Enhance Eye Disease Detection using Learnable Probabilistic Discrete Latents in Machine Learning Architectures

Anonymous authors
Paper under double-blind review

Abstract

Ocular diseases, including diabetic retinopathy and glaucoma, present a significant public health challenge due to their high prevalence and potential for causing vision impairment. Early and accurate diagnosis is crucial for effective treatment and management. In recent years, deep learning models have emerged as powerful tools for analysing medical images, such as retina imaging. However, challenges persist in model relibability and uncertainty estimation, which are critical for clinical decision-making. This study leverages the probabilistic framework of Generative Flow Networks (GFlowNets) to learn the posterior distribution over latent discrete dropout masks for the classification and analysis of ocular diseases using fundus images. We develop a robust and generalizable method that utilizes GFlowOut integrated with ResNet18 and ViT models as the backbone in identifying various ocular conditions. This study employs a unique set of dropout masks - none, random, bottomup, and topdown - to enhance model performance in analyzing these fundus images. Our results demonstrate that our learnable probablistic latents significantly improves accuracy, outperforming the traditional dropout approach. We utilize a gradient map calculation method, Grad-CAM, to assess model explainability, observing that the model accurately focuses on critical image regions for predictions. The integration of GFlowOut in neural networks presents a promising advancement in the automated diagnosis of ocular diseases, with implications for improving clinical workflows and patient outcomes.

1 Introduction

The world faces considerable challenges in terms of eye care. Research indicates that in 2020, the estimated global cases of age-related macular degeneration stood at 196 million, and this figure is anticipated to escalate to 288 million by 2040 Wong et al. (2014). According to the World Health Organization, in its report of "World Report on Vision", more than 2.2 billion people suffer from vision impairment or blindness. Importantly, it is estimated that over 1 billion of these cases could potentially have been avoided with proper prevention or effective treatment WTO (2019). The World Vision Report indicates that primary causes of blindness include Glaucoma, Age-Related Macular Degeneration, and Diabetic Retinopathy. Diagnosing these conditions typically involves an ophthalmologist evaluating a patient's symptoms, analyzing various eye and retina images, and conducting a manual examination. This process is thorough but can be time-consumingWTO (2019). Other researchers highlighted that the prevalence of Age-Related Macular Degenerations (AMDs) is notably higher in Africa and the Eastern Mediterranean regions compared to other areas of the world Xu et al. (2020). The lack of and unequal distribution of medical resources means that preventable and treatable cases of blindness and low vision predominantly affect people in less developed countries and regions. Vision impairment stems from various factors, notably the retina, which is a key element in disorders like glaucoma, diabetic retinopathy, and age-related macular degeneration. Properly addressing eye health requires not only accurate diagnosis but also effective prevention and treatment strategies for these conditions Yang et al. (2021).

Ophthalmology heavily depends on imaging for diagnosis, as the majority of eye conditions are identified through image analysis. However, traditional screening involves handling large volumes of data, is highly

