validation further benefits from assessing multiple downstream tasks. Finally, our findings highlight that progress in uncertainty-aware learning is driven not only by representational advances but also by the design of theoretically grounded and computationally efficient uncertainty measures.

Limitations. Alternative second-order representations for classification, such as Dirichlet-based models and random sets, were not included in this study. While these frameworks are conceptually related, they lack a general, representation-agnostic construction from arbitrary EU-aware predictors and do not admit straightforward translation across formalisms. This complicates controlled, like-for-like comparisons under a shared experimental protocol. Consequently, we restrict attention to representations that can be consistently derived from a common finite set of predictive distributions. For similar reasons, we focus on probability-interval credal sets induced by finite predictive samples and do not consider alternative credal constructions explored in other credal classification frameworks.

Computational considerations further constrain the scale of the empirical analysis. Extending the evaluation to a wider range of uncertainty measures, datasets, second-order predictors, and benchmark settings (e.g., active learning) remains an important direction for future work.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback. This work was supported by the Start-up Grant from Nanyang Technological University, Singapore, and by the European Union's Horizon 2020 research and innovation programme under grant agreement No 964505 (E-pi) and under the Marie Sklodowska-Curie grant agreement No 955768 (AUTOBarge). This research was also partially supported by Flanders Make, the strategic research center for the manufacturing industry.

References

Taiga Abe, Estefany Kelly Buchanan, Geoff Pleiss, Richard Zemel, and John P Cunningham. Deep ensembles work, but are they necessary? *Advances in Neural Information Processing Systems*, 35:33646–33660, 2022.

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

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

Neil Band, Tim GJ Rudner, Qixuan Feng, Angelos Filos, Zachary Nado, Michael W Dusenberry, Ghassen Jerfel, Dustin Tran, and Yarin Gal. Benchmarking Bayesian deep learning on diabetic retinopathy detection tasks. In *Proceedings of Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.

Michele Caprio, Souradeep Dutta, Kuk Jin Jang, Vivian Lin, Radoslav Ivanov, Oleg Sokolsky, and Insup Lee. Credal Bayesian deep learning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

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

Siu Lun Chau, Michele Caprio, and Krikamol Muandet. Integral imprecise probability metrics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Siu Lun Chau, Antonin Schrab, Arthur Gretton, Dino Sejdinovic, and Krikamol Muandet. Credal two-sample tests of epistemic uncertainty. In *International Conference on Artificial Intelligence and Statistics*, pages 127–135. PMLR, 2025b.

Siu Lun Chau, Soroush H Zargarbashi, Yusuf Sale, and Michele Caprio. Quantifying epistemic predictive uncertainty in conformal prediction. *arXiv preprint arXiv:2602.01667*, 2026.

Giorgio Corani, Alessandro Antonucci, and Marco Zaffalon. Bayesian networks with imprecise probabilities: Theory and application to classification. *Data Mining: Foundations and Intelligent Paradigms: Volume 1: Clustering, Association and Classification*, pages 49–93, 2012.

Luis M. De Campos, Juan F. Huete, and Serafin Moral. Probability intervals: A tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 02(02):167–196, June 1994. ISSN 0218-4885, 1793-6411. doi: 10.1142/S0218488594000146.

Didier Dubois, Henri Prade, and Philippe Smets. Representing partial ignorance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 26 (3):361–377, 2002.

Nikita Durasov, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Masksembles for uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13539–13548, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.

Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

Ralph VL Hartley. Transmission of information 1. *Bell System technical journal*, 7(3):535–563, 1928.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Jens Christian Hühn and Eyke Hüllermeier. FR3: A fuzzy rule learner for inducing reliable classifiers. *IEEE Transactions on Fuzzy Systems*, 17(1):138–149, 2008.

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

Eyke Hüllermeier, Sébastien Destercke, and Mohammad Hossein Shaker. Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In *Uncertainty in Artificial Intelligence*, pages 548–557. PMLR, 2022.

Paul F Jaeger, Carsten Tim Lüth, Lukas Klein, and Till J. Bungert. A call to reflect on evaluation practices for failure detection in image classification. In *The Eleventh International Conference on Learning Representations*, 2023.

Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on Bayesian neural networksa tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17 (2):29–48, 2022.

Mahendra Khened, Avinash Kori, Haran Rajkumar, Ganapathy Krishnamurthi, and Balaji Srinivasan. A generalized deep learning framework for whole-slide image segmentation and analysis. *Scientific reports*, 11(1):11579, 2021.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. CIFAR-10 (Canadian Institute For Advanced Research). Technical report, CIFAR, 2009.

David Krueger, Chin-Wei Huang, Riashat Islam, Ryan Turner, Alexandre Lacoste, and Aaron Courville. Bayesian hypernetworks. *ArXiv Preprint ArXiv:1710.04759*, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

Olivier Laurent, Adrien Lafage, Enzo Tartaglione, Geoffrey Daniel, Jean-marc Martinez, Andrei Bursuc, and Gianni Franchi. Packed ensembles for efficient uncertainty estimation. In *The Eleventh International Conference on Learning Representations*, 2022.

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

Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim GJ Rudner, C Bayan Bruss, and Andrew Gordon Wilson. Position: Supervised classifiers answer the wrong questions for ood detection. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Timo Löhr, Paul Hofman, Felix Mohr, and Eyke Hüllermeier. Credal prediction based on relative likelihood. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

Shireen Kudukkil Manchingal, Muhammad Mubashar, Kaizheng Wang, and Fabio Cuzzolin. A unified evaluation framework for epistemic predictions. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025a.

Shireen Kudukkil Manchingal, Muhammad Mubashar, Kaizheng Wang, Keivan Shariatmadar, and Fabio Cuzzolin. Random-set neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Hendrik A Mehrtens, Alexander Kurz, Tabea-Clara Bucher, and Titus J Brinker. Benchmarking common uncertainty estimation methods with histopathological images under domain shift and label noise. *Medical image analysis*, 89:102914, 2023.

Aryan Mobiny, Pengyu Yuan, Supratik K Moulik, Naveen Garg, Carol C Wu, and Hien Van Nguyen. Dropconnect is effective in modeling uncertainty of Bayesian deep networks. *Scientific Reports*, 11(1):1–14, 2021.

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24384–24394, 2023.

Yusuf Sale, Viktor Bengs, Michele Caprio, and Eyke Hüllermeier. Second-order uncertainty quantification: A distance-based approach. In *Forty-first International Conference on Machine Learning*, 2024a.

Yusuf Sale, Paul Hofman, Timo Löhr, Lisa Wimmer, Thomas Nagler, and Eyke Hüllermeier. Label-wise aleatoric and epistemic uncertainty quantification. In *The Fortieth Conference on Uncertainty in Artificial Intelligence*, 2024b.

Mohammad Hossein Shaker and Eyke Hüllermeier. Ensemble-based uncertainty quantification: Bayesian versus credal inference. In *Proceedings of the Tirty-first Workshop Computational Intelligence*, volume 25, page 63, 2021.

Rui Tuo and Wenjia Wang. Uncertainty quantification for Bayesian optimization. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 2862–2884. PMLR, 28–30 Mar 2022.

Kaizheng Wang, Fabio Cuzzolin, Shireen Kudukkil Manchingal, Keivan Shariatmadar, David Moens, and Hans Hallez. Credal deep ensembles for uncertainty quantification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a.

Kaizheng Wang, Fabio Cuzzolin, Keivan Shariatmadar, David Moens, and Hans Hallez. Credal wrapper of model averaging for uncertainty estimation in classification. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Kaizheng Wang, Keivan Shariatmadar, Shireen Kudukkil Manchingal, Fabio Cuzzolin, David Moens, and Hans Hallez. CreINNs: Credal-set interval neural networks for uncertainty estimation in classification tasks. *Neural Networks*, 185:107198, 2025b. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2025.107198.

Kaizheng Wang, Fabio Cuzzolin, David Moens, and Hans Hallez. Credal ensemble distillation for uncertainty quantification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(31):26319–26327, 2026a.

Kaizheng Wang, Fabio Cuzzolin, Keivan Shariatmadar, David Moens, and Hans Hallez. A review of uncertainty representation and quantification in neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2476–2495, 2026b. doi: 10.1109/TPAMI.2025.3626645.

Kaizheng Wang, Ghifari Adam Faza, Fabio Cuzzolin, Siu Lun Chau, David Moens, and Hans Hallez. Learning credal ensembles via distributionally robust optimization. In *Forty-third International Conference on Machine Learning*, 2026c.

Yunjia Wang, Kaizheng Wang, Zihao Zhang, Jeroen Boydens, Davy Pissoort, and Mathias Verbeke. Navigating the waters of object detection: Evaluating the robustness of real-time object detection models for autonomous surface vehicles. In *2024 IEEE conference on artificial intelligence (CAI)*, pages 985–992. IEEE, 2024b.

Yunjia Wang, Zihao Zhang, Kaizheng Wang, Holger Caesar, Jeroen Boydens, Davy Pissoort, and Mathias Verbeke. Enhancing the dependability of autonomous surface vehicles through robustness benchmarking of real-time object detection models. *Expert Systems with Applications*, page 129151, 2025c.

Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*, 2020.

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.

Dengyong Zhou, Sumit Basu, Yi Mao, and John Platt. Learning from the wisdom of crowds by minimax entropy. *Advances in Neural Information Processing Systems*, 25, 2012.

Set-based v.s. Distribution-based Representations of Epistemic Uncertainty: A Comparative Study (Supplementary Material)

Kaizheng Wang[‡1,2] **Yunjia Wang**[2,5] **Fabio Cuzzolin**[3] **David Moens**[4,5] **Hans Hallez**[2] **Siu Lun Chau**[1]

[1]College of Computing and Data Science, Nanyang Technological University, Singapore
[2]Department of Computer Science, KU Leuven, Belgium
[3]School of Engineering, Computing, and Mathematics, Oxford Brookes University, U.K.
[4]Department of Mechanical Engineering, KU Leuven, Belgium
[5]Flanders Make@KU Leuven, Belgium

A IMPLEMENTATION DETAILS

A.1 EXPERIMENTAL CONFIGURATIONS

Configurations of predictive models. In our comparative study, we consider several second-order predictive models: **i)** Stochastic variational inference (SVI) [Blundell et al., 2015, Graves, 2011], a classical BNN approach that approximates the parameter posterior with a Gaussian distribution. **ii)** Monte Carlo Dropout (MCDO) [Gal and Ghahramani, 2016], which estimates the posterior through stochastic forward passes with dropout enabled at inference time. **iii)** Deep ensembles (DE) [Lakshminarayanan et al., 2017], which approximate Bayesian inference by marginalizing predictions from independently trained models.

In addition, we include three recent computationally efficient variants of DE: **iv)** Batch ensembles (BatchEns) [Wen et al., 2020], which factorize each weight matrix into a shared component and a rank-one, member-specific term; **v)** Masked ensembles (MaskEns) [Durasov et al., 2021], which use fixed binary masks to control correlations among ensemble members; and **vi)** Packed ensembles (PackEns) [Laurent et al., 2022], which leverage grouped convolutions to parallelize ensemble members within a shared backbone.

All predictors use a ResNet architecture [He et al., 2016] as the backbone (ResNet-34 for the SeaShip and Camelyon17 datasets, and ResNet-18 for CIFAR10), with an input size of $(3, 224, 224)$. For DE, we train $M = 5$ independent neural networks with different random seeds. The implementations of the ensemble members, SVI, and MCDO follow the guidelines of the repository `https://github.com/DBO-DKFZ/uncertainty-benchmark`. BatchEns, MaskEns, and PackEns are implemented according to the repository `https://torch-uncertainty.github.io/`, using the default ensemble size $M = 4$. For SVI and MCDO, inference is performed using $M = 10$ stochastic forward passes.

Training configurations. The training batch size is set to 128. We use the Adam optimizer with an initial learning rate of 0.001, which is reduced by a factor of 10 if the validation cross-entropy loss does not improve for three consecutive epochs. Data preprocessing, augmentation, and other training procedures follow the protocol described in Mehrtens et al. [2023]. Specifically, for the Camelyon17 and SeaShip datasets, we apply the strong augmentation mode, while for the relatively simpler CIFAR10 dataset, we use the crop mode.

