Membership Inference Attacks Against Thermal Image Classification Models for Stress Detection

Amin Fallahi

Introduction

- My research project
- Stress detection using thermal images
- Health monitoring
- Data sensitivity
- Membership inference attacks
    - Detects whether a patient data has been used for training the classifier
    - Health applications

Membership Inference Attacks

- Training dataset unknown => Access to similar dataset
  - Shadow networks
  - Assumes that classification probability vector from the input that has been used in training data is distinguishable

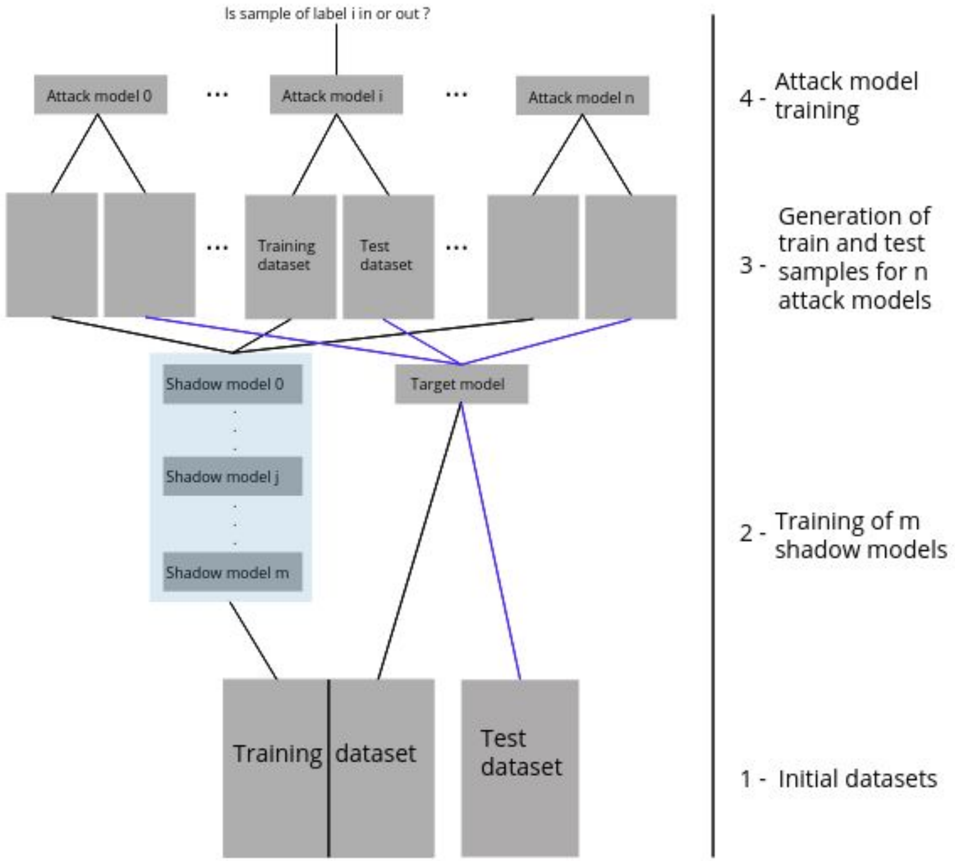
$$f_{target} : D^{train} \longrightarrow \mathbb{R}$$

$$f_{attack} : D^{train} \cup D^{test} \times \mathbb{R} \longrightarrow 0, 1$$

Membership Inference Attacks

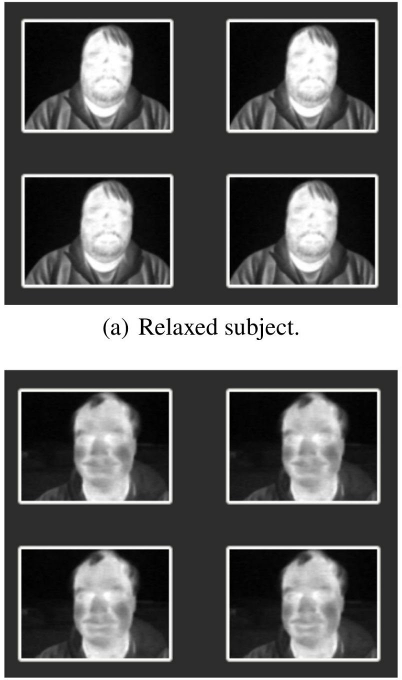
- Target Network
    - Purpose: some classification task
    - Input: samples from multiclass dataset
    - Output: classification probability vector
- Shadow Network
    - Purpose: produce 2 sets of probability vectors (training and non-training)
    - Input: samples from multiclass dataset
    - Output: classification probability vector
- Attack Network
    - Purpose: classify training vs non-training data
    - Input: probability vector
    - Output: probability whether the input is in the training dataset

Membership Inference Attacks



Dataset

- Data collection
- Terravic Facial IR Database
- Labeled ThermalAI dataset



(a) Relaxed subject.

