Debiased Noise Editing on Foundation Models for Fair Medical Image Classification

Ruinan Jin[1,2], Wenlong Deng[1,2], Minghui Chen[1,2], Xiaoxiao Li[1,2]
[1]The University of British Columbia [2]Vector Institute

Paper Code

I. Overview

Background: Foundation models (FMs), show promise in adapting to various medical imaging tasks, such as using FMs as APIs to generate representations for linear probing. However, this approach carries inherent biases in using FMs.

Objective: Improve the fairness of using FM APIs for medical image classification.

Challenge: Existing debiasing methods become impractical due to:

1. Computational inefficiency in updating FMs.

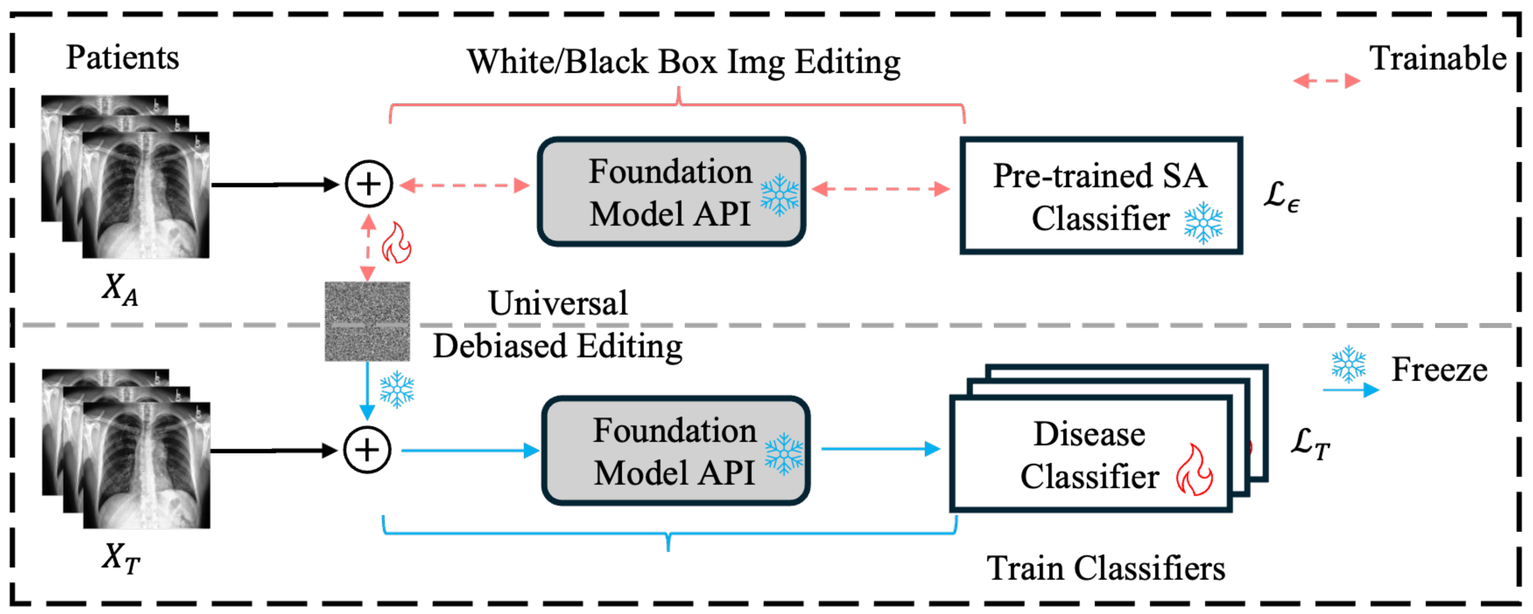
2. Infeasibility with black-box FM APIs, restricting access to model internals.

Solution:

