# Evidential Deep Learning for Uncertainty-Aware Mobile Health

*Tong Xia, Dr. Jing Han, Dr. Lorena Qendro, Prof. Cecilia Mascolo*

*Department of Computer Science and Technology*

UNIVERSITY OF CAMBRIDGE

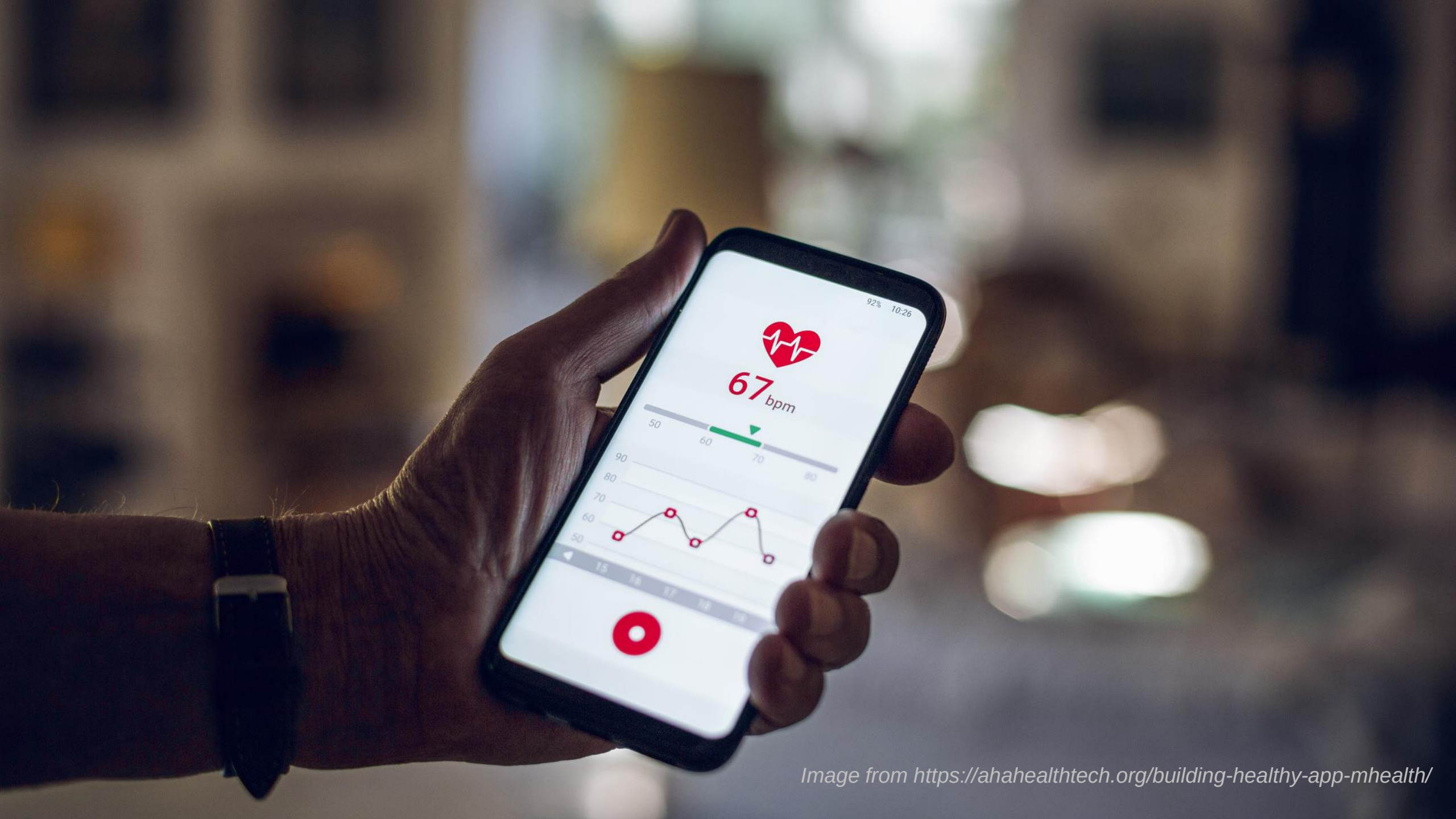
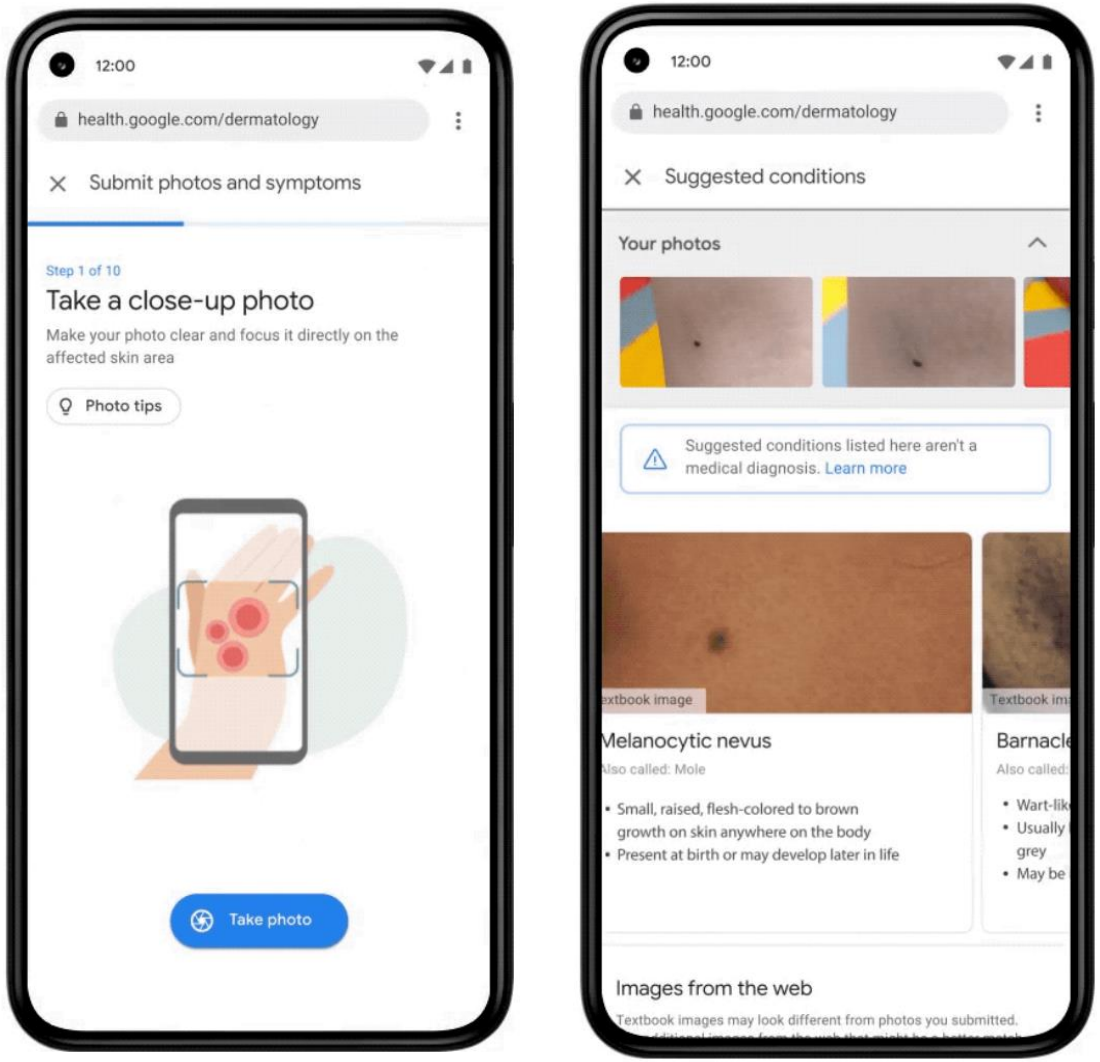
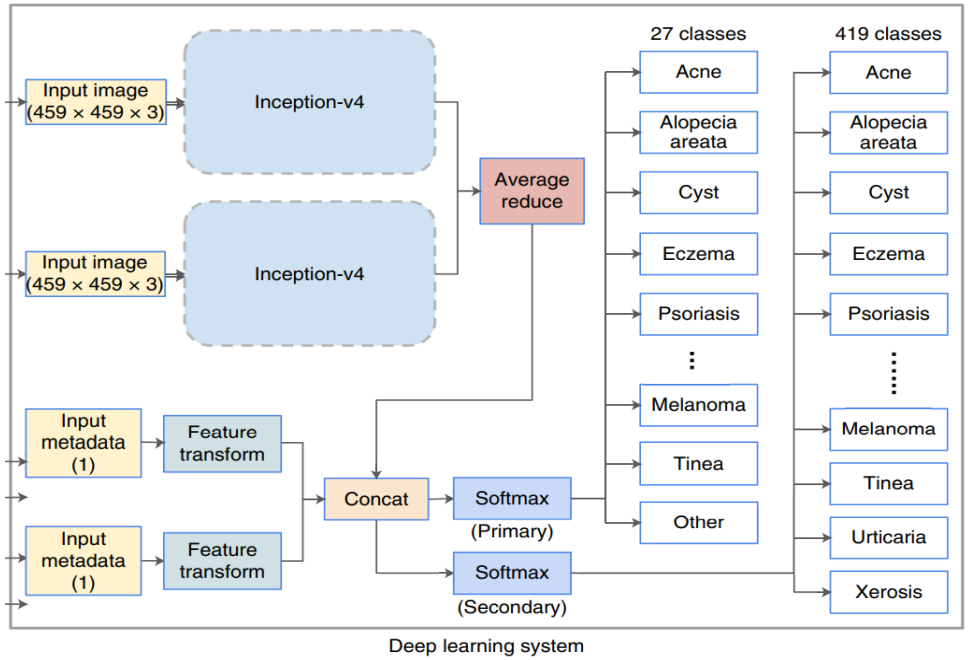



