A Appendix

Model evaluation

Table A and Table B are extended version of Table 1 and Table 2. These tables include results for the augmented and non-augmented versions of training data. We observe that the augmented version is most of the time better than the non-augmented version in terms of scoring metrics.

Model	Accuracy	Precision	Recall	F1 Score
VGG16	0.948	0.753	**0.835**	0.792
VGG16-A	0.955	0.816	0.794	0.805
ResNet50	**0.956**	**0.883**	0.726	**0.797**
ResNet50-A	**0.964**	**0.880**	**0.808**	**0.842**
DenseNet121	0.948	0.830	0.698	0.761
DenseNet121-A	0.959	0.864	0.780	0.820
EfficientNetB0	0.943	0.788	0.712	0.748
EfficientNetB0-A	0.961	0.865	0.795	0.829

Table A: Performance metrics for different models using Bangladesh dataset for training and testing.

Model	Accuracy	Precision	Recall	F1 Score
VGG16	**0.893**	0.805	**0.810**	**0.807**
VGG16-A	0.837	0.920	0.451	0.605
ResNet50	0.863	0.943	0.536	0.683
ResNet50-A	0.906	0.963	0.682	0.801
DenseNet121	0.840	**0.985**	0.431	0.600
DenseNet121-A	**0.907**	0.939	**0.712**	**0.810**
EfficientNetB0	0.870	0.966	0.549	0.700
EfficientNetB0-A	0.905	**1.000**	0.660	0.795

Table B: Performance metrics for different vanilla and augmented models using Bangladesh for training and Indian dataset for testing

Active Learning

In Figure A, we show the active learning plot for all metrics for training and testing on the Indian dataset. This figure is an extended version of Figure 3.

In Figure B and C, we present the performance outcomes of active learning within the transfer context. In this approach, we initially fine-tune the model using datasets from both India and Bangladesh. The model then undergoes querying using the Indian dataset. Notably, during the initial stages of active learning iterations, This model outperforms a model fine-tuned solely on a limited set of Indian dataset images shown in Figure A. However, the performance gain via active learning is not much compared to the random baseline.

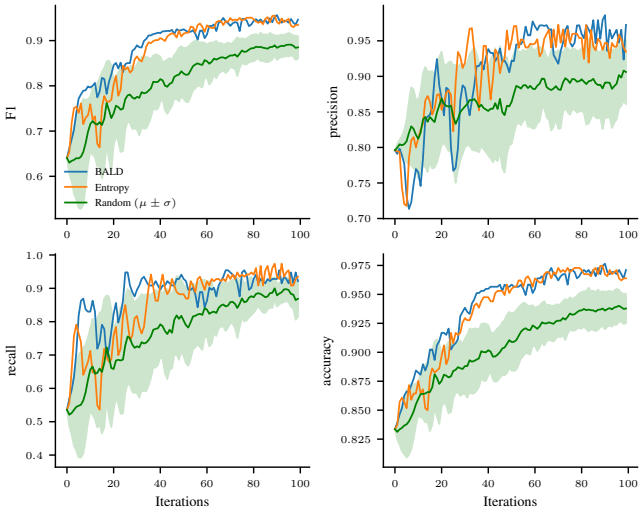

Figure A: Different metric scores with active learning iterations for training and testing on Indian dataset.

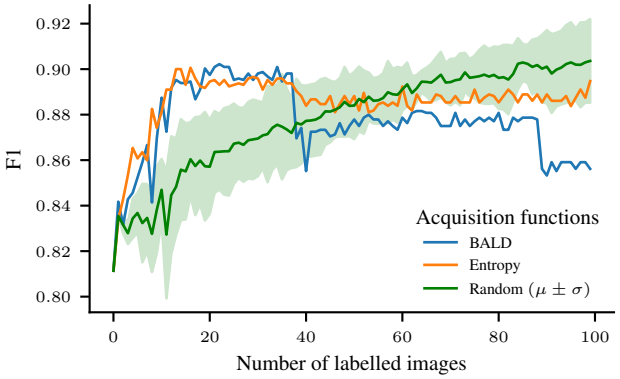

Figure B: F1 score trend with the increasing number of labeled images for training on Bangladesh and Indian datasets and testing on Indian datasets. We observe that the performance of BALD and entropy does not gain much over random after 40 iterations.

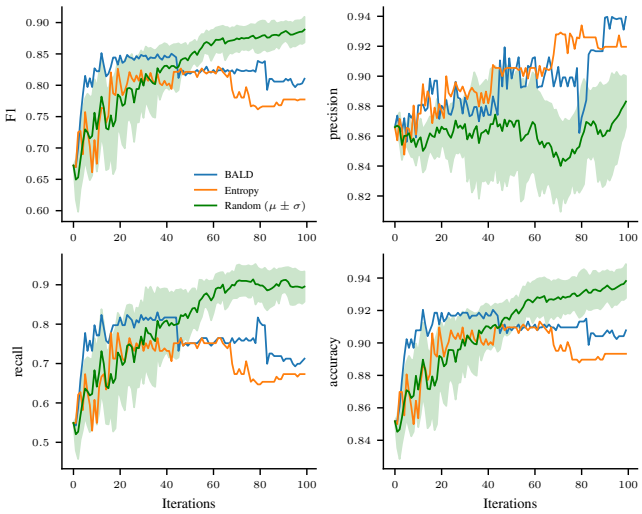

Figure C: Different metric scores with an increasing number of active learning iterations for training on Bangladesh and Indian datasets and testing on Indian dataset.