2024 Yonsei Digital Healthcare Cybersecurity Competition

Robust Medical Image Classification Against Data Contamination and Poisoning

Department of Digital Analytics, Yonsei University

Min Kyoon Yoo, Jihae So, Shiwon Kim, Ho Seung Kang, Donghyeok Seo

Introduction

Objective

- Develop a model that remains **robust against data contamination** in medical imaging.
- Ensure stable classification of normal / abnormal and subtyping even with corrupted data.
- Data contamination is a key factor that lowers model reliability and hinders accurate clinical decisions.

Types of Data Contamination

- **Noise Injection** – Unintended noise degrades image quality and obscures diagnostic features.
- **Label Error** – Incorrect labels mislead the model, causing inaccurate predictions.
- **Poisoning Attack** – Malicious data intentionally inserted into training sets to reduce performance.

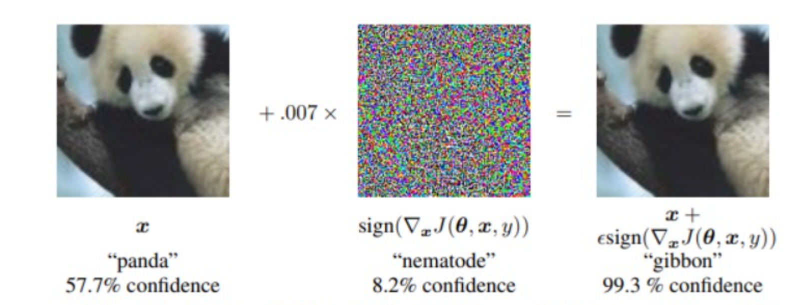


Fig 1. An example of noise injection.

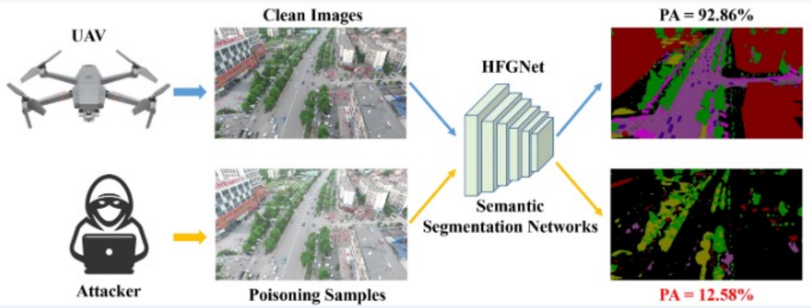


Fig 2. An example of poisoning attack.

Introduction

Poisoning Attacks in the Medical Domain

- **Finlayson, et al. (2019)** [1] experimentally demonstrated that label flipping attacks cause cancer diagnosis models to misclassify normal as malignant or vice versa.
- As medical systems become increasingly digitalized (e.g., telemedicine), such attacks can be executed **more easily and at larger scales.**

Related Work – Defense Approaches

- **Steinhardt, et al. (2017)** [2] proposed a model that maintains stability even with partially poisoned datasets.
- **Alzubaidi, et al. (2024)** [3] introduced model ensemble feature fusion (MEFF) for robust medical imaging. Trained multiple models through *adversarial training*, which intentionally generates various types of attacks and train models to become robust against them.

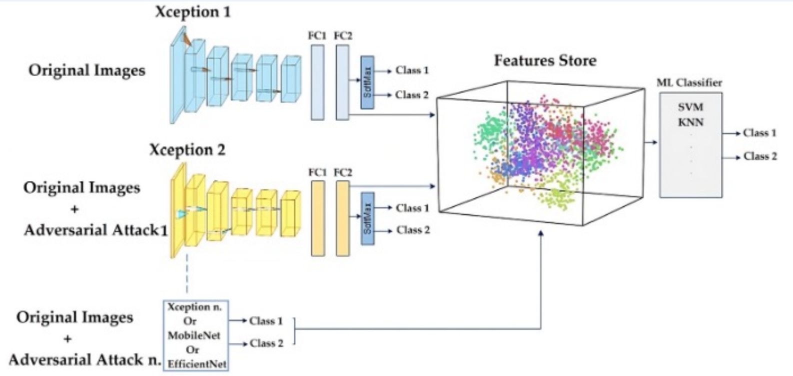


Fig 3. Workflow of the MEFF framework.

