

# Open Set Recognition Through Deep Neural Network Uncertainty: Does Out-of-Distribution Detection Require Generative Classifiers?

Martin Mundt, Iuliia Pliushch, Sagnik Majumder and Visvanathan Ramesh  
 Goethe University, Frankfurt, Germany

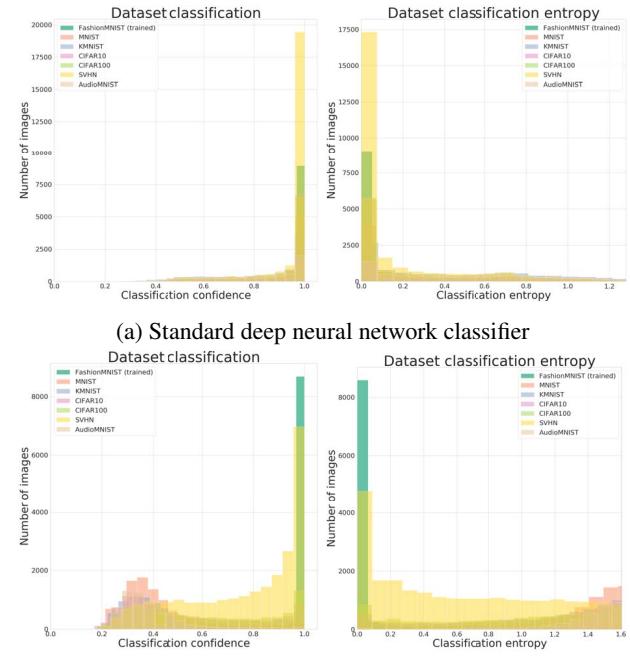
{mmundt, pliushch, vramesh}@em.uni-frankfurt.de majumder@ccc.cs.uni-frankfurt.de

## Abstract

We present an analysis of predictive uncertainty based out-of-distribution detection for different approaches to estimate various models' epistemic uncertainty and contrast it with extreme value theory based open set recognition. While the former alone does not seem to be enough to overcome this challenge, we demonstrate that uncertainty goes hand in hand with the latter method. This seems to be particularly reflected in a generative model approach, where we show that posterior based open set recognition outperforms discriminative models and predictive uncertainty based outlier rejection, raising the question of whether classifiers need to be generative in order to know what they have not seen.

## 1. Introduction

A particular challenge of modern deep learning based computer vision systems is a neural network's tendency to produce outputs with high confidence when presented with task unrelated data. Early works have identified this issue and have shown that methods employing forms of thresholding a neural network's softmax confidence are generally not enough for rejection of unknown inputs [15]. Recently, deep learning methods for approximate Bayesian inference [12, 5, 10, 5], such as deep latent variable models [12] or Monte Carlo dropout (MCD) [5], have opened the pathway to capturing neural network uncertainty. Access to these uncertainties comes with the promise of allowing to separate what a model is truly confident about through output variability. However, misclassification is not prevented and in a Bayesian approach uncertain inputs are not necessarily unknown and vice versa unknowns do not necessarily appear as uncertain [3]. This has recently been observed on a large empirical scale [19] and figure 1 illustrates this challenge. Here we show the prediction confidence and entropy of two deep residual neural networks [7, 23] trained on FashionMNIST [22] as obtained through a standard feed-forward pass and variational inference using 50 MCD samples. Neither



(b) Approximate variational inference with average over 50 Monte Carlo dropout stochastic forward passes

Figure 1: Classification confidence and entropy for deep neural network classifiers with and without approximate variational inference. Models have been trained on Fashion-MNIST and are evaluated on out-of-distribution datasets.

of the approaches is able to avoid over-confident predictions on previously unseen datasets, even if MCD fares much better in separating the distributions.