(b) Stressed subject.

Dataset

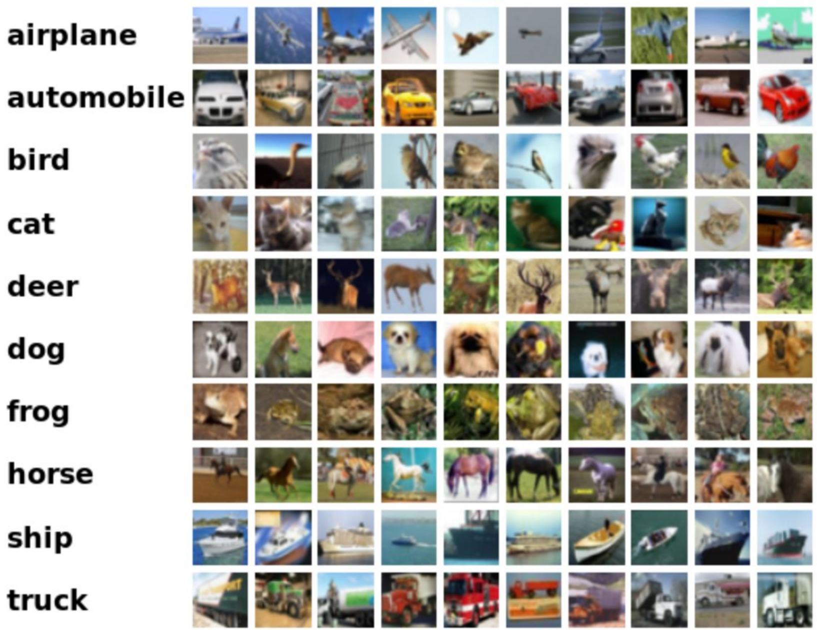
- Baseline and Stress tasks
- 23 presentation and 19 relaxation
- Thermal images captured at 10fps
- Depth images
- Data from distance
- 320x240px

Implementation

- MIA library by Kulynych and Yaghini
- Implements the original shadow model attack
- Is customizable, can use any scikit learn's Estimator-like object as a shadow or attack model
- Is tested with Keras and PyTorch
- Fine tuning and modifications

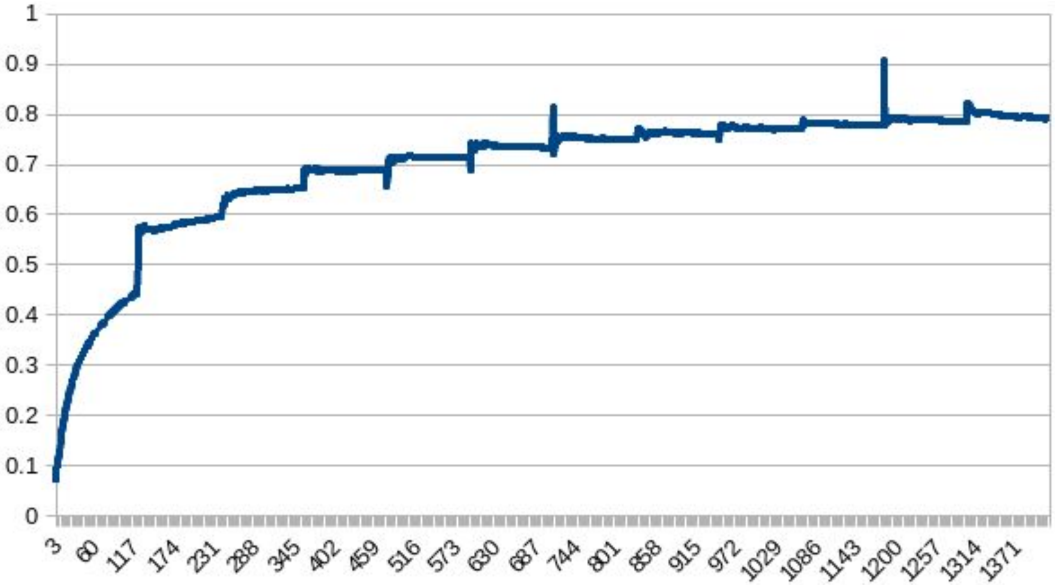
Initial results on CIFAR-10

- 60000 32x32 colour images
- 10 classes
- 6000 images per class
- 50000 training images
- 10000 test images
- 55-65% accuracy