Each predictive model is trained for up to 30, 60, and 100 epochs on Camelyon17, SeaShip, and CIFAR10, respectively. For evaluation, we select the checkpoint with the best balanced validation accuracy. All models are trained on a single NVIDIA P100 SXM2@1.3, GHz GPU. An exception is BatchEns on Camelyon17, which requires an NVIDIA V100 SXM2@1.5,GHz GPU due to higher memory consumption.

To support reproducibility, the core implementation code for running and analyzing the experiments will be released publicly upon publication under a license that allows free use for research.

[*]This work was initiated at KU Leuven and primarily completed at Nanyang Technological University. Corresponding author: Kaizheng Wang (`kaizheng.wang@ntu.edu.sg`).

Additional information for experiments on Camelyon17 dataset. In our evaluation on medical classification, we focus on a challenging real-world medical diagnosis task involving whole-slide images (WSIs). The main difficulties arise from **i)** the enormous size of WSIs combined with the limited availability of annotated data, and **ii)** distribution shifts due to differences in image acquisition across institutions and scanners, where deployment data often deviates from the training distribution [Mehrtens et al., 2023].

For our experiments, we use the Camelyon17 dataset [Bandi et al., 2018], which consists of 50 breast lymph node WSIs with metastatic tissue, collected from five medical centers in the Netherlands and scanned on three different devices. An example WSI is shown in Figure A.1. To simulate a strong domain shift [Mehrtens et al., 2023], we partition the dataset

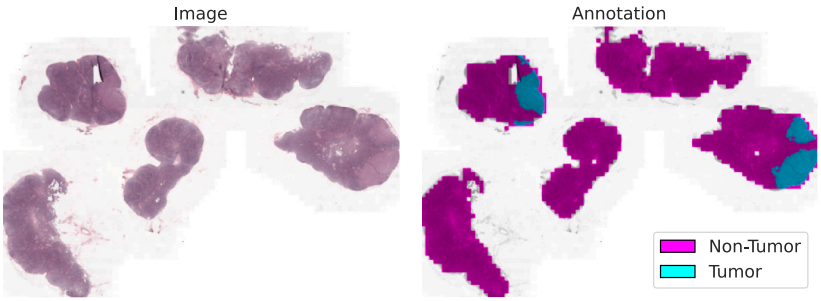

Figure A.1: An example of the whole slide image (referring to node 2 of patient 017) with ground-truth annotations.

as follows: centers 0, 1, and 3scanned with 3DHistech devicesare grouped as in-distribution (ID) data, while centers 2 and 4scanned with Philips and Hamamatsu devicesserve as domain-shifted data. Following the protocol described in [Mehrtens et al., 2023, Khened et al., 2021], lesion-level tile instances are extracted from the WSIs, which are then used for network training, validation, and testing. The resulting dataset statistics are summarized in Table A.1.

Table A.1: Number of tile instances in each dataset split.

ID dataset			Domain-shifted test dataset	
Training	Validation	Testing	Center 2	Center 4
383406	110561	109060	89351	166607

A.2 ALGORITHMIC IMPLEMENTATIONS

Algorithm 1 GH Calculation Procedure

Input: $[\boldsymbol{p}_L, \boldsymbol{p}_U] := \{[p_{L_k}, p_{U_k}]\}_{k=1}^{K}$; Target space $\mathcal{Y}$
Output: $\mathrm{GH}(\mathcal{K})$
Initialize: $\mathrm{GH}(\mathcal{K}) = 0$
for all $\mathcal{A} \subseteq \mathcal{Y}$ **and** $|\mathcal{A}| \geq 2$ **do**
  Initialize: $\mathrm{m}_{\mathcal{K}}(\mathcal{A}) = 0$
  for all $\mathcal{Q} \subseteq \mathcal{A}$ **do**
    Compute $\nu_{\mathcal{K}}(\mathcal{Q})$ using (15)
    $\mathrm{m}_{\mathcal{K}}(\mathcal{A}) = \mathrm{m}_{\mathcal{K}}(\mathcal{A}) + (-1)^{|\mathcal{A}\setminus\mathcal{Q}|} \cdot \nu_{\mathcal{K}}(\mathcal{Q})$ using (14)
  end for
  $\mathrm{GH}(\mathcal{K}) = \mathrm{GH}(\mathcal{K}) + \mathrm{m}_{\mathcal{K}}(\mathcal{A}) \cdot \log_2(|\mathcal{A}|)$ using (13)
end for

Algorithm 2 Selective prediction procedure

Input: Test dataset $\mathcal{D}_{\text{test}}$; rejection rate β; second-order predictor $f(\cdot)$
1. Obtain second-order representations from $f(\cdot)$ on $\forall \boldsymbol{x}_n \in \mathcal{D}_{\text{test}}$
$\mathcal{B}_n \leftarrow f(\boldsymbol{x}_n)$ (practical Bayesian) or $\mathcal{K}_n \leftarrow \mathcal{B}_n$ using (4) (credal)
2. Quantify epistemic predictive uncertainty given each sample, denoted by $u_{\text{EU},n}$
$u_{\text{EU},n} \forall n$ using (6)/(7)/(8) for $\mathcal{B}_n$ or (10)/(13)/(16) for $\mathcal{K}_n$
3. Sort uncertainty estimates in ascending order
let $\boldsymbol{\pi} = \{1, 2, ..., N_t\}$ with $N_t = |\mathcal{D}_{\text{test}}|$ so that $u_{\text{EU},\pi(1)} \leq u_{\text{EU},\pi(2)} \leq ... \leq u_{\text{EU},\pi(N_t)}$
4. Select top $\lfloor (1-\beta)N_t \rfloor$ certain averaged probabilities
$\tilde{\boldsymbol{p}}_{\pi(1)}, ..., \tilde{\boldsymbol{p}}_{\pi(\lfloor (1-\beta)N_t \rfloor)}$ for class prediction

Algorithm 3 Out-of-distribution detection procedure

Input: ID and OOD test samples: $\{\boldsymbol{x}_{\text{id},1}, ..., \boldsymbol{x}_{\text{id},N_{\text{id}}}\}$ and $\{\boldsymbol{x}_{\text{ood},1}, ..., \boldsymbol{x}_{\text{ood},N_{\text{ood}}}\}$; second-order predictor $f(\cdot)$
1. Set labels to build the full detection test data
$\mathcal{D}_{\text{detect}} = \{(\boldsymbol{x}_{\text{id},n}, y_n = 0)\}_{n=1}^{N_{\text{id}}} \cup \{(\boldsymbol{x}_{\text{ood},n}, y_n = 1)\}_{n=1}^{N_{\text{ood}}}$
2. Obtain second-order representations from $f(\cdot)$ on $\forall \boldsymbol{x}_n \in \mathcal{D}_{\text{detect}}$
$\mathcal{B}_n \leftarrow f(\boldsymbol{x}_n)$ (practical Bayesian) or $\mathcal{K}_n \leftarrow \mathcal{B}_n$ using (4) (credal)
3. Quantify epistemic predictive uncertainty given each sample, denoted by $u_{\text{EU},n}$
$u_{\text{EU},n}$ $\forall n$ using (6)/(7)/(8) for $\mathcal{B}_n$ or (10)/(13)/(16) for $\mathcal{K}_n$
4. Compute the AUROC using EU estimates for all test samples
auroc_score $\big((u_{\text{EU},1}, ..., u_{\text{EU},N_{\text{detect}}}), (y_1, ..., y_{N_{\text{detect}}})\big)$

A.3 SHIP CLASSIFICATION DATASETS

SeaShip dataset. To derive a classification dataset from existing object detection benchmarks. The target includes six classes of ships, including bulk cargo carriers, container ships, fishing boats, general cargo ships, ore carriers, and passenger ships. We are designing a cropping pipeline that transforms each annotated bounding box into an independent image crop. The procedure is summarized below:

1. *Input format.* We are assuming standard YOLO-style annotations, where each object is represented as $(c, x_{ctr}, y_{ctr}, w, h)$ in normalized coordinates.
2. *Bounding-box expansion.* For each object, we are randomly expanding its bounding box by a multiplicative factor. The horizontal and vertical expansion ratios are being independently sampled from $[1+\alpha_{\min},\ 1+\alpha_{\max}]$, with $\alpha_{\min} = 0.25$ and $\alpha_{\max} = 0.50$ in our experiments. This step is ensuring that the classification model is seeing the object together with limited contextual background, thereby avoiding overly tight crops.
3. *Cropping and clamping.* The expanded box is being converted into pixel coordinates and is being clamped to the image boundaries. The resulting crops are being extracted and saved as individual images, each inheriting the original class label.
4. *Filtering based on area.* The derived images are being filtered using a threshold of 256×256 pixels. Tiny images are being removed.
5. *Manifest generation.* Alongside the image crops, we are generating a JSON manifest containing, for each sample, the file path, source image identifier, class ID, original and expanded bounding boxes (in COCO (x, y, w, h) format), the new crop area, and the ambiguity flag. This manifest is enabling reproducibility and facilitating downstream training pipelines.

Overall, this procedure is converting every annotated object in the detection dataset into one or more classification samples, while preserving traceability to the source image and the original detection labels. The train/validation/test split is being kept consistent with the experimental settings presented in Wang et al. [2024b, 2025c].

SeaShip-C dataset. In addition to the clean train/validation/test classification splits described above, we are further constructing corrupted classification test sets aligned with the SeaShip-C benchmark in Wang et al. [2025c]. SeaShip-C is defining 25 synthesized corruption types that are being applied to clean images and evaluated across multiple datasets and models. The box labels remain unchanged after corruption.

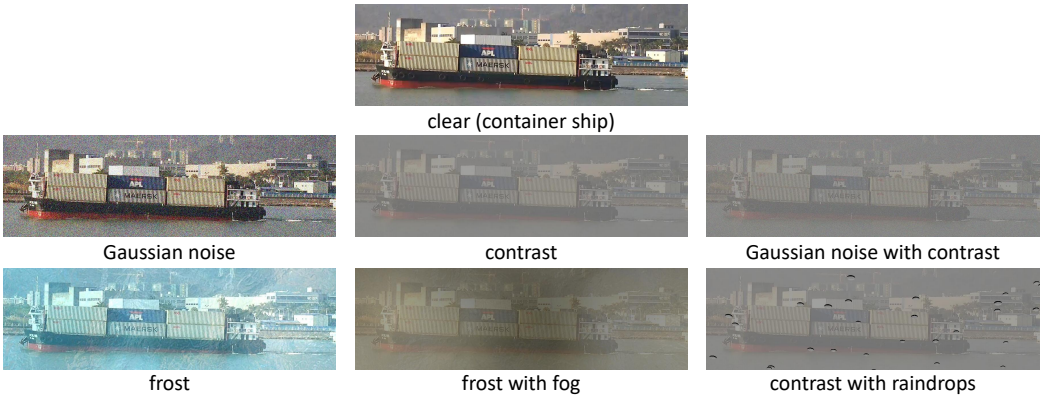

Figure A.2: Examples of the SeaShip dataset v.s. instances from the Seaship-C at severity level 3.

In each experiment (i.e., a given model on a given dataset), the corruptions are being categorized as mild, moderate, or severe depending on the observed degradation in model performance. For our study, we are focusing on the subset of corruptions that are being identified as severe in at least one experimental setting reported in Wang et al. [2025c]. This selection is yielding 6 corruption types: Gaussian noise, frost, contrast, Gaussian noise with contrast, contrast with raindrops, and frost with fog. For each corruption type and severity level ($\{1, 3, 5\}$), the classification crops are being re-extracted from the corresponding corrupted detection images, ensuring one-to-one alignment with the clean test set, as shown in Figure A.2.

SeaShip-O dataset cropped from SSAVE and SMD datasets. In addition to the classification datasets derived from SeaShip, we are also constructing two out-of-distribution (OOD) classification datasets using the SMD and SSAVE datasets in Wang et al. [2025c]. The procedure is following the same cropping strategy described above, using identical expansion parameters for bounding-box augmentation. After cropping the training partitions of both datasets, we are applying two additional steps:

1. *Filtering.* Crops with an effective area smaller than 128×128 pixels are being discarded in order to exclude tiny instances.
2. *Downsampling.* To address the repeated appearance of vessels across frames, we are downsampling the cropped datasets: for SMD we are retaining one crop out of every 8 (8:1), and for SSAVE one out of every 4 (4:1).

