Uncertainty-Aware Vision Transformers for Medical Image Analysis

Franciskus Xaverius Erick[1], Mina Rezaei[2], Johanna Paula Müller[1], and Bernhard Kainz[1,3]

[1] Friedrich-Alexander University Erlangen-Nürnberg, Schloßplatz 4, 91054 Erlangen, Germany

[2] Ludwig Maximilian University of Munich, Geschwister-Scholl-Platz 1, 80539 München, Germany

[3] Department of Computing, Imperial College London, London SW7 2AZ, United Kingdom

Abstract. Vision transformers (ViTs) have emerged as strong alternatives to conventional convolutional neural networks (CNNs), owing to their scalability, enhanced generalization, and superior performance in out-of-distribution (OOD) scenarios. Despite their strengths, ViTs are prone to significant overfitting with scarce training data. This issue severely limits their reliability in critical applications, such as biomedical image analysis, where accurate uncertainty estimation is crucial. The challenge lies in the inherent lack of insight into the transformer network's confidence and uncertainty levels. To tackle this issue, we propose a novel stochastic vision transformer characterized by three components: 1) Stochastic elliptical Gaussian embedding which encodes uncertainty into the embedding of image patches, 2) a Fréchet Inception Distance (FID)-based attention mechanism for the Gaussian embeddings and 3) a FID-based regularization term, which imposes distance and uncertainty awareness into the learning of stochastic representations. We demonstrate the effectiveness of our method for in-distribution calibration and OOD detection experiments on the skin cancer dataset ISIC2019.

Keywords: Vision Transformers · Out-of-Distribution Detection.

1 Introduction

Recently, vision transformers (ViTs) [3] have emerged as a competitive alternative to Convolutional Neural Networks (CNNs), owing to their scalability and generalization ability when trained with large-scale natural image datasets [22]. Nevertheless, training ViTs with scarcely available medical datasets proves to be challenging, as a significant number of training samples is important for the performance of ViTs [18]. The tendency of overfitting and inferring unreliable, overconfident predictions severely hinders wide-scale applications of ViTs in safety-critical downstream applications. The development of reliable neural networks, which are robust against potential real-life distribution shifts, is thus important.

A key application with potential for broad adoption is the detection of skin cancer, which could be facilitated by everyday devices such as smartphones, allowing for patient-initiated screenings. According to a 2018 report by the WHO [21], over 14 million new cases of cancer were identified worldwide, leading to more than 9.6 million fatalities [12]. These figures underscore cancer as a predominant cause of mortality globally. Skin cancer, in particular, begins in the epidermis, the skin's outermost layer, making it visible to the naked eye and, therefore, one of the more detectable cancers. It remains a significant factor in global mortality rates. The accuracy and reliability of automated tests are paramount; early detection can be lifesaving, yet a high rate of false positives could cause undue anxiety and overwhelm healthcare systems with unnecessary consultations. However, to enable widely used automated tests, robust machine learning models are indispensable

Despite the availability of many available methods for enhancing neural networks' robustness, direct application of these methods on large transformer architectures with millions of parameters is challenging. Uncertainty estimation methods such as Deep Ensembles [11] and Bayesian Neural Networks [13] necessitate rigorous training procedures, which, combined with the large ViT architecture and the higher embedding dimensions of image modalities, render them computationally infeasible. One simple solution is to directly inject stochasticity into the model's parameters, thereby yielding diverse sets of solutions in place of the conventional deterministic point solutions, encouraging robust, uncertainty-aware training.