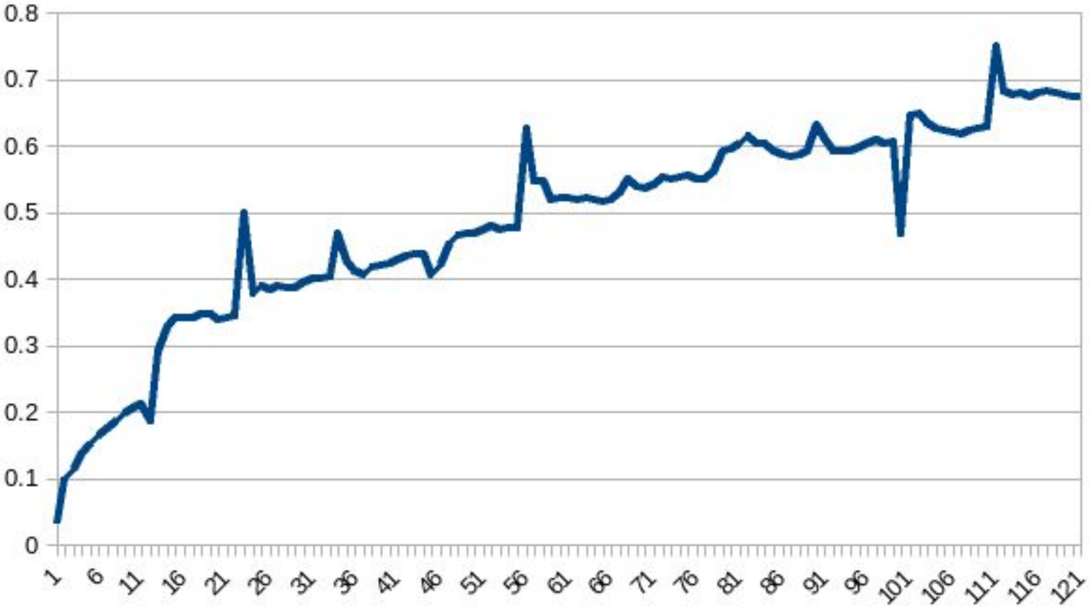
Initial results on CIFAR-10

- Target model training 45000 samples 12 epochs



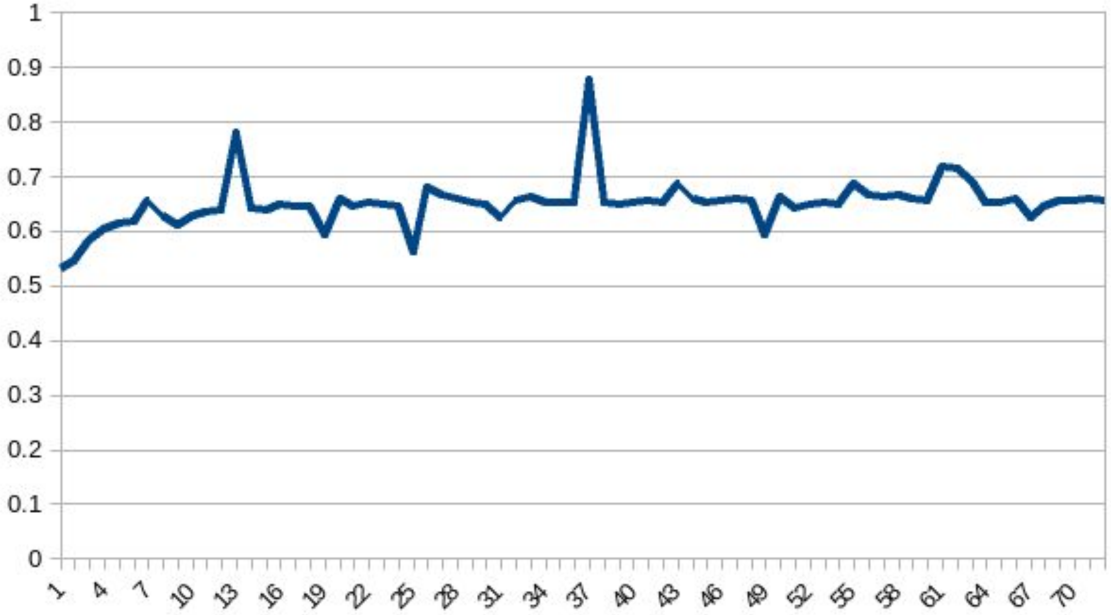
Initial results on CIFAR-10

- Shadow model training 4000 samples 12 epochs



Initial results on CIFAR-10

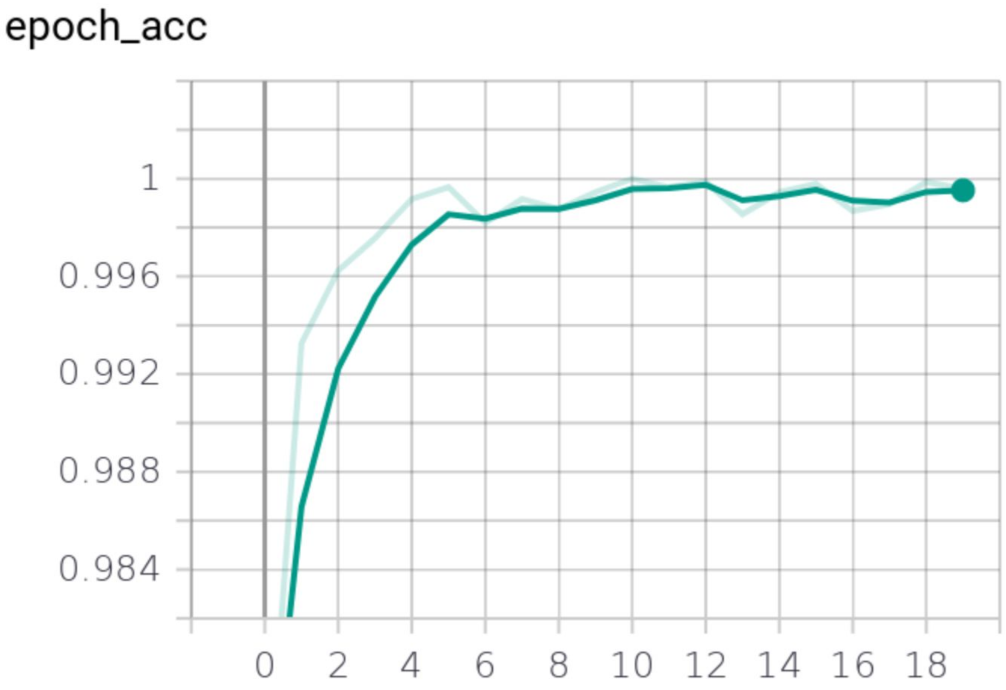
- Attack model training 2300 samples 12 epochs



Results

- 2 classes
- 288x216px