For the label space, we are following Wang et al. [2025c] and using the original class taxonomy in SMD. For SSAVE, we are restricting the dataset to the ship class only, since other categories are not relevant to our study. The resulting datasets are serving as OOD testbeds for classification, complementing the in-distribution dataset derived from SeaShip.

B ADDITIONAL EXPERIMENTAL RESULTS

Table A.2: AUARC scores for selective prediction tasks on various datasets.

	DE	PackEns	MaskEns	BatchEns	SVI	MCDO	Overall
In-distribution Camelyon17 test data							
MI	0.9737±0.0024	0.9725±0.0021	0.9722±0.0033	0.9742±0.0033	0.9643±0.0161	0.9041±0.1289	0.9602±0.0568
LWV	0.9749±0.0020	0.9750±0.0020	0.9748±0.0028	0.9765±0.0025	0.9655±0.0139	0.9082±0.1195	0.9625±0.0532
WD	0.9756±0.0018	0.9754±0.0020	0.9753±0.0028	0.9770±0.0025	0.9659±0.0139	0.9106±0.1200	0.9633±0.0530
GH	0.9754±0.0019	0.9754±0.0020	0.9753±0.0028	0.9769±0.0026	0.9656±0.0144	0.9111±0.1183	0.9633±0.0523
H_{diff}	0.9726±0.0020	0.9705±0.0026	0.9708±0.0036	0.9731±0.0035	0.9616±0.0179	0.9080±0.1164	0.9594±0.0517
MMI	0.9754±0.0019	0.9754±0.0020	0.9753±0.0028	0.9769±0.0026	0.9656±0.0144	0.9111±0.1183	0.9633±0.0523
Distribution-shift Camelyon17 test data							
MI	0.9575±0.0043	0.9686±0.0039	0.9704±0.0024	0.9694±0.0033	0.9658±0.0093	0.9507±0.0394	0.9637±0.0176
LWV	0.9595±0.0037	0.9715±0.0040	0.9728±0.0025	0.9730±0.0027	0.9672±0.0094	0.9573±0.0271	0.9669±0.0131
WD	0.9620±0.0034	0.9724±0.0037	0.9738±0.0022	0.9738±0.0025	0.9685±0.0089	0.9594±0.0263	0.9683±0.0125
GH	0.9614±0.0036	0.9722±0.0037	0.9736±0.0023	0.9735±0.0025	0.9683±0.0089	0.9592±0.0264	0.9680±0.0126
H_{diff}	0.9558±0.0041	0.9663±0.0032	0.9685±0.0025	0.9668±0.0037	0.9616±0.0102	0.9490±0.0359	0.9613±0.0164
MMI	0.9614±0.0036	0.9722±0.0037	0.9736±0.0023	0.9735±0.0025	0.9683±0.0089	0.9592±0.0264	0.9680±0.0126
In-distribution SeaShip test data							
MI	0.9962±0.0006	0.9952±0.0015	0.9931±0.0018	0.9929±0.0023	0.9639±0.0267	0.9910±0.0058	0.9887±0.0156
LWV	0.9968±0.0005	0.9962±0.0011	0.9941±0.0019	0.9939±0.0022	0.9660±0.0260	0.9922±0.0053	0.9899±0.0151
WD	0.9971±0.0005	0.9964±0.0010	0.9945±0.0017	0.9943±0.0020	0.9676±0.0249	0.9924±0.0051	0.9904±0.0144
GH	0.9970±0.0005	0.9963±0.0011	0.9944±0.0016	0.9943±0.0020	0.9675±0.0248	0.9924±0.0051	0.9903±0.0144
H_{diff}	0.9947±0.0005	0.9942±0.0019	0.9917±0.0022	0.9914±0.0031	0.9479±0.0423	0.9885±0.0070	0.9848±0.0237
MMI	0.9969±0.0004	0.9963±0.0011	0.9943±0.0017	0.9941±0.0021	0.9667±0.0259	0.9924±0.0051	0.9901±0.0149
In-distribution CIFAR10 test data							
MI	0.9828±0.0013	0.9784±0.0029	0.9851±0.0018	0.9823±0.0039	0.9763±0.0078	0.9794±0.0038	0.9807±0.0050
LWV	0.9836±0.0011	0.9796±0.0027	0.9855±0.0016	0.9832±0.0035	0.9767±0.0078	0.9801±0.0035	0.9815±0.0049
WD	0.9845±0.0012	0.9805±0.0026	0.9864±0.0015	0.9840±0.0033	0.9777±0.0072	0.9810±0.0034	0.9823±0.0046
GH	0.9843±0.0011	0.9805±0.0025	0.9863±0.0016	0.9839±0.0034	0.9776±0.0072	0.9807±0.0034	0.9822±0.0046
H_{diff}	0.9790±0.0020	0.9734±0.0041	0.9823±0.0026	0.9791±0.0052	0.9707±0.0105	0.9739±0.0057	0.9764±0.0068
MMI	0.9840±0.0011	0.9801±0.0026	0.9860±0.0016	0.9837±0.0034	0.9771±0.0074	0.9803±0.0035	0.9819±0.0048

Table A.3: AUROC scores for OOD detection tasks on various dataset pairs.

	DE	PackEns	MaskEns	BatchEns	SVI	MCDO	Overall
SeaShip v.s. SeaShip-O							
MI	0.9153±0.0026	0.9347±0.0127	0.9145±0.0106	0.9072±0.0139	0.7129±0.0678	0.8315±0.0379	0.8694±0.0840
LWV	0.9085±0.0021	0.9265±0.0136	0.9066±0.0107	0.8986±0.0139	0.7147±0.0672	0.8378±0.0355	0.8655±0.0797
WD	0.9235±0.0028	0.9373±0.0136	0.9202±0.0108	0.9129±0.0128	0.7449±0.0681	0.8545±0.0322	0.8822±0.0739
GH	0.9240±0.0025	0.9403±0.0125	0.9235±0.0104	0.9161±0.0126	0.7519±0.0654	0.8577±0.0314	0.8856±0.0720
H_{diff}	0.9023±0.0077	0.9220±0.0159	0.8939±0.0136	0.8913±0.0174	0.6841±0.0725	0.8162±0.0397	0.8516±0.0893
MMI	0.9168±0.0024	0.9339±0.0131	0.9159±0.0104	0.9086±0.0136	0.7356±0.0656	0.8503±0.0323	0.8768±0.0751
SeaShip v.s. SeaShip-C							
MI	0.8815±0.0063	0.8892±0.0146	0.8709±0.0101	0.8720±0.0141	0.6734±0.0576	0.7629±0.0414	0.8250±0.0859
LWV	0.8782±0.0061	0.8868±0.0139	0.8696±0.0101	0.8696±0.0137	0.6870±0.0565	0.7886±0.0382	0.8300±0.0776
WD	0.8990±0.0048	0.9033±0.0128	0.8864±0.0080	0.8898±0.0112	0.7214±0.0534	0.8071±0.0365	0.8512±0.0720
GH	0.9028±0.0050	0.9077±0.0114	0.8908±0.0074	0.8943±0.0103	0.7299±0.0495	0.8110±0.0361	0.8561±0.0703
H_{diff}	0.8496±0.0088	0.8632±0.0195	0.8418±0.0121	0.8353±0.0216	0.6248±0.0626	0.7425±0.0400	0.7929±0.0912
MMI	0.8926±0.0058	0.8993±0.0124	0.8827±0.0082	0.8848±0.0119	0.7128±0.0514	0.8036±0.0365	0.8460±0.0729
CIFAR10 v.s. SVHN							
MI	0.7818±0.0094	0.7561±0.0119	0.7940±0.0159	0.7894±0.0200	0.7435±0.0272	0.7292±0.0192	0.7657±0.0301
LWV	0.7806±0.0099	0.7557±0.0136	0.7968±0.0147	0.7927±0.0214	0.7467±0.0346	0.7440±0.0221	0.7694±0.0295
WD	0.8326±0.0082	0.8046±0.0101	0.8366±0.0128	0.8321±0.0168	0.8043±0.0246	0.7944±0.0168	0.8174±0.0227
GH	0.8445±0.0073	0.8191±0.0081	0.8463±0.0120	0.8435±0.0145	0.8218±0.0189	0.8068±0.0134	0.8303±0.0197
H_{diff}	0.7950±0.0132	0.7579±0.0192	0.8121±0.0214	0.8002±0.0270	0.7622±0.0397	0.7509±0.0268	0.7797±0.0343
MMI	0.8150±0.0081	0.7927±0.0103	0.8246±0.0124	0.8229±0.0167	0.7883±0.0255	0.7795±0.0169	0.8038±0.0235
CIFAR10 v.s. FMNIST							
MI	0.8852±0.0097	0.8675±0.0096	0.8929±0.0111	0.8740±0.0105	0.8271±0.0231	0.8366±0.0117	0.8639±0.0276
LWV	0.8653±0.0103	0.8456±0.0101	0.8736±0.0109	0.8580±0.0123	0.8149±0.0231	0.8224±0.0134	0.8466±0.0256
WD	0.8907±0.0082	0.8736±0.0084	0.8992±0.0098	0.8804±0.0103	0.8438±0.0204	0.8508±0.0105	0.8731±0.0232
GH	0.8968±0.0075	0.8816±0.0079	0.9059±0.0098	0.8867±0.0089	0.8512±0.0200	0.8584±0.0096	0.8801±0.0226
H_{diff}	0.8912±0.0096	0.8716±0.0128	0.9043±0.0120	0.8835±0.0118	0.8416±0.0286	0.8542±0.0154	0.8744±0.0266
MMI	0.8799±0.0087	0.8644±0.0085	0.8885±0.0100	0.8732±0.0096	0.8322±0.0208	0.8400±0.0106	0.8630±0.0237

Table A.4: Net wins for intra-representation comparisons on selective prediction tasks across datasets. The first rank according to net wins is shown in bold red, and the second rank in bold blue.

	DE	PackEns	MaskEns	BatchEns	SVI	MCDO	Total	DE	PackEns	MaskEns	BatchEns	SVI	MCDO	Total
Distribution-based measures														
	In-distribution Camelyon17 test data							Distribution-shift Camelyon17 test data						
MI	-2	-2	-2	-2	-1	-2	-11	-2	-2	-2	-2	-2	-2	-12
LWV	0	0	0	0	0	0	**0**	0	0	0	0	0	0	**0**
WD	2	2	2	2	1	2	**11**	2	2	2	2	2	2	**12**
	In-distribution SeaShip test data							In-distribution CIFAR10 test data						
MI	-2	-2	-2	-2	-2	-2	-12	-2	-2	-2	-2	-2	-2	-12
LWV	0	0	0	0	0	0	**0**	0	0	0	0	0	0	**0**
WD	2	2	2	2	2	2	**12**	2	2	2	2	2	2	**12**
Credal-based measures														
	In-distribution Camelyon17 test data							Distribution-shift Camelyon17 test data						
GH	1	1	1	1	1	1	**6**	1	1	1	1	1	1	**6**
H_{diff}	-2	-2	-2	-2	-2	-2	-12	-2	-2	-2	-2	-2	-2	-12
MMI	1	1	1	1	1	1	**6**	1	1	1	1	1	1	**6**
	In-distribution SeaShip test data							In-distribution CIFAR10 test data						
GH	2	1	2	2	2	1	**10**	2	2	2	2	2	2	**12**
H_{diff}	-2	-2	-2	-2	-2	-2	-12	-2	-2	-2	-2	-2	-2	-12
MMI	0	1	0	0	0	1	**2**	0	0	0	0	0	0	**0**

Table A.5: Net wins for inter-representation comparisons on selective prediction tasks across datasets. The first rank according to net wins is shown in bold red, and the second rank in bold blue.