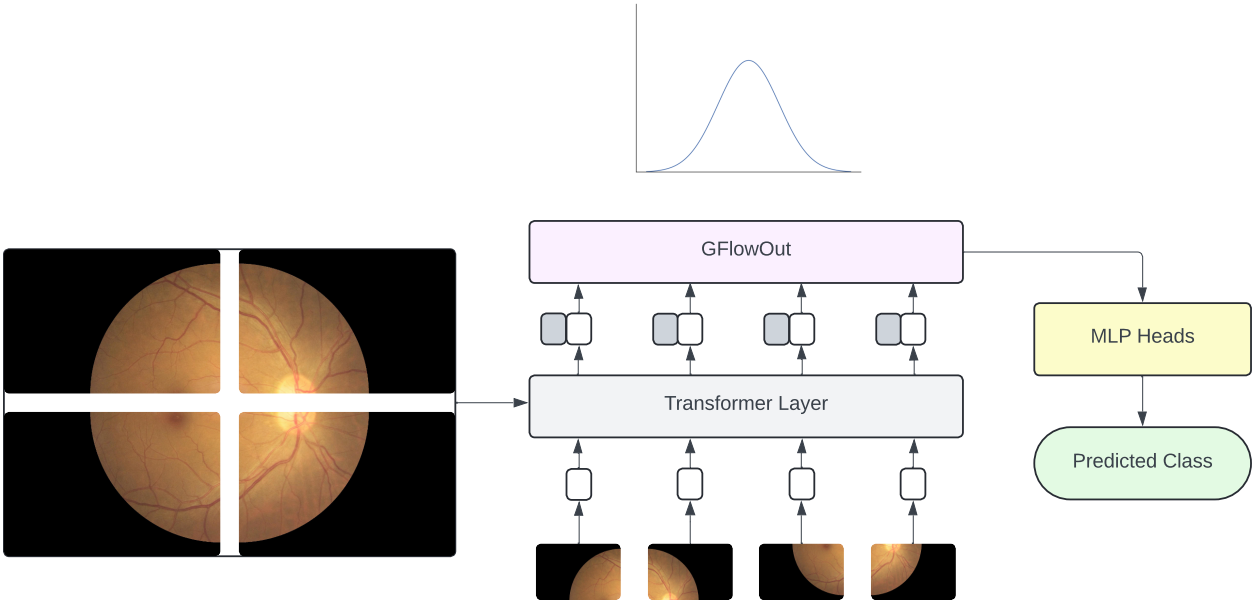

Figure 1: In the vision transformer architecture, we apply GFlowOut, a learnable dropout technique, in the transformer encoder. This allows us to learn posterior distribution over dropout masks tailored to our dataset, improving performance of the model.

subjective, and requires complex data analysis. This presents a significant burden for both patients and ophthalmologists, complicating long-term follow-ups Besenczi et al. (2016). The incorporation of artificial intelligence(AI), particularly machine learning and deep learning, into this field has significantly boosted the efficiency of clinical eye specialists. AI technology processes and analyzes ophthalmic images, thereby streamlining diagnostic procedures Padhy et al. (2019); Yu et al. (2018); Rajpurkar et al. (2022); Zhou et al. (2023). Currently, there has been considerable research on artificial intelligence-assisted diagnosis in diseases such as glaucoma, diabetic retinopathy, retinopathy of prematurity, and age-related macular degeneration (AMD) Ting et al. (2018). However, we found the majority of the models primarily focuse on diagnosing a single ophthalmic disease Li et al. (2021). There are multiple works shown that deep learning algorithms are promising in the diagnosing diabetic retinopathy through retinal fundus image grading Oh et al. (2021); Wang et al. (2022); Son et al. (2020). However, the high performance of these methods often comes with a significant increase in time complexity. Additionally, their performance can be limited by using uniform image sizes, leading to less robust classifications Li et al. (2022).

Along with these issues, a significant limitation of current deep neural networks is their tendency to exhibit *overconfidence in predictions* and lack a mechanism for capturing uncertainty, particularly when there is a shift in the data distribution between training and testing datasets Folgoc et al. (2021). This issue is especially prominent in medical imaging, where variability in data can impact diagnostic accuracy. While methods such as standard dropout exist to address this, they often fail to capture the multi-modality of posterior distributions over dropout masks. To mitigate these challenges, GFlowOut Liu et al. (2023) has been recently proposed, leveraging Generative Flow Networks (GFlowNets) Bengio et al. (2023) to model the posterior distribution over dropout masks. However, its potential in real-world medical applications remains underexplored. In this work, we bridge this gap by applying GFlowOut to neural networks for ocular disease classification using the Ocular Disease Intelligent Recognition (ODIR) dataset. Our key contribution is demonstrating the utility of GFlowOut in improving model uncertainty estimation and diagnostic accuracy across a diverse set of ocular conditions, providing a robust solution to variability in medical imaging datasets.

2 Related Work

2.1 Generative Flow Networks

Generative Flow Networks (GFlowNets) have recently emerged as a compelling framework for generating complex, high-dimensional objects by modeling the flow of probability through sequences of states. GFlowNets address the challenge of sampling objects in proportion to a predefined reward function by adopting a control problem formulation, where objects are constructed sequentially via probabilistic steps. This methodology enables GFlowNets to efficiently explore and sample from multimodal distributions, making them particularly well-suited for applications requiring diverse and high-quality solutions, such as drug discovery and protein design Bengio et al. (2023).

The versatility of GFlowNets has been demonstrated across various domains, including drug discovery Bengio et al. (2021a), biological sequence design Jain et al. (2022), robust combinatorial optimization Zhang et al. (2022), causal discovery Deleu et al. (2022), and neural network structure learning Pan et al. (2023a). Foundational work has highlighted the ability of GFlowNets to generalize effectively to complex distributions and reduce gradient variance relative to traditional policy gradient methods, thereby establishing a robust framework for probabilistic modeling Malkin et al. (2023b); Bengio et al. (2021b).

Subsequent advancements have further extended the capabilities of GFlowNets. For instance, Pan et al. (2023b) introduced Stochastic GFlowNets to address the challenges posed by stochastic environments, incorporating intrinsic exploration rewards to enhance training. Additionally, Deleu et al. (2022); Nishikawa-Toomey et al. (2022) applied GFlowNets to the generative modeling of discrete and composite objects, with a particular focus on Bayesian structure learning of complex causal graphs. The framework has also been leveraged in approximate maximum-likelihood training of energy-based models, as demonstrated by Zhang et al. (2022), without the need for a predefined target reward. Moreover, GFlowNets have been applied to tackle NP-hard combinatorial optimization problems, providing a promising approach to these computationally intensive tasks Zhang et al. (2022). In the realm of biological sequence design, Jain et al. (2022) employed GFlowNets within an active learning loop to optimize sequence generation. Furthermore, Zimmermann et al. (2023) offered a variational perspective on GFlowNets by formulating variational objectives through the use of KL divergences. Collectively, these studies underscore the adaptability and potential of GFlowNets in addressing a wide array of generative modeling challenges across diverse fields.

