# Self-Assessment and Robust Anomaly Detection with Bayesian Deep Learning

Giuseppina Carannante[†,1], Dimah Dera[*,1], Orune Aminul[*], Nidhal C. Bouaynaya[†] and Ghulam Rasool[‡]
[*]*Department of Electrical and Computer Engineering*, University of Texas Rio Grande Valley, Brownsville, TX 78520
[†]*Department of Electrical and Computer Engineering*, Rowan University, Glassboro, NJ 08028
[‡]*Machine Learning Department*, Moffit Cancer Center, Tampa, FL 12902

***Abstract*—Deep Learning (DL) models have achieved or even surpassed human-level accuracy in several areas, including computer vision and pattern recognition. The state-of-art performance of DL models has raised the interest in using them in real-world applications, such as disease diagnosis and clinical decision support systems. However, the challenge remains the lack of trustworthiness and reliability of these DL models. The detection of incorrect decisions or flagging suspicious input samples is essential for the reliability of machine learning models. Uncertainty estimation in the output decision is a key component in establishing the trustworthiness and reliability of these models. In this work, we use Bayesian techniques to estimate the uncertainty in the model's output and use this uncertainty to detect distributional shifts linked to both input perturbations and labels shifts. We use the learned uncertainty information (i.e., the variance of the predictive distribution) in two different ways to detect anomalous input samples: 1) a static threshold based on average uncertainty of a model evaluated on the clean test data, and 2) a statistical threshold based on the significant increase in the average uncertainty of the model evaluated on corrupted (anomalous) samples. Our extensive experiments demonstrate that both approaches can detect anomalous samples. We observe that the proposed thresholding techniques can distinguish misclassified examples in the presence of noise, adversarial attacks, anomalies or distributional shifts. For example, when considering corrupted versions of MNIST and CIFAR-10 datasets, the rate of detecting misclassified samples is almost twice as compared to Monte-Carlo-based approaches.**

***Index Terms*—Bayesian Deep Learning, Uncertainty Estimation, Anomaly Detection and Distributional Shift.**

## I. Introduction

Despite the recent flourishing integration of Deep Learning (DL) models in many research areas, the successful deployment of these models remains limited in safety-critical real-world applications, e.g., autonomous driving or medical diagnoses [1]. The hindrance is linked to the lack of trustworthiness and reliability of the models' predictions. The absence of intrinsic quantitative methods to evaluate the model performance after deployment, i.e., self-assessment or self-evaluation, makes them unreliable and unsafe for use in critical application areas [2].

It is well known that DL models can deliver remarkable results when dealing with "*familiar*" inputs that are "*close*" to the training data points (in-distribution inputs) [3]. However, DL models tend to associate high soft-max output values (which are erroneously considered as model confidence), to incorrect predictions or unrecognizable inputs [4], [5]. A trustworthy system should be able to identify unknown input samples and alert human users for safe handling.

All machine learning models are built on the mathematical assumption that the data samples are independently and identically distributed (i.i.d.), i.e., the data samples are drawn from the same distribution during training, testing and after deployment. Unfortunately, this assumption rarely holds in real-world scenarios where testing data samples are noisy. In many domains, the distribution shift arises from the multi-modal nature of the data distribution [6], which makes the accurate prediction of all instances challenging. For example, solid tumors detected from magnetic resonance imaging (MRI) data are known to be heterogeneous and differ in their visual characteristics in every patient. Similarly, an autonomous vehicle may encounter an unusual object that is not similar to any object available in the training samples.

In the context of Artificial Intelligence (AI) safety, being able to deal with "*diverse*" data represents an important step toward promoting reliability [7]. Identifying incorrect predictions due to a perturbation in the input space, i.e., covariate shifts, or detecting unknown labels, i.e., semantic shifts, is of utmost importance [7]. The model should be able to evaluate its own predictions or be self-aware under the various distributional shifts rather than failing blindly without any warning.

Several methods have been proposed in the literature that focus on the problems of detecting misclassification and/or out-of-distribution (OOD) examples, identifying distributional shifts, and the generalizability and robustness of DL models [6], [8]–[19]. Depending on the research domain, application area, and the motivations, similarities and differences exist among various approaches proposed in the literature.

In this work, we consider the problem of anomaly detection in computer vision, including both input perturbations (covariate shifts) and label shifts (known as OOD or semantic shifts). We propose two automated anomaly detection approaches by capitalizing on the recent work in Bayesian neural networks (BNN) [16]–[20]. We leverage uncertainty information available in Bayesian convolutional neural networks (CNNs) to detect anomalies and OOD examples. In recent work on

[1] Giuseppina Carannante and Dimah Dera contributed equally to this work as first authors.

Bayesian CNNs, we have shown that the model's output uncertainty, attained by the variance of the predictive distribution, monotonically increases when inputs are corrupted with noise or adversarial attacks [16]–[18]. In this paper, we present two automated anomaly detection techniques based on the learned uncertainty in Bayesian CNNs: 1) a static threshold based on the average uncertainty of the model evaluated on clean (familiar) test data, and 2) a statistical threshold based on the statistically-significant increase in the average uncertainty of the model evaluated on corrupted (unfamiliar) test datasets. The specific contributions are summarized as follows,

- In light of the recent research in Bayesian uncertainty estimation, we propose two anomaly detection and self-assessment approaches that exploit the predictive variance learned by the Bayesian CNNs. Unlike prior work, we propose extensive simulations on both covariate and semantic shifts. We compare our proposed methods with Monte Carlo (MC)-based uncertainty estimation approaches [21], [22].
- Our first anomaly detection method is based on a static threshold defined over the predictive variance. For the second method, we conduct an exhaustive statistical analysis of the average output uncertainty under noisy conditions and select a statistical threshold based on the significant increase in the output uncertainty. We provide a detailed comparison of both approaches.
- Under covariate shifts, there is always an irreducible part of data to be misclassified by the model. We demonstrate the benefit of the two proposed approaches by identifying misclassified samples under covariate distributional shifts.

