CONDITIONAL DIFFUSION MODELS ARE MEDICAL IMAGE CLASSIFIERS THAT PROVIDE EXPLAINABILITY AND UNCERTAINTY FOR FREE

PREPRINT.

Gian Mario Favero[*]
McGill University
Mila – Quebec AI Institute

Parham Saremi[*]
McGill University
Mila – Quebec AI Institute

Emily Kaczmarek
McGill University
Mila – Quebec AI Institute

Brennan Nichyporuk
McGill University
Mila – Quebec AI Institute

Tal Arbel
McGill University
Mila – Quebec AI Institute

ABSTRACT

Discriminative classifiers have become a foundational tool in deep learning for medical imaging, excelling at learning separable features of complex data distributions. However, these models often need careful design, augmentation, and training techniques to ensure safe and reliable deployment. Recently, diffusion models have become synonymous with generative modeling in 2D. These models showcase robustness across a range of tasks including natural image classification, where classification is performed by comparing reconstruction errors across images generated for each possible conditioning input. This work presents the first exploration of the potential of class conditional diffusion models for 2D medical image classification. First, we develop a novel majority voting scheme shown to improve the performance of medical diffusion classifiers. Next, extensive experiments on the CheXpert and ISIC Melanoma skin cancer datasets demonstrate that foundation and trained-from-scratch diffusion models achieve competitive performance against SOTA discriminative classifiers without the need for explicit supervision. In addition, we show that diffusion classifiers are intrinsically explainable, and can be used to quantify the uncertainty of their predictions, increasing their trustworthiness and reliability in safety-critical, clinical contexts. Further information is available on our project page: med-diffusion-classifier.github.io/.

Keywords diffusion, classification, explainability, uncertainty

[1]* Contributed equally

1 Introduction

Deep learning applications in medicine have received significant attention in recent years due to their potential to revolutionize healthcare outcomes. For instance, the ability to accurately classify disease pathology from medical images using discriminative classifiers (e.g., ResNet [1], ViT [2]) is central to advancing early diagnosis, personalized treatment, and overall patient care. In the ideal scenario, discriminative classifiers are robust and generalizable; however, state-of-the-art performance often relies heavily on data augmentation and hyperparameter tuning, which can be time- and computation-expensive, and may still be prone to overfitting and/or learning shortcuts [3]. Even with strong classification performance, models must be explainable and provide uncertainty estimates to ensure reliable and trustworthy predictions for safe clinical deployment. Current explainability and uncertainty methods depend largely on post-hoc analysis or model modifications. For example, explainability often relies on gradient-based analysis after training [4] or counterfactual generation with a separate model [5], whereas uncertainty methods range from simple model modifications, like Monte Carlo (MC) dropout, to expensive ensembling methods. Thus, there remains limitations to the safe use of discriminative classifiers in medical imaging, particularly due to the lack of built-in explainability and uncertainty analysis.

Diffusion models [6] make up one class of generative models that has shown remarkable flexibility and robustness across various deep learning tasks, achieving state-of-the-art performance in image [7], video [8], and audio [9] generation tasks. Recently, generative models have been used directly for image classification [10, 11, 12, 13] in natural imaging, showing that large pre-trained models like Stable Diffusion [14] can be used as classifiers that are competitive with state-of-the-art supervised discriminative classifiers [1, 2]. Diffusion models are increasingly being used in the medical domain for data augmentation [15], segmentation [16, 17], anomaly detection [18], and probabilistic classification [19]. However, despite many conditional diffusion models developed for medical image analysis, they have yet to be explored as classifiers that can provide explainability and uncertainty-estimation for free.

In this work, we present a comprehensive evaluation of how conditional diffusion models can be re-purposed and leveraged for image classification, explainability, and uncertainty estimation in the medical domain. First, we propose a novel majority voting-based method that improves the performance of diffusion classifiers in medical imaging. We then demonstrate that classifiers derived from foundation and trained-from-scratch diffusion models perform competitively with state-of-the-art medical image discriminative classifiers through extensive experiments on the publicly available CheXpert [20] and ISIC Melanoma skin cancer [21] datasets, despite not being trained for classification. Next, we show that diffusion classifiers offer explainability (via counterfactual generation) and uncertainty quantification (via entropy) out-of-the-box. We validate the uncertainty by showing that when the model is confident, it is correct, and vice versa. This is shown as model accuracy drastically improves as its uncertainty threshold increases. For example, Stable Diffusion reaches classification accuracies of 100% and 95% on ISIC and CheXpert, respectively, with only 45% of its most uncertain samples filtered out.

2 Methodology

In this work, we present diffusion classifiers for medical imaging classification tasks. We first present an overview of diffusion models in Section 2.1. Next, in Section 2.2, we define conditional diffusion models and demonstrate how they can perform classification. Section 2.3 introduces all extensions to the diffusion classifier, including: our novel algorithm for improving classification

performance through majority voting, as well as the ability to perform counterfactual explainability and uncertainty quantification without any modifications.

2.1 Diffusion Models

Diffusion models (DM) are likelihood-based models that learn to approximate a data distribution through a process of iterative noising and denoising involving two key phases: a fixed forward process and a learned backward process. In the forward process, Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathrm{I})$ is gradually added to data in a controlled manner, destroying its structure until it is pure Gaussian noise. This process, which is done on a sample over time, can be expressed by its marginal for all t on a continuous interval, $[0, 1]$:

$$q(\boldsymbol{z}_t|\boldsymbol{x}) = \alpha_\lambda \boldsymbol{x} + \sigma_\lambda \boldsymbol{\epsilon} \text{ where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathrm{I}). \tag{1}$$

The forward process is defined to be variance-preserving, imposing the constraint $\alpha_\lambda^2 = \text{sigmoid}(\lambda)$, $\sigma_\lambda^2 = \text{sigmoid}(-\lambda)$, where λ is the log-SNR given by $\lambda = \log \alpha_\lambda^2 / \sigma_\lambda^2$. The noise schedule is a monotonically decreasing and invertible function, $f_\lambda(t)$, that connects the time variable, t, with the log-SNR, λ: $\lambda = f_\lambda(t)$. During training, t is sampled from $\mathcal{U}(0, 1)$ which is then used to compute λ. The resulting distribution over noise levels can be defined as $p(\lambda) = -1/f'_\lambda(t)$ [22].