In this paper, we formalize a comprehensive method for robust, uncertainty-aware stochastic vision transformer encoders. Our contributions can be summarised as:

1. We propose a stochastic ViT encoder backbone featuring distributional Gaussian embeddings. Our approach ensures that stochasticity is propagated throughout all layers of the encoder, thereby encoding uncertainty at every stage. The interaction between the stochastic embeddings is assessed using an attention mechanism based on the Fréchet Inception Distance (FID).
2. We incorporate a novel regularization term based on FID into our model's training approach. This term is designed to foster uncertainty- and distance-awareness, compelling the network to embed similar embedding representations closer together. This addition not only enhances the model's ability to discern and represent the underlying data structure but also significantly improves its robustness by embedding a deeper understanding of the data's inherent variability and uncertainty.
3. We conduct comprehensive experiments to validate our method's capacity. We evaluate predictive accuracy and uncertainty quantification across three distinct scenarios: 1) In-distribution detection; 2) Out-of-distribution detection; 3) Few-shot detection; Our approach demonstrates a practicable trade-off between predictive performance and uncertainty estimation compared to the other baselines.

2 Method

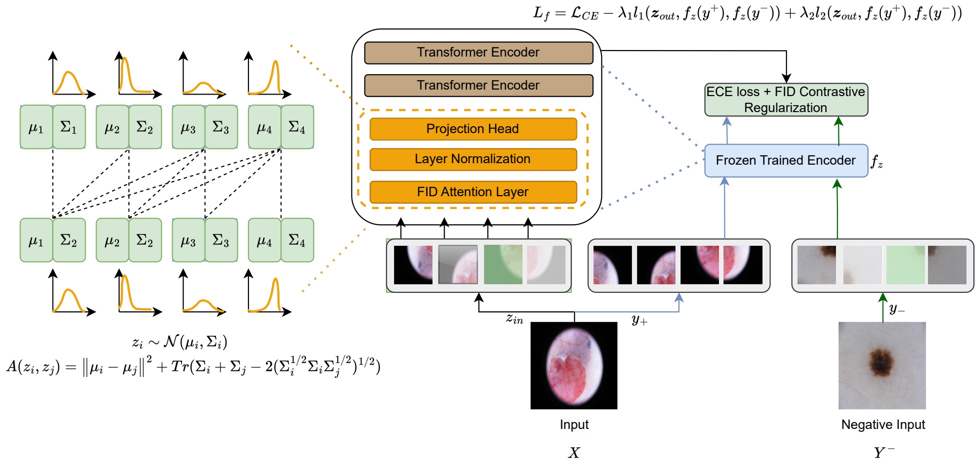

$L_f = \mathcal{L}_{CE} - \lambda_1 l_1(\boldsymbol{z}_{out}, f_z(y^+), f_z(y^-)) + \lambda_2 l_2(\boldsymbol{z}_{out}, f_z(y^+), f_z(y^-))$
μ_1 Σ_1 μ_2 Σ_2 μ_3 Σ_3 μ_4 Σ_4
μ_1 Σ_2 μ_2 Σ_2 μ_3 Σ_3 μ_4 Σ_4
$z_i \sim \mathcal{N}(\mu_i, \Sigma_i)$
$A(z_i, z_j) = \|\mu_i - \mu_j\|^2 + Tr(\Sigma_i + \Sigma_j - 2(\Sigma_i^{1/2}\Sigma_i\Sigma_j^{1/2})^{1/2})$
Transformer Encoder
Transformer Encoder
Projection Head
Layer Normalization
FID Attention Layer
ECE loss + FID Contrastive Regularization
Frozen Trained Encoder
f_z
z_{in}
y_+
y_-
Input
X
Negative Input
Y^-

Fig. 1: Overview of the stochastic vision transformer training pipeline. The input sample batch $\boldsymbol{X}$ is split into patches z_{in}, undergoing subsequent augmentations. The original non-augmented copy of the patches from the same input batch is taken as the positive examples y_+. Negative example batch y_- is obtained from sampling from the other classes.