2.2 GFlowOut - Dropout with Generative Flow Networks

Liu et al. (2023) introduced GFlowOut, a novel solution to the challenges inherent in traditional dropout techniques used within neural networks. These challenges include the multi-modality of the posterior distribution over dropout masks and the difficulty in fully utilizing sample-dependent information and the correlation among dropout masks. GFlowOut leverages the principles of Generative Flow Networks (GFlowNets) to enhance dropout regularization by learning the posterior distribution over dropout masks. Traditional dropout methods often struggle to accurately capture the posterior due to the multimodal and discrete nature of dropout masks Liu et al. (2023); Jain et al. (2022).

GFlowOut addresses these limitations by employing GFlowNets to generate layer-wise dropout masks that are conditioned on previous layer activations and labels, thus improving the estimation of uncertainty and robustness to distributional shifts. Empirical evaluations have demonstrated that GFlowOut significantly outperforms standard methods, such as Random Dropout and Contextual Dropout, across a variety of tasks, including image classification under deformations, visual question answering, and real-world clinical predictions Liu et al. (2023). By utilizing the Trajectory Balance objective during training, GFlowOut ensures that the generated masks are proportionate to the reward function, providing a robust framework for improving posterior estimation and effectively leveraging sample-dependent information in neural networks. This results in enhanced generalization and superior performance in downstream tasks Liu et al. (2023); Malkin et al. (2023a).

3 Method

3.1 Model Structure

In our approach, we integrate learnable probabilistic discrete latent variables into established vision models by implementing GFlowOut within the architectures of ResNet18 and Vision Transformer, which serve as the backbone models. To achieve this, we modify specific layers of these models to incorporate GFlowOut functionality.

For the ResNet18 model, the standard dropout layers present after every residual block were removed by setting the dropout probability to 0. In their place, we introduced GFlowOut layers to manage the dropout process. This modification was consistently applied across all 12 residual blocks, though the implementation is flexible and can be customized to target specific blocks while omitting others as needed.

In the Vision Transformer architecture, we implemented dropout after every Attention-MLP block. Similar to our approach with ResNet18, the dropout probability for the standard dropout layers was set to 0, and GFlowOut layers were inserted to manage the dropout.

Both backbone models, ResNet18 and Vision Transformer, were pre-trained on the ImageNet dataset. Following pre-training, the final dense layers of these models were fine-tuned on the specific dataset utilized in this study. The GFlowOut layers are implemented as multi-layer perceptron (MLP) layers, which compute dropout probability distributions based on the context provided by previous layers and the input to the current layer, contingent on the masks used.

3.2 GFlowOut Masks

In this study, we employ four types of masks: `none`, `random`, `bottomup`, and `topdown`. The `none` mask indicates the absence of any applied mask. The `random` mask functions similarly to traditional dropout layers, applying a randomly generated mask, thereby mimicking the behavior of standard random dropout.

The `bottomup` mask generates dropout masks based on both the input data and the contextual information from previous layers, allowing for a more data-driven computation of the dropout probability distribution. In contrast, the `topdown` mask creates dropout masks solely based on the contextual information from preceding layers, without incorporating any direct input data.

We hypothesize that the `bottomup` masks will outperform the others, as they leverage additional data input to inform the computation of the dropout probability distribution, potentially leading to more effective regularization and improved model performance.

3.3 Eye Disease Dataset

Table 1: Summary of diseases in the Ocular Disease Intelligent Recognition (ODIR) dataset

Disease Class	**Count**
Normal	2873
Diabetes	1608
Glaucoma	284
Cataract	293
Age-related Macular Degeneration	266
Hypertension	128
Pathological Myopia	232
Other diseases/abnormalities	708

The Ocular Disease Intelligent Recognition (ODIR) dataset Maranhão (2020) is a comprehensive ophthalmic database consisting of records from 5,000 patients, including age information, color fundus photographs of both eyes, and diagnostic keywords provided by medical professionals. This dataset reflects a "real-life"

