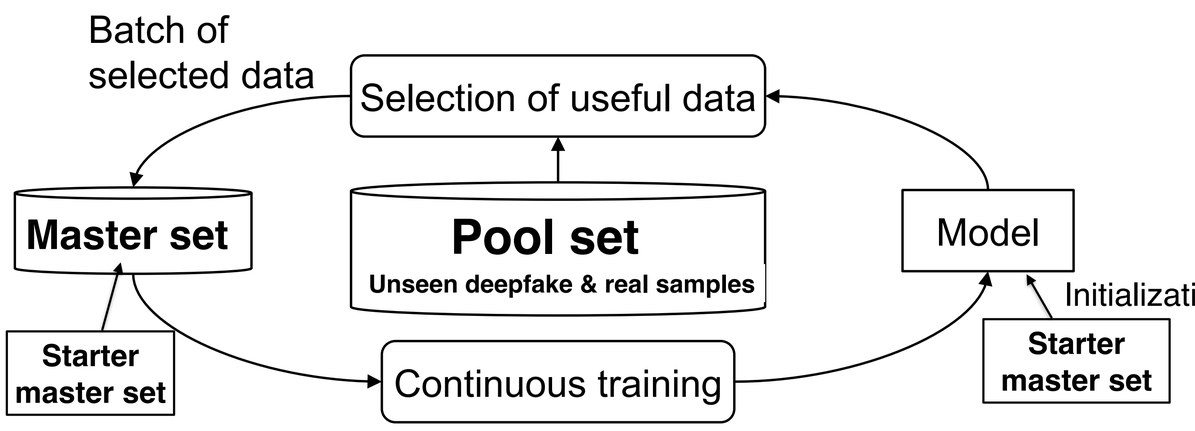
Exploring Active Data Selection Strategies for Continuous Training in Deepfake Detection

Yoshihiko Furuhashi, Junichi Yamagishi*, Xin Wang, Huy H. Nguyen, and Isao Echizen
National Institute of Informatics, Tokyo, Japan

Abstract and Take-home Messages

- **Main research question**: *How do we adjust the deepfake detection models to new unseen deepfake methods while preventing catastrophic forgetting on old deepfakes?*
- **Our proposed idea in this poster**: ***Automatically and actively select the small amount of additional useful data suitable for the continuous training of deepfake detection models***
- **Take-home messages**: *The automatic selection of small additional useful data from a redundant large pool set using the confidence score of the detection model is simple but useful and* ***continuously trained deepfake detection models have better performance than one without active data selection and than one using random selection***

Active data selection and continuous training strategy of deepfake detection models



How to choose useful data: **Confidence score** c_m **where** $l_{m,j}$ is a logit value of the softmax operation

$$c_m = -T \log \sum_{j=1}^{J} \exp(\frac{l_{m,j}}{T}),$$

Starter master set (i.e. an initial dataset to train the first detection model from scratch)

Data set containing multiple spoofing methods

Pool set

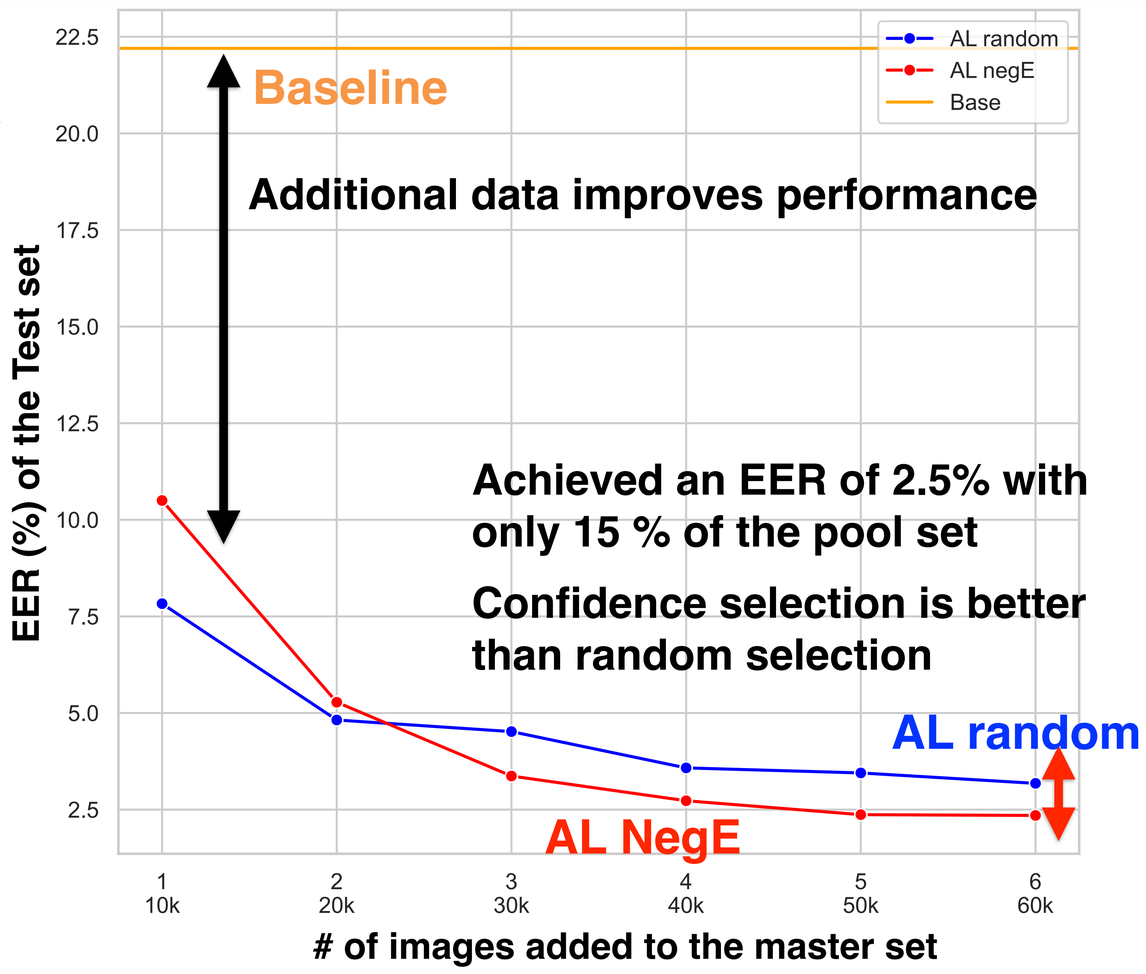
Redundant dataset containing real and fake data