In the backward process, a neural network attempts to learn how to remove the added noise and recover an approximate sample from the original data distribution. Kingma et al. show that the variational lower bound objective (VLB) function for training diffusion models can be derived in continuous time with respect to its log-SNR, λ, noise sampling distribution, $p(\lambda)$ and weighting function, $w(\lambda)$ [23]. This VLB is:

$$\log p(x) = \mathcal{L}_x + \mathcal{L}_T - \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\mathrm{I}), \lambda \sim p(\lambda)} \left[\frac{w(\lambda)}{p(\lambda)} ||\boldsymbol{x} - \hat{\boldsymbol{x}}_\theta(\boldsymbol{z}_\lambda; \lambda)||_2^2 \right]. \tag{2}$$

Where $\mathcal{L}_x = -\log p(\boldsymbol{x}|\boldsymbol{z}_0) \approx 0$ for discrete $\boldsymbol{x}$ and $\mathcal{L}_T = D_{KL}(q(\boldsymbol{z}_T|x)||p(\boldsymbol{z}_T)) \approx 0$ for a well-defined forward process. We use a min-SNR weighing function [24], a shifted-cosine noise schedule [25], and v-prediction parameterization for greater stability during training and sampling [26].

2.2 Conditional Diffusion Models as Classifiers

Conditional diffusion models incorporate text or categorical inputs, such that the prediction becomes $\hat{\boldsymbol{x}}_\theta(\boldsymbol{z}_\lambda, c)$ where c is a conditioning embedding. In this paper, we implement conditioning through cross-attention in a UNet-based diffusion model [14], and adaptive layer normalization in DiTs [27].

Recent works [10, 11, 12, 13] have explored using conditional diffusion models as discriminative classifiers. As shown in Figure 1, classification is performed by comparing reconstruction errors across images generated for each possible conditioning input. Specifically, using the labels, $\boldsymbol{C} = \{c_i\}$, and Bayes' theorem on model predictions, $p(\boldsymbol{x}|c_i)$, we can derive $p(c_i|\boldsymbol{x})$:

$$p(c_i|\boldsymbol{x}) \approx \frac{\exp\{\mathbb{E}_{\epsilon \sim \mathcal{N}(0,\mathrm{I}), \lambda \sim p(\lambda)} \left[||\boldsymbol{x} - \hat{\boldsymbol{x}}_\theta(\boldsymbol{z}_\lambda, c_i)||_2^2 \right]\}}{\exp\{\sum_j \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\mathrm{I}), \lambda \sim p(\lambda)} \left[||\boldsymbol{x} - \hat{\boldsymbol{x}}_\theta(\boldsymbol{z}_\lambda, c_j)||_2^2 \right]\}}. \tag{3}$$

A Monte Carlo estimation of the expectation for an arbitrary class, c_j, can be computed by sampling N noise level pairs, $(\boldsymbol{\epsilon}, \lambda)$ and averaging the reconstruction error:

$$\frac{1}{N} \sum_{k=1}^{N} \left[||\boldsymbol{x} - \hat{\boldsymbol{x}}_\theta(\alpha_{\lambda_k} \boldsymbol{x} + \sigma_{\lambda_k} \boldsymbol{\epsilon}_k, c_j)||_2^2 \right]. \tag{4}$$

Equation 4 shows that classifying one sample requires N many steps per condition, where the Monte Carlo estimate becomes more accurate as the number of steps increases. To reduce the variance of prediction for a given image, $\boldsymbol{x}$, an identical set of $(\boldsymbol{\epsilon}_k, \lambda_k) \in S\{(\boldsymbol{\epsilon}_k, \lambda_k)\}_{k=1}^{N}$ is used for every condition, which increases the accuracy of the prediction $p(\boldsymbol{C}|\boldsymbol{x})$. In practice, Eq. 3 is equivalent to choosing the class with the minimum average reconstruction error.