## II. Related Work

Detecting distributional shifts has attracted great attention in the literature, given its important relationship with AI safety and model reliability. Several authors have investigated the problems related to noisy datasets, adversarial attacks, and various OOD-related detection tasks, e.g., novelty detection and open set recognition [6]. On the one hand, some of the work that falls under "anomaly or OOD detection" focuses on the label or semantic shifts, i.e., the case when test examples are drawn from new classes that were not seen during the training [8]. On the other hand, some works discuss adversarial perturbations or noise as covariate shifts [13]. Our proposed work addresses the general problem of anomaly detection in computer vision that covers both covariate and semantic shifts.

In the context of model trustworthiness, it is important to detect all possible sources of misclassifications and errors. A similar goal is shared among several DL areas: identifying the deviations, anomalies, outliers, novelties, i.e., OOD samples, and making more *aware* predictions. Early work on anomaly detection used heuristic thresholds directly set on the maximum softmax value [8]. The intuition is that the pre-trained networks tend to produce lower softmax values for misclassified and unusual inputs. Some work increased the softmax gap between in-distribution and OOD by using a temperature scaling or adding adversarial perturbations or geometric transformations to the training inputs [9], [23], [24]. Other approaches obtained confidence scores by introducing an additional network output with one or more fully-connected layers or using Mahalanobis distance [10], [25]. Another line of work focused on generating a gap between in-distribution and anomalous data using auto-encoders and generative models [11], [26]–[29]. These approaches try to learn the distribution of "unusual" data by generating such inputs and exposing the models to these unusual inputs during training.

Bayesian approaches are well-suited for detecting distributional shifts and misclassified inputs since a confidence score is delivered in the form of uncertainty or equivalently the variability (variance or the second central moment) of the posterior distribution. Lakshminarayanan *et al.* proposed an ensemble model to increase robustness and studied the output uncertainty for the OOD data [14]. Malinin and Gales proposed an alternative training procedure to expose the model to the distribution of unusual inputs using Prior Networks [30], [31]. Wang and Aitchison [32] trained a BNN and used the outlier exposure technique proposed earlier by [24]. Leveraging uncertainty estimates to detect misclassification and OOD samples have been studied in [12]. Some authors investigated the relationship between uncertainty estimates and misclassification [33], while others focused solely on the response to distributional shifts [15]. To the best of our knowledge, none of the recent work explicitly targeted the detection of system failures in the face of both misclassification and distributional shifts (both covariate and semantic) using Bayesian approaches.

We develop an anomaly detection framework based on learning uncertainty in the model prediction in Bayesian CNNs. We propose two strategies for setting up thresholds on the learned uncertainty (variance of the predictive distribution) that successfully detect misclassifications and OOD samples under covariate and semantic shifts.

## III. Learning Uncertainty in Bayesian Neural Networks

In BNNs, the network parameters (weights and biases), $\mathcal{W}$, are interpreted as random variables with a prior distribution $\mathcal{W} \sim p(\mathcal{W})$. Once we observe the training dataset $\mathcal{D} = \{\mathbf{X}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{n}$, we use Bayes' rule to infer the posterior distribution $p(\mathcal{W}|\mathcal{D})$. By inferring the posterior, we can compute the predictive distribution, i.e., the distribution of any unseen data point $\tilde{\mathbf{X}}$ with the corresponding output $\tilde{\mathbf{y}}$,

$$p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \mathcal{D}) = \int p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \mathcal{W})\ p(\mathcal{W}|\mathcal{D})\ d\mathcal{W}. \tag{1}$$

The predictive distribution contains all information about the prediction. The mean represents the network prediction, while the variance is the uncertainty or confidence of the model attached to the same prediction. The exact Bayesian inference on neural networks is mathematically intractable due to the nonlinear functional form of neural networks and the high dimensionality of the parameter space [21], [34]. Variational

inference is a popular approach that approximates the posterior distribution with a parametrized variational distribution, $q_{\boldsymbol{\theta}}(\boldsymbol{\mathcal{W}})$, by minimizing the Kullback-Leibler (KL) divergence between the variational and true posterior distributions. The optimization objective function is known as the evidence lower bound that is given as follows,

$$\mathcal{L}(\boldsymbol{\theta}) = -\,\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{\mathcal{W}})}[\log(p(\mathcal{D}|\boldsymbol{\mathcal{W}})] + \mathrm{KL}(q_{\boldsymbol{\theta}}(\boldsymbol{\mathcal{W}})\big|\big|p(\boldsymbol{\mathcal{W}})). \quad (2)$$

Recently, Dera *et al.* proposed a variational density propagation (VDP) framework that propagates the mean and covariance matrix of the variational posterior distribution through layers of CNNs. The VDP framework approximates the mean and covariance after non-linear activation functions using either the first-order Taylor series approximation (termed as extended VDP or exVDP model) or unscented transformation (known as unVDP) [16], [18]. Later, Carannante *et al.* introduced ensemble density propagation (enVDP) that propagates random samples from the variational distribution across layers of CNNs and estimates the mean and covariance of the variational posterior after passing through each layer, including non-linear activation functions [17]. The propagated covariance at the output layer of a CNN provides a measure of learned confidence or uncertainty attached to each output prediction during and after training [16]–[18]. VDP CNNs have significantly improved the robustness of these models to noise and adversarial attacks. VDP models possess a self-assessment ability, i.e., they associate increasing values of predictive variance to increasing levels of noise or adversarial attacks [16]–[18]. The predictive variance is computed from the covariance matrix of the predictive distribution by considering the diagonal element corresponding to the predicted class. In this work, we establish that the Bayesian VDP models, i.e., exVDP, unVDP and enVDP, can detect anomalies using two proposed methods based on the learned uncertainty (variance of the predictive distribution) in Bayesian CNNs.