Dataset design for continuous training

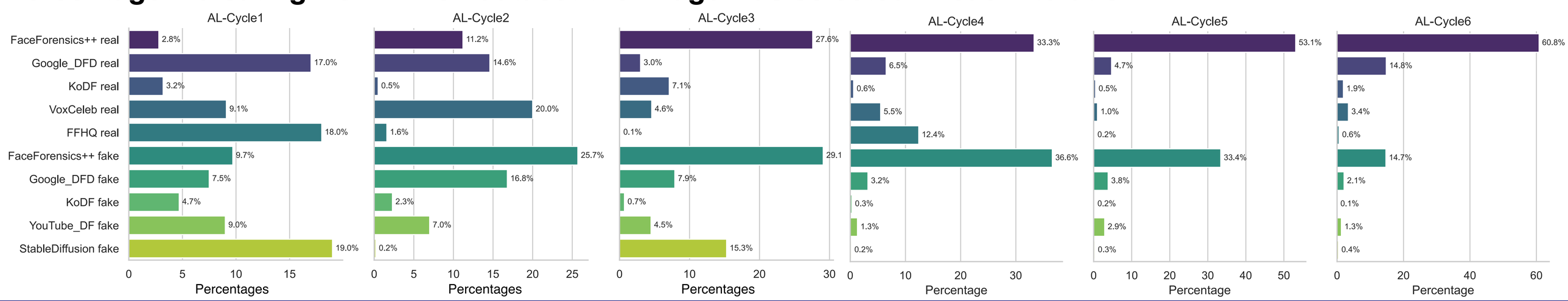
Database	Type	Initial	AL Pool	Val.	Test
Starter master set					
ForgeryNet [He21]	Real	163,200		1,000	1,000
ForgeryNet [He21]	Fake	163,200		1,000	1,000
Pool set					
FF++ [Ro19]	Real		40,000	1,000	1,000
FF++ (5 types) [Ro19]	Fake		40,000	1,000	1,000
Google DFD [DG19]	Real		40,000	1,000	1,000
Google DFD [DG19]	Fake		40,000	1,000	1,000
VoxCeleb [CNZ18]	Real		40,000	1,000	1,000
YouTube DF [Ku20]	Fake		40,000	1,000	1,000
KoDF [Kw21]	Real		40,000	1,000	1,000
KoDF [Kw21]	Fake		40,000	1,000	1,000
FFHQ [KLA19]	Real		40,000	1,000	1,000
Stable Diffusion 2.1 [Ro22]	Fake		40,000	1,000	1,000

Experiments: Continuous training of facial deepfake detection models

- **Facial deepfake detectors used in the experiments**
  - EfficientNet V2-M architecture pre-trained by ImageNet21k
  - A head layer for binary prediction
  - Data augmentation similar to DeepfakeBench [Ya 23]
- **Baseline (BASE)**
  - Baseline system using the starter master set only
- **Confidence-score based selection (AL negE)**
  - Add 10,000 selected images to the master set in each iteration
  - Fine-tune a model at the previous iteration for 3 epochs
- **Random selection (AL random)**
  - Add 10,000 images randomly selected from the (balanced) pool set in each iteration
  - Fine-tune a model at the previous iteration for 3 epochs



Percentage indicating from which dataset the image was selected in each iteration



Contact: jyamagis@nii.ac.jp