# Abnormality Detection in Histopathology via Density Estimation with Normalising Flows

**Nick Pawlowski**[1] NP716@IMPERIAL.AC.UK
**Ben Glocker**[1] B.GLOCKER@IMPERIAL.AC.UK
[1] *Imperial College London*

**Editors:** Under Review for MIDL 2021

## Abstract

Diagnosis of cancer often relies on the time-consuming examination of histopathology slides by expert pathologists. Automation via supervised deep learning methods require large amounts of pixel-wise annotated data that is costly to acquire. Unsupervised density estimation methods that rely only on the availability of healthy examples could cut down the cost of annotation. We propose to use residual flows as density estimator and compare different tests for out-of-distribution (OOD) detection. Our results suggest that unsupervised OOD detection is a viable approach for detecting suspicious regions in histopathology slides.

## 1. Out-of-distribution detection for histopathology

Current state-of-the-art methods in machine learning for histopathology use deep learning methods and are often trained on extracted patches from large pixel-wise annotated datasets (Ehteshami Bejnordi et al., 2017). Methods that work with image-level labels only have to overcome challenges which arise from large image size, as well as the low ratio of objects of interest (cancerous cells) to background in those images (Katharopoulos and Fleuret, 2019; Pawlowski et al., 2019). We aim to reframe this task as an out-of-distribution (OOD) detection task that detects pathologies as outliers under a statistical model of healthy data (Chen et al., 2018). We show that recent deep learning-based density estimation methods achieve competitive performance to fully supervised methods.

Recent work on normalising flows (Dinh et al., 2016) allows for density estimation on high dimensional image data. Normalising flows model a complex probability density $p(x)$ using a bijective transformation $f$ of a base distribution $\pi(u)$ as $x = f(u) \mid u \sim \pi(u)$. However, it has been shown that the estimated likelihood is not guaranteed to be a reliable estimate for detecting OOD samples (Nalisnick et al., 2018; Le Lan and Dinh, 2020; Kirichenko et al., 2020) and other OOD scoring metrics have been proposed (Choi et al., 2018; Nalisnick et al.; Ren et al., 2019). However, these methods either require to train multiple density estimation models (Choi et al., 2018; Ren et al., 2019) or can only handle batch-wise OOD detection. Instead we propose to cut down compute requirements during training by interpreting different points along the training trajectory as different models, similar to (Huang et al., 2017; Pawlowski et al., 2017; Maddox et al., 2019).

Given multiple density estimators $p_{\phi_1}, \dots, p_{\phi_n}$, we consider the following OOD scores: log-likelihood: $\log p_{\phi_i}$; expected log-likelihood: $\mathbb{E}_i[\log p_{\phi_i}]$, WAIC (Choi et al., 2018): $\mathbb{E}_i[\log p_{\phi_i}(x)] - \mathrm{Var}_i[\log p_{\phi_i}(x)]$; a variation on the typicality test from (Nalisnick et al.): $|\mathbb{E}_i[-\log p_{\phi_i}(x) - \mathbb{E}_{x' \sim X_{train}}[-\log p_{\phi_i}(x')]]|$; and lastly the variance of the log-likelihood

$\mathrm{Var}_i[\log p_{\phi_i}]$. Note that, different to the other scores, we expect the variance of inliers to be higher than that of outliers as we expect training of the models to mainly impact the behaviour for inlier samples, whereas the likelihood of outlier samples will mainly depend on the inductive biases of the model (Kirichenko et al., 2020).

## 2. Experiments & Discussion

We use the PatchCamelyon (PCam) dataset (Veeling et al., 2018) to test our concept of using normalising flows for OOD detection on histopathology images. We train our density estimator on all negative examples from the training set. We then calculate the area under the ROC curve (AUROC) to estimate the separability and classification performance of positive and negative patches. We train Residual Flows (Chen et al., 2019) using the original codeas a density estimator on the $32 \times 32$ px centre patches for 60 epochs and use the checkpoints at epochs 52-60 as the different density estimators. We compare our proposed method to a statistical baseline as well as a fully supervised learning method with varying amounts of positive patches. The statistical baseline estimates the probability of an inlier as $p(x) = \mathcal{N}(x[:,1] \mid \mu_1, \sigma_1)\mathcal{N}(x[:,2] \mid \mu_2, \sigma_2)\mathcal{N}(x[:,3] \mid \mu_3, \sigma_3)$, where $x[:,i]$ denotes the $i$th colour channel of the patch $x$ and $\mu_i, \sigma_i$ the corresponding empirical mean and variance.