- 20 target and shadow epochs
- 20 attack epochs
- 3 shadow models
- 10000 baseline
- 10000 stress
- Shadow train on 800 samples val on 400 samples

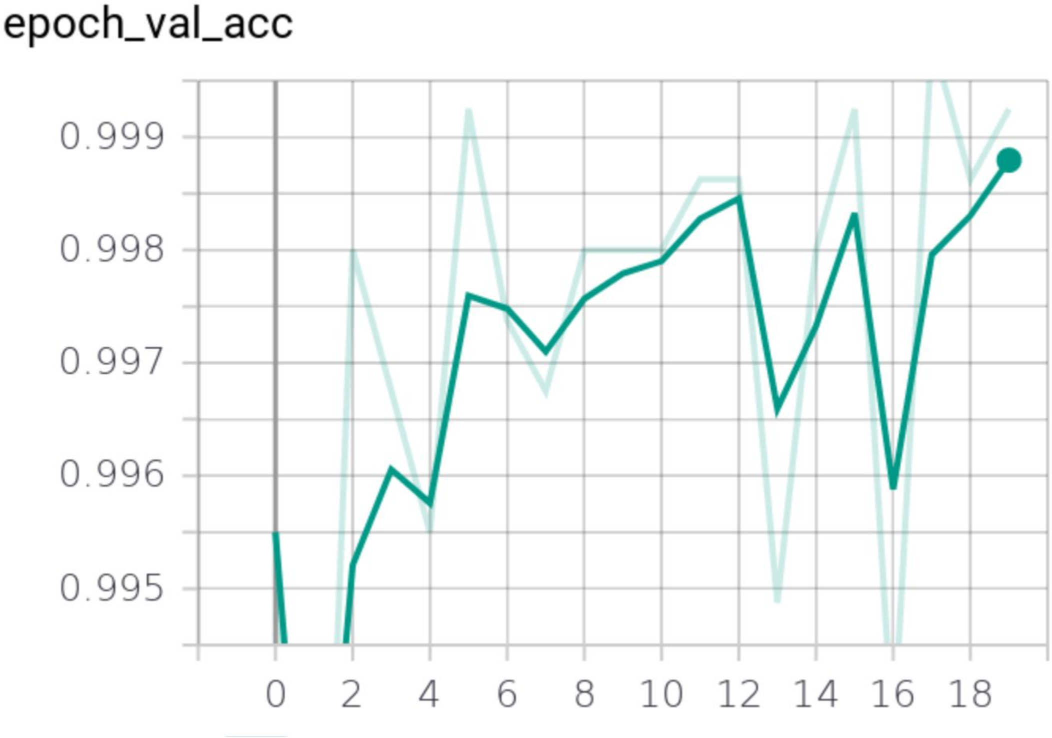
Target Training Accuracy



Results

- 2 classes
- 288x216px
- 20 target epochs
- 20 attack epochs
- 3 shadow models
- 10000 baseline
- 10000 stress
- Shadow train on 800 samples val on 400 samples