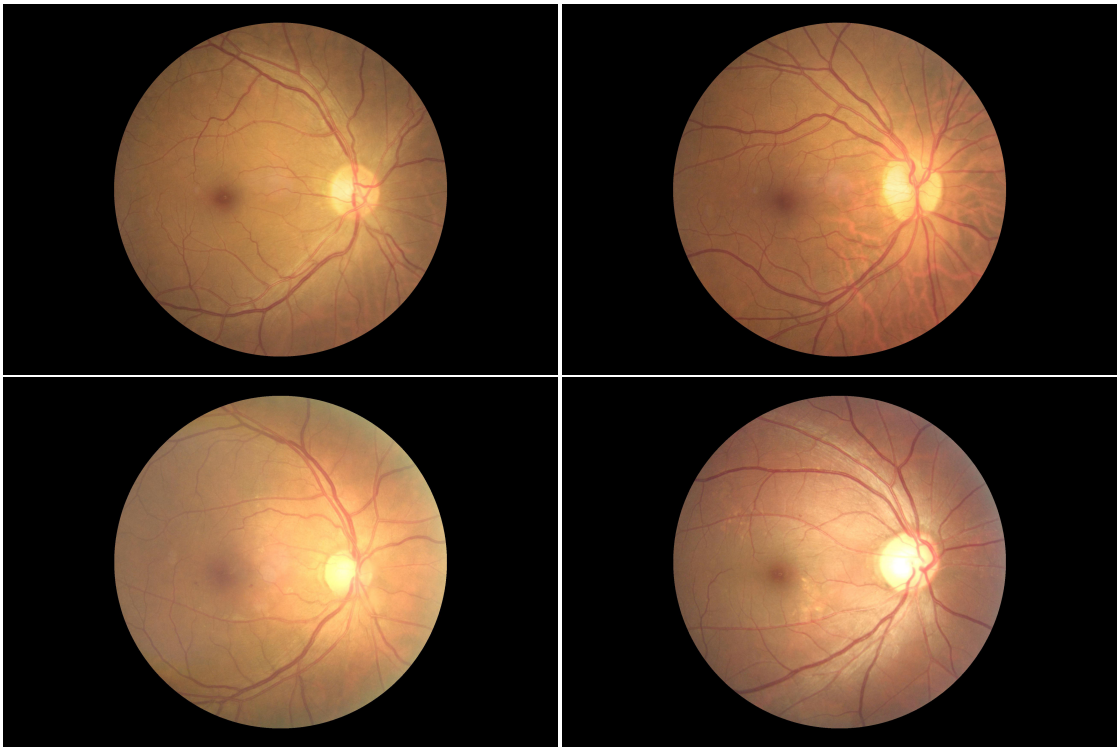

Figure 2: Sample images from the ODIR dataset. The top row displays left and right eye fundus photographs of a normal individual, i.e., a person not diagnosed with any ocular disease. The bottom row shows left and right eye fundus photographs of a patient with diabetes.

patient cohort, collected by Shanggong Medical Technology Co., Ltd., from multiple hospitals and medical centers across China. The fundus images in the database were captured using various commercially available cameras, such as Canon, Zeiss, and Kowa, leading to variations in image resolution.

The dataset categorizes images into several classes: normal, diabetes, glaucoma, cataract, age-related macular degeneration, hypertension, pathological myopia, and other diseases. For the purposes of this study, we focused on images labeled as normal and diabetes. These images were randomly shuffled and then split into training and testing datasets with an 80% and 20% allocation, respectively. Due to data constraints, we limited our study to these two classes, though the methodology employed can be extended to other diseases and multi-class classification problems. The pixel values of the images were normalized to lie within the range $[0, 1]$.

Prior to being fed into the model, the images underwent several pre-processing steps. Initially, the images were cropped to a size of $224 \times 224 \times 3$ pixels. They were then resized to $256 \times 256 \times 3$ pixels, normalized using means of $\mu = [0.485, 0.456, 0.406]$ and standard deviations of $\sigma = [0.229, 0.224, 0.225]$, and finally, bi-linear interpolation was applied. These pre-processing steps are consistent with the standard procedures for preparing inputs to ResNet and Vision Transformer models.

For experiments involving noise, Gaussian, Salt, and Speckle noise were added following the completion of the image pre-processing steps. These types of noise were introduced to simulate various noise conditions that could occur in clinical settings, where noisy data might be provided as input to the model. This approach aims to mimic such real-world conditions.

4 Experiments and Results

4.1 Eye Disease Detection Experiment

The models were trained using NVIDIA Tesla P100 GPUs for 100 epochs. The dataset was divided into training and testing subsets with a split ratio of `0.2`, ensuring a robust evaluation framework. During the training process, both models were subjected to all four different map types, with the results tabulated for comparative analysis. Our findings indicate that the Vision Transformer generally outperforms the ResNet model. However, when focusing on the same backbone model, the `bottomup` mask emerges as the superior performer, delivering the highest accuracy among the tested configurations. Conversely, the model with no mask applied exhibited the lowest accuracy levels, underscoring the critical role of appropriate masking strategies.

We also performed experiments with noise added to the images, which revealed insightful results. Models equipped with GFlowOut showed enhanced performance compared to their standard counterparts, even under noisy conditions. Remarkably, the accuracy of these models with GFlowOut remained comparable to scenarios involving non-noisy data, underscoring the robustness of the model against different types of noise. This robustness is a significant finding, highlighting the model's potential for practical applications where data imperfections are common.