### A. Bayesian Learned Uncertainty

Figure 1 illustrates three different flavors of the VDP framework. VDP propagates the mean and covariance of the variational distribution through a non-linear activation function, $\psi$, in a CNN. The mean and covariance of the input feature map $\mathbf{z}$, i.e., $\boldsymbol{\mu}_{\mathbf{z}}$ and $\boldsymbol{\Sigma}_{\mathbf{z}}$, are propagated through $\psi$ to find the output feature map $\mathbf{g}$. All three types of VDP, i.e., exVDP, unVDP and enVDP, are built using a mathematical formulation of probability density function tracking, as used in extended, unscented, and ensemble Kalman filters, respectively [35], [36]. The exVDP model approximates the mean and covariance using the first-order Taylor series as follows,

$$\boldsymbol{\mu}_{\mathbf{g}} \approx \psi(\boldsymbol{\mu}_{\mathbf{z}}), \qquad \boldsymbol{\Sigma}_{\mathbf{g}} \approx \mathbf{J}_{\psi}\boldsymbol{\Sigma}_{\mathbf{z}}\mathbf{J}_{\psi}^{T}, \quad (3)$$

where $\mathbf{J}_{\psi}$ is the Jacobian matrix of $\psi$ with respect to $\mathbf{z}$ evaluated at $\boldsymbol{\mu}_{\mathbf{z}}$ [35].

The unVDP uses unscented transformation (UT) to propagate the mean and covariance after the non-linear function $\psi$ by carefully choosing a set of samples called *sigma points*. The total number of sigma points is $2d$, where $d$ indicates the dimension of the random vector $\mathbf{z}$. The *sigma points* are generated as follows [35],

$$\mathbf{z}_i = \boldsymbol{\mu}_{\mathbf{z}} + \tilde{\mathbf{z}}_i, \qquad i = 1, \cdots, 2d$$
$$\tilde{\mathbf{z}}_i = \left(\sqrt{d\ \boldsymbol{\Sigma}_{\mathbf{z}}}\right)_i^T \text{ and } \tilde{\mathbf{z}}_{i+d} = -\left(\sqrt{d\ \boldsymbol{\Sigma}_{\mathbf{z}}}\right)_i^T, \qquad i = 1, \cdots, d,$$

where, $\left(\sqrt{d\ \boldsymbol{\Sigma}_{\mathbf{z}}}\right)_i$ is the $i^{\text{th}}$ row of the matrix square root. The non-linear activation function $\psi$ transforms every single *sigma point*, i.e., $\mathbf{g}_i = \psi[\mathbf{z}_i]$, where $i = 1, \cdots, 2d$. Let $N = 2d$, the approximate mean and covariance of $\mathbf{g}$ are computed as,

$$\boldsymbol{\mu}_{\mathbf{g}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathbf{g}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{g}_i - \boldsymbol{\mu}_{\mathbf{g}})(\mathbf{g}_i - \boldsymbol{\mu}_{\mathbf{g}})^T. \quad (4)$$

The enVDP model performs stochastic sampling of $N$ random samples, $\mathbf{z}_i$, $i = 1, \cdots, N$. We pass each ensemble member $\mathbf{z}_i$ through the activation function $\psi$ and obtain $\mathbf{g}_i = \psi[\mathbf{z}_i]$. The approximate sample mean and covariance are computed using eq. 4.

Unlike other approaches, the three VDP frameworks do not rely on MC sampling to approximate the ELBO cost function (eq. 2). Additionally, the computational complexity of these approaches is comparable to the deterministic setting [18].

### B. Uncertainty Estimation Using Monte Carlo Methods

In the literature, there are other Bayesian approaches that use variational inference for estimating uncertainty in neural networks. Bayes-by Backprop (BBB) introduces a fully factorized Gaussian distribution over the parameters of fully-connected neural networks [21]. On the other hand, Monte Carlo Dropout (MC-Drop) by Gal and Ghahramani interprets dropout as a Bernoulli variational distribution over convolutional kernels [22]. The shared idea between these two approaches is the requirement of MC sampling during the testing/inference phase to estimate uncertainty. During training, one sample is drawn randomly from the variational posterior

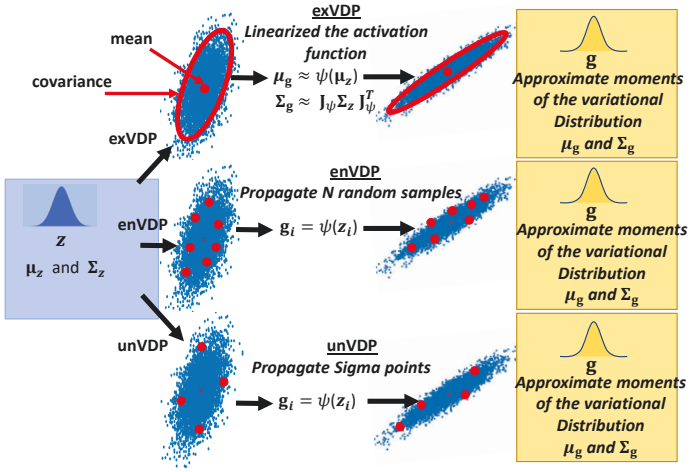


Fig. 1. A schematic illustration of the VDP framework by propagating the mean and covariance of the variational distribution through a non-linear activation function $\psi$ using: 1) the first-order Taylor series (exVDP) [16], 2) the unscented transformation (unVDP) [18], or 3) the ensemble sampling propagation (enVDP) [17].

and is passed forward through the network layers. Essentially, the uncertainty in the prediction is not learned during the training but rather estimated using MC sampling. This is in contrast to the VDP models, where we learn uncertainty during the training.

### C. Anomaly Detection Using Uncertainty

In Bayesian VDP models, the variance of the predictive distribution is the diagonal element of the propagated covariance matrix corresponding to the predicted class. These variance values increase when we add random noise or adversarial attacks to the test samples [16]–[18]. By considering this behavior of increasing uncertainty values under noisy conditions, we develop two threshold-based techniques to detect anomalous and OOD samples.

*1) Detection Using a Static Threshold:* We analyze the predictive variance during testing by computing the average predictive variance or uncertainty of correctly classified as well as misclassified clean test samples (familiar or in-distributional inputs). We notice that the average variance of the correctly classified samples is always considerably lower than that of the misclassified ones. Based on this analysis, we select the average predictive variance of correctly classified test samples as a "*Static*" threshold for detecting anomalies. Thus, any new input samples at the test time that produce variance values above the static threshold will be considered anomalous.