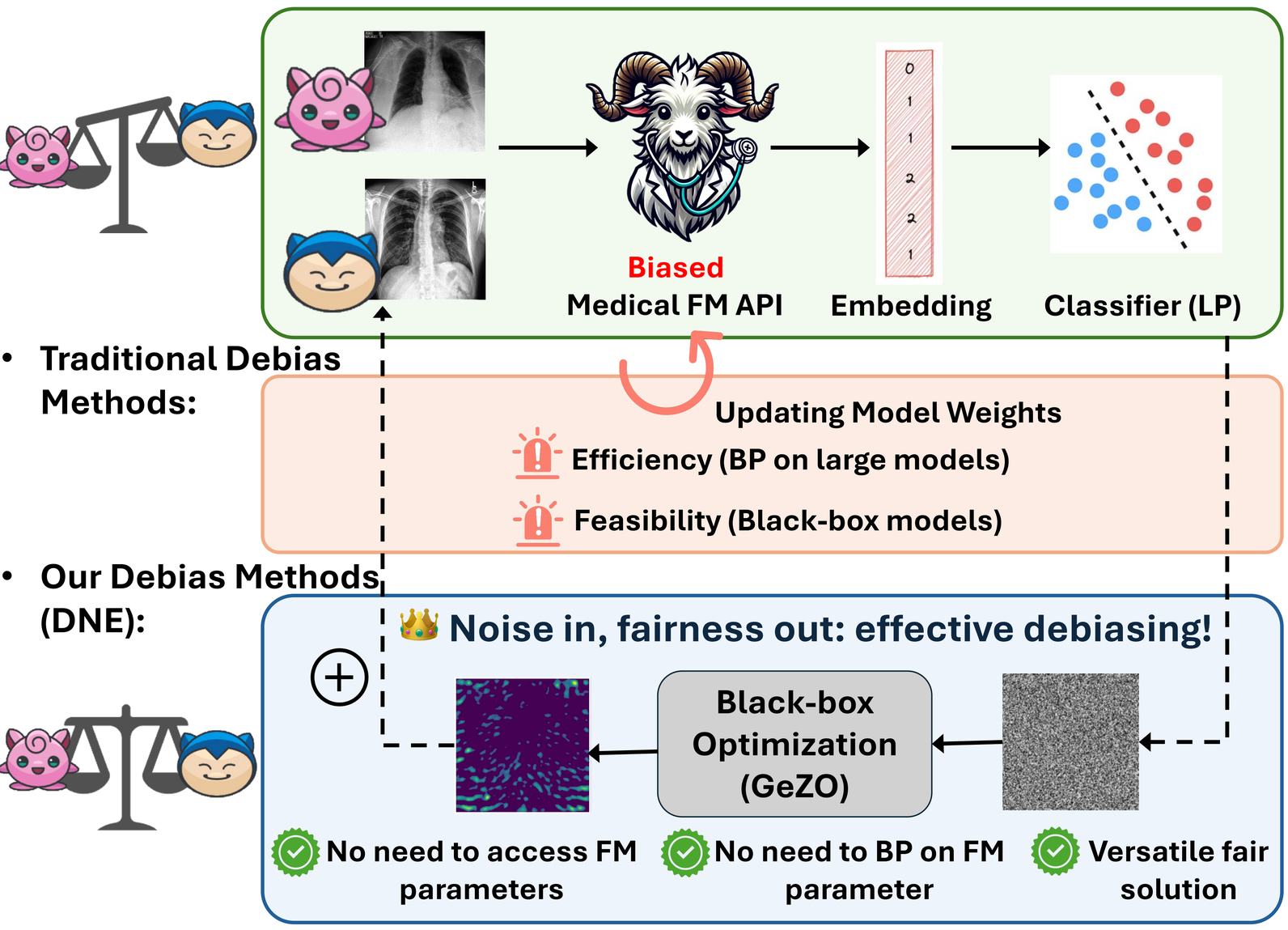
1. **D**ebiased **N**oise **E**diting (**DNE**), learnable noises added to input data to reduce bias.

2. **Gre**edy **Z**eroth-order **O**ptimization (**GeZO**), to update the DNE for black-box FM APIs.

II. Our Proposed Debiasing Pipeline (DNE)



Stage 1 (upper) - Utilizing the FM API's embedding to train a sensitive attribute (SA) classifier.
Stage 2 (middle) - Initialize the DNE noise (a vector with the size of the image) and update it to fool the SA classifier in previous step.
Stage 3 (bottom) - Train the disease classifier with the FM API's embedding with the DNE noise added to every input images.