	DE	PackEns	MaskEns	BatchEns	SVI	MCDO	Total	DE	PackEns	MaskEns	BatchEns	SVI	MCDO	Total
	In-distribution Camelyon17 test data							Distribution-shift Camelyon17 test data						
MI	-3	-3	-3	-3	-2	-4	-18	-3	-3	-3	-3	-3	-4	-19
LWV	-1	-1	-1	-1	1	-2	-5	-1	-1	-1	-1	-1	-1	-6
WD	5	3	3	5	2	3	**21**	5	5	5	5	5	5	**30**
GH	2	3	3	2	2	3	**15**	2	2	2	2	2	2	**12**
H_{diff}	-5	-5	-5	-5	-5	-3	-28	-5	-5	-5	-5	-5	-4	-29
MMI	2	3	3	2	2	3	**15**	2	2	2	2	2	2	**12**
	In-distribution SeaShip test data							In-distribution CIFAR10 test data						
MI	-3	-3	-3	-3	-3	-3	-18	-3	-3	-3	-3	-3	-3	-18
LWV	0	-1	-1	-1	-1	0	-4	-1	-1	-1	-1	-1	-1	-6
WD	5	5	4	4	4	3	**25**	5	4	5	4	5	5	**28**
GH	3	2	4	4	4	2	**19**	3	4	3	4	3	3	**20**
H_{diff}	-5	-5	-5	-5	-5	-5	-30	-5	-5	-5	-5	-5	-5	-30
MMI	0	2	1	1	1	3	8	1	1	1	1	1	1	6

Table A.6: Net wins for intra-representation comparisons on OOD detection tasks across datasets. The first rank according to net wins is shown in bold red, and the second rank in bold blue.

	DE	PackEns	MaskEns	BatchEns	SVI	MCDO	Total	DE	PackEns	MaskEns	BatchEns	SVI	MCDO	Total
Distribution-based measures														
	SeaShip v.s. SeaShip-O							SeaShip v.s. SeaShip-C						
MI	0	0	0	0	-1	-2	**-3**	0	0	-1	0	-2	-2	**-5**
LWV	-2	-2	-2	-2	-1	0	-9	-2	-2	-1	-2	0	0	-7
WD	2	2	2	2	2	2	**12**	2	2	2	2	2	2	**12**
	CIFAR10 v.s. SVHN							CIFAR10 v.s. FMNIST						
MI	0	-1	-2	-2	-1	-2	-8	0	0	0	0	0	0	**0**
LWV	-2	-1	0	0	-1	0	**-4**	-2	-2	-2	-2	-2	-2	-12
WD	2	2	2	2	2	2	**12**	2	2	2	2	2	2	**12**
Credal-based measures														
	SeaShip v.s. SeaShip-O							SeaShip v.s. SeaShip-C						
GH	2	2	2	2	2	2	**12**	2	2	2	2	2	2	**12**
H_{diff}	-2	-2	-2	-2	-2	-2	-12	-2	-2	-2	-2	-2	-2	-12
MMI	0	0	0	0	0	0	**0**	0	0	0	0	0	0	**0**
	CIFAR10 v.s. SVHN							CIFAR10 v.s. FMNIST						
GH	2	2	2	2	2	2	**12**	2	2	1	1	2	1	**9**
H_{diff}	-2	-2	-2	-2	-2	-2	-12	0	0	1	1	0	1	**3**
MMI	0	0	0	0	0	0	**0**	-2	-2	-2	-2	-2	-2	-12

Table A.7: Net wins for inter-representation comparisons on OOD detection tasks across datasets. The first rank according to net wins is shown in bold red, and the second rank in bold blue.

	DE	PackEns	MaskEns	BatchEns	SVI	MCDO	Total	DE	PackEns	MaskEns	BatchEns	SVI	MCDO	Total
	SeaShip v.s. SeaShip-O							SeaShip v.s. SeaShip-C						
MI	-1	0	0	0	-2	-3	-6	-1	-1	-2	-1	-3	-3	-11
LWV	-3	-3	-3	-3	-2	-1	-15	-3	-3	-2	-3	-1	-1	-13
WD	4	3	3	3	3	3	**19**	3	3	3	3	3	3	**18**
GH	4	5	5	5	5	5	**29**	5	5	5	5	5	5	**30**
H_{diff}	-5	-5	-5	-5	-5	-5	-30	-5	-5	-5	-5	-5	-5	-30
MMI	1	0	0	0	1	1	3	1	1	1	1	1	1	6
	CIFAR10 v.s. SVHN							CIFAR10 v.s. FMNIST						
MI	-3	-3	-5	-5	-4	-5	-25	-1	-1	-1	-2	-3	-3	-11
LWV	-5	-3	-3	-3	-4	-3	-21	-5	-5	-5	-5	-5	-5	-30
WD	3	3	3	3	3	3	**18**	2	2	1	1	2	2	10
GH	5	5	5	5	5	5	**30**	5	5	4	4	5	4	**27**
H_{diff}	-1	-3	-1	-1	-1	-1	-8	2	2	4	4	2	3	**17**
MMI	1	1	1	1	1	1	6	-3	-3	-3	-2	-1	-1	-13

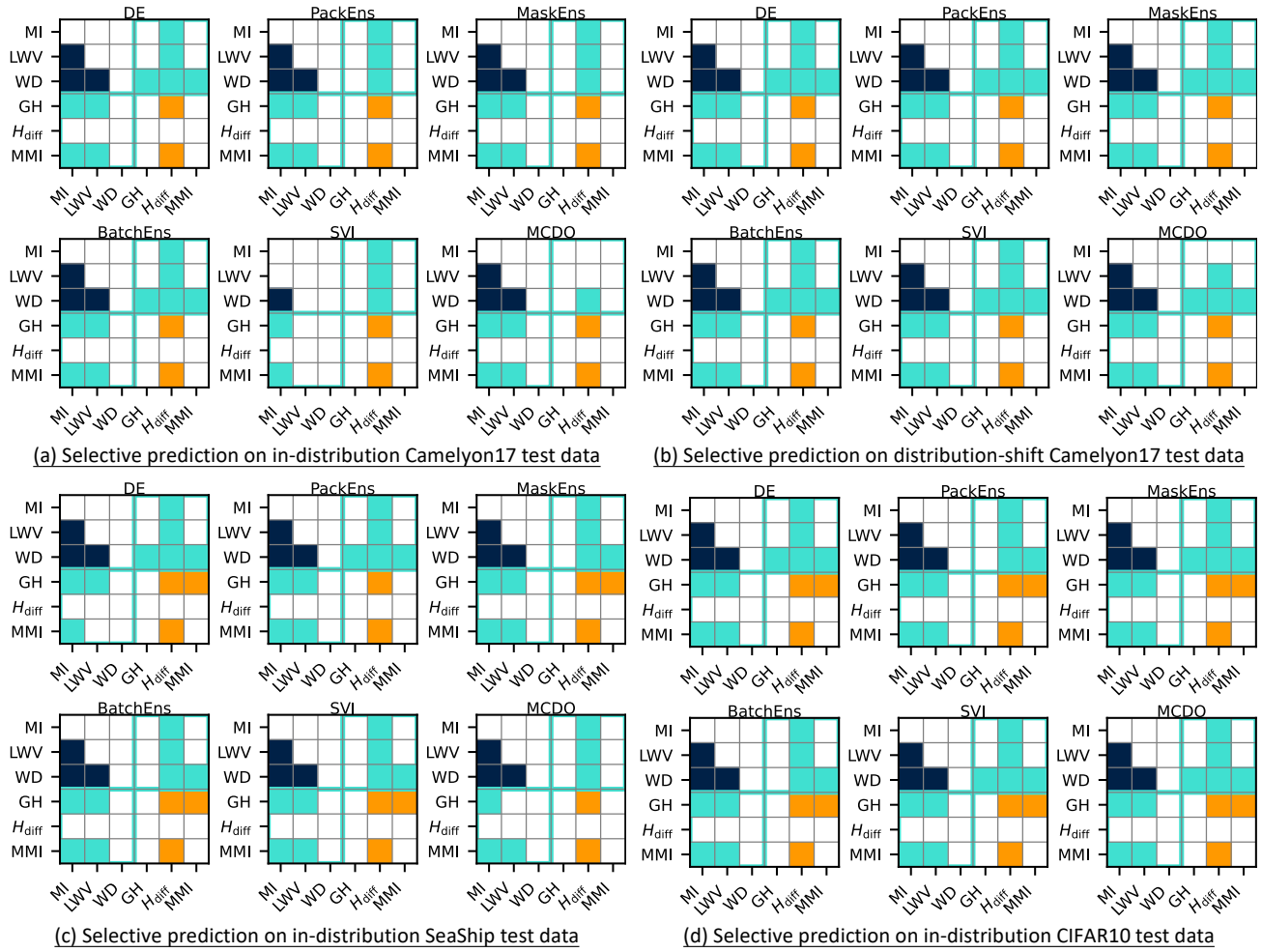

Figure A.3: Statistical significance plots on different selective prediction (a-d) benchmarks across different underlying predictive models. A cell is shaded if the measure in the i-th row is statistically significantly better than that in the j-th column according to a pairwise one-sided Wilcoxon signed-rank test at the 5% significance level. Intra-representation comparisons are shown in blue (distribution-based measures) and orange (credal-based measures), while inter-representation comparisons are shown in green.

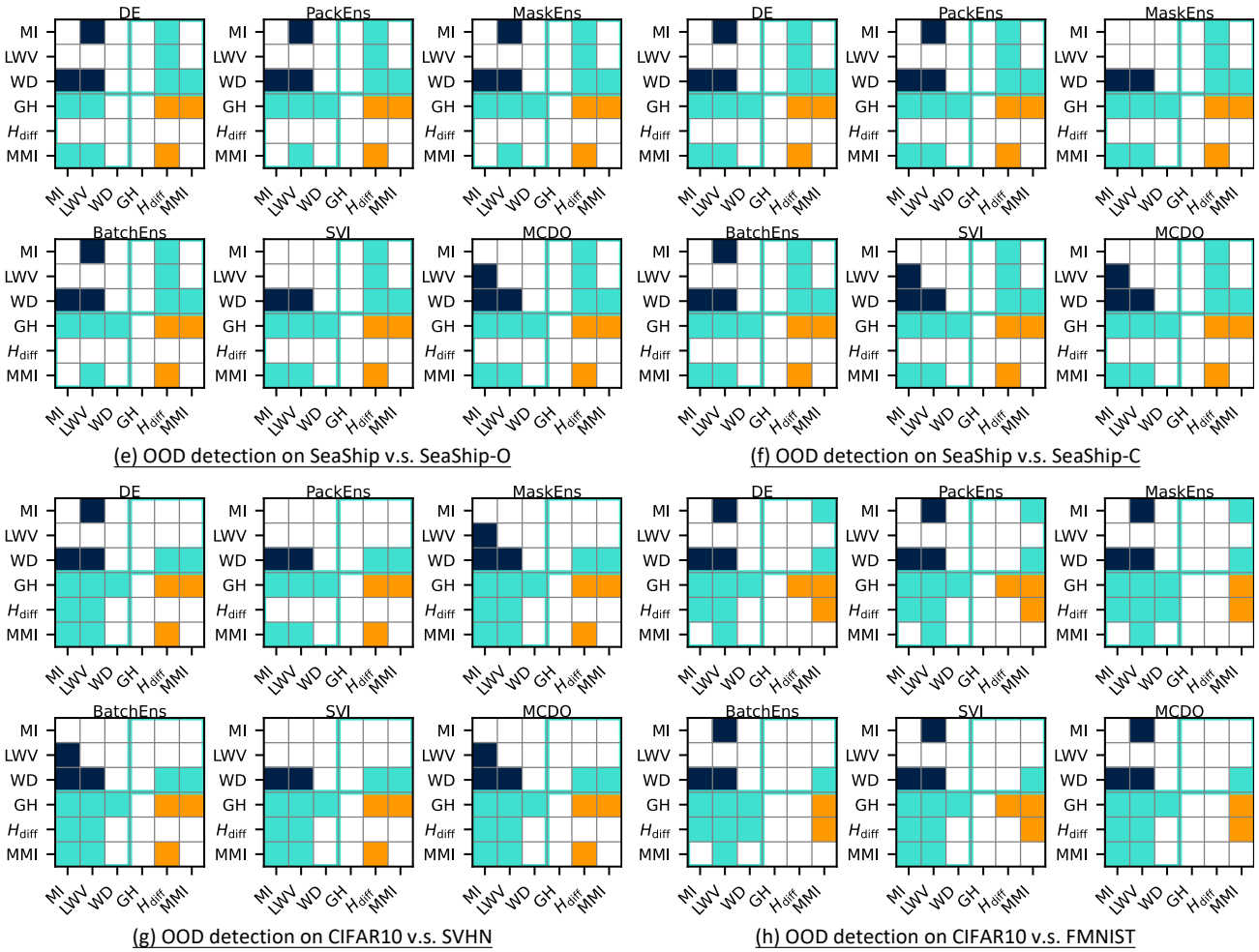

Figure A.4: Statistical significance plots on different OOD detection (e-h) benchmarks across different underlying predictive models. A cell is shaded if the measure in the i-th row is statistically significantly better than that in the j-th column according to a pairwise one-sided Wilcoxon signed-rank test at the 5% significance level. Intra-representation comparisons are shown in blue (distribution-based measures) and orange (credal-based measures), while inter-representation comparisons are shown in green.