*Image from https://ahahealthtech.org/building-healthy-app-mhealth/*

# Google AI-powered Dermatology Tool



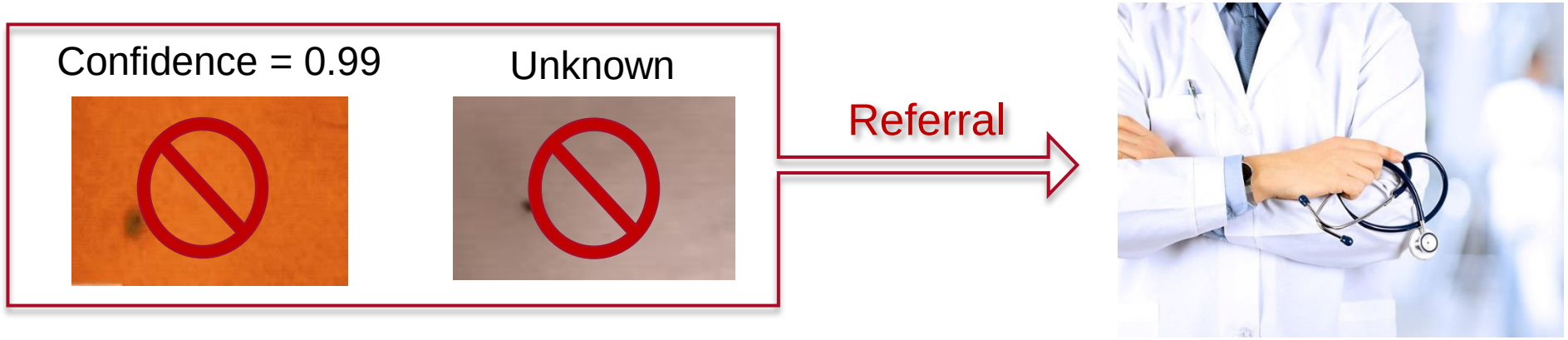

Deep learning system

**Misdiagnosis Risk**:

- Unexpected environment/lighting
- Difference among cameras
- Uncommon skin condition
- …

*https://blog.google/technology/health/ai-dermatology-preview-io-2021/*

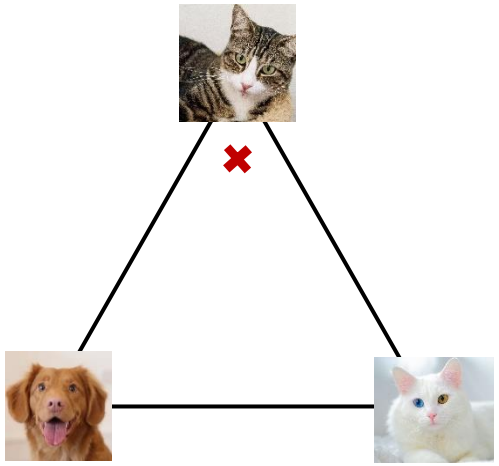
# Why uncertainty is important for mHealth



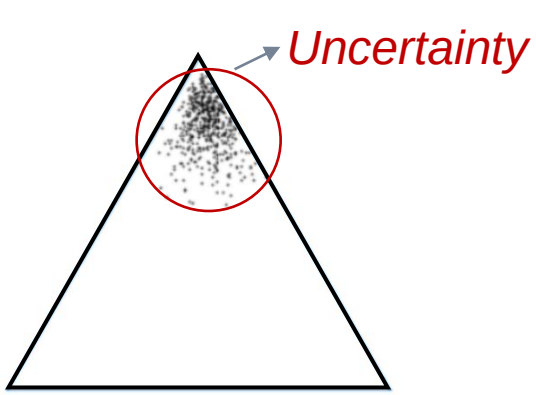

**Uncertainty Quantification for Health**:

- Calibration: the confidence of prediction associated with a trained condition
- OOD aware: the likelihood of an input belonging to the out-of-distribution regime of training data
- Low cost: no extra inference cost, no extra training data required