Table 2: Experimental results of disease diagnosis . The above metrics mentioned are weighed averages. We note that the `bottomup` mask based on GFlowOut outperforms the other methods.

		Precision	**Recall**	**F1-Score**	**Accuracy**
ResNet18	`none`	0.66	0.58	0.61	52.72
	`random`	0.70	0.64	0.66	55.50
	`bottomup`	**0.85**	**0.83**	**0.83**	**69.94**
	`topdown`	0.73	0.69	0.70	64.67
Vision Transformer	`none`	0.64	0.68	0.65	69.04
	`random`	0.70	0.66	0.67	75.52
	`bottomup`	**0.91**	**0.89**	**0.89**	**83.26**
	`topdown`	0.75	0.71	0.72	79.89

Table 3: Robustness to noise experiments, with **Gaussian** noise applied to the images. The above metrics mentioned are weighed averages. We note that the `bottomup` mask based on GFlowOut outperforms the other methods.

		Precision	**Recall**	**F1-Score**	**Accuracy**
ResNet18	`none`	0.64	0.58	0.60	49.72
	`random`	0.67	0.66	0.65	52.66
	`bottomup`	**0.82**	**0.80**	**0.80**	**68.89**
	`topdown`	0.71	0.67	0.68	61.48
Vision Transformer	`none`	0.62	0.67	0.64	68.62
	`random`	0.69	0.66	0.67	71.65
	`bottomup`	**0.90**	**0.86**	**0.87**	**82.98**
	`topdown`	0.75	0.69	0.71	76.53

These results are in line with our expectations. The Vision Transformer, being both a larger and transformer-based model as compared to the ResNet-50 model, is expected to learn more features from the datasets and perform better at the task out of the models in consideration. Similarly, for a fixed backbone model, we expect the observed pattern in the various masks. The `none` mask performs the worst, since it is behaving as though there is no dropout. The `random` mask performs like a regular dropout layer, which is slightly better than having no dropout in these large models. `topdown` and `bottomup` perform better and the best

respectively, since they take into consideration the previous layer's context, and in the case of `bottomup` mask, the input data as well, to compute the probability distribution that is to be used for dropout.

4.2 Out of Distribution Evaluation and Entropy Calculations

To thoroughly evaluate the performance of our model, we tested it on out-of-distribution (OOD) datasets and calculated the entropy of the forward pass results. Specifically, we utilized the JSIEC dataset (JSIEC) as our OOD dataset for evaluation. The JSIEC dataset, recognized for its comprehensive and diverse collection of eye images, presents significant challenges, making it an ideal benchmark for assessing the robustness and generalization capabilities of the model.

Table 4: Robustness to noise experiments, with **Salt** noise applied to the images. The above metrics mentioned are weighed averages. We note that the `bottomup` mask based on GFlowOut outperforms the other methods.

		Precision	**Recall**	**F1-Score**	**Accuracy**
ResNet18	`none`	0.64	0.51	0.56	50.79
	`random`	0.68	0.66	0.66	48.20
	`bottomup`	**0.80**	**0.81**	**0.80**	**67.50**
	`topdown`	0.73	0.64	0.68	62.15
Vision Transformer	`none`	0.60	0.65	0.62	68.08
	`random`	0.66	0.66	0.66	75.99
	`bottomup`	**0.84**	**0.88**	**0.85**	**79.44**
	`topdown`	0.70	0.68	0.68	77.13

Table 5: Robustness to noise experiments, with **Speckle** noise applied to the images. The above metrics mentioned are weighed averages. We note that the `bottomup` mask based on GFlowOut outperforms the other methods.

		Precision	**Recall**	**F1-Score**	**Accuracy**
ResNet18	`none`	0.58	0.59	0.58	51.22
	`random`	0.69	0.60	0.64	52.22
	`bottomup`	**0.85**	**0.79**	**0.81**	**69.42**
	`topdown`	0.70	0.66	0.67	63.85
Vision Transformer	`none`	0.62	0.65	0.63	67.15
	`random`	0.69	0.68	0.68	75.36
	`bottomup`	**0.90**	**0.85**	**0.87**	**81.04**
	`topdown`	0.77	0.77	0.77	79.13

In our evaluation process, we performed multiple forward passes on both the training and evaluation datasets. By calculating the entropy of the outputs from these forward passes, we quantified the uncertainty in the model's predictions. Typically, higher entropy values indicate greater uncertainty, while lower entropy values suggest more confident predictions. By analyzing these entropy values, we identified patterns and differences in the model's performance on in-distribution versus out-of-distribution data. This analysis also enabled us to pinpoint specific images within the datasets associated with high or low entropy. Images with high entropy often highlight areas where the model struggles to make confident predictions, revealing potential weaknesses. Conversely, images with low entropy indicate areas where the model excels, making accurate and confident predictions.