2.1 Stochastic Gaussian Embedding

Conventional transformers embed input tokens into deterministic vector points. Alternatively, the input tokens can also be embedded as Gaussian distributions [20, 16, 5], leveraging uncertainty into the embedding representations of the data. Motivated by this, we embed each image patch as an elliptical Gaussian distribution with mean μ and variance σ vectors. In addition, we introduce separate positional encoding vectors for the mean μ and variance σ vectors. We formalize the stochastic Gaussian embeddings as $\boldsymbol{z}_\mu^0, ..., \boldsymbol{z}_\mu^L$ and $\boldsymbol{z}_\sigma^0, ..., \boldsymbol{z}_\sigma^L$. The stochastic embeddings are passed into our specialized stochastic encoder blocks, consisting of normalization layers, stochastic Fréchet Inception Distance attention layers, and a projection head. In this manner, we instill distributional uncertainty information throughout the whole architecture.

2.2 Fréchet Inception Distance Attention

The attention mechanism facilitates transformers to evaluate contextual correlations between the embedded vector components within each batch. In vision transformers, the input image is tokenized into tokens with the embedded dimension d and the sequence length l. The tokens are subsequently linearly projected into the query vectors $Q \in \mathbb{R}^{l \times h \times \frac{d}{h}}$, key $K \in \mathbb{R}^{l \times h \times \frac{d}{h}}$ vectors, and value $V \in \mathbb{R}^{l \times h \times \frac{d}{h}}$ vectors. The self-attention matrix between the vectors is evaluated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{1}$$

$$Q = W_Q\boldsymbol{x}, K = W_K\boldsymbol{x}, V = W_V\boldsymbol{x}; \quad W_Q, W_K, W_V \in \mathbb{R}^{d \times d} \tag{2}$$

The vanilla attention mechanism returns a single dot vector output and, thus is deterministic and provides no further insights into the uncertainty of the transformers. Our proposed Fréchet inception distance-based attention mechanism accommodates effective attending of the stochastic Gaussian embeddings. The stochastic embeddings are passed through linear layers of the projection head, maintaining information on the uncertainty through the Gaussian Q(Query), K(Key), and V(Value) representations $\boldsymbol{z}_{qkv} \sim N(\mu_{qkv}, \sigma_{qkv})$, formulated as follows,

$$\begin{aligned} \mu_{qkv} &= z_\mu W^\mu_{qkv} \\ \sigma_{qkv} &= \mathbf{ELU}(diag(\boldsymbol{z}_\sigma W^\sigma_{qkv})) + 1. \end{aligned} \tag{3}$$

The ELU activation enforces the positive definiteness of the covariance vectors. In place of the dot-product operation used in conventional deterministic transformers, the attention scores of the distributional embeddings Q and K embeddings are correspondingly calculated from the negative Fréchet Inception Distance between the embeddings. The Fréchet inception distance is a formulation of the 2-Wasserstein distance between Gaussian distributions [4][14][10][8]. This is formulated as follows:

$$\begin{gathered} A_{Q,K} = -(W_2^2(Q, K)) = -(\left\|\mu_Q - \mu_K\right\|^2 + Tr(\Sigma_Q + \Sigma_K - 2(\Sigma_Q^{1/2}\Sigma_Q\Sigma_K^{1/2})^{1/2})), \\ A_{\boldsymbol{z}} = \text{softmax}\left(\frac{A_{Q,K}}{\sqrt{d}}\right). \end{gathered} \tag{4}$$

for Gaussian distributional embeddings of $Q \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $K \sim \mathcal{N}(\mu_2, \Sigma_2)$ on $\mathbb{R}^d$. Tr represents the trace operator of the covariance matrices. The final attention scores for both the mean and value embeddings are evaluated by multiplying the attention scores with the value embeddings denoted in Equation 5 and Equation 6. This operation is iterated through the block depth, thus promoting effective learning of the spatial correlations between the embedded distributions.

