Assessing Risk of Stealing Proprietary Models for Medical Imaging Tasks

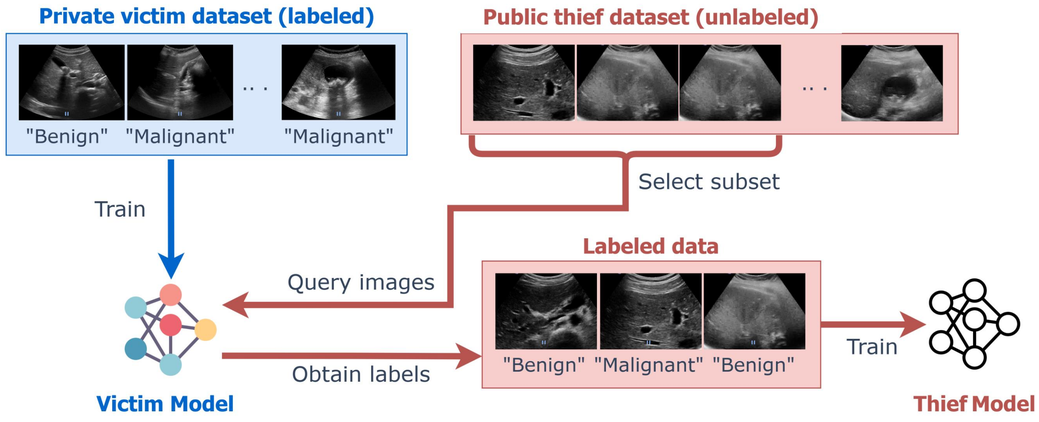
Ankita Raj[1], Harsh Swaika[1], Deepankar Varma[2], Chetan Arora[1]
[1] Indian Institute of Technology Delhi, New Delhi, India, [2] Thapar Institute of Engineering and Technology, Patiala, India

Model Stealing Attack

What is Model Stealing?

Companies like *Qure.ai, SkinVision* are adopting the **Machine Learning as a Service** (MLaaS) setup to monetize from their proprietary medical imaging models.

Thieves can replicate the functionality of black-box models by repeatedly querying the model and training a thief model on the acquired predictions.



Threat Model

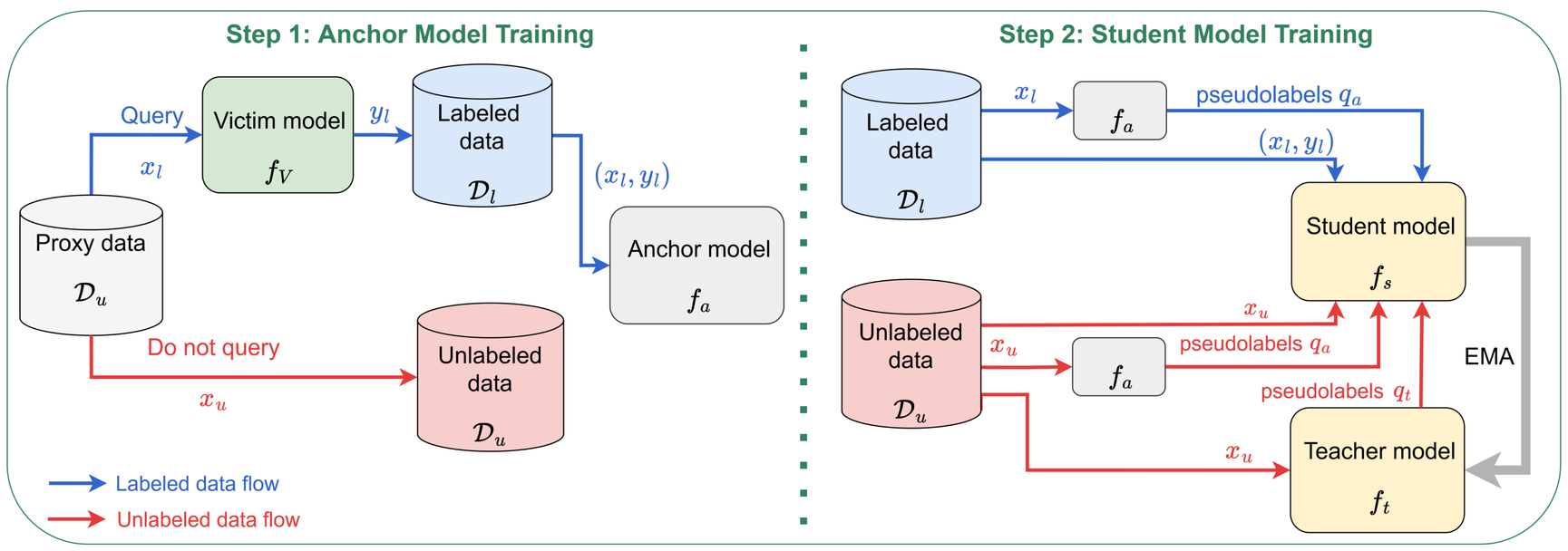
- Unknown victim architecture
- Private victim dataset
- Thief uses unlabeled publicly available proxy dataset