Methods

Overview

- The figure illustrates the **deep mutual learning (DML)** process using two models (feature extractors) and a shared simplex equiangular tight frame (ETF) classifier.

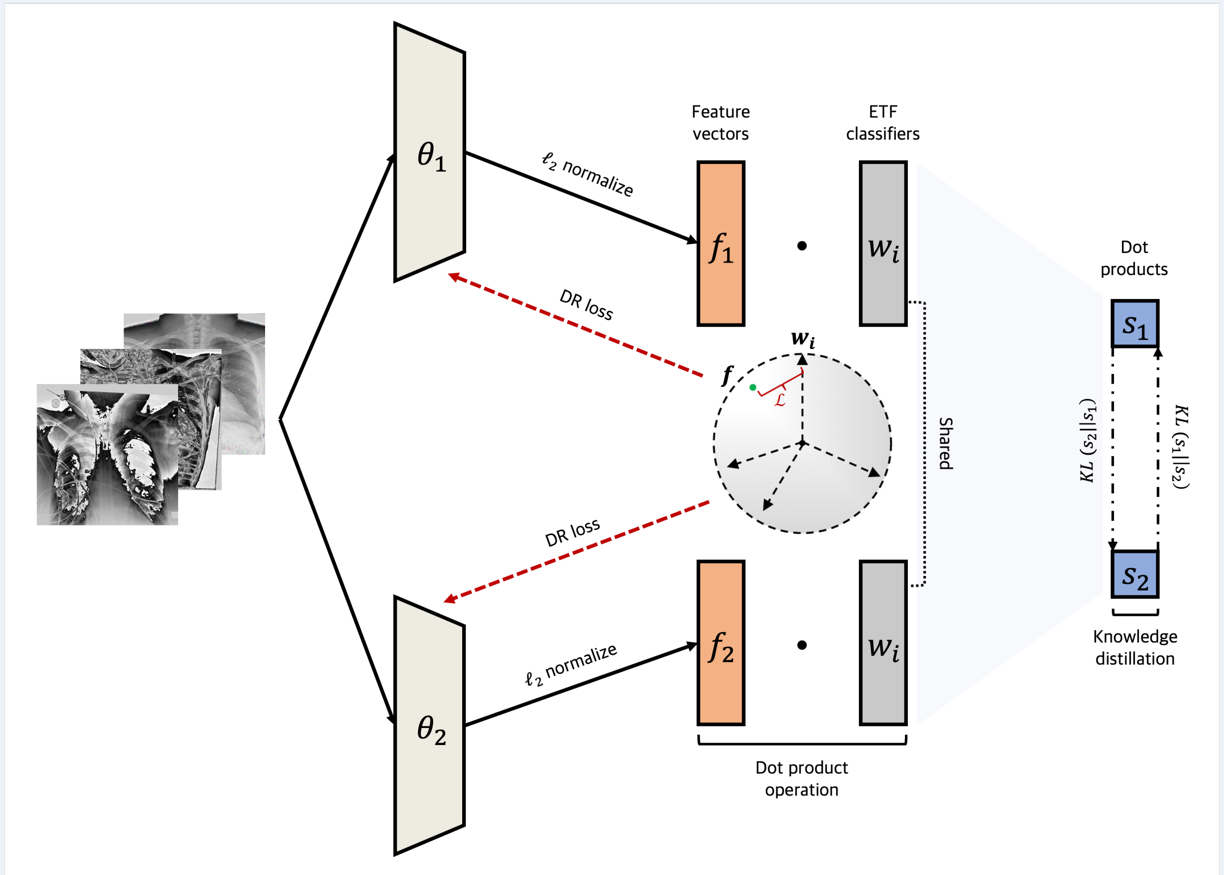


Fig 4. Overall framework of the proposed method.