$$\mathrm{A}_\mu = A_{\boldsymbol{z}} V_\mu, \tag{5}$$

$$\mathrm{A}_\sigma = A_{\boldsymbol{z}}^2 V_\sigma. \tag{6}$$

2.3 Fréchet Inception Distance Regularization

The training objective is divided into two primary elements: the deterministic component and the stochastic component. The deterministic component enforces the standard classification performance, while the novel stochastic component encourages the transformer to encode the image patches into distinct, robust representations through a Fréchet inception distance-based regularization. Unaugmented image patches are taken as the positive example images, while images from the other classes are considered as the negative example images. We consider one mini-batch of the output embeddings $\boldsymbol{z}_{out} \sim N(\mu_{out}, \sigma_{out})$, the trained stochastic encoder function f_z, the positive example y^+, and the negative example y^-. The resulting learning objective is denoted as follows:

$$L_f = \mathcal{L}_{CE} - \lambda_1 l_1(\boldsymbol{z}_{out}, y^+, y^-) + \lambda_2 l_2(\boldsymbol{z}_{out}, y^+, y^-), \tag{7}$$

whereby the distributional regularization terms l_1 and l_2 are given as follows:

$$l_1(\boldsymbol{z}_{out}, y^+, y^-) = \log(\sigma(W_2^2(\boldsymbol{z}_{out}, f_z(y^+)) - W_2^2(\boldsymbol{z}_{out}, f_z(y^-)))), \tag{8}$$

$$l_2(\boldsymbol{z}_{out}, y^+, y^-) = [W_2^2(\boldsymbol{z}_{out}, f_z(y^+)) - W_2^2(f_z(y^+), f_z(y^-))]_+. \tag{9}$$

$\mathcal{L}_{CE}$ denotes the deterministic cross-entropy classification loss term, $[x]_+ = \max(x, 0)$ denotes the hinge loss operator, y^+ and y^- denote the positive and negative examples. The regularization terms are tuned by the parameters λ_1 and λ_2, with the former regulating the distance between the stochastic embeddings of the input images and the examples, and the latter encouraging more distinctive embedding space separation between the positive and negative examples.

3 Experimental Settings

Network architecture. We conduct all our experiments with the ViT-B backbone, using the default ViT-B model parameters while optimizing the training hyperparameters through grid search. We train the network with a batch size of 256, a learning rate of 1×10^{-3}, and the stochastic regularization terms λ_1 of 0.1 and λ_2 of 0.01. The input images are resized to 224×224 and augmented with the following operations: horizontal and vertical flips, color jitters, and rotation. In addition, we perform further augmentation with AugMix [9]. We pre-train our models with the ImageNet-1k dataset.

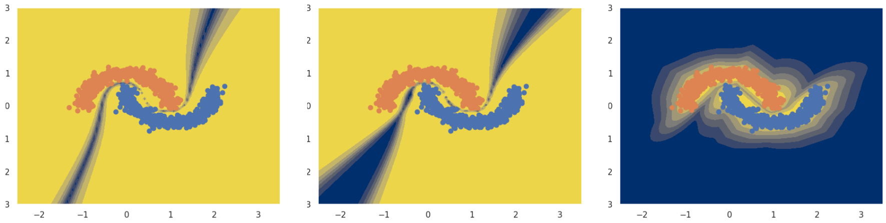
Data. We finetune our models with the skin lesion ISIC2019 dataset [19][2][1], consisting of eight distinct classes. We split the ISIC2019 dataset into the ID dataset and the near OOD dataset by allocating images from two classes with the least number of samples as the near OOD set. Images from the remaining six classes are allocated into the ID set. We perform 5-fold validations of our experiments, with a train-validation split of 80% training and 20% validation. We assess the performance of our method in two distinct tasks. In the first task, we evaluate the model's ID calibration and uncertainty quantification performance together with the OOD inference performance. In the second task, we perform few-shot training with respectively 1% and 10% of the training data available and investigate the ID calibration performance of our model. We performed further experiments with the DermaMNIST dataset with results summarized in the Appendix.