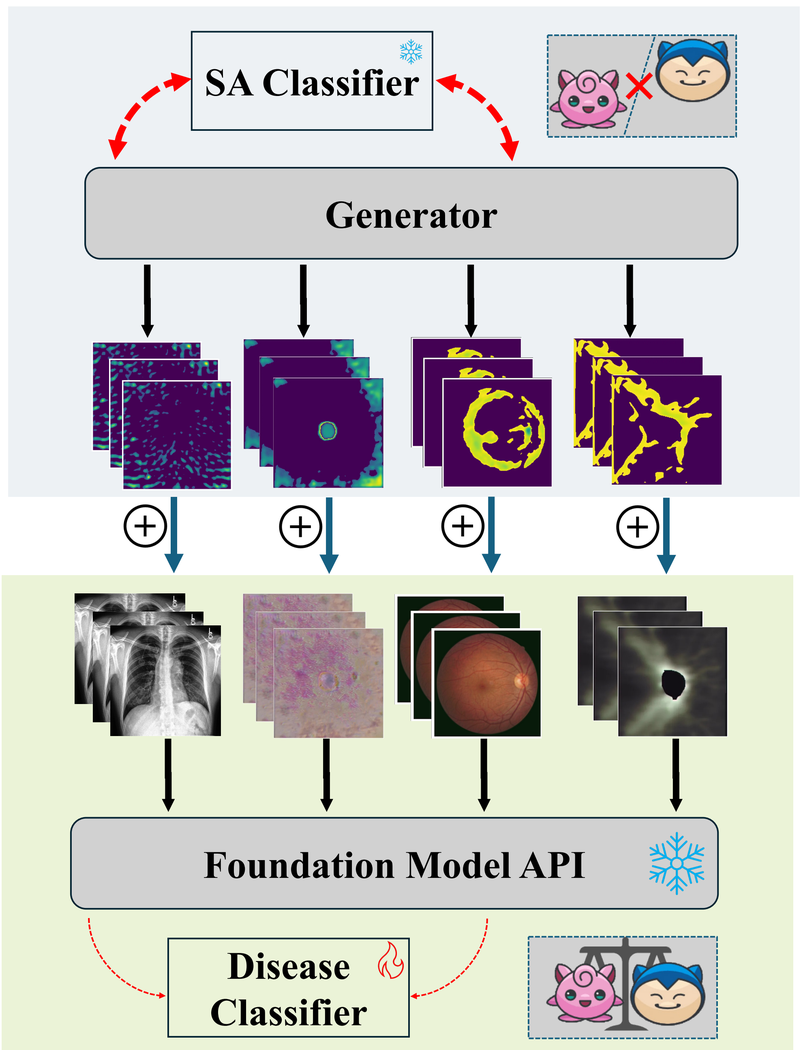


III. DNE ++ (ongoing extension)

1- Condition the generation of DNE based on the input image.

2- An updated black-box optimization algorithm for this version of DNE.

3- Benchmark the effectiveness of DNE on ***diverse*** medical modalities.



IV. Comparative Results

Comparison of binary prediction (% ACC) of Pleural Effusion, Pneumonia, and Edema. We label the best performance in bold and the second-best performance with underline.

Diseases	Pleural Effusion				Pneumonia				Edema			
	$EO_n \downarrow$	$EO_p \downarrow$	|1-DI| $\downarrow$	Acc $\uparrow$	$EO_n \downarrow$	$EO_p \downarrow$	|1-DI| $\downarrow$	Acc $\uparrow$	$EO_n \downarrow$	$EO_p \downarrow$	|1-DI| $\downarrow$	Acc $\uparrow$
ERM	40.5	57.0	58.5	<u>72.9</u>	70.0	70.0	74.5	59.6	42.0	39.0	42.0	74.5
Sketch [27]	44.0	52.0	57.5	66.3	64.0	61.0	72.6	56.3	44.0	**15.0**	**15.5**	67.0
Group DRO [18]	41.0	58.0	59.5	72.3	60.0	56.0	64.4	61.0	40.5	40.5	43.3	74.8
Batch Samp. [17]	41.5	50.5	51.8	74.3	64.0	72.0	76.6	60.5	39.0	43.0	44.8	**76.0**
BiasAdv [12]	39.0	54.5	58.2	71.1	<u>36.0</u>	54.0	77.1	61.0	39.0	33.5	36.6	74.9
DNE	**28.5**	**23.0**	**25.6**	**75.1**	38.0	**25.0**	<u>34.7</u>	<u>61.3</u>	**27.5**	17.0	19.7	<u>75.6</u>
DNE-GeZO	<u>37.0</u>	<u>25.0</u>	<u>28.5</u>	72.8	**35.0**	<u>27.0</u>	**33.7**	**61.5**	<u>34.0</u>	<u>15.5</u>	<u>17.2</u>	75.1

VI. Conclusion and Future Directions

- DNE is the first attempt to introduce an interpretable image editing strategy for mitigating the biases of FM APIs in black-box scenarios.
- Another effort (**FairMedFM,** Jin et al., arXiv:2407.00983**)** is to benchmarking fairness and performance of FMs in medical image analysis.