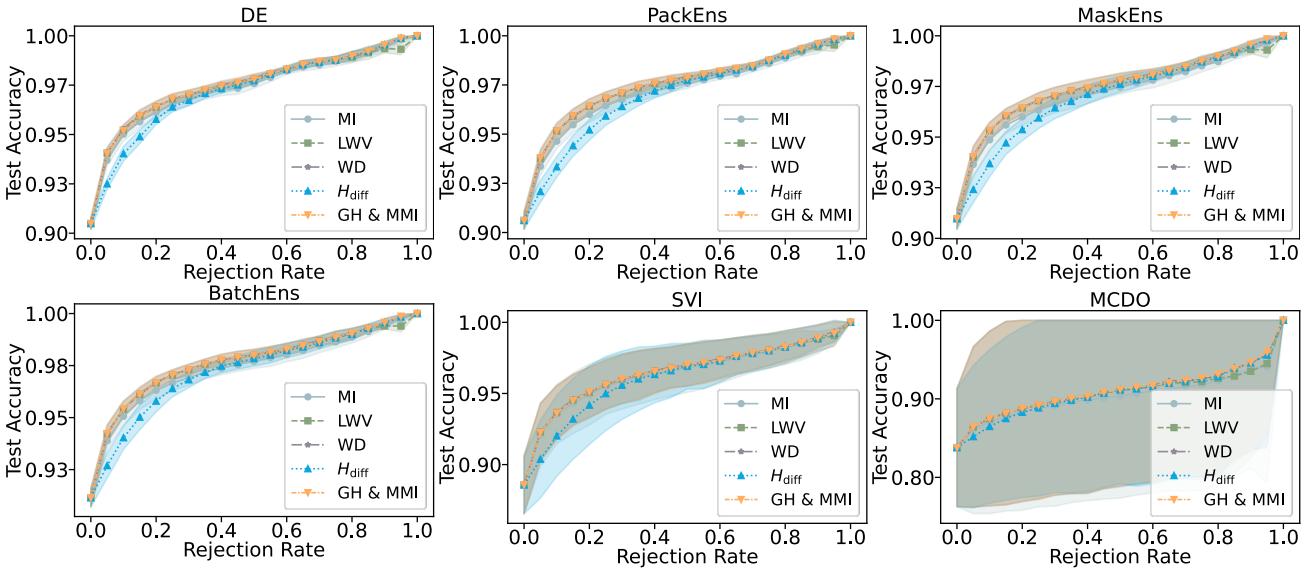

Figure A.5: Accuracy rejection curves using different uncertainty representations and measures on in-distribution Camelyon17 test data across different underlying predictive models.

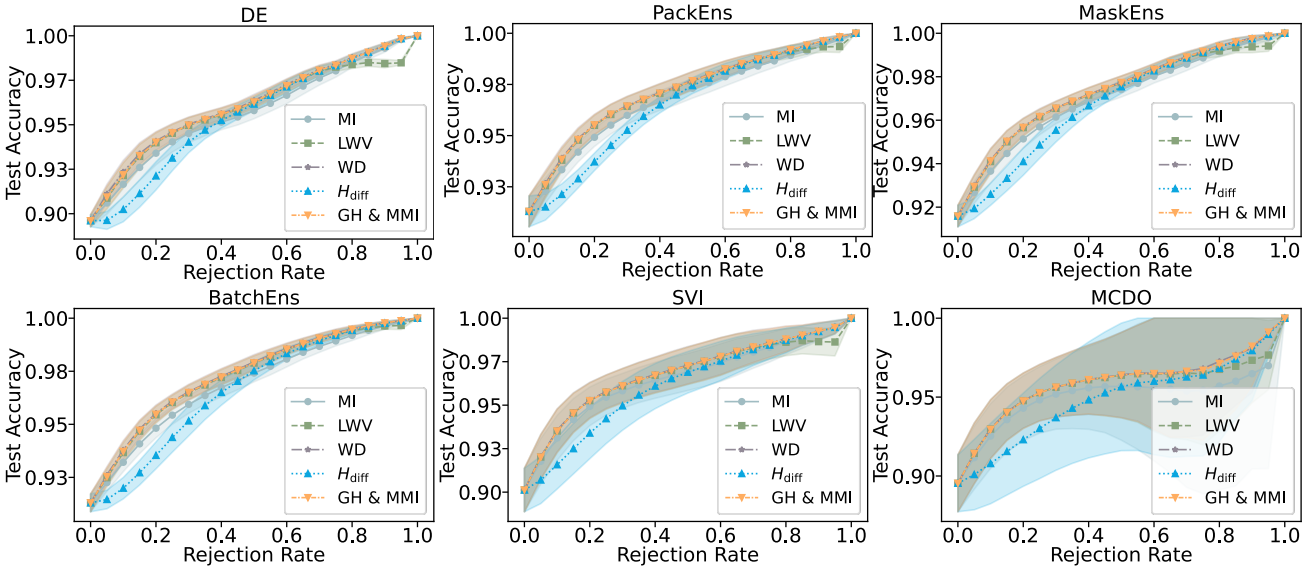

Figure A.6: Accuracy rejection curves using different uncertainty representations and measures on distribution-shift Camelyon17 test data across different underlying predictive models.

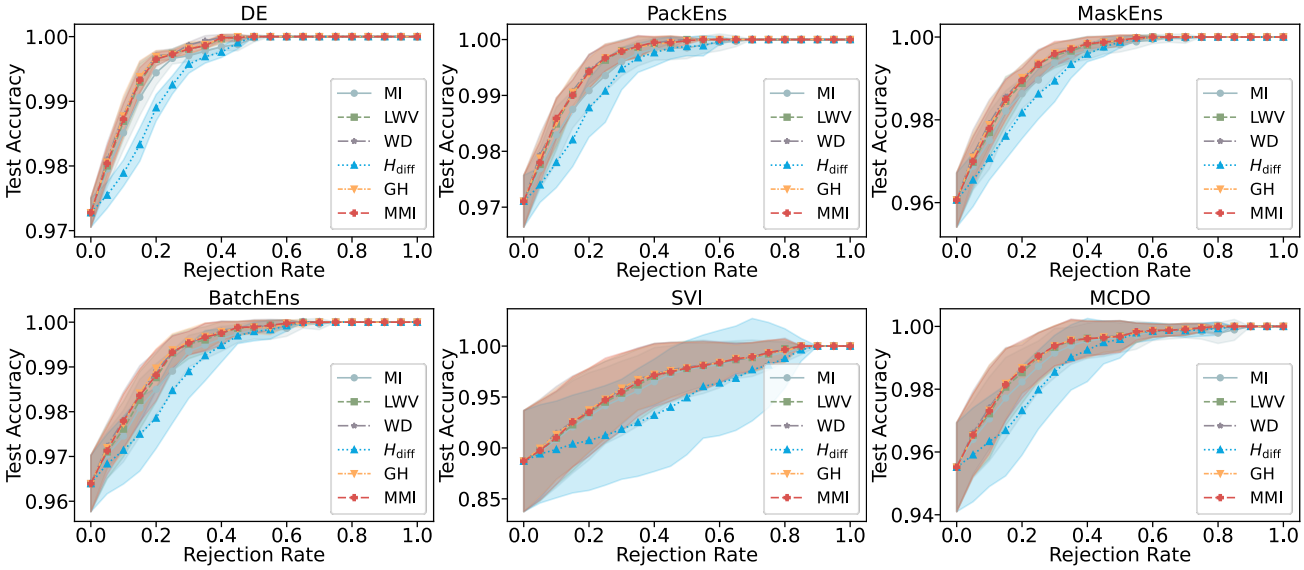

Figure A.7: Accuracy rejection curves using different uncertainty representations and measures on in-distribution SeaShip test data across different underlying predictive models.

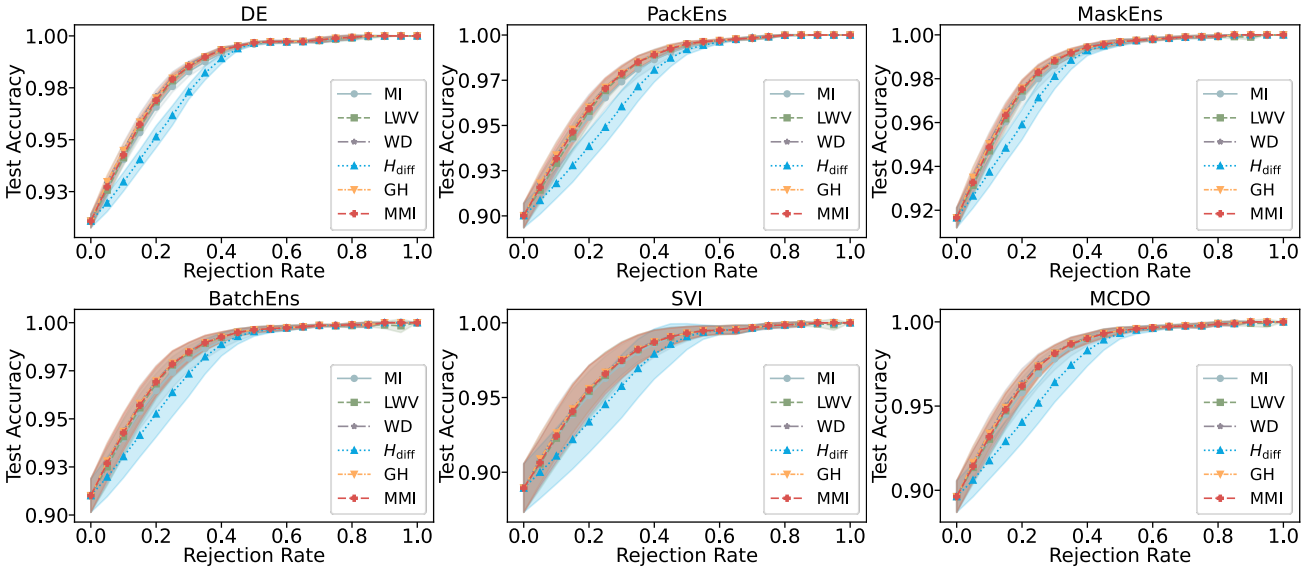

Figure A.8: Accuracy rejection curves using different uncertainty representations and measures on in-distribution CIFAR10 test data across different underlying predictive models.

