Supplementary Material: Modeling and Understanding Uncertainty in Medical Image Classification

1 More Experimental Details

Table 1. Data information of **ISIC 2018**. ISIC 2018 consists of 7 skin diseases: *melanoma*, *melanocytic nevus*, *basal cell carcinoma*, *actinic keratosis / Bowen's disease (intraepithelial carcinoma)*, *benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis)*, *dermatofibroma*, and *vascular lesion*.

# of total samples	# of training samples	# of validation samples	# of test samples
10,015	6,409	2,003	1,603

Table 2. Data information of **BloodMNIST**. BloodMNIST contains 8 classes of normal cells: *basophil*, *eosinophil*, *erythroblast*, *immature granulocytes (myelocytes, metamyelocytes, and promyelocytes)*, *lymphocyte*, *monocyte*, *neutrophil*, and *platelet*.

# of total samples	# of training samples	# of validation samples	# of test samples
17,092	11,959	1,712	3,421

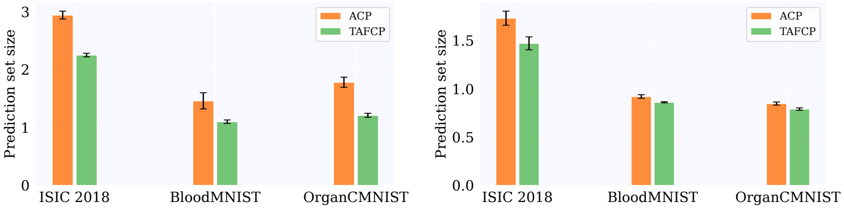
Table 3. Data information of **OrganCMNIST**. OrganCMNIST has 11 categories of organs: *bladder*, *femur-left*, *femur-right*, *heart*, *kidney-left*, *kidney-right*, *liver*, *lung-left*, *lung-right*, *pancreas*, and *spleen*.

# of total samples	# of training samples	# of validation samples	# of test samples
23,583	12,975	2,392	8,216

2 More Experimental Results

Table 4. Test accuracy of medical image classification tasks using ResNet-18 and a 2-layer convolutional neural network (CNN).

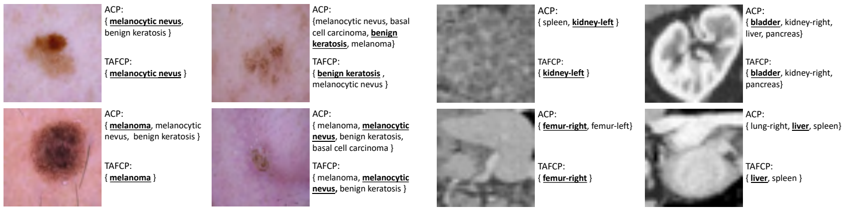
Classifier	BloodMNIST	OrganCMNIST	ISIC 2018
ResNet-18	96.08%	89.04%	72.92%
CNN	89.71%	73.39%	67.68%



(a) Significance level ε = 0.05

(b) Significance level ε = 0.15

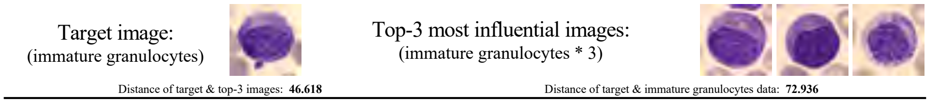
Fig. 1. Efficiency comparison under different significance levels. A smaller size implies better efficiency. We observe that TAFCP consistently outperforms the ACP baseline by yielding smaller prediction set sizes.


ACP:
{ melanocytic nevus, benign keratosis }
TAFCP:
{ melanocytic nevus }
ACP:
{melanocytic nevus, basal cell carcinoma, benign keratosis, melanoma}
TAFCP:
{ benign keratosis , melanocytic nevus }
ACP:
{ melanoma, melanocytic nevus, benign keratosis }
TAFCP:
{ melanoma }
ACP:
{ melanoma, melanocytic nevus, benign keratosis, basal cell carcinoma }
TAFCP:
{ melanoma, melanocytic nevus, benign keratosis }
ACP:
{ spleen, kidney-left }
TAFCP:
{ kidney-left }
ACP:
{ bladder, kidney-right, liver, pancreas}
TAFCP:
{ bladder, kidney-right, pancreas}
ACP:
{ femur-right, femur-left}
TAFCP:
{ femur-right }
ACP:
{ lung-right, liver, spleen}
TAFCP:
{ liver, spleen }

(a) ISIC 2018

(b) OrganCMNIST

Fig. 2. Conformal sets comparison. Bold and underlined phrases mean true labels. As depicted, TAFCP produces more efficient conformal sets that include the true diseases or organ categories.


Target image:
(immature granulocytes)
Top-3 most influential images:
(immature granulocytes * 3)
Distance of target & top-3 images: 46.618
Distance of target & immature granulocytes data: 72.936

	Before update	After update
Prediction set	{ **immature granulocytes**, monocyte }	{ **immature granulocytes**, monocyte, erythroblast }

Fig. 3. Our proposed prediction uncertainty explanations on the BloodMNIST dataset. As shown, after deleting the top-3 most influential images from the training set, an extra cell category is included. This is due to the close proximity of these excluded images to the target sample (classified as “immature granulocytes”), resulting in insufficient training in this region. Therefore, the “erythroblast” label is included in the prediction set, and the uncertainty is amplified.