Specifically, we conducted five forward passes using the ViT-GFN model on both the training and evaluation datasets. For each pass, we computed the minimum, maximum, and average entropy values. These results are systematically presented in Table 6. By examining high and low entropy images, we gained a deeper understanding of the types of data our model handles effectively and the types that pose challenges. This

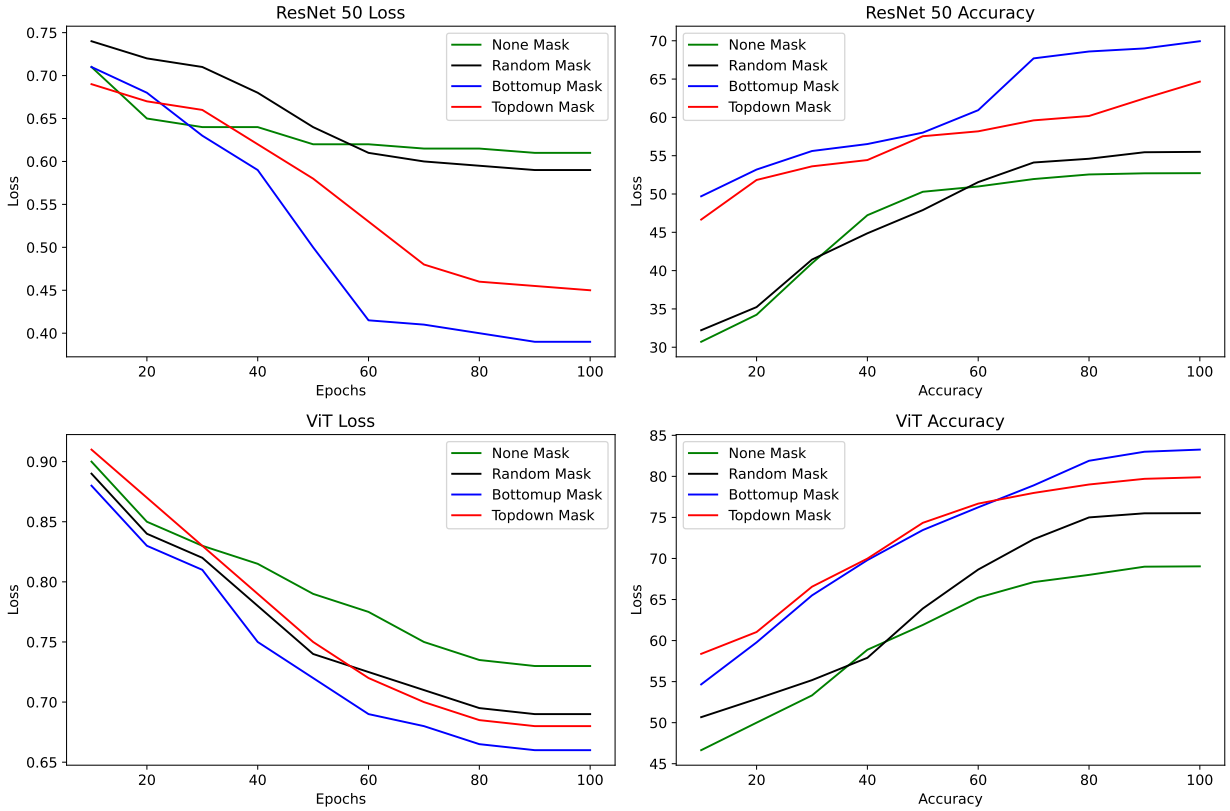

Figure 3: These plots show the loss curves and accuracy curves for the different models used. The top row has the metrics for ResNet18 model, and the bottom row has the metrics for the Vision Transformer model. We also plot metrics for each of the masks evaluated: `none`, `random`, `topdown` and `bottomup`.

information is crucial for guiding future improvements and fine-tuning the model to enhance its overall performance.

To further explore the explainability of our model, we visualized the attention maps of the Vision Transformer model. Using the PyTorch GradCAM implementation Gildenblat & contributors (2021), we generated attention maps and overlaid these maps on the original sample images. This visualization highlights the regions of the image deemed important by the model, thereby enhancing our understanding of the model's decision-making process.

Table 6: Entropy values on training and evaluation dataset

Dataset	**Min Entropy**	**Max Entropy**	**Avg Entropy**
ODIR	0.408967	0.663598	0.506282
JSIEC	0.470288	0.69244	0.564921

The entropy data provides significant insights into the model's performance. By analyzing the entropy values, we can identify which images our model handles well and which ones it struggles with. Images that exhibit the lowest entropy values, as shown in Figure 4, typically perform better. These images are often clear and well-centered, facilitating more accurate model predictions. Conversely, images with the highest entropy values, depicted in Figure 4, tend to perform worse. These problematic images are frequently either too bright or too dark, complicating the model's ability to make accurate predictions. Additionally, unclear or blurry images significantly degrade the model's performance, leading to lower accuracy rates.