Each model fits its feature vectors to **a shared ETF classifier** and mutually performs **knowledge distillation** to collaboratively learn robust feature representations using the fixed ETF as an anchor.

Methods

Simplex Equiangular Tight Frame (ETF) Classifier

- A simplex ETF is a mathematical structure that **maximizes the pair-wise angles** between all vectors [4].
- The ETF classifier is introduced to enforce well-separated and uniformly distributed class representations in the feature space [5].
- It minimizes confusion between different classes by ensuring **equal angular distances** among class vectors.

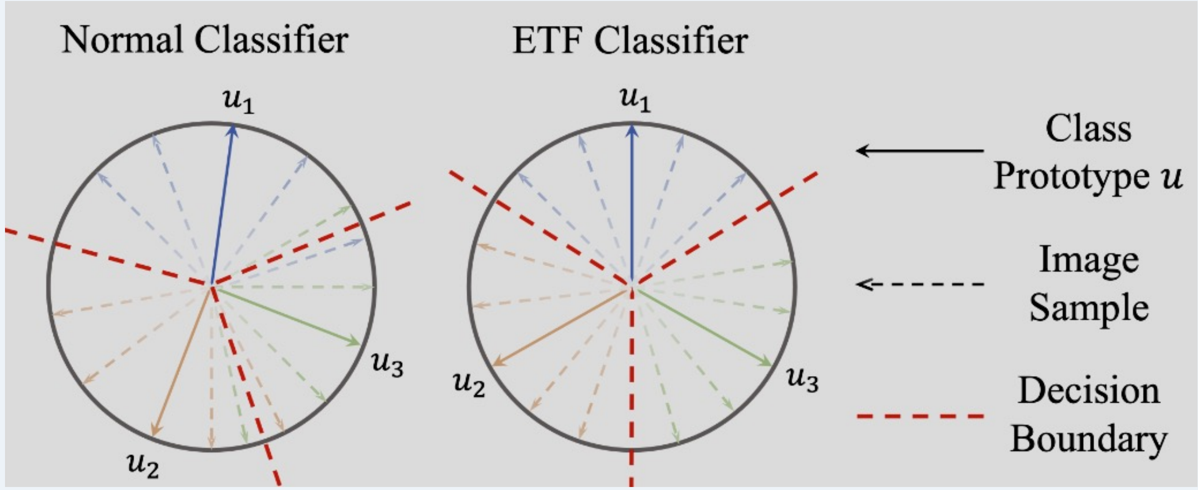


Fig 5. Illustration of a normal classifier and an ETF classifier.

Methods

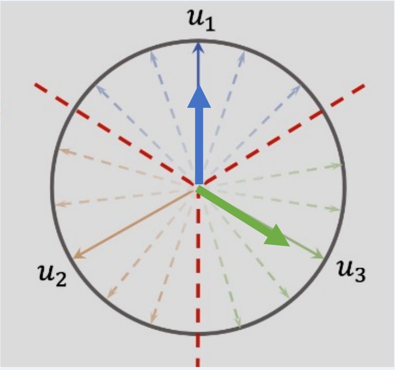
Rectification (Rect)

- Rect [6] operates on ETF vectors to address **class imbalance** and achieve finer inter-class boundaries.
- Classes with fewer samples are assigned longer feature vectors.

$$\gamma = \frac{B}{U} \qquad \mathrm{rect}(i) = \sqrt{\frac{\gamma}{c_i}}$$

B: batch size
U: number of classes in a batch
$\boldsymbol{c_i}$: number of samples for class i

- The formula assigns weights inversely proportional to class frequency (c_i)
- It assigns higher weights to the minority classes thereby promoting balanced learning across all categories.



Angle: class separation
Length: class imbalance

Methods

Deep Mutual Learning (DML)

- DML [7] is a training paradigm where multiple models learn simultaneously while sharing knowledge with each other. Instead of learning independently, each model **incorporates the peer's predictions** into its own learning process, improving generalization for all models.
- To maximize ensemble effects, DML allows models to overfit to the training data.
- We perform **knowledge distillation** between the inner-products of the model output vectors and the ground-truth ETF vectors for each model.