Challenges

- Extremely low query budget
- Victim outputs hard labels

Proposed Method: QueryWise

Key Idea: Use unlabeled proxy dataset in addition to the labeled (queried) data.



1. Train anchor model on labeled subset using existing model stealing methods (KnockOff Nets [1], ActiveThief [2]).

2. Train student model on both labeled and unlabeled data.

   Compute **labeled loss** using hard labels from victim model, soft pseudolabels from anchor model.

   Compute **unlabeled loss** using soft pseudolabels from anchor and teacher models.

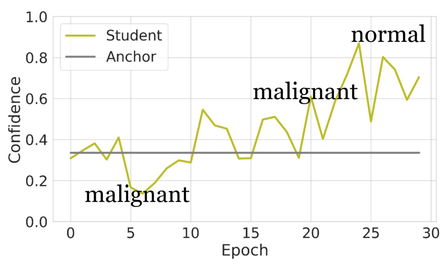
Experiments and Results

Steal a Gall Bladder Cancer classification Model

Victim
RadFormer [3]
Dataset: GBCU dataset [4] (1255 images)

Thief
ResNet-50 / Inception / ViT / DeiT
Proxy Dataset: GBC US videos [5] (64 videos)
Query Budget: 5000

GT: Normal
Anchor: Malignant
Student: Normal



	Accuracy	Specificity	Sensitivity	Agreement
Victim Model	90.16	90.00	92.86	-
Radiologist A	70.00	87.30	70.70	-
Radiologist B	68.30	81.10	73.20	-
ResNet-50 Thief				
Random (KnockOff Nets [1])	66.39	85.00	61.90	71.31
K-Center (ActiveThief [2])	71.31	**87.50**	71.43	68.85
Random + FixMatch [6]	65.57	82.00	62.00	66.39
Random + QueryWise	71.31	80.00	**81.00**	74.59
K-Center + QueryWise	**72.95**	79.00	**81.00**	**80.33**
DeiT Thief				
Random	71.31	**81.25**	78.57	74.59
Random + QueryWise	**77.05**	76.00	**90.00**	**77.05**

Conclusion

Our thief models surpass radiologists' accuracy!!

Medical Imaging Models are indeed vulnerable to model stealing, necessitating robust defenses.

References

1. Orekondy, et al. *CVPR* 2019.
2. Pal et al. *AAAI* 2020.
3. Basu et al. *Medical Image Analysis* 83 (2023).
4. Basu et al. *CVPR* 2022.
5. Basu et al. *MICCAI*, 2022.
6. Sohn et al. *NeurIPS* 2020.

Code

https://github.com/rajankita/QueryWise

Acknowledgement

We acknowledge and thank the funding support from AIIMS Delhi-IIT Delhi Center of Excellence in AI funded by Ministry of Education, government of India, Central Project Management Unit, IIT Jammu with sanction number IITJMU/CPMU-AI/2024/0002.