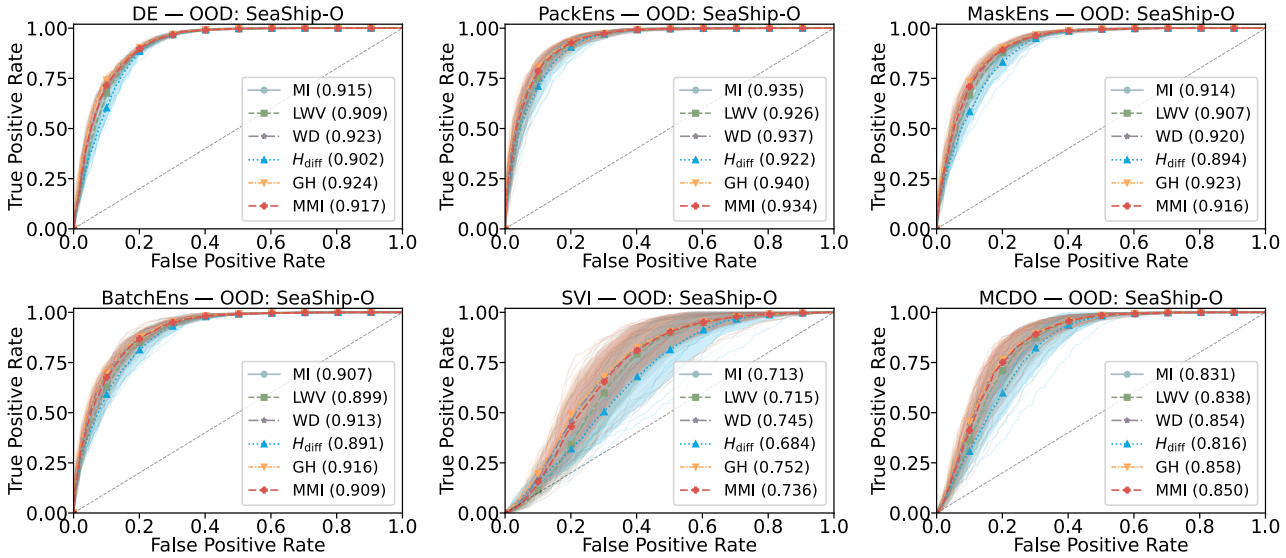

Figure A.9: ROC curves for OOD detection (SeaShip v.s. Seachip-O) across uncertainty measures and backbone methods. Each panel corresponds to one backbone, with all uncertainty measures plotted together. Solid lines represent the mean TPR over 10 independent runs, shaded regions indicate ± 1 standard deviation, and faint lines show individual runs. Mean AUROC values are reported in the legend.

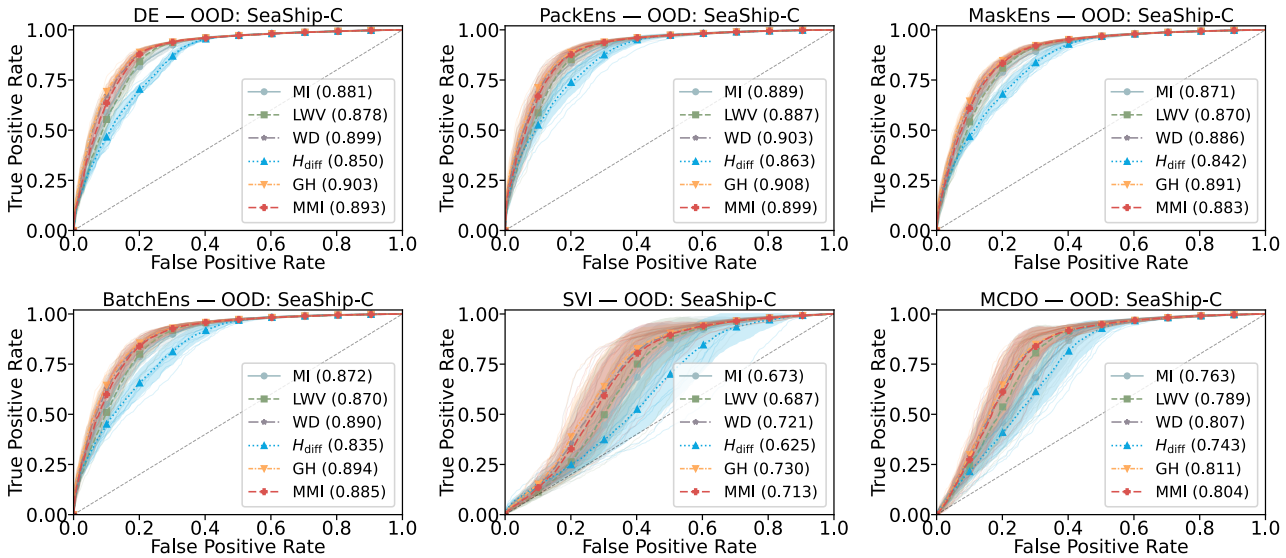

Figure A.10: ROC curves for OOD detection (SeaShip v.s. Seachip-C) across uncertainty measures and backbone methods. Each panel corresponds to one backbone, with all uncertainty measures plotted together. Solid lines represent the mean TPR over 10 independent runs, shaded regions indicate ± 1 standard deviation, and faint lines show individual runs. Mean AUROC values are reported in the legend.

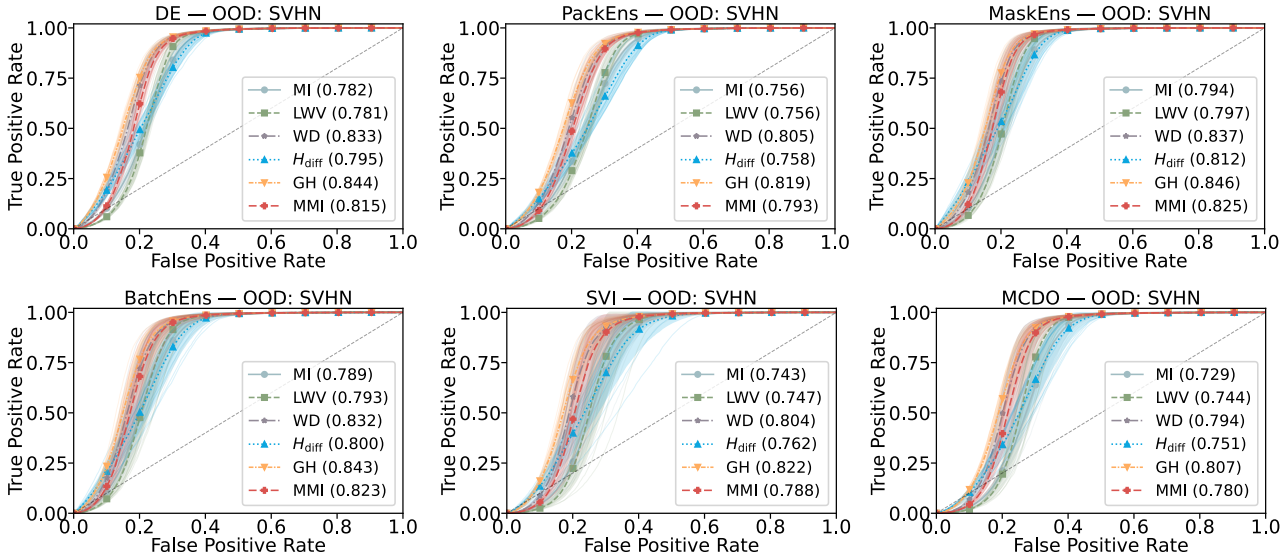

Figure A.11: ROC curves for OOD detection (CIFAR10 v.s. SVHN) across uncertainty measures and backbone methods. Each panel corresponds to one backbone, with all uncertainty measures plotted together. Solid lines represent the mean TPR over 10 independent runs, shaded regions indicate ± 1 standard deviation, and faint lines show individual runs. Mean AUROC values are reported in the legend.

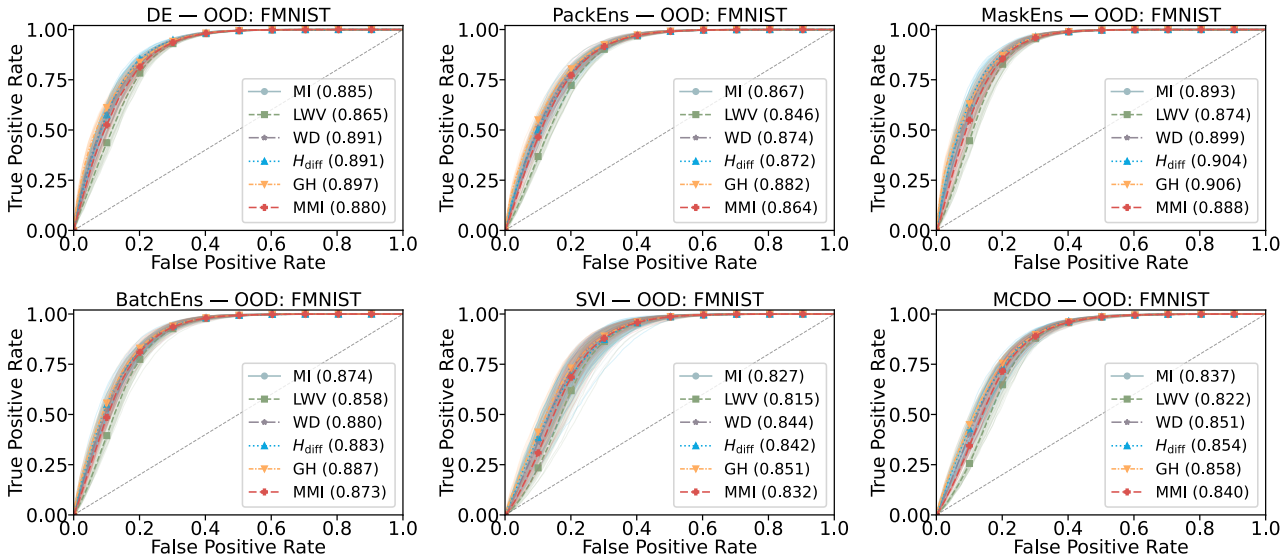

Figure A.12: ROC curves for OOD detection (CIFAR10 v.s. FMNIST) across uncertainty measures and backbone methods. Each panel corresponds to one backbone, with all uncertainty measures plotted together. Solid lines represent the mean TPR over 10 independent runs, shaded regions indicate ± 1 standard deviation, and faint lines show individual runs. Mean AUROC values are reported in the legend.

C FURTHER EXPERIMENTS

Table A.8: Average AUARC on the selective prediction task and AUROC on OOD detection tasks for models trained on **FMNIST**, averaged across six distinct predictive models. Intra- and inter-representation net wins are reported alongside. The first rank is shown in **bold red**, the second in **bold blue**.

	Selective Prediction			**OOD Detection (SVHN)**			**OOD Detection (CIFAR10)**		
	Average scores	Intra	Inter	Average scores	Intra	Inter	Average scores	Intra	Inter
MI	0.9901±0.0017	-12	-18	0.9148±0.0713	**-1**	-8	0.9074±0.0701	**6**	3
LWV	0.9910±0.0016	**0**	-6	0.9026±0.0604	-9	-27	0.8943±0.0609	-10	-28
WD	0.9912±0.0015	**12**	**25**	0.9323±0.0511	**10**	**14**	0.9126±0.0536	**4**	-2
GH	0.9912±0.0015	**12**	**23**	0.9414±0.0473	**12**	**30**	0.9173±0.0523	**1**	**12**
H_{diff}	0.9874±0.0026	-12	-30	0.9215±0.0612	-7	-4	0.9306±0.0554	**11**	**29**
MMI	0.9911±0.0015	**0**	6	0.9272±0.0525	**-5**	-5	0.9094±0.0551	-12	-14