*2) Detection Using a Statistical Threshold:* We conduct a statistical analysis of the average predictive variance for both correctly and misclassified test samples during testing under Gaussian noise and adversarial attacks. The noise is measured by the signal-to-noise ratio (SNR) [37]. We perform pair-wise comparisons between the average predictive variance of test samples at each noise level and the variance at zero noise (clean test data) using the Wilcoxon signed-rank test. The variance value that is significantly higher (with a $99\%$ confidence level) than the variance at zero noise is selected as the "*Statistical*" threshold for detecting anomalies.

## IV. Methods and Experiments

We evaluate both anomaly detection approaches in three uncertainty propagation models, i.e., exVDP, unVDP and enVDP [16]–[18]. We compare these models with BBB and MC-Drop [21], [22]. We consider two benchmark datasets, MNIST and CIFAR-10 [38], [39]. We train all five models on the MNIST dataset and exVDP, unVDP, enVDP and MC-Drop models on the CIFAR-10 dataset. We use the same network architecture and hyper-parameters. BBB model is a fully-connected network that performs poorly on the CIFAR-10 dataset and is tested only on MNIST. For the MNIST dataset, we use a CNN with one convolution (32 kernels) layer and one fully-connected layer. The rectified linear unit is used as an activation function. We train models for 20 epochs, with a batch size of 50 and Adam optimizer. For the CIFAR-10 dataset, we use CNNs with 11 layers, 10 convolutional, 1 fully-connected, and 5 max-poling. The numbers of convolutional kernels are set to $32, 32, 32, 32, 64, 64, 64, 128, 128$ and $128$, respectively. We use 500 epochs, 50 batch-size, an exponential linear unit (ELU) activation and Adam optimizer. We use 20 MC samples to compute the uncertainty in the BBB and MC-Drop models.

We assess the performance of the proposed anomaly detection approaches, i.e., the static and statistical thresholds, under the covariate shift (by adding Gaussian noise and adversarial perturbations to the image input space during testing) and semantic shift (detecting an unknown label) with the Fashion MNIST [40] and SVHN [41] datasets. Besides testing under noise and adversarial attacks, we analyze the model behavior using two datasets consisting of corrupted MNIST (MNIST-C) [42] and corrupted CIFAR-10 (CIFAR-10-C) [43] samples to demonstrate the performance under covariate shift.

### A. Covariate Shifts

*1) Gaussian Noise and Adversarial attacks:* We establish the detection proficiency of the proposed anomaly detection methods under covariate shifts by adding different levels of Gaussian noise and adversarial attacks to the test datasets. We measure the noise level using SNR, where increasing the noise level results in a decreasing SNR. We used the fast gradient sign method (FGSM) to generate the attacks [44].

*2) Corrupted images:* We evaluate both proposed anomaly detection techniques on OOD samples of MNIST-C [42] and CIFAR-10-C datasets [43]. These datasets have 15 different types of data corruption. In MNIST-C, these corruptions are: 1) Glass Blur, 2) Motion Blur, 3) Zigzag, 4) Dotted Line, 5) Scale, 6) Spatter, 7) Brightness, 8) Shear, 9) Shot, 10) Stripe, 11) Translate, 12) Fog, 13) Rotate, 14) Canny Edge, 15) Impulse noise. In CIFAR-10-C, the corruptions are: 1) Brightness, 2) Contrast, 3) Defocus blur, 4) Elastic transform, 5) Fog, 6) Frost, 7) Gaussian blur, 8) Gaussian noise, 9) Glass blur, 10) Impulse noise, 11) JPEG compression, 12) Motion blur, 13) Pixelate, 14) Saturate, 15) Shot noise.

In this experiment, we are interested in detecting misclassified samples under covariate shifts. So, in the first simulation, we consider all test samples for each corruption type from both MNIST-C and CIFAR-10-C. Then, we randomly select 1000 samples from the test data under each type of corruption, shuffle them, and test the models on this new pooled corrupted test set. We define average true-misclassified rate (TMR) as the number of detected misclassified samples out of the total number of misclassified samples. Higher values of TMR refer to better detection efficiency. In contrast, the false-misclassified rate (FMR) is the number of correctly classified samples detected by the static or statistical threshold as misclassified out of the total number of correctly classified samples. Lower FMR values represent better performance.

### B. Semantic shift

The proposed anomaly detection approaches are evaluated under semantic shifts for detecting unknown labels. For the first experiment, the in-distribution data is MNIST, and we consider Fashion MNIST [40] as OOD. We shuffle the two test datasets and evaluate each model (initially trained on MNIST)