A different thread for open-set recognition in deep neural networks is through extreme-value theory (EVT) based meta-recognition [21, 2]. When applied to a neural network's penultimate feature representation, it has originally been shown to improve out-of-distribution (OOD) detection in contrast to simply relying on a neural network's output values. We have recently extended this approach by adapting EVT to each class' approximate posterior in a latent

variable model for continual learning [16]. However, EVT based open set recognition and capturing epistemic uncertainty need not be seen as separate approaches. In this work we thus empirically demonstrate that:

1. combining the benefit of capturing a model’s uncertainty with EVT based open set recognition outperforms out-of-distribution detection using prediction uncertainty on a variety of classification tasks.
2. moving to a generative model, which in addition to the label distribution  $p(\mathbf{y})$  also approximates the data distribution  $p(\mathbf{x})$ , results in similar prediction entropy but further improves the latent based EVT approach.

## 2. Variational open set neural networks

We consider three different models for which we investigate open set detection based on both prediction uncertainty as well as the EVT based approach. The simplest model is a standard deep neural network classifier. Such a model however doesn’t capture epistemic uncertainty. We thus consider variational Bayesian inference with neural networks consisting of an encoder with variational parameters  $\theta$  and a linear classifier  $p_\xi(\mathbf{y}|\mathbf{z})$  that gives the probability density of target  $\mathbf{y}$  given a sample  $\mathbf{z}$  from the approximate posterior  $q_\theta(\mathbf{z}|\mathbf{x})$ . We optionally also consider the addition of a probabilistic decoder  $p_\phi(\mathbf{x}|\mathbf{z})$  that returns the probability density of  $\mathbf{x}$  under the generative model. With the added decoder we thus learn a joint generative model  $p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{y}|\mathbf{z})p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ . These models are trained by optimizing the following variational evidence lower-bound:

$$\mathcal{L}(\theta, \phi, \xi) = \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})} [\log p_\phi(\mathbf{x}|\mathbf{z}) + \log p_\xi(\mathbf{y}|\mathbf{z})] - \beta KL(q_\theta(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})) \quad (1)$$

Here  $\beta$  is an additional parameter that weighs the contribution of the Kullback-Leibler divergence between approximate posterior  $q_\theta(\mathbf{z}|\mathbf{x})$  and prior  $p(\mathbf{z})$  as suggested by the authors of  $\beta$ -Variational Autoencoder [8]. We can summarize the considered models as follows:

1. Standard discriminative neural network classifier that maximizes  $\log p_\theta(\mathbf{y}|\mathbf{x})$  (not described by equation 1).
2. Variational discriminative classifier with graph  $\mathbf{x} \rightarrow \mathbf{z} \rightarrow \mathbf{y}$ . Maximizes the lower-bound to  $p(\mathbf{y})$  as given by equation 1 without the  $\phi$  dependent (blue) term.
3. Variational generative model as described by equation 1 with generative process  $p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{y}|\mathbf{z})p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ . In addition to  $p(\mathbf{y})$ , also jointly maximizes the variational lower-bound to  $p(\mathbf{x})$ .

Following a variational formulation, the second and third model have natural means to capture epistemic uncertainty,

---

**Algorithm 1 Open set recognition calibration for deep variational neural networks.** A Weibull model fit of tail-size  $\eta$  is conducted to bound the per class approximate posterior. Per class  $c$  Weibull models  $\rho_c$  with their respective shift  $\tau_c$ , shape  $\kappa_c$  and scale  $\lambda_c$  parameters are returned.

---

**Require:** Trained encoder  $q_\theta(\mathbf{z}|\mathbf{x})$  and classifier  $p_\xi(\mathbf{y}|\mathbf{z})$   
**Require:** Classifier probabilities  $p_\xi(\mathbf{y}|\mathbf{z})$  and samples from the approximate posterior  $\mathbf{z}(\mathbf{x}^{(i)}) \sim q_\theta(\mathbf{z}|\mathbf{x}^{(i)})$  for each training dataset example  $\mathbf{x}^{(i)}$   
**Require:** For each class  $c$ , let  $\mathbf{S}_c^{(i)} = \mathbf{z}(\mathbf{x}_c'^{(i)})$  for each correctly classified training example  $\mathbf{x}_c'^{(i)}$   
1: **for**  $c = 1 \dots C$  **do**  
2:     **Get per class latent mean**  $\bar{\mathbf{S}}_c = \text{mean}(\mathbf{S}_c^{(i)})$   
3:     **Weibull model**  $\rho_c = \text{Fit Weibull} (||\mathbf{S}_c - \bar{\mathbf{S}}_c||, \eta)$   
4: **Return** means  $\bar{\mathbf{S}}$  and Weibull models  $\rho$

---

**Algorithm 2 Open set probability estimation for unknown inputs.** Data points are considered statistical outliers if a Weibull model’s cumulative distribution function’s (CDF) probability value exceeds a task specific prior  $\Omega_t$ .