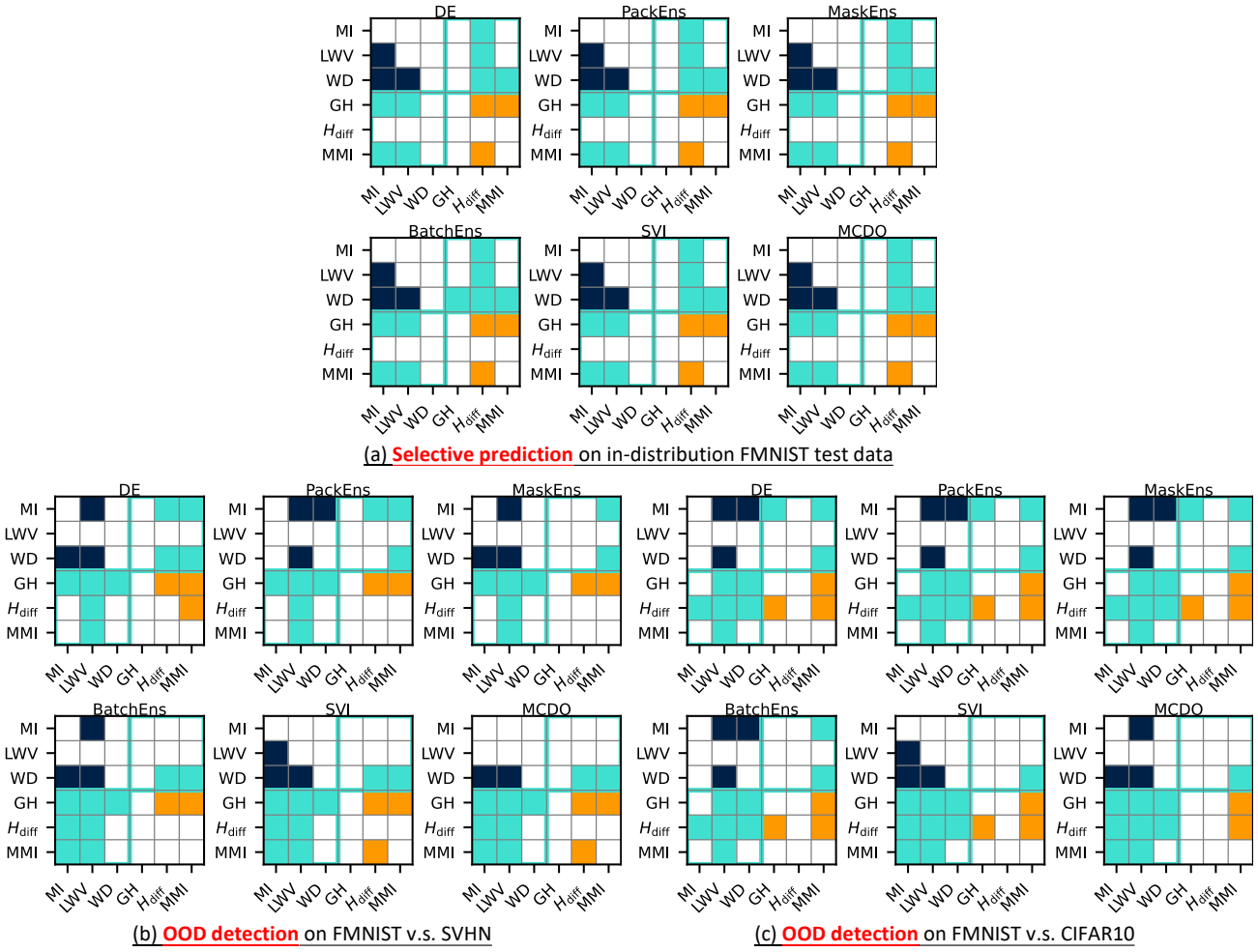

Figure A.13: Statistical significance plots on different selective prediction (a) and OOD detection (b-c) benchmarks across different underlying predictive models trained on the **FMNIST** dataset. A cell is shaded if the measure in the i-th row is statistically significantly better than that in the j-th column according to a pairwise one-sided Wilcoxon signed-rank test at the 5% significance level. Intra-representation comparisons are shown in blue (distribution-based measures) and orange (credal-based measures), while inter-representation comparisons are shown in green.

Table A.9: Average AUARC on the selective prediction task and AUROC on OOD detection tasks for models trained on **SVHN**, averaged across six distinct predictive models. Intra- and inter-representation net wins are reported alongside. The first rank is shown in **bold red**, the second in **bold blue**.

	Selective Prediction			**OOD Detection (FMNIST)**			**OOD Detection (CIFAR10)**		
	Average scores	Intra	Inter	Average scores	Intra	Inter	Average scores	Intra	Inter
MI	0.9950±0.0007	-10	-16	0.9704±0.0108	**6**	**13**	0.9765±0.0074	**6**	**12**
LWV	0.9952±0.0006	**2**	5	0.9570±0.0122	-12	-30	0.9629±0.0097	-12	-28
WD	0.9953±0.0007	**8**	**22**	0.9704±0.0092	**6**	12	0.9769±0.0062	**6**	**12**
GH	0.9953±0.0007	**10**	**15**	0.9741±0.0082	**12**	**29**	0.9806±0.0051	**12**	**30**
H_{diff}	0.9944±0.0010	-12	-30	0.9649±0.0118	-9	-15	0.9653±0.0100	-12	-20
MMI	0.9953±0.0007	**2**	4	0.9665±0.0100	**-3**	-9	0.9730±0.0072	**0**	-6

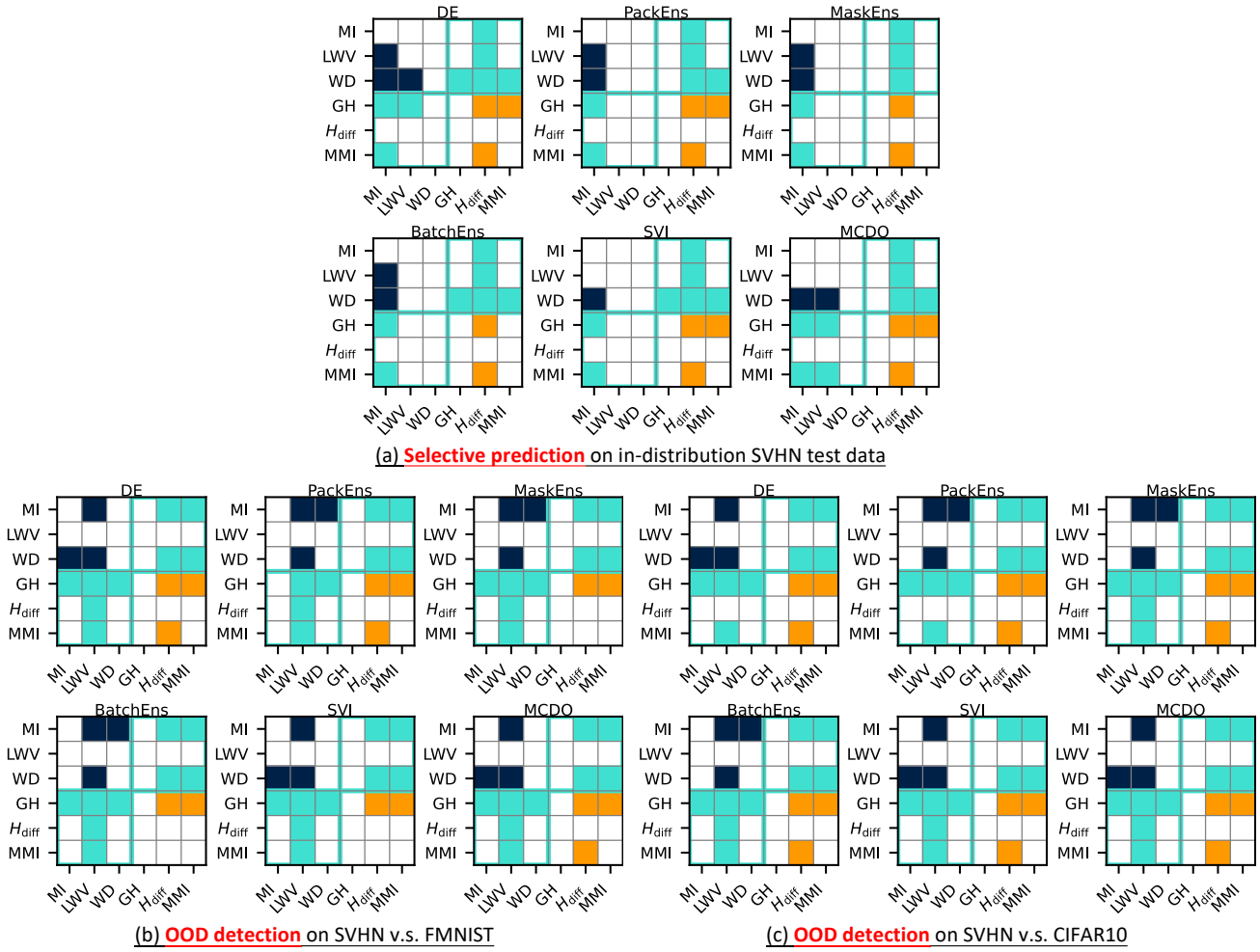

Figure A.14: Statistical significance plots on different selective prediction (a) and OOD detection (b-c) benchmarks across different underlying predictive models trained on the **SVHN** dataset. A cell is shaded if the measure in the i-th row is statistically significantly better than that in the j-th column according to a pairwise one-sided Wilcoxon signed-rank test at the 5% significance level. Intra-representation comparisons are shown in blue (distribution-based measures) and orange (credal-based measures), while inter-representation comparisons are shown in green.

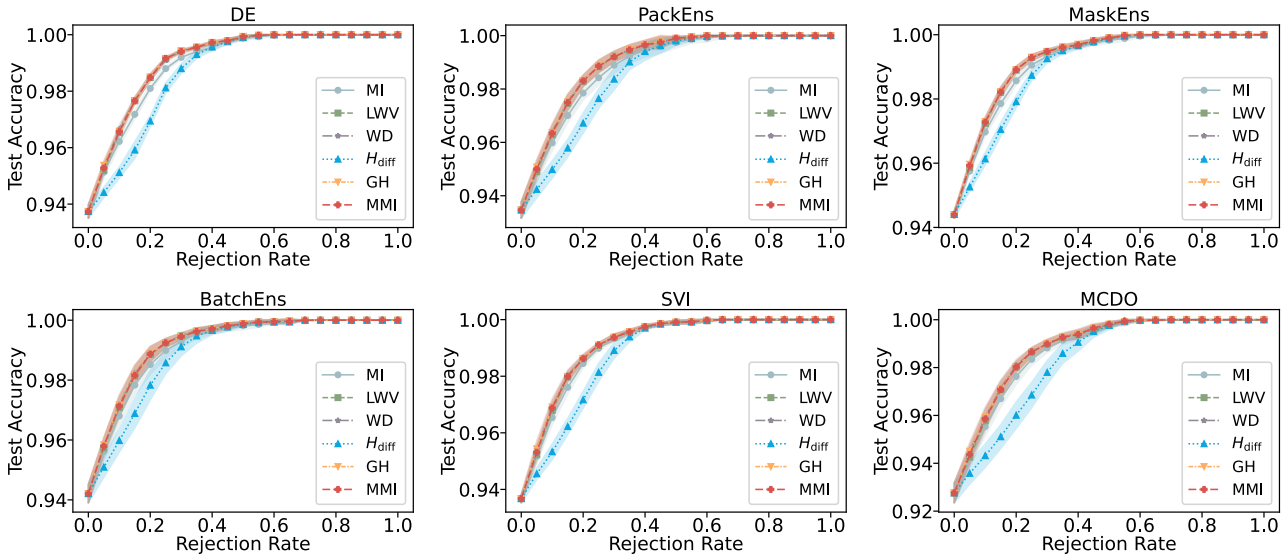

Figure A.15: Accuracy rejection curves using different uncertainty representations and measures on in-distribution **FMNIST** test data across different underlying predictive models.