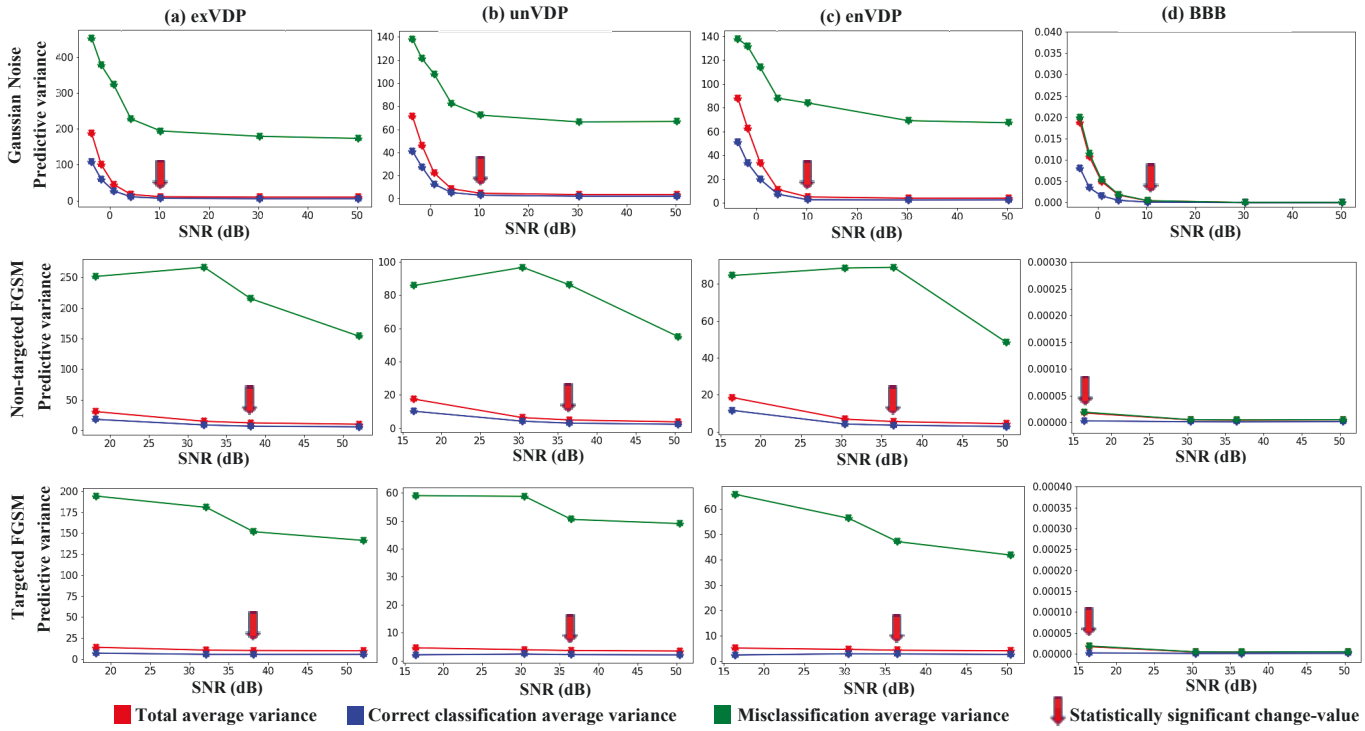


Fig. 2. Average predictive variance versus signal-to-noise ratio (SNR) for: 1) total number of test samples (red curves), 2) correctly classified samples (blue curves), and 3) misclassified samples (green curves). The variance-vs-SNR curves are plotted for exVDP, unVDP, enVDP and BBB models using the MNIST dataset corrupted with Gaussian noise as well as targeted and non-targeted adversarial attacks.

TABLE I
THE STATIC AND STATISTICAL THRESHOLDS ON MNIST AND CIFAR-10, AND THE AVERAGE VARIANCE ON FASHION MNIST AND SVHN.

| | MNIST | | F-MNIST | CIFAR-10 | | SVHN |
|---|---|---|---|---|---|---|
| Bayesian Models | Static Thre. | Statistical Thre. | Average Variance | Static Thre. | Statistical Thre. | Average Variance |
| exVDP | 5.52 | 10.21 | 163.57 | .036 | .043 | .269 |
| unVDP | 2.03 | 3.63 | 80.93 | .036 | .043 | .196 |
| enVDP | 2.68 | 4.12 | 75.47 | .004 | .0059 | .012 |
| MC-Drop | $6.6\times10^{-5}$ | $1.7\times10^{-4}$ | .0019 | .009 | .01 | .018 |
| BBB | $9.8\times10^{-6}$ | $1.7\times10^{-5}$ | $5.9\times10^{-5}$ | – | – | – |

on them. Similarly, in the second experiment, we consider the CIFAR-10 dataset as in-distribution and the SVHN test dataset [41] as OOD.

## V. RESULTS AND DISCUSSION

### A. Evaluation Under Covariate Shifts

*1) Gaussian noise and Adversarial attacks:* We compute the average predictive variance of all test samples, correctly classified samples and misclassified samples, and plot them against SNR for Gaussian perturbation as well as targeted and non-targeted adversarial attacks. Figures 2 and 3 show the variance-vs-SNR curves (interpreted from right to left) of exVDP, unVDP, enVDP, BBB, and MC-Drop for: 1) total number of test samples (red curves), 2) correctly classified samples (blue curves), and 3) misclassified samples (green curves) for MNIST and CIFAR-10, respectively. The red arrows refer to the noise levels where the variance becomes significantly higher (statistical threshold). We observe from the figures that the variance of misclassified samples (green curves) is considerably higher than that of the correctly classified or total number of test samples. The gap between the green curve (variance of misclassified samples) and the other two curves (correctly classified and the total number of test examples) is large for exVDP, unVDP and enVDP, making it feasible to detect misclassified samples under noise or attacks. Furthermore, the variance of correctly classified samples is always the smallest. The variance value corresponding to the statistical threshold (red arrows in Figs. 2 and 3) is also much lower than the variance of misclassified samples. Thus, both the static and statistical thresholds detect misclassified samples under these covariate shifts.

On the contrary, the gap between misclassification and correct classification variance is negligible for both BBB and MC-Drop models, as evident from Figs. 2(d) and 3(d). We believe that this observation may be linked to the propagation of uncertainty (covariance matrix of the variational distribution) in exVDP, unVDP and enVDP models. Variance propagation helps both the learning process and the proposed detection approaches. This results in self-aware models that are capable of detecting anomalies and perturbed samples.

*2) Corrupted MNIST and CIFAR-10:* In the first experiment, we plot the predictive variance for all test samples under the 15 corruption types described above for MNIST and CIFAR-10. We demonstrate in Fig. 4 that the average variance