---

**Require:** Trained encoder  $q_\theta(\mathbf{z}|\mathbf{x})$   
**Require:** Per class latent mean  $\bar{\mathbf{S}}_c$  and Weibull model  $\rho_c$ , each with parameters  $(\tau_c, \kappa_c, \lambda_c)$   
**For a novel input example  $\hat{\mathbf{x}}$  sample  $\mathbf{z} \sim q_\theta(\mathbf{z}|\hat{\mathbf{x}})$**   
2: **Compute distances to  $\bar{\mathbf{S}}_c$ :**  $d_c = ||\bar{\mathbf{S}}_c - \mathbf{z}||$   
**for**  $c = 1 \dots C$  **do**  
4:     **Weibull CDF**  $\omega_c(d_c) = 1 - \exp \left( -\frac{||d_c - \tau_c||}{\lambda_c} \right)^{\kappa_c}$   
**Reject input** if  $\omega_c(d_c) > \Omega_t$  for any class  $c$ .

---

i.e. uncertainty that could be lowered by training on more data. Drawing multiple samples  $\mathbf{z} \sim q_\theta(\mathbf{z}|\mathbf{x})$  from the approximate posterior yields a distribution over the models’ outputs as specified by the expectation in 1. For all above approaches we can additionally place a prior distribution over the models’ weights to find a distribution  $q_\theta(\mathbf{W})$  for the weights posterior. This can be achieved by performing a dropout operation [20] at every weight layer and conducting approximate variational inference through multiple stochastic forward passes during evaluation. We do not consider variational autoencoders [12] that only maximize the variational lower-bound to  $p(\mathbf{x})$  (i.e. equation 1 without the blue term), as these models have been shown to be incapable of separating seen from unseen data in previous literature [17].

### 2.1. Open set meta-recognition

For a standard deep neural network classifier we follow the EVT based approach based on the features of the penultimate layer [2]. To bound the open-space risk of our variational models we follow the adaptation of this method to operate on the latent space and thus on the basis of the approx-

imate posterior in Bayesian inference [16]. In the Bayesian interpretation we obtain a Weibull distribution fit on the distances from the approximate posterior  $\mathbf{z}(\mathbf{x}) \sim q_\theta(\mathbf{z}|\mathbf{x})$  of each correctly classified training example. This leads to a bound on the regions of posterior high density as the tail of the Weibull distribution limits the amount of allowed low density space around these regions. Given such an estimate of the regions where the posterior has high density and the model can thus be trusted to make an informed decision, a novel unseen input example can be rejected according to the statistical outlier probability given the Weibull cumulative distribution function (CDF) between the unseen example’s posterior samples and their distances to the high density regions. The corresponding procedures to obtain the Weibull fits and estimate an unseen data-point’s outlier probability are outlined in algorithms 1 and 2.

### 3. Experiments and results

We base our encoder and optional decoder architecture on 14-layer wide residual networks [7, 23], in the variational cases with a latent dimensionality of 60. The classifier always consists of a single linear layer. We optimize all models using a mini-batch size of 128 and Adam [11] with a learning rate of 0.001, batch normalization [9] with a value of  $10^{-5}$ , ReLU activations and weight initialization according to He et. al [6]. For each convolution we include a dropout layer with a rate of 0.2 that we can use for MCD. We train all our model variants for 150 epochs until full convergence on three datasets: FashionMNIST [22], MNIST [14] and SVHN [18]. We do not apply any pre-processing or data augmentation. For the EVT based outlier rejection we fit Weibull models with a tail-size set to 5% of training data examples per class. The used distance measure is the cosine distance. After training we evaluate out of distribution detection on the other two datasets and additionally the KMNIST [4], CIFAR10 and 100 [13] and the non-image based AudioMNIST [1] datasets. For the latter we follow the authors’ steps to convert the audio data into spectrograms. To make this cross-dataset evaluation possible, we repeat all gray-scale datasets to a three channel representations and resize all images to  $32 \times 32$ .

