XTranPrune: eXplainability-aware Transformer Pruning for Bias Mitigation in Dermatological Disease Classification

Ali Ghadiri[1](✉), Maurice Pagnucco[1], and Yang Song[1]

University of New South Wales, Sydney, Australia
a.ghadiri@unsw.edu.au

A Supplementary Material

A.1 Pseudocode of XTranPrune

Algorithm 1 : One Iteration of XTranPrune

Input: *Main Model, Sensitive Attribute (SA) Model*
Parameter: $RetainRate(RR), PruningRate(PR), NumBatch$
Output: $PruningMask$

```
# Initialize main_Attr as an all-zero matrix.
# Initialize SA_Attr as an all-zero matrix.

for i = 1 in range(NumBatch) do
    main_Attr ← main_Attr + ExplainabilityMethod(Main Model, batch(i))
    SA_Attr ← SA_Attr + ExplainabilityMethod(SA Model, batch(i))
end for
# Get the average of the attributions over the batches
main_Attr ← main_Attr/NumBatch
SA_Attr ← SA_Attr/NumBatch

for block in Main Model do
    for head in block do
        # Setting the mask to zero for the RR% of the most important nodes
        PerformanceMask ← SelectNodes(main_Attr)
        SA_Attr ← SA_Attr * PerformanceMask

        # Setting the mask to zero for the PR% nodes with the highest SA_Attr
        PruningMaskHead ← SelectNodes(SA_Attr)
    end for
    PruningMaskBlock ← Concatenate(PruningMaskHead)
end for
PruningMask ← Concatenate(PruningMaskBlock)
return PruningMask
```

A.2 More Information about datasets

Table 1. More Information about the datasets. In the Fitzpatrick17k dataset, we excluded the pictures without the Fitzpatrick scale, leaving us with 16,012 images in total in our experiments. Similarly, in the PAD-UFES-20 dataset, the subgroup with skin tone 6 has been excluded due to having only one record.

Dataset	Skin Condition	Skin Type						Total
		T1	T2	T3	T4	T5	T6	
Fitzpatrick17k	Benign	444	671	475	367	159	44	2160
	Malignant	453	742	456	301	147	61	2160
	Non-neoplastic	2050	3395	2377	2113	1227	530	11692
	Total	2947	4808	3308	2781	1533	635	16012
PAD-UFES-20	ACK	25	156	89	10	3	-	283
	BCC	101	502	217	22	3	-	845
	MEL	3	35	10	4	0	-	52
	NEV	4	34	24	12	1	-	75
	SCC	16	126	45	3	2	-	192
	SEK	4	23	7	11	1	-	46
	Total	153	876	392	62	10	-	1493

Fig. 1. Sample images of different skin tones in the Fitzpatrick17K dataset.

Fig. 2. Sample images of different conditions in the Fitzpatrick17K dataset.

Fig. 3. Sample images of different skin tones in the PAD-UFES-20 dataset.

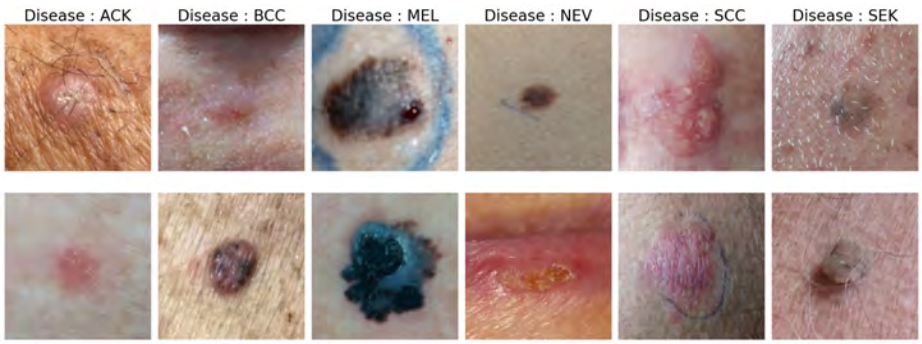

Fig. 4. Sample images of different conditions in the PAD-UFES-20 dataset.