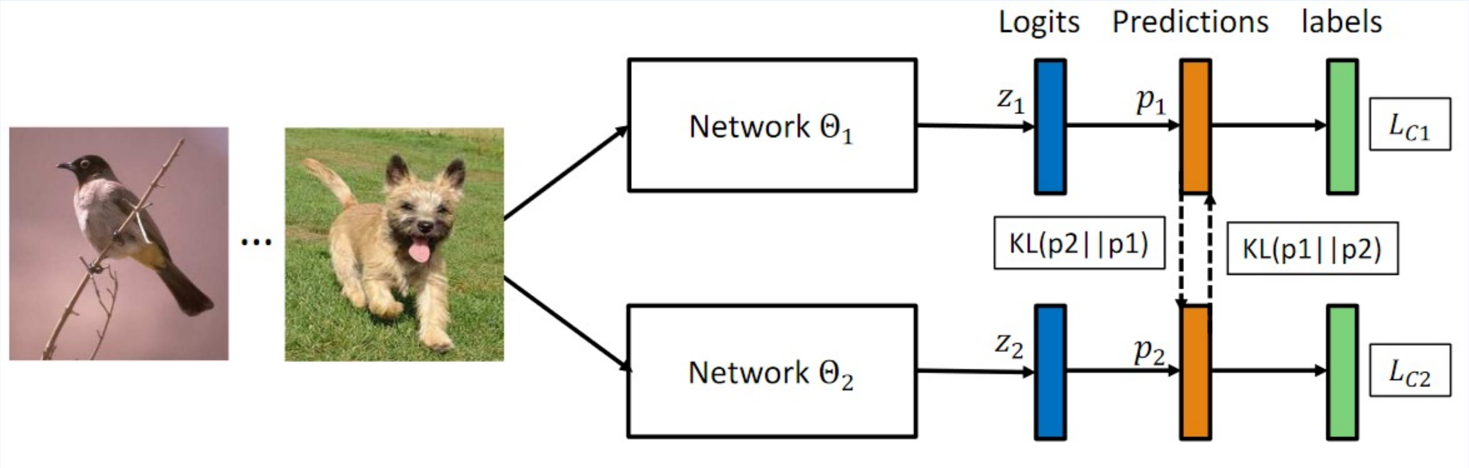


Fig 6. The deep mutual learning process.

Evaluation Metric

Combined Performance Drop

- Prior works measured robustness by comparing model performance **only under poisoned conditions** [8, 9], which does not accurately reflect the model's robustness considering that higher scores may simply indicate stronger baseline model performance.
- Therefore, we evaluate each model on **both clean and poisoned datasets** and compute the drop ratio.
- For each metric M, the percentage drop between the model's performance on **clean data** M_{normal} and on **attacked or noisy data** M_{attack} is calculated as follows:

$$\text{Drop}_M = \frac{M_{\text{normal}} - M_{\text{attack}}}{M_{\text{normal}}} \times 100$$

- After calculating the percentage drop for each performance metric $M_1, M_2, \ldots, M_n$, we take their average to obtain the combined performance drop:

$$\text{Combined Drop} = \frac{1}{n} \sum_{i=1}^{n} \text{Drop}_{M_i}$$

Experiments

Experimental Settings

- Data split = 7 : 2 : 1 (train : validation : test)
- Optimizer = Adam (LR: 0.0001, weight decay: 0.001)
- Patience (early stopping) = 20
- Scheduler = ReduceLROnPlateau

- For performance drop calculation:
  - **Experiment 1**: Train and test on the original (clean) dataset
  - **Experiment 2**: Train on the original dataset / add Gaussian noise to the test set (Mean: 0.0, Std: 0.01)

Experiments

Results – Experiment 1

Model	Accuracy	F1 Score	Precision
ResNet-50 (Baseline)	0.384	0.314	0.288
ResNet-50 + ETF	0.443	0.365	0.357
ResNet-50 + ETF +Rect	0.463	0.399	0.420
ResNet-50 + ETF +Rect +DML	**0.467**	**0.449**	**0.469**

Table 1. Performance comparison of different models on medical image classification task (clean dataset).