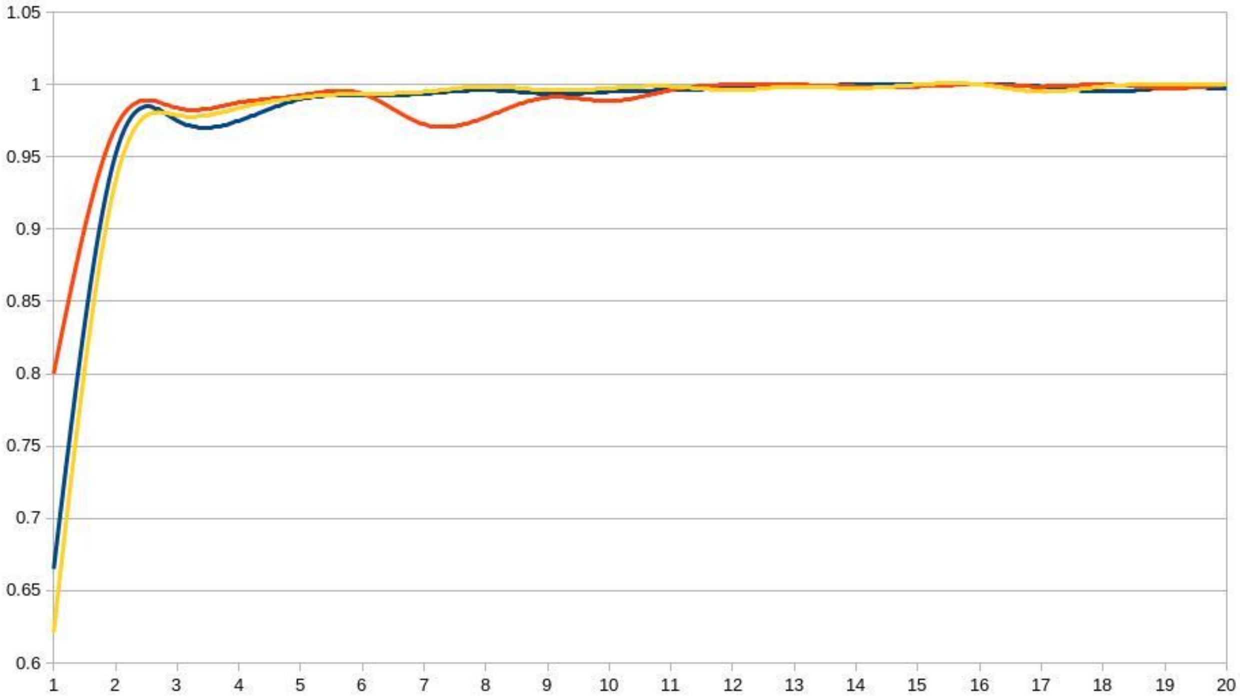
Target Training Val Accuracy



Results

- 2 classes
- 288x216px
- 20 target epochs
- 20 attack epochs
- 3 shadow models
- 10000 baseline
- 10000 stress
- Shadow train on 800 samples val on 400 samples