A.3 Fairness Metrics

Table 2. Metrics used fairness evaluation.

Metric	Formula	Description
DPM	$\frac{1}{M}\sum_{i=1}^{M}\frac{\min[p(\hat{y}=i \mid s=j), j\in S]}{\max[p(\hat{y}=i \mid s=j), j\in S]}$	Measures the ratio of the worst-case positive outcome probability to the best-case according to the sensitive attribute across all subclasses in the classification task.
EOM	$\frac{1}{M}\sum_{i=1}^{M}\frac{\min[p(\hat{y}=i \mid y=i, s=j), j\in S]}{\max[p(\hat{y}=i \mid y=i, s=j), j\in S]}$	Computes the same ratio as DPM, but considers the true positive rate.
EOpp0	$\sum_{m=1}^{M}\left\lvert TNR_m^1 - TNR_m^0\right\rvert$	Calculates the True Negative Rate disparity between subgroups.
EOpp1	$\sum_{m=1}^{M}\left\lvert TPR_m^1 - TPR_m^0\right\rvert$	Measures the True Positive Rate disparity between subgroups.
EOdd	$\sum_{m=1}^{M}\left\lvert TPR_m^1 - TPR_m^0 + FPR_m^1 - FPR_m^0\right\rvert$	Focuses on the overall positive rate disparity.
NFR	$\frac{F1\ score_{\max} - F1\ score_{\min}}{\text{mean}(F1\ score)}$	By lower NFR we promote minimal disparity among subgroups while the overall performance is high.

A.4 Extended Results on the Fitzpatrick17k dataset

Table 3. Additional results on the Fitzpatrick17k dataset.

Model	F1 score (%)	Worst-case F1 score (%)↑	DPM↑	EOM↑	EOpp0↓	EOpp1↓	EOdd↓	NFR↓
FairTune[1]	66.80	54.44	0.538	0.686	0.114	0.104	0.195	0.238
DomainInd[2]	69.06	64.26	0.571	0.714	**0.055**	0.128	0.139	0.119
XTranPrune	**73.51**	**69.13**	**0.586**	**0.790**	0.086	**0.066**	**0.095**	**0.114**

Table 4. Additional results on the PAD-UFES-20 dataset.

Model	F1 score (%)	Worst-case F1 score (%)↑	DPM↑	EOM↑	EOpp0↓	EOpp1↓	EOdd↓	NFR↓
DomainInd[2]	62.51	41.58	0.005	0.462	0.764	1.440	1.796	1.578
XTranPrune	62.01	**57.03**	**0.009**	**0.624**	**0.389**	**1.141**	**0.909**	**0.587**

A.5 Ablation Study on the PAD-UFES-20 dataset

Table 5. Ablation study on node attribution calculation method for the PAD-UFES-20 dataset.

Method	F1 score (%)	Worst-case F1 score (%)↑	DPM↑	EOM↑	EOpp0↓	EOpp1↓	EOdd↓	NFR↓
Attention Map	61.91	57.02	0.009	0.560	0.453	1.420	1.303	0.603
Gradients	48.19	26.54	0.002	0.352	0.455	1.217	1.600	1.412
LRP	61.88	53.28	0.009	0.524	0.391	1.224	1.051	0.665
Our method	**62.01**	**57.03**	**0.009**	**0.624**	**0.389**	**1.141**	**0.909**	**0.587**

References

1. Dutt, R., Bohdal, O., Tsaftaris, S.A., Hospedales, T.: Fairtune: Optimizing parameter efficient fine tuning for fairness in medical image analysis. arXiv preprint arXiv:2310.05055 (2023)
2. Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., Russakovsky, O.: Towards fairness in visual recognition: Effective strategies for bias mitigation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8919–8928 (2020)