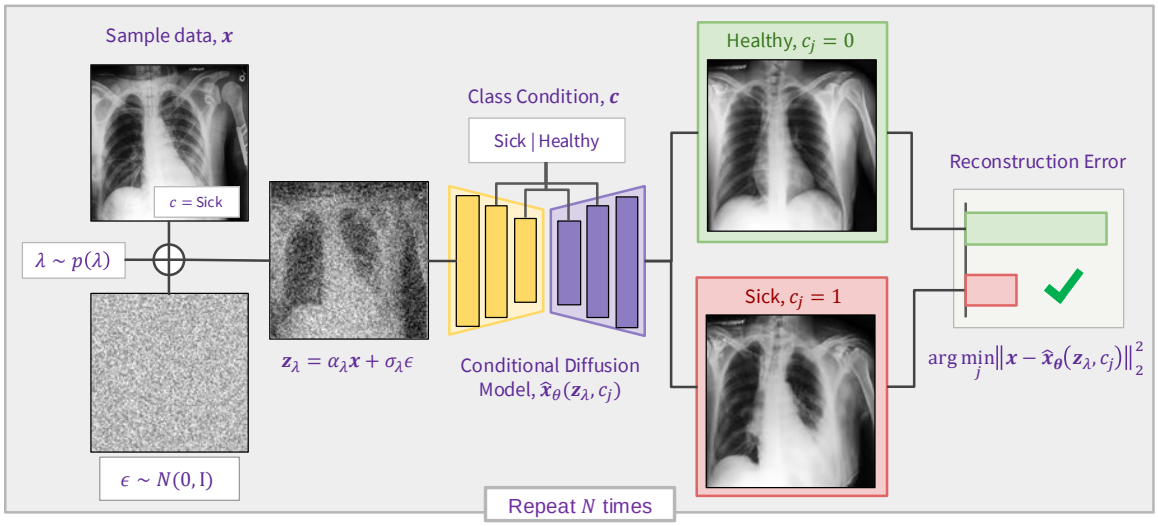


Figure 1: **Diffusion classifiers** are conditional diffusion models repurposed as classifiers. First, a sample, $\boldsymbol{x}$, is noised at a randomly chosen noise level, $(\boldsymbol{\epsilon}_k, \lambda_k)$. The noised sample is then denoised by the diffusion network with each possible conditioning input, c_j. The conditioning variable, c_j, that results in the denoised output, $\hat{\boldsymbol{x}}_\theta(\boldsymbol{z}_\lambda, c_j)$, with the smallest reconstruction error is selected as the class. This process is repeated for a set of N noise levels $(\boldsymbol{\epsilon}, \lambda)$ with the reconstruction errors aggregated (e.g., average/majority voting) for a more accurate prediction.

2.3 Extensions on Diffusion Classifiers

Majority Voting: In this work, we introduce a novel majority-vote-based algorithm for determining the predicted class. Here, we tally votes across all $(\boldsymbol{\epsilon}, \lambda)$ pairs by identifying the class with the lowest error as the prediction for that pairing, and take the final predicted class as the one with the majority of individual votes (see Appendix A). We posit that averaging reconstruction error over all noise levels inherently weights higher values of λ more, which is not always beneficial (i.e., reconstructions at higher values of λ are naturally much harder and thus have greater error, introducing more noise into the average reconstruction error).

Intrinsic Explainability: Diffusion classifiers use Classifier-Free Guidance (CFG) [28] to understand which features are most influential in generating certain classes. CFG is a common approach in which a conditional diffusion model is simultaneously trained for an unconditional task by randomly dropping out c ($\sim$10% of the time). In doing so, sampling can be guided towards an intended class with a guidance-scale, w:

$$\tilde{\boldsymbol{x}}_\theta(\boldsymbol{z}_\lambda, c) = (1 + w)\hat{\boldsymbol{x}}_\theta(\boldsymbol{z}_\lambda, c) - w\hat{\boldsymbol{x}}_\theta(\boldsymbol{z}_\lambda, \varnothing). \tag{5}$$

At inference, the model permits explainability for free, through the conditional generation of the factual and counterfactual images of the input image. First, noise is added to obscure the input

images, while preserving enough image structure to make reconstruction possible. Then, by varying the condition at inference, the model can shift its generation process to any possible conditional class. These generated images represent the reconstruction of the image provided by the true class, and any counterfactual image(s), where difference maps can be created to highlight class-specific regions modified by the network.

Uncertainty Quantification: Diffusion classifiers are also able to produce uncertainty estimates without any additional modifications to the model. The set of N (ϵ, λ) pairs required for accurate classification results (Eq. 4) results in numerous predictions generated for each sample and thus inherently resembles the uncertainty estimation strategy via MC dropout or ensemble methods. As explained in Section 2.2, to achieve accurate classification, a single sample requires N steps (repeated per condition) where reconstruction errors for that sample are calculated at different (ϵ, λ) noise levels (Eq. 4). To quantify the uncertainty of the overall predicted class, we construct a Bernoulli distribution from each of the N predictions. This creates a probability density from which uncertainty can be computed.

3 Experiments

We first evaluate the performance of the average reconstruction error objective (Section 2.2) against a majority voting alternative, demonstrating that the latter yields superior results in our tasks. We then compare the classification performance of diffusion classifiers with state-of-the-art discriminative baselines, and furthermore, show that conditional diffusion models are interpretable out-of-the-box, and capable of producing uncertainty quantifications.

3.1 Datasets

ISIC Melanoma: A publicly available dataset [21] containing over 35,000 images of skin lesions and corresponding labels for the presence of melanoma. We balance the dataset for our experiments, resulting in 10,212 total images. The data are randomly split into an 80/10/10 train/validation/test set.

CheXpert: A publicly available dataset [20] containing over 200,000 chest X-ray images with binary labels for 14 diseases and the presence of support devices. For our experiments, we use "Pleural Effusion" and "No Findings" as mutually exclusive labels and filter for frontal views of the chest, resulting in a balanced dataset of 20,404 samples. The data are randomly split into an 80/10/10 train/validation/test set.

3.2 Baseline Models

To establish a comparative baseline, we evaluate the performance of both convolutional and transformer-based architectures. We use `torchvision` implementations of ResNet-18 and ResNet-50 [1], and `timm` implementations of ViT-S/16 and ViT-B/16 [2], EfficientNet-B0 and EfficientNet-B4 [29], and Swin-B Transformer [30]. More details can be found in Appendix B.

3.3 Conditional Diffusion Models

We implement a UNet backbone based on the ADM architecture [7] at 256^2 resolution, incorporating improvements from simple diffusion [25], such as scaling the number of ResBlocks at lower resolutions to save memory at higher resolutions. For transformer-based diffusion models, we include the DiT-B/4 variant from [27]. Unless otherwise noted, all images are compressed with a single-stage discrete wavelet transform (DWT) using a Haar wavelet. More details can be found in Appendix B.

3.4 Foundation Models

Ideally, foundation models like Stable Diffusion can be repurposed as zero-shot classifiers. However, we find that such models are not trained on enough medical data to perform adequately by default. Thus, to ensure a fair comparison, we fine-tune Stable Diffusion v2-base [14] on an amalgamation of our CheXpert and ISIC Melanoma training splits. Given that the model is designed for text-to-image generation, we replace labels in the datasets with text prompts, e.g., "a benign skin lesion", or, "a frontal chest xray of a sick patient with pleural effusion". More details can be found in Appendix D.