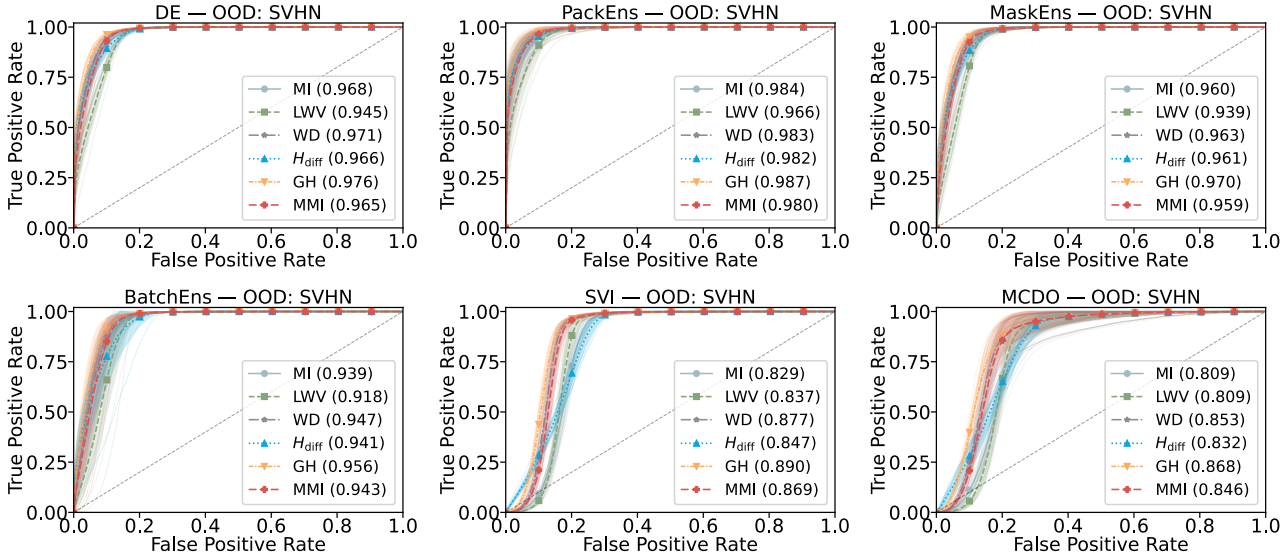

Figure A.16: ROC curves for OOD detection (**ID: FMNIST v.s. OOD: SVHN**) across uncertainty measures and backbone methods. Each panel corresponds to one backbone, with all uncertainty measures plotted together. Solid lines represent the mean TPR over 10 independent runs, shaded regions indicate ± 1 standard deviation, and faint lines show individual runs. Mean AUROC values are reported in the legend.

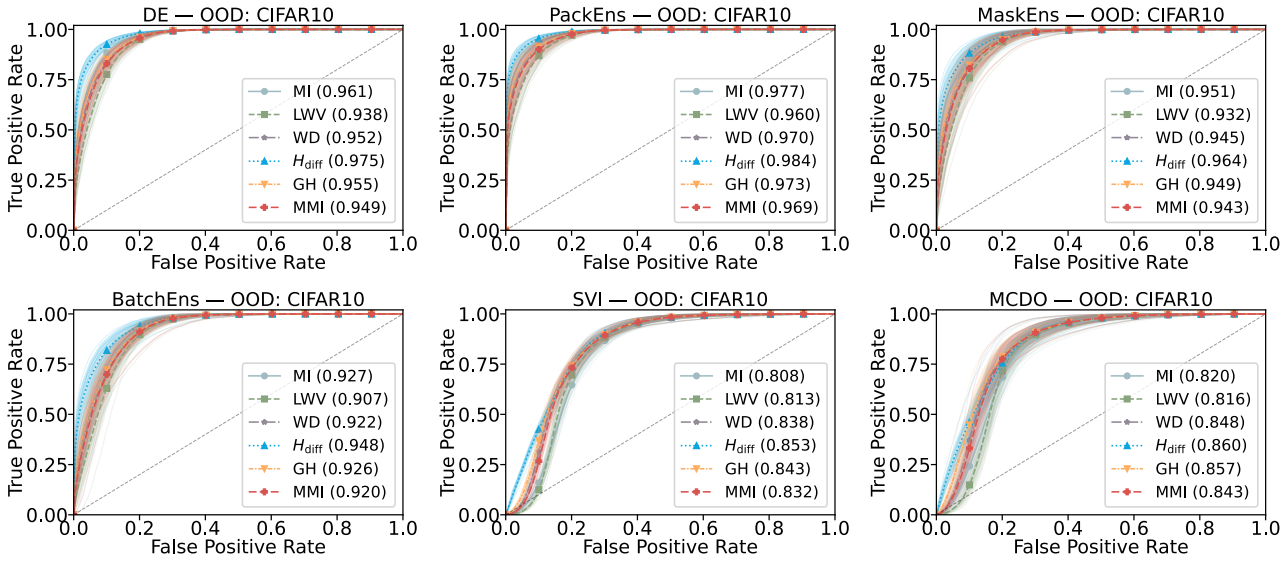

Figure A.17: ROC curves for OOD detection (**ID: FMNIST v.s. OOD: CIFAR10**) across uncertainty measures and backbone methods. Each panel corresponds to one backbone, with all uncertainty measures plotted together. Solid lines represent the mean TPR over 10 independent runs, shaded regions indicate ± 1 standard deviation, and faint lines show individual runs. Mean AUROC values are reported in the legend.

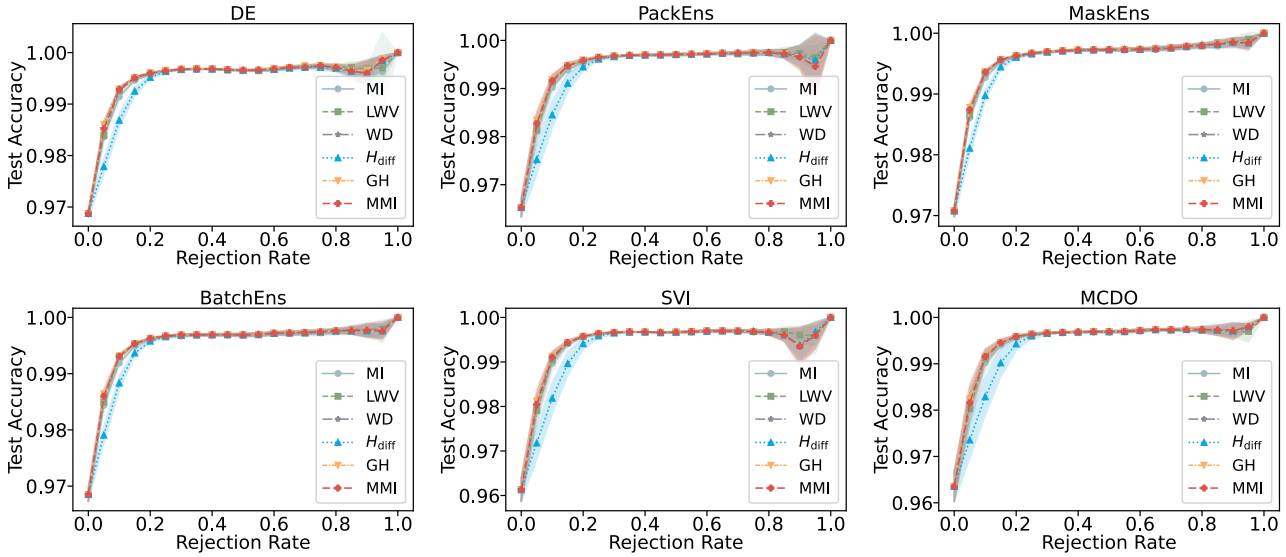

Figure A.18: Accuracy rejection curves using different uncertainty representations and measures on in-distribution **SVHN** test data across different underlying predictive models.

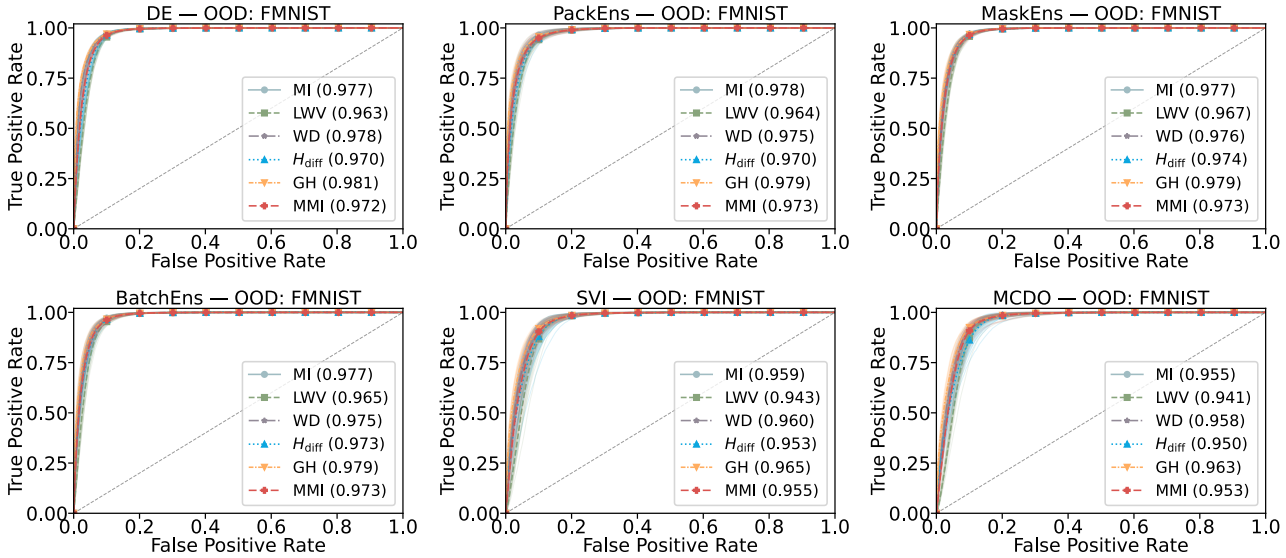

Figure A.19: ROC curves for OOD detection (**ID: SVHN v.s. OOD: FMNIST**) across uncertainty measures and backbone methods. Each panel corresponds to one backbone, with all uncertainty measures plotted together. Solid lines represent the mean TPR over 10 independent runs, shaded regions indicate ± 1 standard deviation, and faint lines show individual runs. Mean AUROC values are reported in the legend.

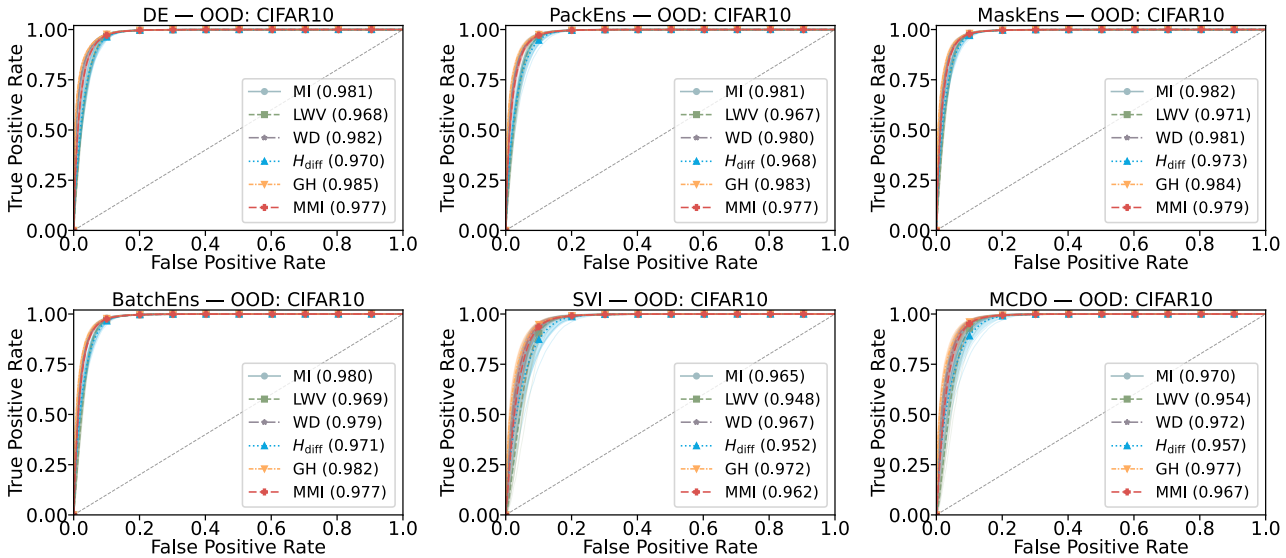

Figure A.20: ROC curves for OOD detection (**ID: SVHN v.s. OOD: CIFAR10**) across uncertainty measures and backbone methods. Each panel corresponds to one backbone, with all uncertainty measures plotted together. Solid lines represent the mean TPR over 10 independent runs, shaded regions indicate ± 1 standard deviation, and faint lines show individual runs. Mean AUROC values are reported in the legend.