Evaluation metrics We report the performance metrics using the following notation: upward arrows signify that higher values are considered more optimal while downward arrows indicate the opposite. **Selective accuracy** ↑: Selective prediction allows the model to reject samples during inference, specifically samples with confidence levels below a specified confidence rejection threshold. The accuracy values are accumulated over the area under the curve for varying threshold values. A larger selective accuracy denotes the model's ability to perform confident uncertainty-aware inference in real-life safety-critical tasks. **ECE** ↓: Expected calibration error denotes the sum of the differences in the model's accuracy and confidence values for differing bin values. A lower ECE value denotes a more well-calibrated model that returns higher accuracy predictions with higher confidence. **NLL** ↓: Negative log likelihood between the predicted logit distribution and ground truth targets. A lower NLL value implies that the model returns logit distributions closer to the targets. We evaluate the near Out-of-Distribution robustness of our method with the **AUROC** ↑ metric: Area Under Receiver Operating Characteristic curve. This metric assesses the model's ability to discriminate positive and negative classes across various thresholds. A higher OOD AUROC value signifies the model's increased capacity to perform robust inference in OOD inference cases.

Compared methods For comparison, we perform experiments with the following baselines: 1) conventional ViT-B [3], 2) Deep ensembles of ViT-B with k=10 ensemble members [11], 3) MC-Dropout of ViT-B with a dropout rate of 0.1 [7], 4) Evidence Reconciled Neural Network (ERNN) for OOD detection [6], 5) Function Space Empirical Bayes (FSEB) [17]. Deep ensembles and MC-Dropout are considered state-of-the-art methods for uncertainty quantification, whereby stochasticity is induced from the ensemble inference for the former and from multiple inference runs with dropout for the latter. ERNN introduces a novel Evidence Reconciled Block in place of the conventional Softmax normalization at the output of the ResNet-18 encoders. FSEB incorporates function and parameter space regularization into the training of ResNet-18 encoders.

4 Results and Discussions

Qualitative analysis. To illustrate the quality of our embedding, we investigate the embedding representation quality of our method with the Two Moons dataset [15]. Figure 2 shows the uncertainty heat maps of the conventional ViT-B, deep ensembles, and our method trained on this dataset. Our method embeds uncertainty following the expectation that the model's uncertainty increases with increasing distance from the trained data points. While deep ensembles improve the uncertainty-awareness of the model in comparison to the ViT-B, the model still possesses low uncertainty at regions far away from the trained data points.



(a) ViT-B (b) Deep Ensembles (c) Our method

Fig. 2: The uncertainty heat map plots of the conventional ViT-B, deep ensembles, and our method on the Two-Moons dataset. The blue regions denote higher uncertainty, while yellow regions depict lower uncertainty.

In-Distribution (IND) and Out-Of-Distribution (OOD) performance. We trained the networks to an equivalent predictive performance and investigated their corresponding in-distribution calibration and out-of-distribution performance. For a fair comparison, we picked the predictive performance threshold of 75% top-1 accuracy, considering the compute requirements and specific hyperparameters required for the differing networks and baseline methods to reach their potential maximum accuracy values. We summarize our findings in Table 1. Our method achieves superior selective accuracy and calibration metrics values, showcasing the benefit of explicitly imparting uncertainty during training to the model's capability to reject predictions with higher estimated uncertainty. Our method also outperforms the other baseline methods in the OOD detection tasks for both datasets, emphasizing the significance of incorporating stochasticity and distance-awareness into ViTs.

Table 1: Results of In-Distribution predictive performance and calibration error for the ISIC2019 dataset. The best score for each metric is shown in **bold**.