4 Results

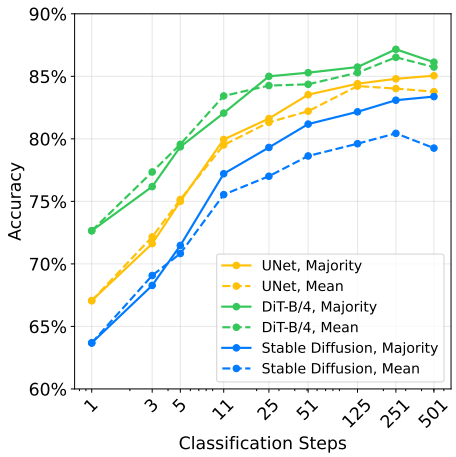


Figure 2: **A majority voting scheme leads to better performance.** We find that tallying the minimum reconstruction error across steps leads to better classification than averaging reconstruction error across steps. Results are seen on the CheXpert dataset.

	Method	CheXpert		ISIC	
		Accuracy	F1	Accuracy	F1
CNN	ResNet-18	90.9	0.910	94.4	0.943
	ResNet-50	91.6	0.914	93.6	0.935
	EfficientNet-B0	90.5	0.907	93.1	0.930
	EfficientNet-B4	90.4	0.904	93.2	0.930
TF	ViT-S/16	86.9	0.869	95.0	0.949
	ViT-B/16	85.1	0.857	94.8	0.948
	Swin-B	86.1	0.863	95.9	0.958
DM	DiT-B/4	86.1	0.860	90.4	0.901
	UNet	84.5	0.854	91.8	0.919
	Stable Diffusion*	85.0	0.839	94.8	0.946
	Stable Diffusion†	48.8	0.656	39.7	0.521

Table 1: **Diffusion classifiers are competitive with discriminative baselines.** * and † denote fine-tuned and zero-shot versions, respectively. Diffusion classifier results (DM) are with 501 classification steps, majority voting, minimal hyperparameter tuning/data augmentation.

4.1 Ablating on the Classification Algorithm

We propose a simple but effective majority voting scheme that, instead of accumulating errors at each timestep, tallies the amount of times a reconstruction error was smaller for each test condition and then chooses the class with the most votes. Figure 2 shows that using majority voting increases classification performance across the board, and specifically so at larger values of N. This result is intuitive: at greater values of N there are more reconstructions attempted from high noise disturbance which can introduce large sources of variance in the average error. Majority voting is used for all diffusion results in this paper.

4.2 Classification Performance on Benchmark Datasets

Table 1 shows the classification accuracy and F1-score of each model on the CheXpert and ISIC Melanoma test sets. Note that the models are grouped by architecture: convolution-based (CNN), transformer-based (TF), and diffusion-based (DM). Our experiments across both datasets demonstrate that diffusion classifiers perform competitively with discriminative classifiers. Notably, diffusion

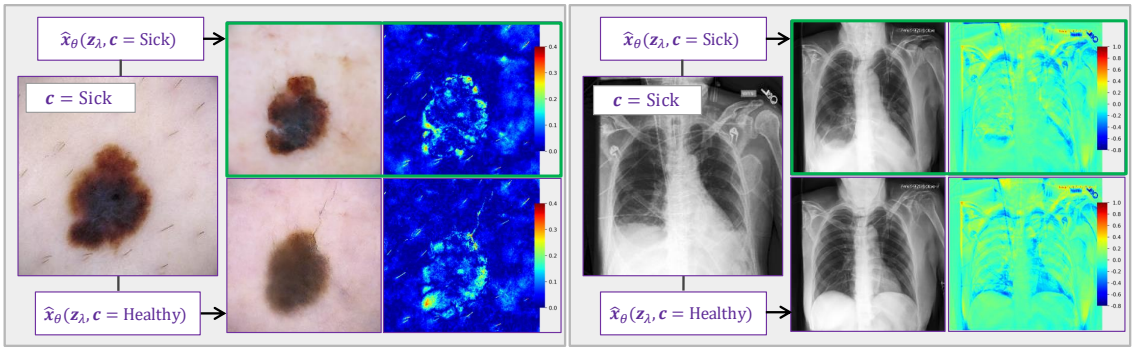


Figure 4: **Diffusion classifiers are naturally explainable** and highlight why they make classification decisions using classifier-free guided sampling. Difference maps show conditional areas of interest (pathology added/removed) during reconstruction. For CheXpert, the heatmap represents a simple difference map, as the images are grayscale and signed differences help capture the directional change. For ISIC, the heatmap represents an absolute difference map, since the images are colored and an absolute difference better reflects overall pixel-wise changes. For CheXpert noise is added at t=0.5. For ISIC: t=0.3. CFG scale is 7.5. All images are generated from the UNet model with 256 sampling steps.

classifiers achieve this performance with minimal hyperparameter tuning, no augmentations, and without being trained on a classification objective.

4.3 Intrinsic Explainability

A key advantage of diffusion classifiers lies in their intrinsic interpretability, which positions diffusion classifiers as not only effective but also transparent. Importantly, diffusion classifiers are able to produce counterfactual explanations, as opposed to other interpretability methods that simply highlight regions of interest. This can be seen in Figure 4: On the left example (skin lesion), the counterfactual of a malignant lesion (melanoma) has changed colour and intensity to become healthy. In the right example (chest X-ray), the counterfactual image of a sick patient (Plural Effusion) shows decreased disease pathology in the left and right lungs. The natural interpretability of diffusion classifiers provides both transparency on how the model is learning (thus allowing the identification of shortcut learning), and specific class information which improves understanding of the disease. In addition to providing disease explainability, the difference maps also reveal how the model makes its decision: the condition with the least reconstruction error is selected as the predicted class.