#### 3.1. Results and discussion

We show outlier rejection curves using both prediction uncertainty as well as EVT based OOD recognition for the three network types trained on FashionMNIST in figure 2. Rejection rates for the variational approaches were computed using 100 approximate posterior samples to capture epistemic uncertainty. When looking at the prediction entropy, we can observe that a standard deep neural network classifier predicts over-confidently for all OOD data. While the EVT based approach alleviates this to a certain extent, the challenge of OOD detection still largely persists. Mov-

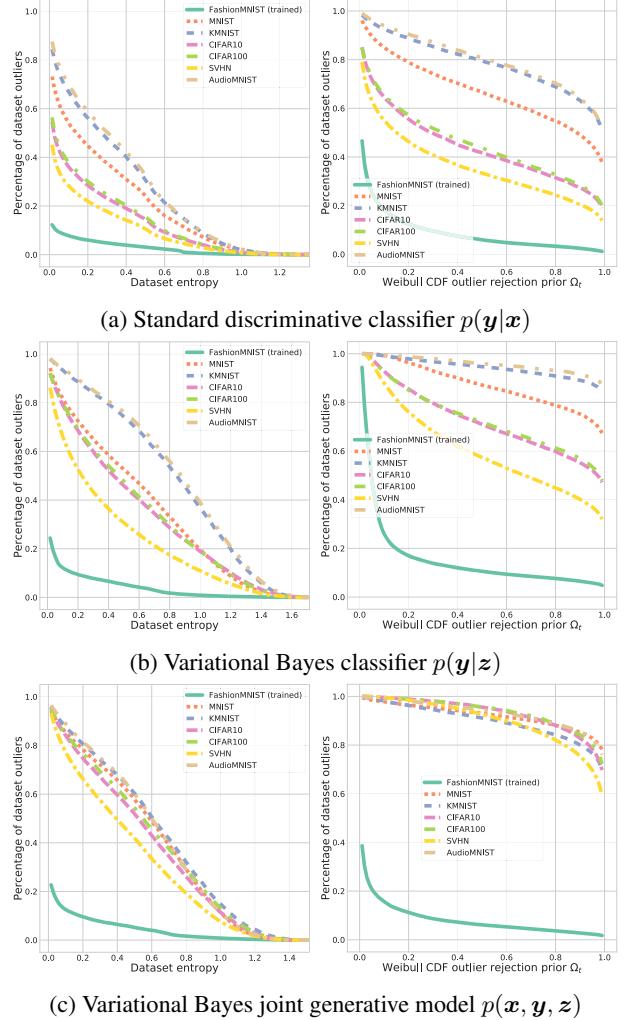


Figure 2: The three different models trained on FashionMNIST and evaluated on unseen datasets. For each model a pair of outlier rejection curves is shown. Left panels depict outlier rejection based on prediction entropy, whereas right panels show the EVT based open set recognition across the range of statistical outlier rejection priors  $\Omega_t$ .

ing to one of the variational models increases the entropy of OOD datasets, although not to the point where a separation from statistically inlying data is possible. Here, the EVT approach fares much better in achieving such separation. Nevertheless, this separation is only consistent across a wide range of rejection priors with the inclusion of the joint generative model. This is particularly important since this rejection prior has to be determined based on the original inlying validation data, as we can assume no access to OOD data upfront. Notice how this choice impacts rejection rates of the joint generative model to a much lesser extent. In addition we show the variational models of figure 2 panels (b) and (c) in figure 3 with 50 Monte Carlo dropout