	Sel. acc. (↑)	**NLL** (↓)	**ECE** (↓)	**OOD AUROC** (↑)
ViT-B	$90.322_{\pm 0.010}$	$0.834_{\pm 0.003}$	$0.106_{\pm 0.003}$	$61.27_{\pm 0.33}$
Ensembles ViT-B	$90.825_{\pm 0.015}$	$0.755_{\pm 0.002}$	$0.098_{\pm 0.002}$	$62.37_{\pm 0.27}$
MC-Dropout	$89.501_{\pm 0.013}$	$0.760_{\pm 0.002}$	$0.100_{\pm 0.003}$	$60.23_{\pm 0.31}$
ERNN	$90.579_{\pm 0.014}$	$1.328_{\pm 0.002}$	$0.450_{\pm 0.002}$	$61.78_{\pm 0.19}$
FSEB	$88.021_{\pm 0.010}$	$0.905_{\pm 0.001}$	$0.115_{\pm 0.002}$	$59.32_{\pm 0.23}$
Our method	$\mathbf{90.976}_{\pm \mathbf{0.012}}$	$\mathbf{0.724}_{\pm 0.001}$	$\mathbf{0.093}_{\pm \mathbf{0.001}}$	$\mathbf{63.52}_{\pm \mathbf{0.25}}$

Few-Shot Learning. To evaluate few-shot learning performance, we trained the models with 1% and 10% of the available training data, thereby emulating possible real-life scenarios of training data scarcity. The predictive and confidence calibration results from Table 2 highlight the enhanced training robustness of our method in the low data regime.

Table 2: Results of In-Distribution predictive performance and calibration error for few-shot learning with 1% and 10% of the ISIC2019 training data available. The best score for each metric is shown in **bold**.

	1% training data			10% training data		
	Sel. Acc.	ECE	NLL	Sel.Acc	ECE	NLL
ViT-B	$70.383_{\pm 0.032}$	$0.061_{\pm 0.007}$	$1.317_{\pm 0.006}$	$82.990_{\pm 0.012}$	$0.076_{\pm 0.003}$	0.936 ± 0.003
Ensembles	$71.297_{\pm 0.021}$	$\mathbf{0.051}_{\pm \mathbf{0.003}}$	$1.246_{\pm 0.003}$	$83.058_{\pm 0.005}$	$0.060_{\pm 0.002}$	0.922 ± 0.002
Ours	$\mathbf{73.146}_{\pm \mathbf{0.025}}$	$0.056_{\pm 0.005}$	$\mathbf{1.193}_{\pm \mathbf{0.004}}$	$\mathbf{83.196}_{\pm \mathbf{0.010}}$	$\mathbf{0.052}_{\pm \mathbf{0.002}}$	$\mathbf{0.917 \pm 0.002}$

Ablation study. We performed experiments with varying stochasticity degrees, incorporating combinations of: Gaussian Embedding(GE), FID-Attention(FID-A), and FID-Regularization(FID-R) into the conventional ViT-B. The results in Table 4 highlight the importance of the combinations of the three components to the training procedure of ViTs. Furthermore, we investigated the influence of the FID regularization coefficients λ_1 and λ_2. Our findings in Table 3 show that the optimal regularization rate hyperparameter combinations ensure the balance of the deterministic main learning objective and the uncertainty-aware regularization effect.

Table 3: Performance of our stochastic transformers with differing FID regularization parameters.

λ_1	λ_2	**Sel. Acc.**(↑)	**ECE** (↓)
$1e^{-1}$	$1e^{-2}$	**90.976**	0.093
$1e^{-1}$	$1e^{-1}$	85.724	**0.090**
$1e^{-2}$	$1e^{-2}$	90.655	0.095
$1e^{-3}$	$1e^{-4}$	90.283	0.097
$1e^{-5}$	$1e^{-5}$	90.115	0.101

Table 4: Performance of our stochastic transformer with varying stochasticity level.