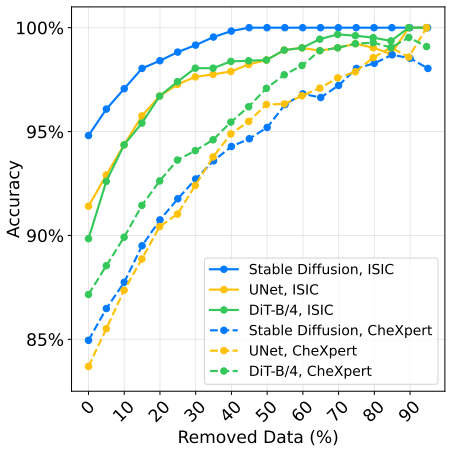


Figure 3: **Diffusion classifiers inherently produce uncertainty estimates.** Filtering uncertain predictions improves performance, showing that when the model is certain, it is correct.

4.4 Uncertainty Quantification

The uncertainty quantification of diffusion classifiers is demonstrated in Figure 3. In addition to competitive classification performance and intrinsic explainability, uncertainty quantifications can be estimated without any model modifications. In medical imaging, uncertainty measures are validated by confirming that when the model is confident, the prediction is correct, and when it is incorrect, it is uncertain [31]. We therefore validate the diffusion model's uncertainty quantification by filtering out the most uncertain predictions and examining the change in performance. Each of the models show accuracy increases as the most uncertain predictions are filtered out for CheXpert (- -) and ISIC (-).

5 Conclusion

In this paper, we provide a comprehensive examination of the benefits of diffusion classifiers in medical imaging. First, we introduce a novel majority voting method to improve the overall performance of diffusion classifiers. We next demonstrate that diffusion classifiers are able to achieve comparable performance to state-of-the-art discriminative classifiers, in addition to providing intrinsic counterfactual explainability and uncertainty quantification.

Limitations: Note that due to the slow inference speed of diffusion, classifying a single image using 501 steps takes 3.72 seconds on an A100 GPU.

We thank Bruno Travouillon, Olexa Bilaniuk, and the Mila IDT team for their support with Mila's HPC. This work was supported by the Natural Sciences and Engineering Research Council of Canada, Fonds de Recherche du Quebec: Nature et Technologies, the Canadian Institute for Advanced Research (CIFAR) Artificial Intelligence Chairs program, Google Research, Calcul Quebec, the Digital Research Alliance of Canada, the Vadasz Scholar McGill Engineering Doctoral Award, and Mila - Quebec AI Institute.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. Eprint arXiv:1512.03385, 2015.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. Eprint arXiv:2010.11929, 2020.

[3] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020.

[4] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vision*, 128(2):336–359, February 2020.

[5] Susu Sun, Stefano Woerner, Andreas Maier, Lisa M. Koch, and Christian F. Baumgartner. Inherently interpretable multi-label classification using class-specific counterfactuals, 2023.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Eprint arXiv:2006.11239, 2020.

[7] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. Eprint arXiv:2105.05233, 2021.

[8] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. Eprint arXiv:2204.03458, 2022.

[9] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. Eprint arXiv:2009.09761, 2020.

[10] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. Eprint arXiv:2303.16203, 2023.

[11] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers, 2023.

[12] Benno Krojer, Elinor Poole-Dayan, Vikram Voleti, Christopher Pal, and Siva Reddy. Are diffusion models vision-and-language reasoners? Eprint arXiv:2305.16397, 2023.

[13] Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model, 2024.

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. Eprint arXiv:2112.10752, 2022.

[15] Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, and Daguang Xu. Maisi: Medical ai for synthetic imaging. Eprint arXiv:2409.11169, 2024.

[16] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. Eprint arXiv:2211.00611, 2023.

[17] Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion based medical image segmentation with transformer. Eprint arXiv:2301.11798, 2023.

[18] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C. Cattin. Diffusion models for medical anomaly detection. Eprint arXiv:2203.04306, 2022.

[19] Xing Shen, Hengguan Huang, Brennan Nichyporuk, and Tal Arbel. Improving robustness and reliability in medical image classification with latent-guided diffusion and nested-ensembles. Eprint arXiv:2310.15952, 2024.

[20] Jeremy Irvin, Doriane Rajan, Dan B. S. D., Neha Gupta, Paul W. C., Avnish Verma, Jin Han, David K. F., Ildefonso, Cormac, et al. Chexpert: A large chest x-ray dataset with uncertainty labels and expert comparison. Eprint arXiv:1901.07031, 2019.

[21] Veronica M Rotemberg, Nicholas R. Kurtansky, Brigid Betz-Stablein, Liam J. Caffery, Emmanouil Chousakos, Noel C. F. Codella, Marc Combalia, Stephen W. Dusza, Pascale Guitera, David Gutman, Allan C. Halpern, Brian Helba, Harald Kittler, Kivanc Kose, Steve G. Langer, Konstantinos Lioprys, Josep Malvehy, Shenara Musthaq, Jabpani Nanda, Ofer Reiter, George Shih, Alexander J. Stratigos, Philipp Tschandl, Jochen Weber, and Hans Peter Soyer. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8, 2020.

[22] Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. Eprint arXiv:2303.00848, 2023.

[23] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. Eprint arXiv:2107.00630, 2023.

[24] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. Eprint arXiv:2303.09556, 2024.

[25] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. Simple diffusion: End-to-end diffusion for high resolution images. Eprint arXiv:2301.11093, 2023.

[26] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. Eprint arXiv:2202.00512, 2022.

[27] William Peebles and Saining Xie. Scalable diffusion models with transformers. Eprint arXiv:2212.09748, 2023.

[28] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. Eprint arXiv:2207.12598, 2022.

[29] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (ICML 2019), 2019.