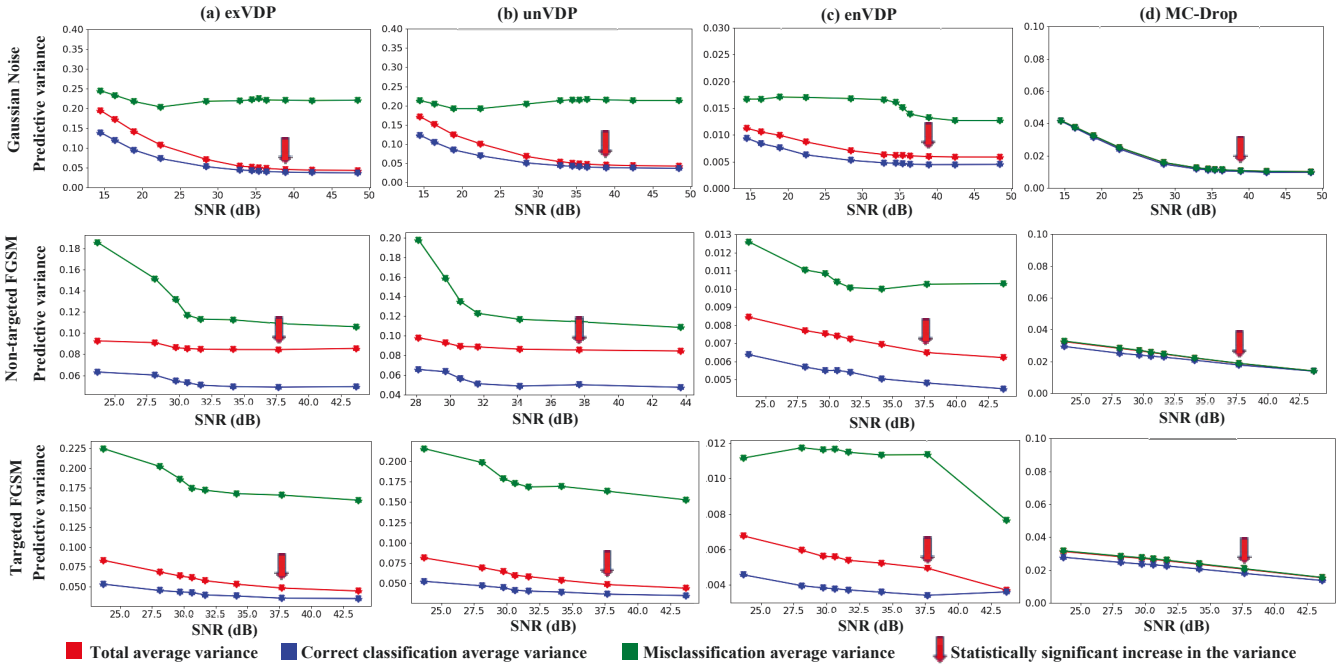


Fig. 3. Average predictive variance versus signal-to-noise ratio (SNR) for: 1) total number of test samples (red curves), 2) correctly classified samples (blue curves), and 3) misclassified samples (green curves). The variance-vs-SNR curves are plotted for exVDP, unVDP, enVDP and MC-Drop models using the CIFAR-10 dataset corrupted with Gaussian noise as well as targeted and non-targeted adversarial attacks.

of misclassified samples is much higher than that of correctly classified samples for exVDP, unVDP and enVDP models. The average variance of misclassified samples is always higher than the static and statistical thresholds. The exact values of the static and statistical thresholds for both MNIST and CIFAR-10 datasets are given in Table I.

From Fig. 4 and Table I, we note that the two proposed anomaly detection techniques are able to detect the misclassified samples by exVDP, unVDP, enVDP, BBB and MC-Drop. It is evident from Figs.4(d) and 4(h) that the average variance values of correctly classified samples for BBB and MC-drop models are high. This can be considered a false alarm from the BBB and MC-Drop models. We consider that the variance produced by BBB and MC-Drop (which was not learned during training) cannot differentiate between correctly classified and misclassified samples. This behavior results in poor generalization to OOD inputs and covariate shifts.

Table II shows the TMR and FMR for all models tested on the pooled corrupted datasets generated from MNIST-C and CIFAR-10-C as described in Section (IV-A2). We report the TMR and FMR values for both the static and the statistical threshold. We notice that exVDP, unVDP, and enVDP present high TMR (above 80%) and low FMR (below 26%) for MNIST and below 33% for CIFAR-10) as opposed to BBB and MC-Drop models.

### B. Evaluation Under Semantic Shift

We count how many test examples from MNIST as well as Fashion MNIST whose predicted uncertainty (or predictive variance) is above the static and statistical thresholds. Similarly, we count the number of samples from CIFAR-10 and

TABLE II
AVERAGE TRUE-MISCLASSIFIED RATE (TMR) AND FALSE-MISCLASSIFIED RATE (FMR) ON CORRUPTED MNIST AND CIFAR-10. THE TMR REFERS TO THE RATE OF DETECTING MISCLASSIFIED SAMPLES, AND THE FMR REFERS TO THE RATE OF DETECTING CORRECTLY CLASSIFIED SAMPLES.

| Bayesian Models | MNIST | | CIFAR-10 | |
|---|---|---|---|---|
| | TMR | FMR | TMR | FMR |
| exVDP | 82% | 25% | 90% | 31% |
| unVDP | 84% | 23% | 90% | 32% |
| enVDP | 84% | 23% | 84% | 32% |
| MC-Drop | 40% | 34% | 51% | 46% |
| BBB | 35% | 26% | – | – |

SVHN datasets that have their variance above the thresholds. The numbers of samples from Fashion MNIST (Fig. 5(a)) or SVHN (Fig. 5(b)) whose variance values are above the thresholds are considered the true positive (TP) counts (OOD with shifted labels detected by the proposed thresholds). The numbers of samples from MNIST (Fig. 5(a)) or CIFAR-10 (Fig. 5(b)) whose variance values are above the thresholds are considered the false positive (FP) counts because they are clean samples (in-distribution). We observe from Fig. 5 that the TP counts are very high for both the static and the statistical thresholds, while the FP counts are very low except for the MC-Drop model (Fig. 5(b)). Thus, the two proposed anomaly detection approaches with the static and statistical thresholds are able to detect semantic shifts with a high detection rate. We compute the average predictive variance (uncertainty) of all test samples from Fashion MNIST and SVHN datasets and compare the values with each model's static and statistical thresholds (Table I). We note that the average variance is