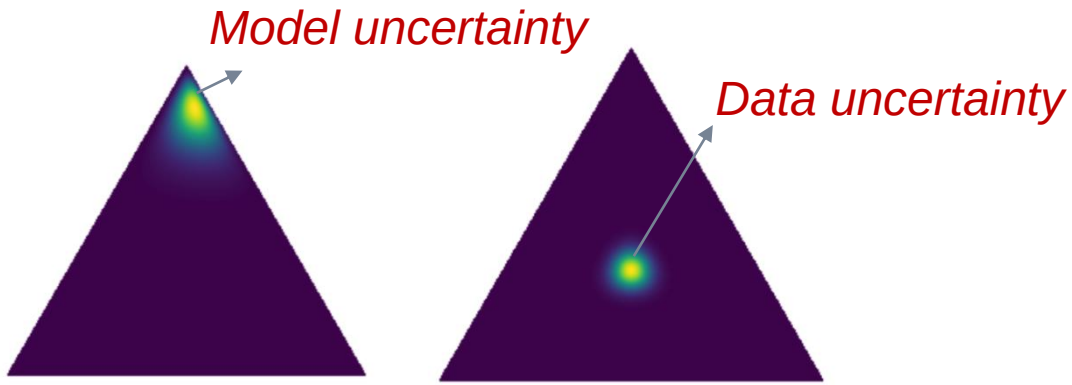
# Evidential Deep Learning (EDL) for Uncertainty Quantification



a) Single Softmax Model




b) Ensemble Softmax Models




c) A single Deep Evidential Model

*Figure 1: 3-classification problem. Brighter colors correspond to higher density.*

*Dennis Ulmer. A Survey on Evidential Deep Learning For Single-Pass Uncertainty Estimation. 2021*

# Evidential Deep Learning (EDL) for Uncertainty Quantification

❑ A Dirichlet distribution $Dir(\boldsymbol{\alpha})$ is parameterised by a group of concentrations denoted by $\boldsymbol{\alpha}$ = [$\alpha$1, $\alpha$2, ..., $\alpha k$ ], corresponding to a $k$-class categorical problem.

- Predictive probability:

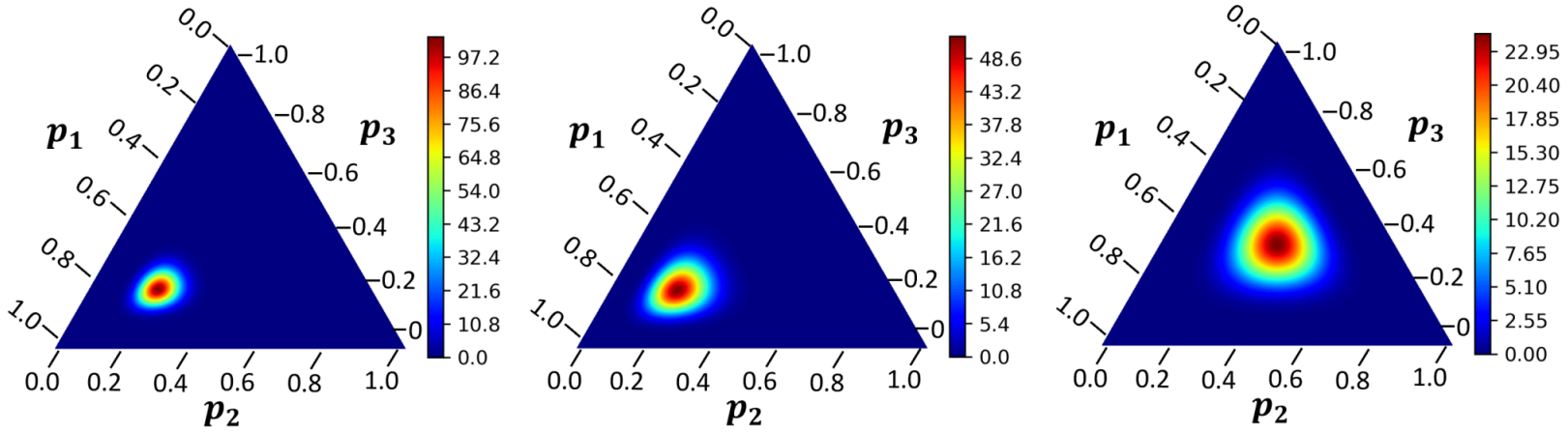
$$\mathbb{E}[p_c] = \frac{\alpha_c}{\alpha_0},$$

- Prediction:

$$\hat{y}^{(i)} = \arg\max_c \mathbb{E}[p_c^{(i)}],$$

- Uncertainty:

$$DE = \mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha}\cdot\boldsymbol{\omega})}[-In(\boldsymbol{p})]$$




(a) $\boldsymbol{\alpha} = [60,20,20]$
$\mathbb{E}[\boldsymbol{p}] = [0.6,0.2,0.2]$
$\boldsymbol{\alpha_0} = 100, \boldsymbol{DE} = -3.66$

(b) $\boldsymbol{\alpha} = [30,10,10]$
$\mathbb{E}[\boldsymbol{p}] = [0.6,0.2,0.2]$
$\boldsymbol{\alpha_0} = 50, \boldsymbol{DE} = -2.99$

(c) $\boldsymbol{\alpha} = [10,10,10]$
$\mathbb{E}[\boldsymbol{p}] = [0.33,0.33,0.33]$
$\boldsymbol{\alpha_0} = 30, \boldsymbol{DE} = -2.27$

Figure 2: Dirichlet probability density functions for a three class setting.