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. Eprint arXiv:2103.14030, 2021.

[31] Tanya Nair, Doina Precup, Douglas L. Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical Image Analysis*, 59:101557, 2020.

A Additional Background on Diffusion Classifiers

We provide the pseudocode for the diffusion classification algorithm used in the experiments. We opt for a majority voting scheme as opposed to the average reconstruction error approach outlined by [10].

```
def classify(x, num_classes, classification_steps):
    errors = fill((x.shape[0], num_classes, classification_steps), inf)
    for step in classification_steps:
        t = rand(0,1)
        z_t, eps_t = diffuse(x, t) # add noise to image at t
        # Get the errors for each class
        for c in range(num_classes):
            pred = model(z_t, t, c) # get noise pred for given class
            error = mse(pred, eps_t)
            errors[:,c,step] = error # store the error
    # Find the class with the lowest error for each step
    end_of_stage_votes = errors[:, :, :classification_steps].argmin(dim=1)

    # Count the votes for each class across all steps
    votes = zeros(x.shape[0], num_classes)
    for b in range(x.shape[0]):
        for step in range(classification_steps):
            class_with_lowest_error = end_of_stage_votes[b, step]
            votes[b, class_with_lowest_error] += 1
    final_classes = votes.argmax(dim=1)

    return final_classes
```

B Experimental Details

B.1 Optimization Settings

```
resnet/efficientnet optimization settings on isic: (3x256x256):
    batch size=64
    optimizer=Adam
    learning_rate=1E-4
    weight_decay=1E-5

resnet/efficientnet optimization settings on chexpert: (3x256x256):
    batch size=64
    optimizer=Adam
    learning_rate=1E-4
    weight_decay=1E-3

vit/swin optimization settings (3x256x256):
    batch size=64,
    optimizer=Adam,
    learning_rate=1E-5
```

```
diffusion optimization settings (3x256x256):
    batch size=128,
    optimizer=Adam,
    learning_rate=1E-4,
    learning_rate_warmup_steps=250,
    grad_clip=1.0,
    ema={beta: 0.999, warmup: 50, update_freq: 5}
```

B.2 Discriminative Baseline Settings

We use official implementations of ResNet-based (`torchvision`), EfficientNet- and ViT-based (`timm`) classifiers in our experiments.

For training, we apply the following augmentations:

- Random rotation with degree range of (-30, 30)
- Random horizontal flip with probability of 0.5
- Random vertical flip with probability of 0.5
- Random Gaussian blur with kernel size of 5 and sigma range of (0.1, 2)

B.3 UNet Settings

The ADM architecture [7] is used as a starting point, with minor alterations based on capacity requirements of each experiment. Class conditions are integrated into the model using cross-attention with a trainable module `nn.encoder`.

```
unet settings (3x256x256):
    prediction_param=v-prediction,
    noise_schedule=shifted cosine, base-64,
    wavelet_transform=single-stage haar wavelet
    sample_size=128,
    channels=12,
    resnet_layers_per_block=2,
    base_channels=128,
    channel_multiplier=(1,1,2,4,8),
    cross_attn_res=16
    encoder_type=nn,
    cross_attn_dim=512,
```

B.4 DiT Settings

The DiT-B/4 architecture is followed as presented in [27].

```
dit settings (3x256x256):
    prediction_param=v-prediction,
    noise_schedule=shifted cosine, base-64,
    wavelet_transform=single-stage haar wavelet,
    sample_size=128,
    channels=12,
    num_attention_heads=12,
    attention_head_dim=64,
```

```
num_layers=12,
patch_size=4,
```

C Training Dynamics

During training, we monitored classification performance using the F1 score and accuracy metrics. Figure 5 illustrates the performance of the UNet and DiT models on the ISIC validation set across different training steps. The results indicate that classification performance can serve as a useful metric for tracking training progress.

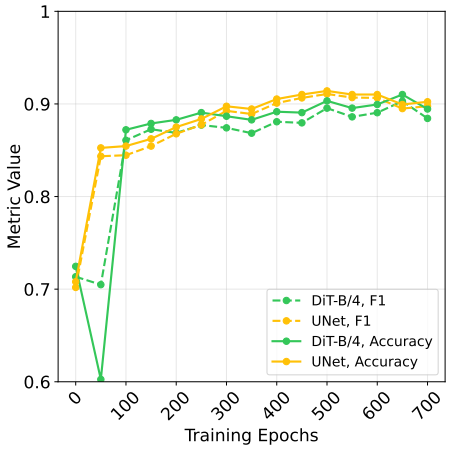


Figure 5: Classification performance of DiT and UNet models on ISIC validation set during training

D Stable Diffusion v2 Fine-Tuning

We fine-tune Stable Diffusion v2-base [14] using the `Hugging Face` training pipeline for a total of 15k iterations. We construct the fine-tuning dataset by amalgamating our CheXpert and ISIC Melanoma training splits. Given that the model is designed for text-to-image generation, we replace labels in the datasets with text prompts, ie. "a benign skin lesion", or, "a frontal chest xray of a sick patient with pleural effusion". Fine-tuning dramatically increased Stable Diffusion's domain knowledge and subsequent classification performance on our benchmark datasets.

E Uncertainty Quantification

We validate our model uncertainty through measuring performance as uncertain predictions are filtered out. For both the pre-trained Stable Diffusion classifier and our diffusion classifiers trained from scratch, accuracy increases for both datasets as the most uncertain predictions are filtered out. This indicates that the model is most uncertain about its incorrect predictions, which is highly valuable across medical applications. See Figure 9 for a breakdown of this quantification in boxplot form.

F More Explainability Results

More explainability results can be found in Figure 10, and Figure 11. Input sick images have been altered to healthy class by adding noise to the input image and denoising with the healhty class. For CheXpert t=0.5 and for ISIC t=0.3 are used.