GE	FID-A	FID-R	**Sel. Acc.**(↑)	**ECE** (↓)
-	-	-	90.322	0.106
✓	-	-	90.283	0.104
✓	✓	-	90.107	0.103
✓	-	✓	85.502	0.108
✓	✓	✓	**90.976**	**0.093**

5 Conclusion

In this paper, we introduced a novel stochastic ViT with stochastic Gaussian embeddings and the associated FID-based attention mechanism, propagating uncertainty and diverse embedding representations throughout the whole architecture. We incorporated the FID-based regularization term to imbue distance- and uncertainty-awareness into the learning process, thereby encouraging robust performance. Our findings from the in-distribution calibration, OOD detection, and few-shot learning studies reveal the potential of our stochastic ViT implementation by providing reliable downstream performance in safety-critical domains such as biomedical imaging diagnosis.In the future we will explore further possibilities to optimize the stochastic learning process with other distributional distance metrics and other possible embedding distributions.

Acknowledgements: The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR projects b143dc and b180dc. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683. Additional support was also received by the ERC - project MIA-NORMAL 101083647, DFG 513220538, 512819079, and by the state of Bavaria.

References

1. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
2. Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic) (2018)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale (2020)
4. Dowson, D.C., Landau, B.V.: The fréchet distance between multivariate normal distributions. Journal of Multivariate Analysis **12**, 450–455 (1982)
5. Fan, Z., Liu, Z., Wang, A., Nazari, Z., Zheng, L., Peng, H., Yu, P.S.: Sequential recommendation via stochastic self-attention (2022)
6. Fu, W., Chen, Y., Liu, W., Yue, X., Ma, C.: Evidence reconciled neural network for out-of-distribution detection in medical images. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023: 26th International Conference, Vancouver, BC, Canada, October 8–12, 2023, Proceedings, Part III. p. 305–315. Springer-Verlag, Berlin, Heidelberg (2023). https://doi.org/10.1007/978-3-031-43898-1_30
7. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning (2016)
8. Givens, C.R., Shortt, R.M.: A class of wasserstein metrics for probability distributions. Michigan Mathematical Journal **31**, 231–240 (1984)
9. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty (2020)

10. Knott, M., Smith, C.S.: On the optimal mapping of distributions. Journal of Optimization Theory and Applications **43**, 39–49 (1984)
11. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles (2017)
12. National Cancer Institute: Cancer Statistics, accessed: 2024-03-03. https://www.cancer.gov/ (2024)
13. Neal, R.M.: Bayesian learning for neural networks, vol. 118. Springer Science & Business Media (2012)
14. Olkin, I., Pukelsheim, F.: The distance between two random vectors with given dispersion matrices. Linear Algebra and its Applications **48**, 257–263 (1982)
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. the Journal of machine Learning research **12**, 2825–2830 (2011)
16. Qian, C., Feng, F., Wen, L., Chua, T.S.: Conceptualized and contextualized gaussian embedding. In: AAAI Conference on Artificial Intelligence (2021)
17. Rudner, T.G.J., Kapoor, S., Qiu, S., Wilson, A.G.: Function-space regularization in neural networks: A probabilistic perspective (2023)
18. Tran, D., Liu, J., Dusenberry, M.W., Phan, D., Collier, M., Ren, J., Han, K., Wang, Z., Mariet, Z., Hu, H., et al.: Plex: Towards reliability using pretrained large model extensions. arXiv preprint arXiv:2207.07411 (2022)
19. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**(1) (Aug 2018). https://doi.org/10.1038/sdata.2018.161, http://dx.doi.org/10.1038/sdata.2018.161
20. Vilnis, L., McCallum, A.: Word representations via gaussian embedding (2015)
21. WHO: Cancer, accessed: 2024-03-03. https://www.who.int/news-room/fact-sheets/detail/cancer (2024)
22. Zhang, C., Zhang, M., Zhang, S., Jin, D., Zhou, Q., Cai, Z., Zhao, H., Liu, X., Liu, Z.: Delving deep into the generalization of vision transformers under distribution shifts (2022)