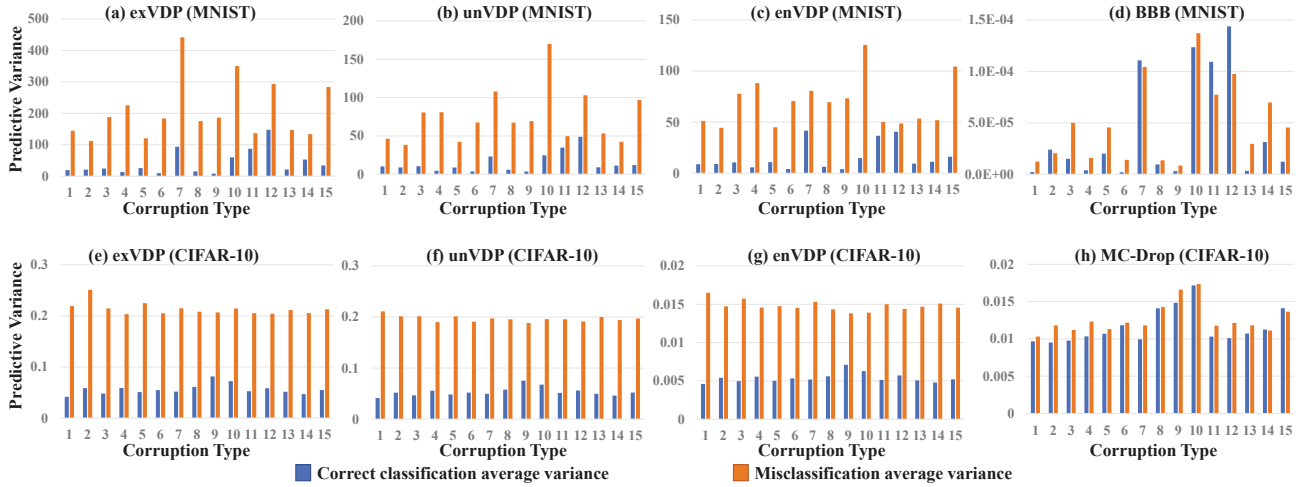


Fig. 4. Average predictive variance of correctly classified samples (blue bars) and misclassified samples (orange bars) for the 15 different corruptions in MNIST-C [42] (the first row) and CIFAR-10-C [43] (the second row).

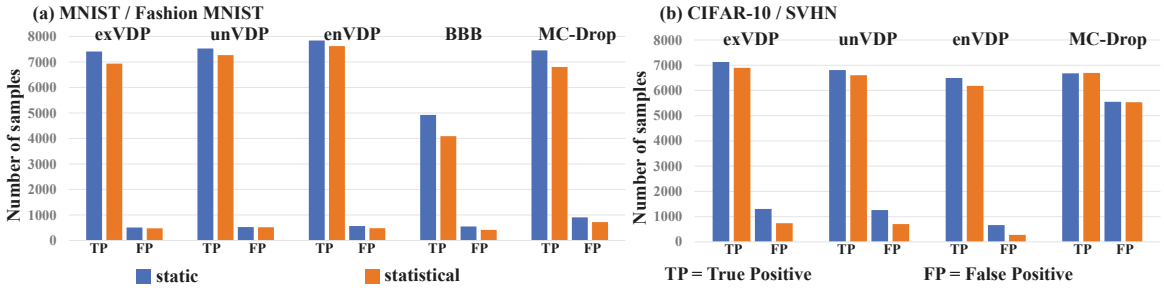


Fig. 5. (a) TP: The numbers of samples of Fashion MNIST dataset detected by the static and statistical thresholds (out-of-distribution or semantic shifts). FP: The number of samples of MNIST dataset detected by the static and statistical thresholds (in-distribution). (b) TP: The numbers of samples of SVHN dataset detected by the static and statistical thresholds (out-of-distribution or semantic shifts). FP: The number of samples of CIFAR-10 dataset detected by the static and statistical thresholds (in-distribution).

notably higher than the thresholds for all models, especially for exVDP, unVDP and enVDP models. Hence, models are clearly detecting the label distributional shift.

## VI. Conclusion and Future Work

This paper proposed two automated anomaly detection approaches based on the learned uncertainty (or predictive variance) of the variational density propagation frameworks in CNNs. We observed from prior work the growing behavior of uncertainty under noisy conditions. Based on this observation, we proposed two automated thresholds: 1) a static threshold based on the average predictive variance of correctly classified clean test samples (in-distribution samples); 2) a statistical threshold based on the Wilcoxon signed-rank test to detect the significant increase in the average variance under noisy conditions (out-of-distribution samples). We demonstrated in our extensive simulation that the two automated detection thresholds were able to recognize the covariate shifts, including Gaussian noise, targeted and non-targeted adversarial attacks, and corrupted MNIST and CIFAR-10 test samples (with 15 different types of corruptions). The misclassified samples by these different types of perturbations were detected with high rates by the static and statistical thresholds. The two thresholds also detected the semantic shifts when we trained the models on the MNIST and CIFAR-10 datasets and then tested them on the Fashion MNIST and SVHN datasets, respectively. In the future, we plan to test the two proposed thresholds with more datasets and larger network architectures. We also plan to investigate the possibility of learning the detection threshold during training, which may result in a higher detection rate.

## Acknowledgment

This work was supported by the National Science Foundation awards NSF CRII-2153413, NSF ECCS-1903466 and NSF OAC-2008690. We are also grateful to UK EPSRC support through EP/T013265/1 project NSF-EPSRC: ShiRAS. Towards Safe and Reliable Autonomy in Sensor Driven Systems, and NJ Health Foundation support through Award number PC 78-21.

## References

[1] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018. 1

[2] Evan Ackerman, "How drive.ai is mastering autonomous driving with deep learning," *IEEE Spectrum Magazine*, Mar. 2017. 1

[3] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021. 1

[4] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1765–1773. 1

[5] Anh Nguyen, Jason Yosinski, and Jeff Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 427–436. 1

[6] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu, "Generalized out-of-distribution detection: A survey," *arXiv preprint arXiv:2110.11334*, 2021. 1, 2

[7] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016. 1

[8] Dan Hendrycks and Kevin Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proceedings of 5th International Conference on Learning Representations, (ICLR)*, 2017. 1, 2

[9] Shiyu Liang, Yixuan Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *Proceedings of the 6th International Conference on Learning Representations, (ICLR)*, 2018. 1, 2

[10] Ryo Kamoi and Kei Kobayashi, "Why is the mahalanobis distance effective for anomaly detection?," *arXiv preprint arXiv:2003.00402*, 2020. 1, 2

[11] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems, (NIPS)*, 2019, vol. 32. 1, 2