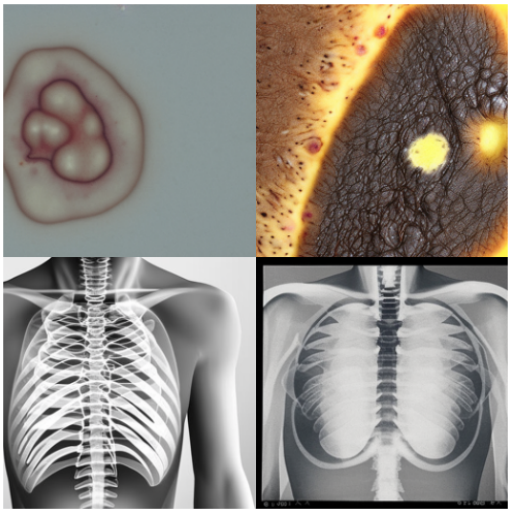


Figure 6: No fine-tuning

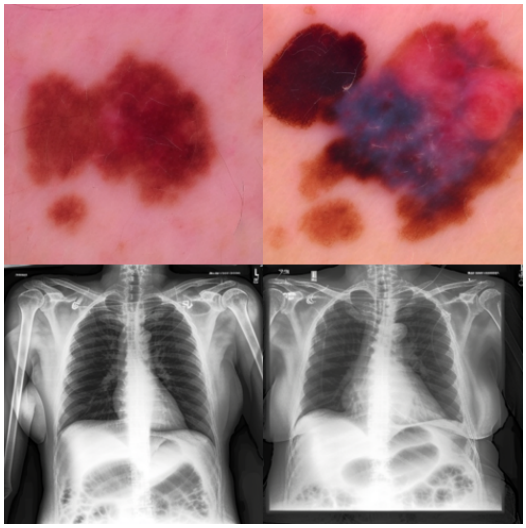
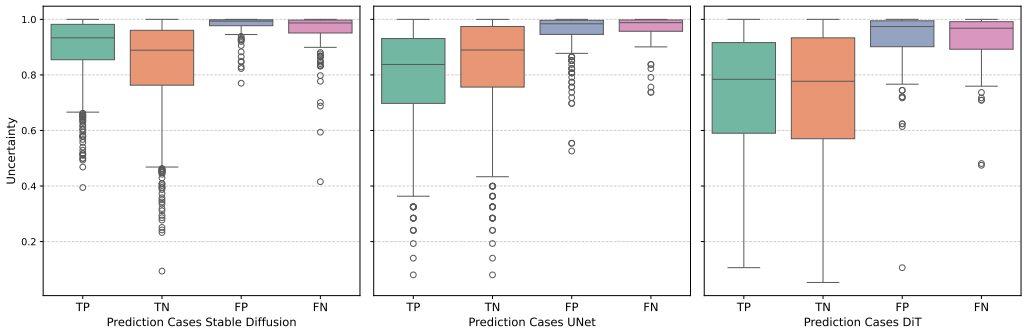


Figure 7: After fine-tuning for 15k steps

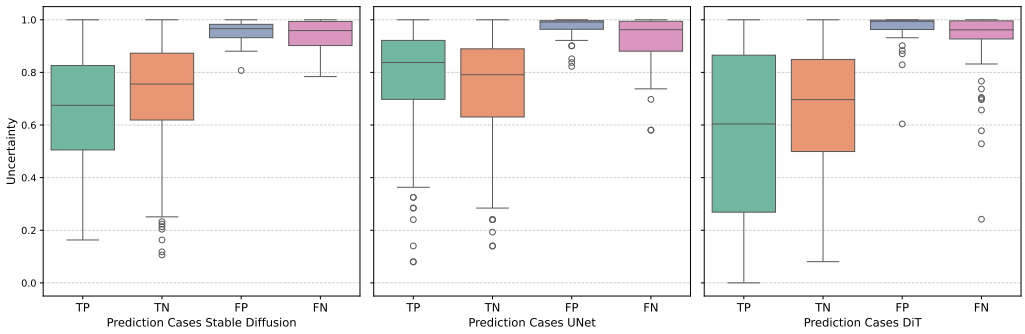
Figure 8: Task-related generation from the Stable Diffusion v2-base model before (left) and after (right) fine-tuning. Training for only a few thousand iterations dramatically increased in-distribution inference and classification performance.

G Computational Resources

All models were trained or fine-tuned a compute cluster of 80 GB A100 GPUs for all experiments in this paper. For inference, a single 80 GB A100 GPU is used.



(a) Uncertainty estimates, CheXpert



(b) Uncertainty estimates, ISIC

Figure 9: We find that all of our diffusion classifier models are more confident about their correct predictions (TP, TN) than their incorrect predictions (FP, FN).