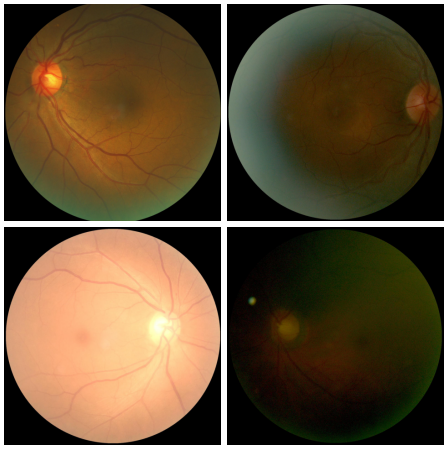

Figure 4: Fundus images from datasets with the minimum and maximum entropy. The top row consists of diabetic and normal fundus images, respectively, which have the minimum entropy. The bottom row consists of diabetic and normal fundus images, respectively, which has maximum entropy. We note that the model has highest confidence in its predictions when the image is clear, and the least confidence when the image is under or over-exposed.

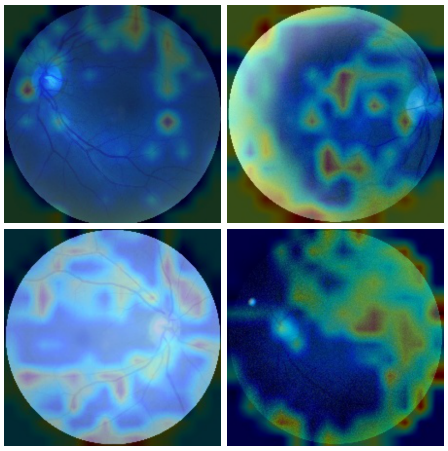

Figure 5: GradCAM analysis of the attention maps of the Vision Transformer. The top row consists of fundus images of diabetic and normal patients with minimum entropy. The bottom row consists of fundus images of diabetic and normal patients with maximum entropy. On top of these images, we apply the attention map computed using GradCAM to understand which parts are considered important by the model.

Finally, we computed the attention maps and superimposed them on the original sample images (Figure 5). We observed that the fundus images of diabetes patients highlighted specific vessels and areas deemed more important by the model. In contrast, the fundus images of normal patients showed more dispersed attention maps, indicating that no specific area of the image contributed predominantly to the classification output.

5 Conclusion

In this study, we present a novel methodology for advancing eye disease detection by integrating learnable probabilistic discrete latents via GFlowOut within ResNet18 and Vision Transformer architectures. Our approach has demonstrated substantial improvements in both accuracy and robustness, particularly under challenging conditions such as noisy data and out-of-distribution scenarios. Empirical evidence reveals

that the use of bottom-up and top-down dropout masks, specifically tailored to the dataset, significantly enhances model performance, surpassing the effectiveness of conventional dropout methods. Additionally, the entropy analysis provided critical insights into the model's predictive confidence, highlighting areas for further optimization.

By enhancing the model's capacity to generalize and manage uncertainty, our approach marks a pivotal advancement in the development of reliable AI-driven diagnostic tools for clinical applications. Future research should investigate the broader applicability of this method across other medical imaging domains and focus on refining the model to improve its interpretability and clinical relevance.

References

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems*, 34:27381–27394, 2021a.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation, 2021b. URL https://arxiv.org/abs/2106.04399.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. Gflownet foundations. *Journal of Machine Learning Research*, 24(210):1–55, 2023.

Renátó Besenczi, János Tóth, and András Hajdu. A review on automatic analysis techniques for color fundus photographs. *Computational and structural biotechnology journal*, 14:371–384, 2016.

Tristan Deleu, António Góis, Chris Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, and Yoshua Bengio. Bayesian structure learning with generative flow networks. In *Uncertainty in Artificial Intelligence*, pp. 518–528. PMLR, 2022.

Loic Le Folgoc, Vasileios Baltatzis, Sujal Desai, Anand Devaraj, Sam Ellis, Octavio E Martinez Manzanera, Arjun Nair, Huaqi Qiu, Julia Schnabel, and Ben Glocker. Is mc dropout bayesian? *arXiv preprint arXiv:2110.04286*, 2021.

Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021.

Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure FP Dossou, Chanakya Ajit Ekbote, Jie Fu, Tianyu Zhang, Michael Kilgour, Dinghuai Zhang, et al. Biological sequence design with gflownets. In *International Conference on Machine Learning*, pp. 9786–9801. PMLR, 2022.

Joint Shantou International Eye Centre (JSIEC). 1000 fundus images with 39 categories, October 2019. URL https://doi.org/10.5281/zenodo.3477553.