Table 1: Comparison of AUROCs of correctly classified patches from the PCam test set. The single log-likelihood result is computed using the last model checkpoint. Typ. refers to our variation on the typicality test. GDensenet refers to the official baseline.

| Method | $\log p_\phi$ | $\mathbb{E}_i[\log p_{\phi_i}]$ | $\mathrm{Var}_i[\log p_{\phi_i}]$ | WAIC | Typ. | Gaussian | GDensenet |
|---|---|---|---|---|---|---|---|
| AUROC [%] | 53.4 | 81.6 | 92.4 | 25.3 | 61.8 | 31.8 | 96.3 |

Consistent with previous work we find that density estimation alone is not a reliable OOD detection metric, as seen with the performance of the Gaussian estimator and the regular log-likelihood. However, more sophisticated OOD scoring metrics achieve superior performance. Specifically, using the variance of the log-likelihood achieves an AUROC of 92.4%, being competitive compared to fully supervised methods such as GDensenet.

The current work is limited as it lacks thorough tuning of the Residual Flow and relies on the PatchCamelyon dataset which is derived from WSI that all contain regions with lesions. Future evaluations will therefore look into training on crops from the CAMELYON17 dataset and examine the performance of methods on whole-slide histopathology images to showcase their real-world applicability. Furthermore, it currently is not clear whether the suggested OOD scoring metric of the likelihood variance during training generalises to other problem domains or is specific to this particular dataset. Initial experiments on synthetic data as well as more common computer vision datasets suggest that this point requires more investigation as the computer vision experiments showed little separation using this metric. Nevertheless, we believe that observing the model behaviour over the course of training warrants future research into new ways of constructing OOD metrics.

## References

RTQ Chen, J Behrmann, D Duvenaud, et al. Residual flows for invertible generative modeling. In *NeurIPS*, 2019.

X Chen, N Pawlowski, M Rajchl, et al. Deep generative models in the real-world: An open challenge from medical imaging. *arXiv preprint arXiv:1806.05452*, 2018.

H Choi, E Jang, and AA Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

L Dinh, J Sohl-Dickstein, and S Bengio. Density estimation using Real NVP. *arXiv preprint arXiv:1605.08803*, 2016.

B Ehteshami Bejnordi, M Veta, PJ van Diest, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, 12 2017.

G Huang, Y Li, G Pleiss, et al. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

A Katharopoulos and F Fleuret. Processing megapixel images with deep attention-sampling models. In *ICML*, pages 3282–3291, 2019.

P Kirichenko, P Izmailov, and AG Wilson. Why normalizing flows fail to detect out-of-distribution data. *NeurIPS*, 33, 2020.

C Le Lan and L Dinh. Perfect density models cannot guarantee anomaly detection. In *"I Can't Believe It's Not Better!"NeurIPS 2020 workshop*, 2020.

WJ Maddox, P Izmailov, T Garipov, et al. A simple baseline for bayesian uncertainty in deep learning. *NeurIPS*, pages 13153–13164, 2019.

E Nalisnick, A Matsukawa, YW Teh, et al. Detecting out-of-distribution inputs to deep generative models using a test for typicality.

E Nalisnick, A Matsukawa, YW Teh, et al. Do deep generative models know what they don't know? In *ICLR*, 2018.

N Pawlowski, M Jaques, and B Glocker. Efficient variational bayesian neural network ensembles for outlier detection. *arXiv preprint arXiv:1703.06749*, 2017.

N Pawlowski, S Bhooshan, N Ballas, et al. Needles in haystacks: On classifying tiny objects in large images. *arXiv preprint arXiv:1908.06037*, 2019.

J Ren, PJ Liu, E Fertig, et al. Likelihood ratios for out-of-distribution detection. *arXiv preprint arXiv:1906.02845*, 2019.

BS Veeling, J Linmans, J Winkens, et al. Rotation equivariant CNNs for digital pathology. June 2018.