# An EDL Framework for mHealth with class imbalance

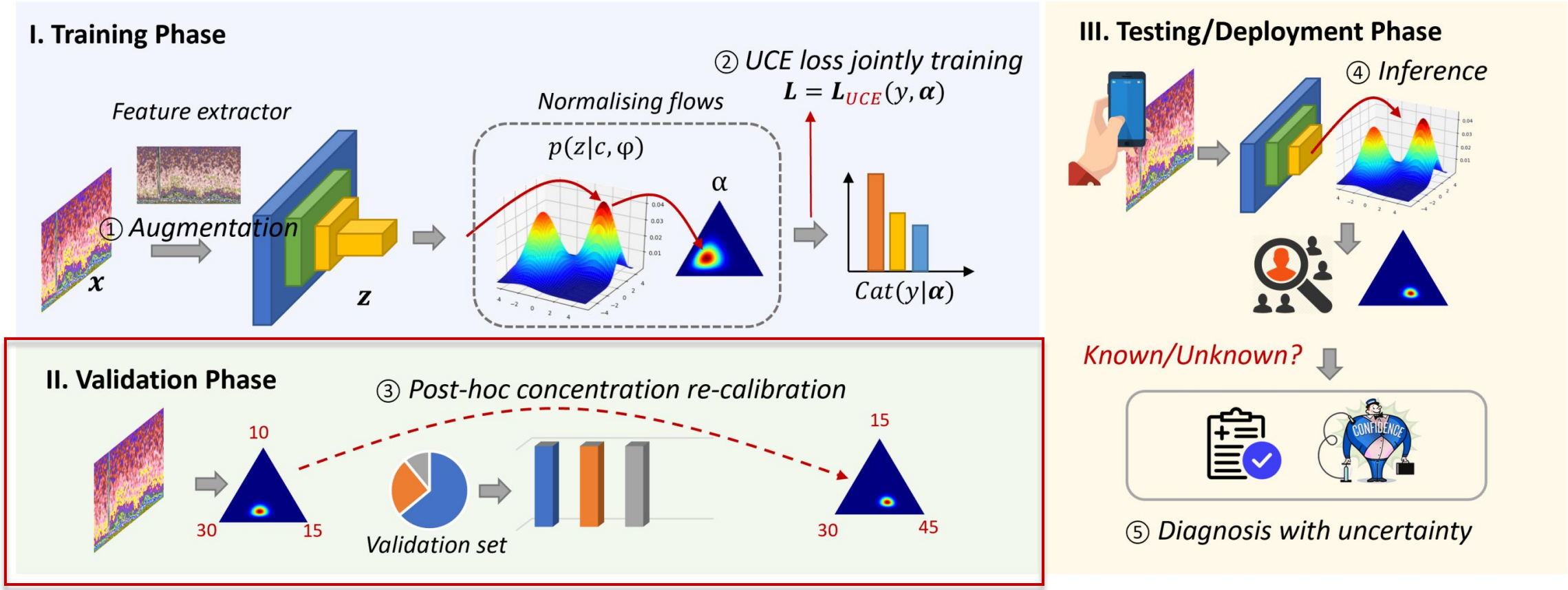



Figure 3: Our framework trains the model with an uncertainty-loss and re-calibrates the model according to the validation data distribution to eliminate the data imbalance bias. At deployment stage, the qualified uncertainty allows the model to recognise unconfident diagnosis and out-of-distribution input.

# Experiments on Real-world Health Data

**Table 1: A summary of tasks and datasets for experiments. #Train is the original data size before augmentation, and it is further split into training and validation folds for five-fold cross-validation. C is the number of classes and D is input data dimension. All experiments are implemented by Pytorch, and * denotes that pre-trained checkpoints were used for model initialisation.**