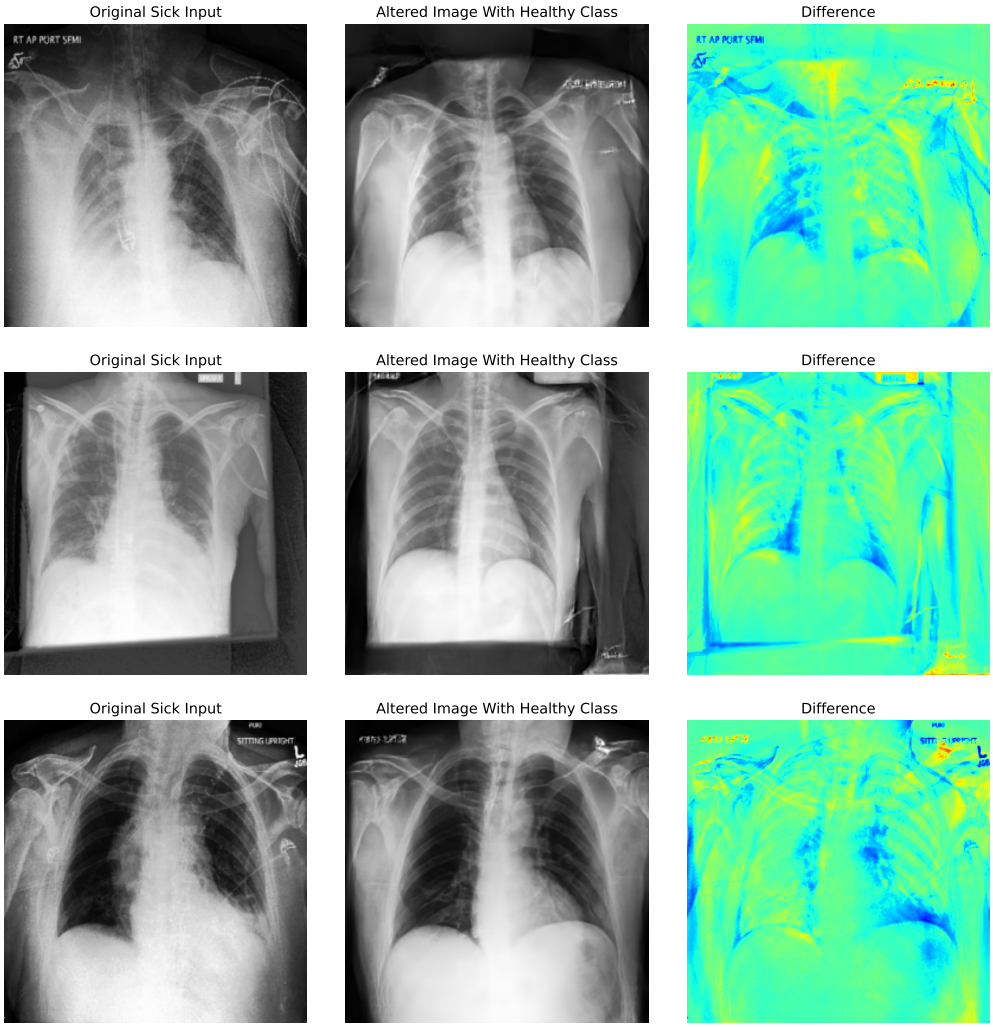


Figure 10: More explainability results for CheXpert by converting input sick images to healthy images. t=0.5 and CFG=7.5 are used for generating these images.

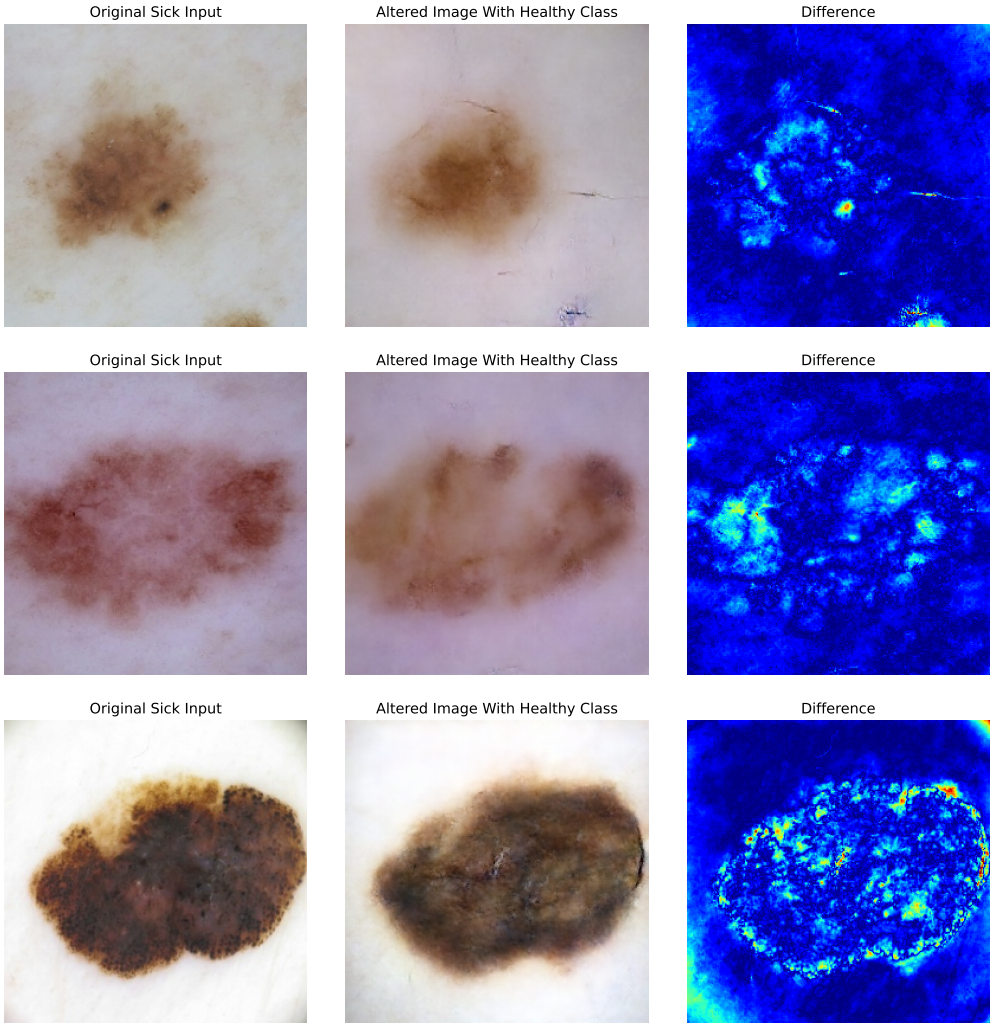


Figure 11: More explainability results for ISIC by converting input sick images to healthy images. t=0.3 and CFG=7.5 are used for generating these images.