Outlier detection at 95% trained dataset inliers (%)			FashionMNIST		MNIST		KMNIST		CIFAR10		CIFAR100		SVHN		AudioMNIST	
Trained	Model variant	Test acc.	Entropy	Latent	Entropy	Latent	Entropy	Latent	Entropy	Latent	Entropy	Latent	Entropy	Latent	Entropy	Latent
Fashion MNIST	standard discriminative	93.36	4.903	4.852	38.36	63.29	48.82	76.97	23.75	38.78	25.27	40.23	18.21	30.65	51.28	77.96
	variational discriminative	93.73	4.911	4.826	50.51	67.42	72.23	84.51	43.64	47.13	45.39	47.87	28.79	32.06	74.03	87.20
	variational generative	93.57	4.878	4.992	54.58	91.13	56.31	88.34	48.69	92.96	53.03	93.36	38.87	88.82	55.87	92.23
	variational discriminative - MCD	93.70	4.864	4.887	91.99	95.24	83.84	88.95	79.27	81.84	72.24	76.86	48.24	58.73	97.01	97.56
MNIST	variational generative - MCD	93.68	4.899	4.908	84.32	95.05	67.24	88.37	68.40	97.16	68.07	97.51	49.98	94.51	75.59	95.11
	standard discriminative	99.43	88.04	90.71	4.968	4.873	85.25	85.40	91.06	87.62	92.39	88.47	86.85	85.59	93.88	93.40
	variational discriminative	99.57	97.55	99.86	4.890	4.871	95.18	99.53	99.76	99.98	99.69	99.97	94.37	97.70	98.61	99.65
	variational generative	99.53	95.12	96.60	4.888	4.954	97.15	98.97	98.60	99.81	98.64	99.65	96.53	96.29	99.65	99.98
SVHN	variational discriminative - MCD	99.55	99.56	99.93	4.879	4.932	98.82	99.66	99.98	99.95	99.99	98.32	98.97	99.86	99.90	99.90
	standard discriminative	97.34	69.67	71.99	18.61	23.48	65.07	74.93	73.96	83.00	72.43	80.34	4.861	4.924	62.75	67.98
	variational discriminative	97.59	75.76	81.00	21.17	24.93	77.14	91.89	82.29	88.68	80.48	88.38	4.879	4.980	72.86	89.36
	variational generative	97.68	75.20	99.13	30.10	70.68	82.88	98.48	81.63	95.14	80.79	93.49	4.893	4.927	72.41	95.26
AudioMNIST	variational discriminative - MCD	97.57	84.97	89.71	95.27	94.97	84.48	90.26	85.86	94.94	85.78	93.46	4.962	4.922	81.66	88.61
	variational generative - MCD	97.58	83.73	93.53	100.0	100.0	98.32	97.57	82.16	93.03	80.40	92.77	4.893	4.910	88.16	94.53

Table 1: Test accuracies and outlier detection values of the three different network types described in section 2 when considering 95% of training validation data is inlying. Additional values are provided with Monte Carlo dropout (MCD). The variational approaches are reported with 100  $z \sim q_\theta(z|x)$  samples and the optional additional 50 MCD samples.

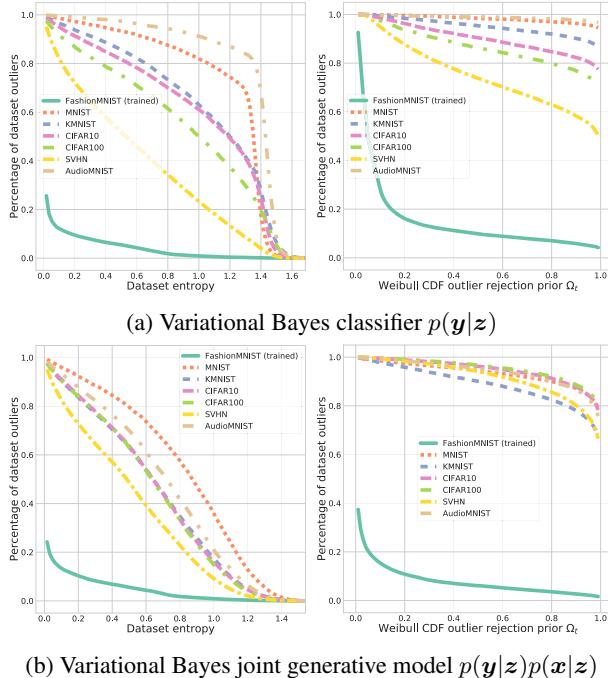


Figure 3: Pair of outlier rejection curves based on prediction entropy (left) and approximate posterior based statistical outlier rejection (right) in analogy to figure 2. Here, panels (a) and (b) correspond to panels (b) and (c) in figure 2 with additional variational Monte Carlo dropout inference.