Feng Li, Yuguang Wang, Tianyi Xu, Lin Dong, Lei Yan, Minshan Jiang, Xuedian Zhang, Hong Jiang, Zhizheng Wu, and Haidong Zou. Deep learning-based automated detection for diabetic retinopathy and diabetic macular oedema in retinal fundus photographs. *Eye*, 36(7):1433–1441, 2022.

Ning Li, Tao Li, Chunyu Hu, Kai Wang, and Hong Kang. A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. In *Benchmarking, Measuring, and Optimizing: Third BenchCouncil International Symposium, Bench 2020, Virtual Event, November 15–16, 2020, Revised Selected Papers 3*, pp. 177–193. Springer, 2021.

Dianbo Liu, Moksh Jain, Bonaventure FP Dossou, Qianli Shen, Salem Lahlou, Anirudh Goyal, Nikolay Malkin, Chris Chinenye Emezue, Dinghuai Zhang, Nadhir Hassen, et al. Gflowout: Dropout with generative flow networks. In *International Conference on Machine Learning*, pp. 21715–21729. PMLR, 2023.

Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in gflownets, 2023a. URL https://arxiv.org/abs/2201.13259.

Nikolay Malkin, Salem Lahlou, Tristan Deleu, Xu Ji, Edward Hu, Katie Everett, Dinghuai Zhang, and Yoshua Bengio. Gflownets and variational inference, 2023b. URL https://arxiv.org/abs/2210.00580.

Andrew Maranhão. Ocular disease intelligent recognition (odir), 2020. URL https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k.

Mizu Nishikawa-Toomey, Tristan Deleu, Jithendaraa Subramanian, Yoshua Bengio, and Laurent Charlin. Bayesian learning of causal structure and mechanisms with gflownets and variational bayes. *arXiv preprint arXiv:2211.02763*, 2022.

Kangrok Oh, Hae Min Kang, Dawoon Leem, Hyungyu Lee, Kyoung Yul Seo, and Sangchul Yoon. Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images. *Scientific reports*, 11(1):1897, 2021.

Srikanta Kumar Padhy, Brijesh Takkar, Rohan Chawla, and Atul Kumar. Artificial intelligence in diabetic retinopathy: A natural step to the future. *Indian journal of ophthalmology*, 67(7):1004, 2019.

Ling Pan, Nikolay Malkin, Dinghuai Zhang, and Yoshua Bengio. Better training of gflownets with local credit and incomplete trajectories. In *International Conference on Machine Learning*, pp. 26878–26890. PMLR, 2023a.

Ling Pan, Dinghuai Zhang, Moksh Jain, Longbo Huang, and Yoshua Bengio. Stochastic generative flow networks. In *Uncertainty in Artificial Intelligence*, pp. 1628–1638. PMLR, 2023b.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.

Jaemin Son, Joo Young Shin, Hoon Dong Kim, Kyu-Hwan Jung, Kyu Hyung Park, and Sang Jun Park. Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. *Ophthalmology*, 127(1):85–94, 2020.

Daniel Shu Wei Ting, Louis R Pasquale, Lily Peng, John Peter Campbell, Aaron Y Lee, Rajiv Raman, Gavin Siew Wei Tan, Leopold Schmetterer, Pearse A Keane, and Tien Yin Wong. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 2018.

Zhaoran Wang, Pearse A Keane, Michael Chiang, Carol Y Cheung, Tien Yin Wong, and Daniel Shu Wei Ting. Artificial intelligence and deep learning in ophthalmology. In *Artificial Intelligence in Medicine*, pp. 1519–1552. Springer, 2022.

Wan Ling Wong, Xinyi Su, Xiang Li, Chui Ming G Cheung, Ronald Klein, Ching-Yu Cheng, and Tien Yin Wong. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *The Lancet Global Health*, 2(2):e106–e116, 2014.

WTO. World report on vision, 2019. `https://www.who.int/publications/i/item/9789241516570`. Accessed: 20 Dec. 2023.

Xiayan Xu, Jing Wu, Xiaoning Yu, Yelei Tang, Xiajing Tang, and Xingchao Shentu. Regional differences in the global burden of age-related macular degeneration. *BMC Public Health*, 20:1–9, 2020.

Jie Yang, Simon Fong, Han Wang, Quanyi Hu, Chen Lin, Shigao Huang, Jian Shi, Kun Lan, Rui Tang, Yaoyang Wu, et al. Artificial intelligence in ophthalmopathy and ultra-wide field image: A survey. *Expert Systems with Applications*, 182:115068, 2021.

Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10):719–731, 2018.

Dinghuai Zhang, Nikolay Malkin, Zhen Liu, Alexandra Volokhova, Aaron Courville, and Yoshua Bengio. Generative flow networks for discrete probabilistic modeling. In *International Conference on Machine Learning*, pp. 26412–26428. PMLR, 2022.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

Heiko Zimmermann, Fredrik Lindsten, Jan-Willem van de Meent, and Christian A Naesseth. A variational perspective on generative flow networks. *Transactions on Machine Learning Research*, 2023.