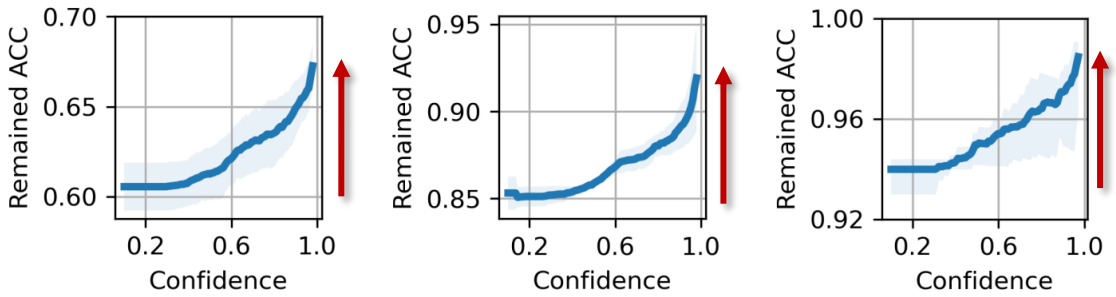
| Task | | | IID Dataset | | | | | OOD Dataset | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Name | Backbone | Modality | Name | #Train | #Test | C | D | Name | Size | Name | Size |
| #1 Respiratory | ResNet-34* | Audio | ICBHI2017 | 4,274 | 2,641 | 4 | 1×32,000 | Stethoscope | 336 | ARCA23K | 2,264 |
| #2 Skin | DenseNet-121* | Image | HAM10000 | 7,206 | 2,809 | 7 | 3×600×450 | ISIC2017 | 1,824 | CIFAR-10 | 10,000 |
| #3 Heart | FCNet | ECG | EGC5000 | 4,500 | 500 | 5 | 1×140 | ECG200 | 200 | FetalECG | 1,965 |

## ❑ RQ 1: Diagnosis accuracy (particular for the minority classes)

| | Task 1 | | | | | | Task 2 | | | | | | Task 3 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC↑ | UAR↑ | SE↑ | NLL↓ | Brier↓ | ECE↓ | ACC↑ | UAR↑ | SE↑ | NLL↓ | Brier↓ | ECE↓ | ACC↑ | UAR↑ | SE↑ | NLL↓ | Brier↓ | ECE↓ |
| **DirichNet** | 0.610 (.014) | 0.442 (.006) | 0.531 (.018) | 1.508 (.185) | 0.642 (.029) | 0.284 (.031) | 0.855 (.011) | 0.750 (.029) | 0.860 (.040) | 0.593 (.010) | 0.220 (.015) | 0.103 (.006) | 0.940 (.007) | 0.743 (.025) | 1.000 (.000) | 0.252 (.043) | 0.115 (.016) | 0.059 (.017) |
| Δ | 0.2% | 2.6% | 8.4% | 35.1% | 3.0% | 11.5% | -2.1% | 1.5% | 3.0% | 6.9% | -2.3% | -1.0% | 0.2% | 2.1% | 0.3% | 10.0% | 5.7% | 4.8% |

# Experiments on Real-world Health Data

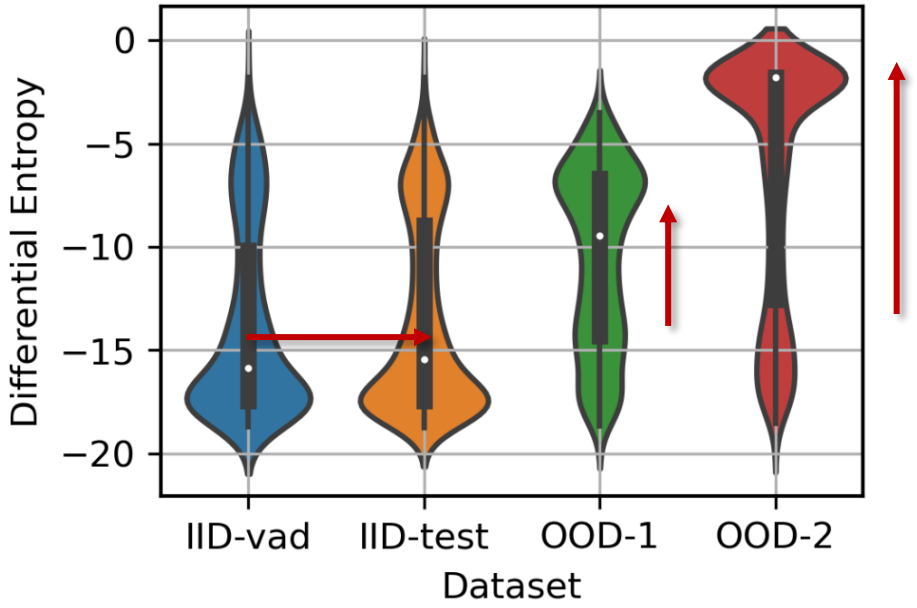
## ❑ RQ 2: Diagnostic confidence




(a) Task1. (b) Task2. (c) Task3.

Performance of selective prediction: After rejecting the predictions with the estimated confidence below a threshold, the diagnosis accuracy can be significantly boosted.

## ❑ RQ 3: Detecting Out-of-distribution Inputs




Uncertainty for IN and OOD sets

# Conclusion

- In this work, we explore evidential deep learning for uncertainty-aware mHealth diagnostic applications from multiple clinical data modalities.  EDL presents its advantage of efficiency and effectiveness.
- With our novel re-calibration approach, the proposed framework shows superior performance in calibrating the predictive confidence and detecting dataset shifts in the presence of imbalanced multi-class heath data.
- Our findings have the potential of playing an important role in facilitating the deployment of mobile health diagnosis models in real-world settings, with a more transparent misdiagnosis risk management mechanism.