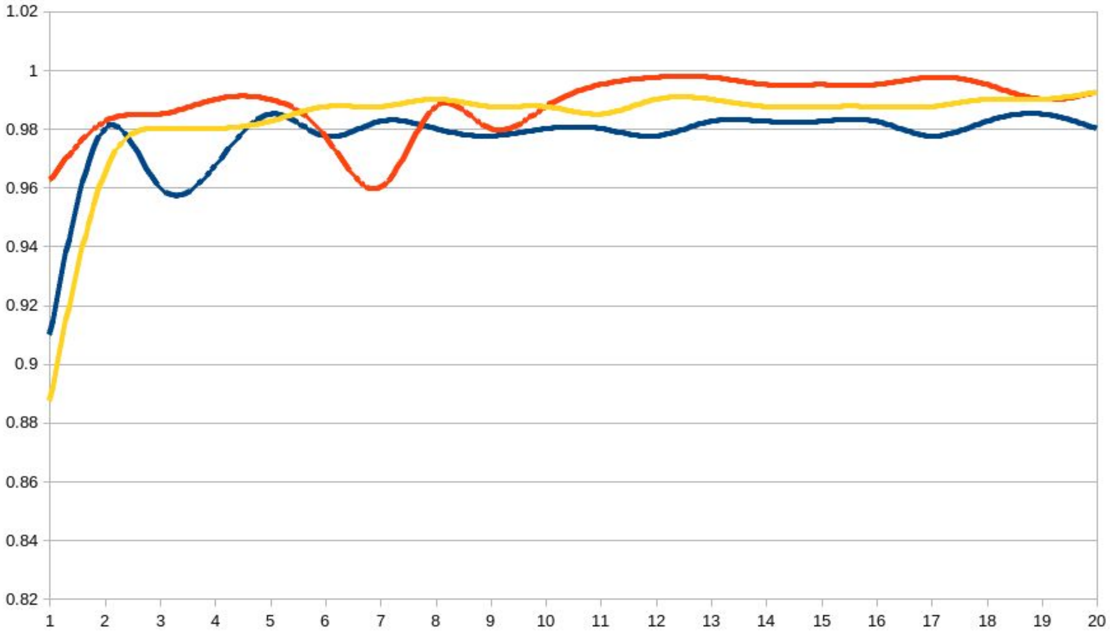
Shadow Training Accuracy



Results

- 2 classes
- 288x216px
- 20 target epochs
- 20 attack epochs
- 3 shadow models
- 10000 baseline
- 10000 stress
- Shadow train on 800 samples val on 400 samples