[12] Bolian Li, Zige Zheng, and Changqing Zhang, "Identifying incorrect classifications with balanced uncertainty," *arXiv preprint arXiv:2110.08030*, 2021. 1, 2

[13] Saikiran Bulusu, Bhavya Kailkhura, Bo Li, Pramod K Varshney, and Dawn Song, "Anomalous example detection in deep learning: A survey," *IEEE Access*, vol. 8, pp. 132330–132347, 2020. 1, 2

[14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proceedings of the 30th International Conference on Neural Information Processing Systems, (NIPS)*, 2017, vol. 30. 1, 2

[15] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," in *Proceedings of the 2nd International Conference on Neural Information Processing Systems, (NIPS)*, 2019, vol. 32. 1, 2

[16] Dimah Dera, Ghulam Rasool, and Nidhal Bouaynaya, "Extended variational inference for propagating uncertainty in convolutional neural networks," in *IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019, pp. 1–6. 1, 2, 3, 4

[17] Giuseppina Carannante, Dimah Dera, Ghulam Rasool, Nidhal C. Bouaynaya, and Lyudmila Mihaylova, "Robust learning via ensemble density propagation in deep neural networks," in *IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2020, pp. 1–6. 1, 2, 3, 4

[18] Dimah Dera, Nidhal Carla Bouaynaya, Ghulam Rasool, Roman Shterenberg, and Hassan M. Fathallah-Shaykh, "Premium-cnn: Propagating uncertainty towards robust convolutional neural networks," *IEEE Transactions on Signal Processing*, vol. 69, pp. 4669–4684, 2021. 1, 2, 3, 4

[19] Giuseppina Carannante, Nidhal C Bouaynaya, and Lyudmila Mihaylova, "An enhanced particle filter for uncertainty quantification in neural networks," in *IEEE 24th International Conference on Information Fusion (FUSION)*, 2021, pp. 1–7. 1

[20] Dimah Dera, Ghulam Rasool, Nidhal C. Bouaynaya, Adam Eichen, Stephen Shanko, Jeff Cammerata, and Sanipa Arnold, "Bayes-SAR Net: Robust SAR image classification with uncertainty estimation using Bayesian convolutional neural network," in *Proceedings of the IEEE International Radar Conference*, 2020. 1

[21] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra, "Weight uncertainty in neural networks," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning, (ICML)*, 2015, vol. 37, pp. 1613–1622. 2, 3, 4

[22] Yarin Gal and Zoubin Ghahramani, "Bayesian convolutional neural networks with Bernoulli approximate variational inference," in *Proceedings of 4th International Conference on Learning Representations, (ICLR) workshop track*, 2016. 2, 3, 4

[23] Izhak Golan and Ran El-Yaniv, "Deep anomaly detection using geometric transformations," in *Proceedings of the 31st International Conference on Neural Information Processing Systems, (NIPS)*, 2018, vol. 31. 2

[24] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich, "Deep anomaly detection with outlier exposure," in *Proceedings of the 7th International Conference on Learning Representations, (ICLR)*, 2019. 2

[25] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira, "Generalized ODIN: Detecting out-of-distribution image without learning from out-of-distribution data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10951–10960. 2

[26] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," *arXiv preprint arXiv:1711.09325*, 2017. 2

[27] Kumar Sricharan and Ashok Srivastava, "Building robust classifiers through generation of confident out of distribution examples," *arXiv preprint arXiv:1812.00239*, 2018. 2

[28] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto, "Generative probabilistic novelty detection with adversarial autoencoders," in *Proceedings of the 31st International Conference on Neural Information Processing Systems, (NIPS)*, 2018, vol. 31. 2

[29] Felix Govaers and Paul Baggenstoss, "On a detection method of adversarial samples for deep neural networks," in *Proceedings of the IEEE 24th International Conference on Information Fusion (FUSION)*, 2021, pp. 1–5. 2

[30] Andrey Malinin and Mark Gales, "Predictive uncertainty estimation via prior networks," in *Proceedings of the 31st International Conference on Neural Information Processing Systems, (NIPS)*, 2018, vol. 31. 2

[31] Andrey Malinin and Mark Gales, "Reverse KL-divergence training of prior networks: Improved uncertainty and adversarial robustness," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems, (NIPS)*, 2019, vol. 32. 2

[32] Xi Wang and Laurence Aitchison, "Bayesian OOD detection with aleatoric uncertainty and outlier exposure," in *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2021. 2

[33] Hamzeh Asgharnezhad, Afshar Shamsi, Roohallah Alizadehsani, Abbas Khosravi, Saeid Nahavandi, Zahra Alizadeh Sani, Dipti Srinivasan, and Sheikh Mohammed Shariful Islam, "Objective evaluation of deep uncertainty predictions for covid-19 detection," *Scientific Reports*, vol. 12, no. 1, pp. 1–11, 2022. 2

[34] Alex Graves, "Practical variational inference for neural networks," in *Proceedings of the 24th International Conference on Systems, (NIPS)*, 2011, pp. 2348–2356. 2

[35] Dan Simon, *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*, Wiley-Interscience, 2006. 3

[36] Simon J Julier and Jeffrey K Uhlmann, "New Extension of the Kalman Filter to Nonlinear Systems," in *Signal processing, sensor fusion, and target recognition VI*. International Society for Optics and Photonics, 1997, vol. 3068, pp. 182–193. 3

[37] Daniel Sage, "Snr, psnr, rmse, mae," Jul 2017. 4

[38] L. Deng, "The MNIST database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, pp. 141–142, 2012. 4

[39] A. Krizhevsky, "Learning multiple layers of features from tiny images," *Master's thesis, Computer Science Department, University of Toronto*, 2009. 4

[40] Han Xiao, Kashif Rasul, and Roland Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017. 4

[41] 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. 4, 5

[42] Norman Mu and Justin Gilmer, "MNIST-C: A robustness benchmark for computer vision," *arXiv preprint arXiv:1906.02337*, 2019. 4, 7

[43] Dan Hendrycks and Thomas Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proceedings of 7th International Conference on Learning Representations, (ICLR)*, 2019. 4, 7

[44] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of 3rd International Conference on Learning Representations, (ICLR)*, 2015. 4