samples. We have observed no substantial further benefits with more samples. Although this sampling can be computationally prohibitively expensive, we have included this comparison to give a better impression of how distributions on a neural network’s weights can aid in capturing uncertainty. In fact, we can observe that in both cases the prediction entropy is further increased, albeit still suffers from the

same challenge as outlined before. On the other hand, the EVT based approach profits similarly from MCD with the generative model still outperforming all other methods and achieving nearly perfect OOD detection.

We have quantified these results in table 1, where we report the network test accuracy as well as the outlier rejection rate with rejection priors and entropy thresholds determined according to categorizing 95 % of the trained dataset’s validation data as inlying. For all values we can observe that capturing epistemic uncertainty with variational Bayes approaches improves upon a standard neural network classifier both slightly in test accuracy as well as in OOD detection. This improvement is further apparent when using the EVT approach that outperforms OOD detection with prediction uncertainty in all cases. Lastly, the joint generative model is apparent to improve the EVT based OOD detection as the posterior now also explicitly captures information about the data distribution  $p(x)$ .

## 4. Conclusion

We have provided an analysis of prediction uncertainty and EVT based out-of-distribution detection approaches for different model types and ways to estimate a model’s epistemic uncertainty. While further larger scale evaluation is necessary, our results allow for two observations. First, whereas OOD detection is difficult based on prediction values even when epistemic uncertainty is captured, EVT based open set recognition based on a latent model’s approximate posterior can offer a solution to a large degree. Second, we might require generative models for open set detection in classification, even if previous work has shown that generative approaches that only model the data distribution seem to fail to distinguish unseen from seen data [17].

## References

[1] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek. Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals. *arXiv preprint arXiv: 1807.03418*, 2018.

[2] A. Bendale and T. E. Boult. Towards Open Set Deep Networks. *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] T. E. Boult, S. Cruz, A. Dhamija, M. Gunther, J. Henrydoss, and W. Scheirer. Learning and the Unknown : Surveying Steps Toward Open World Recognition. *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[4] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha. Deep Learning for Classical Japanese Literature. *Neural Information Processing Systems (NeurIPS), Workshop on Machine Learning for Creativity and Design*, 2018.

[5] Y. Gal and Z. Ghahramani. Dropout as a Bayesian Approximation : Representing Model Uncertainty in Deep Learning. *International Conference on Machine Learning (ICML)*, 48, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *International Conference on Computer Vision (ICCV)*, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *International Conference on Learning Representations (ICLR)*, 2017.

[9] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *International Conference on Machine Learning (ICML)*, 2015.

[10] A. Kendall and Y. Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Neural Information Processing Systems (NeurIPS)*, 2017.

[11] D. P. Kingma and J. L. Ba. Adam: a Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*, 2015.

[12] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *International Conference on Learning Representations (ICLR)*, 2013.

[13] A. Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, Toronto, 2009.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998.

[15] O. Matan, R. Kiang, C. E. Stenard, and B. E. Boser. Handwritten Character Recognition Using Neural Network Architectures. *4th USPS Advanced Technology Conference*, 2(5):1003–1011, 1990.

[16] Martin Mundt, Sagnik Majumder, Iuliia Pliushch, and Visvanathan Ramesh. Unified Probabilistic Deep Continual Learning through Generative Replay and Open Set Recognition. *arXiv preprint arXiv: 1905.12019*, 2019.

[17] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. Do Deep Generative Models Know What They Don't Know? *International Conference on Learning Representations (ICLR)*, 2019.

[18] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. *Neural Information Processing Systems (NeurIPS), Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[19] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *arXiv preprint arXiv: 1906.02530*, 2019.

[20] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research (JMLR)*, 15:1929–1958, 2014.

[21] M. R. P. Thomas, J. Ahrens, and I. Tashev. Probability Models For Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[22] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv: 1708.07747*, 2017.

[23] S. Zagoruyko and N. Komodakis. Wide Residual Networks. *British Machine Vision Conference (BMVC)*, 2016.