# Thank You!

***Tong Xia***

*Welcome to chat with me: tx229@cam.ac.uk*

# An EDL Framework for mHealth with class imbalance

following the Bayesian update,

$$\alpha_c = \alpha_c^{prior} + N_c \cdot \mathbb{P}(z^{(i)}|c;\phi), \quad (2)$$

where $\alpha_c^{prior} = 1$, and $\mathbb{P}(z^{(i)}|c;\phi)$ presents the likelihood of observing $z^{(i)}$. More specifically, $z^{(i)}$ is the latent feature for sample $i$ yielded by the feature extractor $f(\theta)$, and $\phi$ denotes the parameters of normalising flows. The assumption here is: *if the feature extractor $f(\theta)$ brings no additional evidence from sample i for class c, the posterior $\alpha_c$ will approach 1.*

Due to the data imbalance, data augmentation also needs to be implemented to increase the size of minority classes, as shown in step ① in Fig. 1. Finally, on the balanced data, $\theta$ and $\phi$ can be jointly optimised (Fig. 1 ②) via the uncertainty-aware cross-entropy (UCE) loss formulated as below [13],

$$\min_{\theta,\phi} \mathcal{L} = \min_{\theta,\phi} \frac{1}{N} \sum_{i}^{N} \mathbb{E}_{q(p^{(i)})}[CE(p^{(i)}, y^{(i)})] - \lambda \cdot H(q^{(i)}), \quad (3)$$

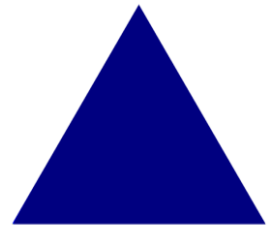
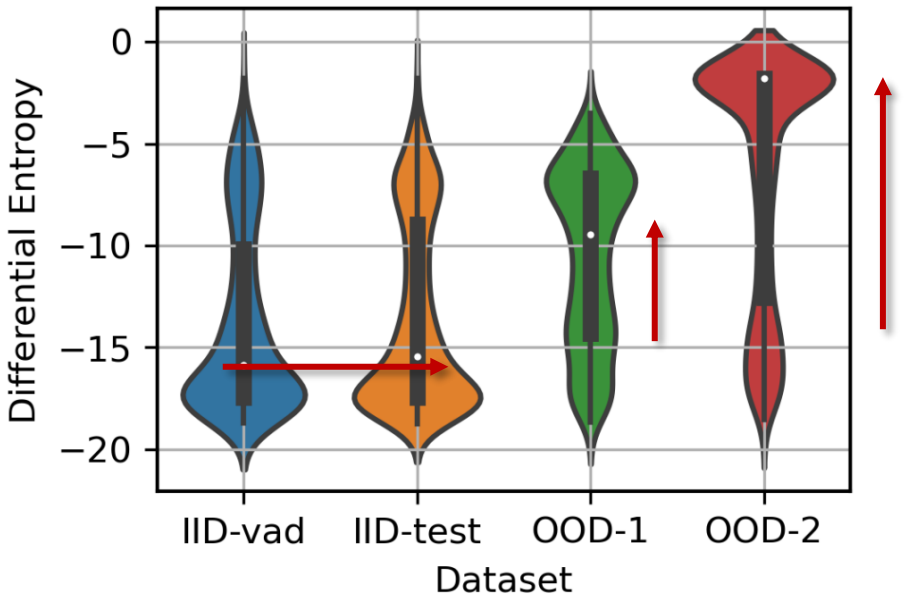


Fig. 1 step ③. Corresponding to $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_c, ...\alpha_K]$, we aim to seek a group of weights $\boldsymbol{\omega} = [\omega_1, ..., \omega_c, ...\omega_K]$ that can eliminate the classification bias, and improve the accuracy for the minority classes, particularly. Let $N_c$ denote the number of samples of class $c$ in the validation set. For ease of notation, we term $\alpha_1$ as the concentration for the majority class $c = 1$ (the class with the largest number of training samples before data augmentation), and we let $\omega_1 = 1$. For $\omega_c$ ($c > 1, N_1 > N_c$), a fine grid search will be conducted to find a value in $[1, \frac{N_1}{N_c})$ to maximise the *unweighted average recall* (UAR) [54] on the validation set by taking the prediction from $\boldsymbol{\alpha} \cdot \boldsymbol{\omega} = [\alpha_1\omega_1, ..., \alpha_c\omega_c, ...]$. We do not use the overall accuracy as