- Combining ETF, Rect, and DML outperforms all other approaches across all metrics.
- Accuracy improved by approximately 9% points, while F1-score and Precision increased by about 15% points.

Experiments

Results – Experiment 2

Model	Accuracy	F1 Score	Precision
ResNet-50 (Baseline)	0.240	0.171	0.260
ResNet-50 + ETF	0.314	0.280	0.390
ResNet-50 + ETF +Rect	**0.332**	0.293	**0.283**
ResNet-50 + ETF +Rect +DML	0.315	**0.376**	0.287

Table 2. Performance comparison of different models on medical image classification task (noisy dataset).

- When noise is added to the test set, the performance *with* and *without* DML is inconsistent across metrics.
- Therefore, the **combined performance drop** is needed to provide a more consistent evaluation of robustness.

Experiments

Results – Experiment 2

Model	Combined Drop	Accuracy Drop	F1 Score Drop	Precision Drop
ResNet-50 (Baseline)	12.57%	37.5%	14.3%	2.8%
ResNet-50 + ETF	9.72 %	29.12%	8.5%	**- 3.3 %**
ResNet-50 + ETF +Rect	**9.51%**	**28.29%**	10.6%	13.7%
ResNet-50 + ETF +Rect +DML	10.93%	32.55%	**7.3%**	18.2%

Table 2. Performance comparison of different models on medical image classification task (noisy dataset).

- The model *without* DML showed the lowest combined drop, indicating the smallest performance drop under attack.
- This demonstrates combining ETF and Rect (*without* DML) provides the highest robustness.

Conclusion

Discussion

- We proposed and evaluated a new framework for efficient and robust medical image classification.
- By combining the **ETF classifier** with **Rect**, our proposed approach showed stronger robustness compared to conventional classification models.
- Incorporating **DML** further enhanced the model's overall performance and resilience.
- We introduced a novel **robustness metric** to evaluate model stability against malicious data perturbations such as poisoning attacks.

Limitations

- First, using only the ETF classifier and Rect sometimes proved more effective than combining it with DML. Therefore, additional experiments depending on the intensity of the poisoning attack are required.
- Second, further validation under diverse real-world clinical attack scenarios is needed.
- Lastly, a standardization process of the proposed robustness metric is required to enable its broader adoption in future studies.

References

[1] Finlayson, Samuel G., et al. "Adversarial attacks against medical deep learning systems." arXiv preprint arXiv:1804.05296 (2018).

[2] Steinhardt, Jacob, Pang Wei W. Koh, and Percy S. Liang. "Certified defenses for data poisoning attacks." *Advances in neural information processing systems* 30 (2017).

[3] Alzubaidi, Laith, et al. "MEFF–A model ensemble feature fusion approach for tackling adversarial attacks in medical imaging." Intelligent Systems With Applications 22 (2024)

[4] Xie, Liang, et al. "Neural collapse inspired attraction–repulsion-balanced loss for imbalanced learning." Neurocomputing 527 (2023): 60-70.

[5] Yang, Yibo, et al. "Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network?." Advances in neural information processing systems 35 (2022)

[6] Yang, Yibo, et al. "Neural collapse inspired feature-classifier alignment for few-shot class incremental learning." arXiv preprint arXiv:2302.03004 (2023).

[7] Zhang, Ying, et al. "Deep mutual learning." Proceedings of the IEEE conference on computer vision and pattern recognition. (2018).

[8] Kim, W. J., Cho, Y., Jung, J., & Yoon, S.-E.. Feature Separation and Recalibration for Adversarial Robustness. 8183–8192 (2023)

[9] Perez, J. C., Alfarra, M., Jeanneret, G., Rueda, L., Thabet, A., Ghanem, B., & Arbelaez, P. (2021).