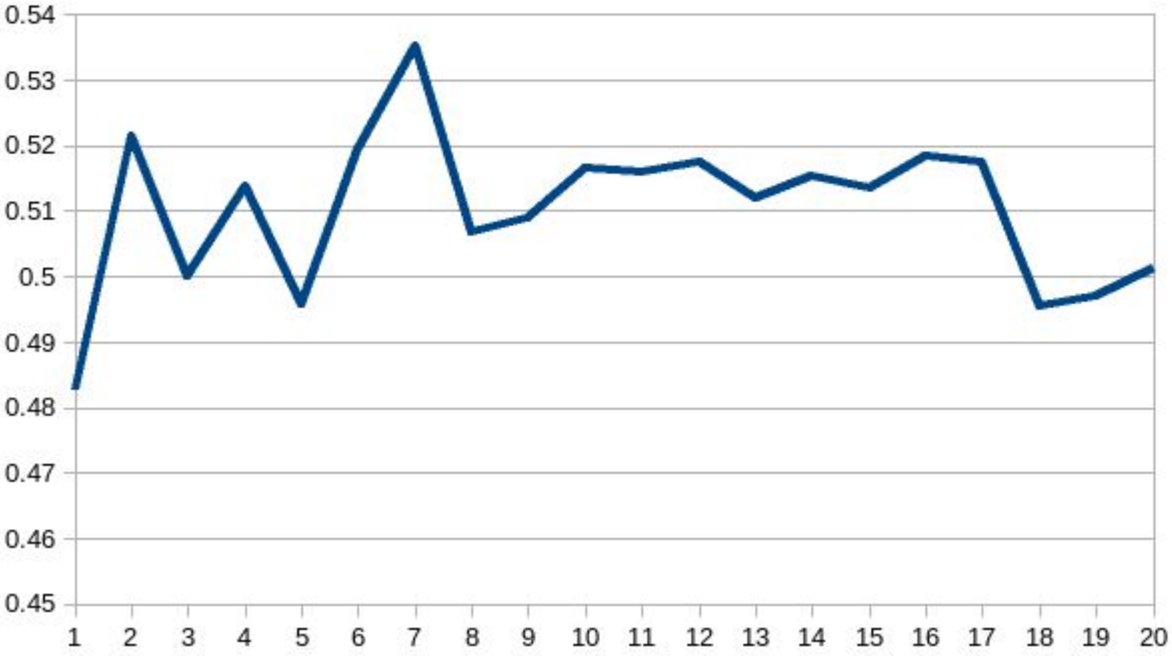
Shadow Val Accuracy



Results

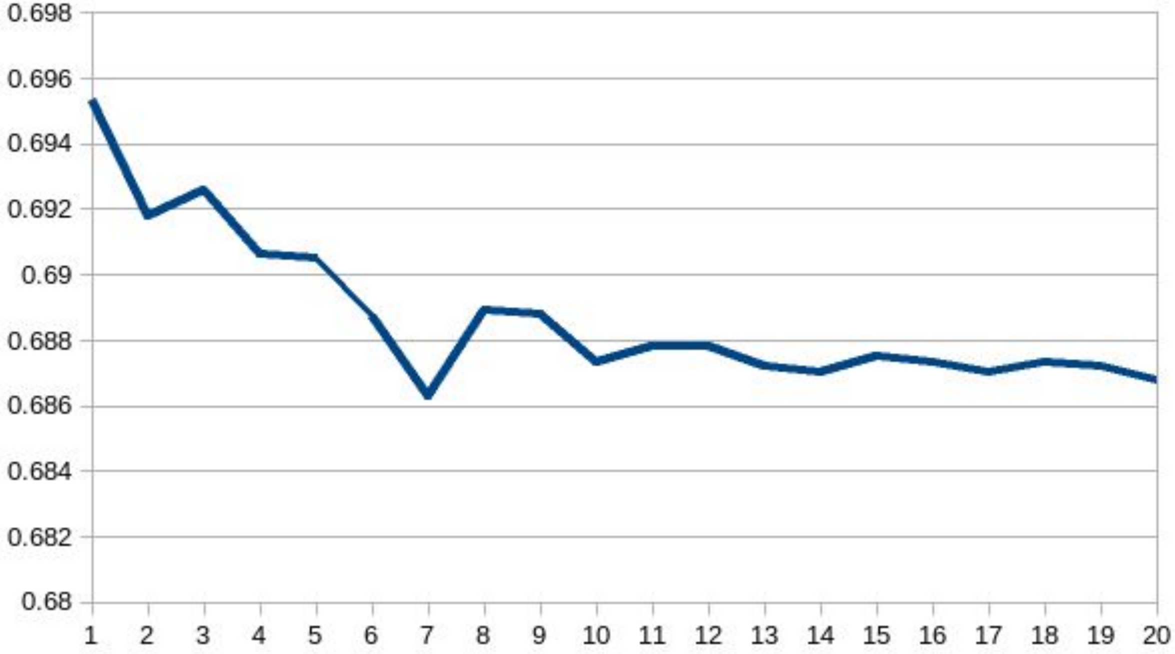
- 51.93%

Attack Testing Accuracy



Results

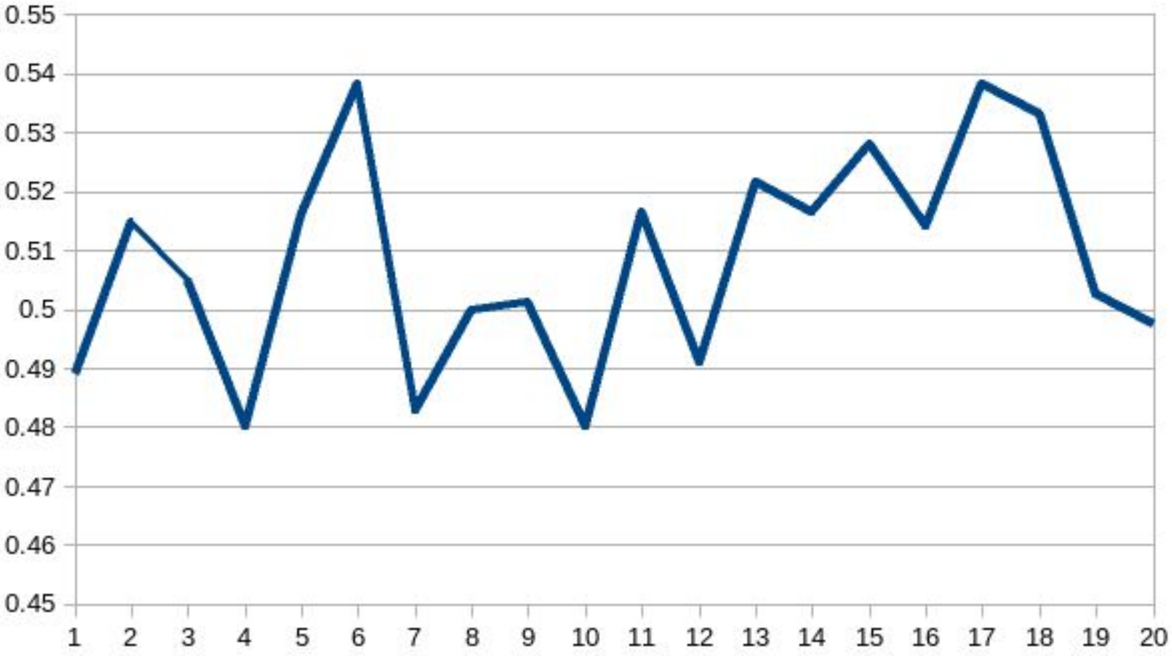
Attack Testing Loss



Results

- 1 shadow
- 20 epochs
- 50%

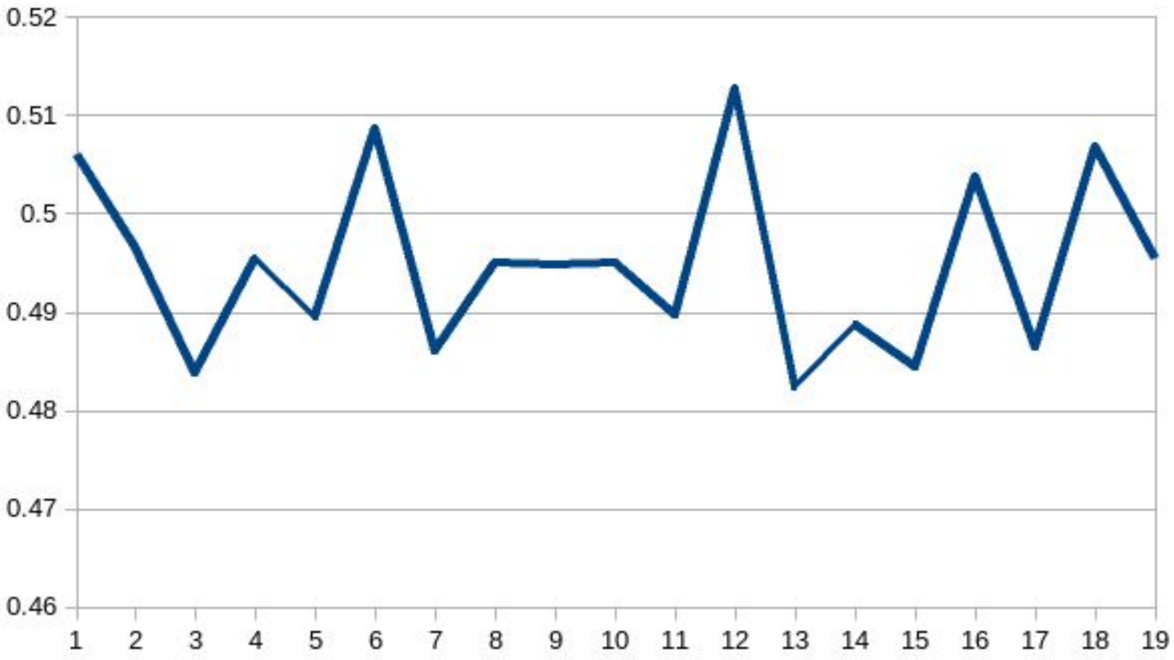
Attack Testing Accuracy



Results

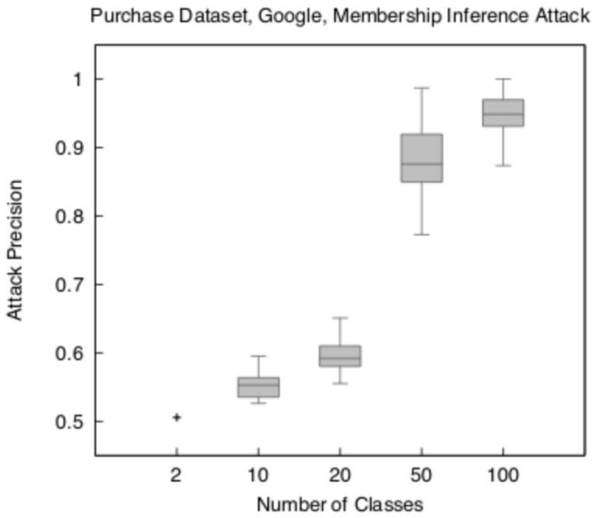
- 32x24
- 20 epochs
- 3 shadows

Attack Testing Accuracy



Discussion

- Low accuracy on high dimensional data
- Number of epochs
- Number of shadow models
- Dataset size
- Number of classes



References

- Cross, C. B.; Skipper, J. A.; and Petkie, D. T. 2013. Thermal imaging to detect physiological indicators of stress in humans. In Stockton, G. R., and Colbert, F. P., eds., Thermosense: Thermal Infrared Applications XXXV, volume 8705, 141 – 155. International Society for Optics and Photonics.
- I., R. M. 2019. An implementation of the paper on detecting facial stress using thermal image processing.
- Irolla, P. 2019. Demystifying the membership inference attack.
- Krizhevsky, A. 2012. Learning multiple layers of features from tiny images. University of Toronto.
- Kulynych, B., and Yaghini, M. 2018. mia: A library for running membership inference attacks against ML models.
- Miezianko, R. Ieee otcbvs ws series bench, terravic research infrared database.
- Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2016. Membership inference attacks against machine learning models.
- Tindall, L. 2019. Membership inference attacks on neural networks.
- Vandersteegen, M. 2018. Seek thermal compact camera driver supporting the thermal compact, thermal compactxr and thermal compactpro.