# Experiments on Real-world Health Data

## ❑ RQ 3: Detecting Out-of-distribution Inputs

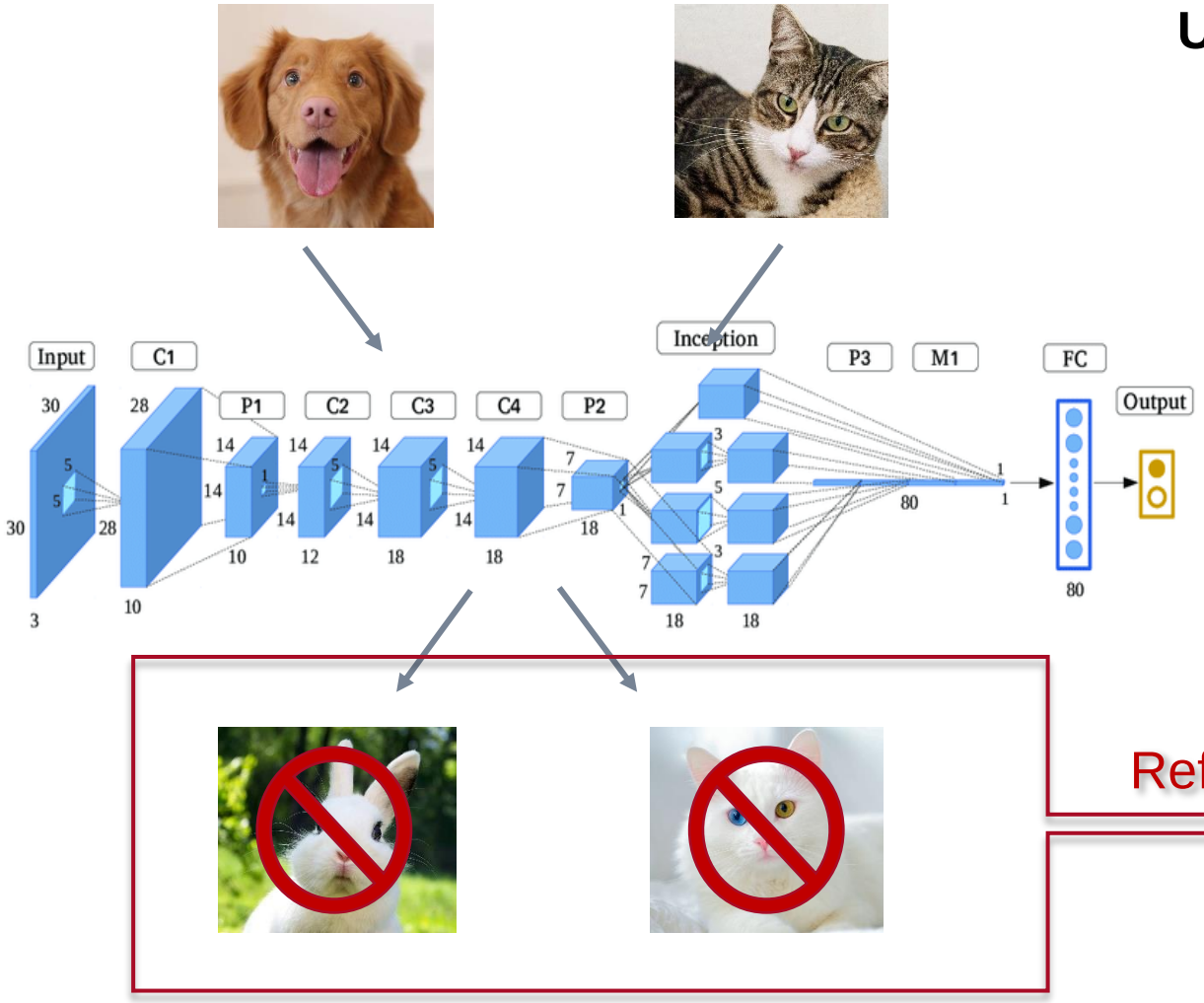
| Task | OOD datasets | *DirichNet (ours)* | | |
|---|---|---|---|---|
| | | $\alpha_0$ | DE | $\Delta$ |
| #1 | OOD 1 (Near) | 0.731(.087) | 0.733(.085) | 13.8% |
| | OOD 2 (Far) | 0.830(.156) | 0.832(.151) | 23.8% |
| #2 | OOD 1 (Near) | 0.740(.096) | 0.744(.095) | 1.1% |
| | OOD 2 (Far) | 0.990(.008) | 0.989(.007) | 0.6% |
| #3 | OOD 1 (Near) | 0.788(.071) | 0.816(.071) | 8.1% |
| | OOD 2 (Far) | 0.951(.029) | 0.958(.021) | 7.2% |




(a) Task 1.

Figure 5: Performance of OOD detection.

# Why uncertainty is important for mHealth




**Uncertainty Quantification for Health**:

- Calibration: the confidence of prediction associated with a trained condition
- OOD aware: the likelihood of an input belonging to the out-of-distribution regime of training data
- Low cost: no extra inference cost, no extra training data required