MACQUARIE University
SYDNEY·AUSTRALIA

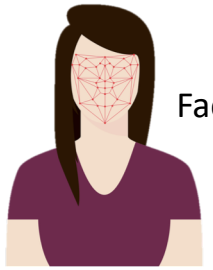
On the Resilience of Biometric Authentication Systems against Random Inputs

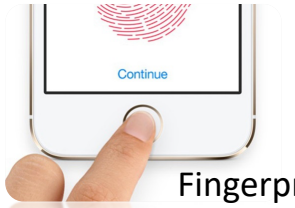
UNSW SYDNEY

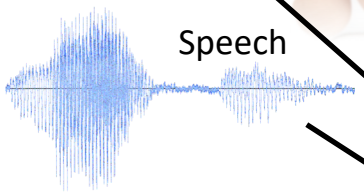
Benjamin Zhao, Hassan Asghar, Dali Kaafar

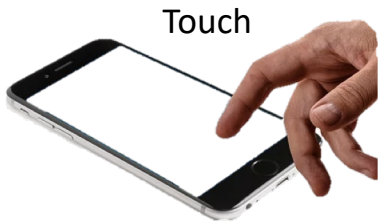
CSIRO DATA 61

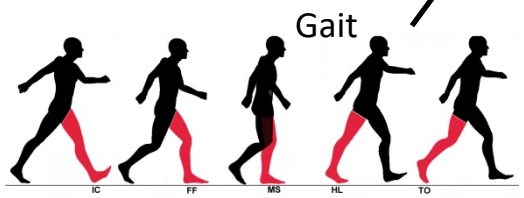
Biometric Authentication: Overview











Feature Extraction
- Engineered
- Embeddings

Registration

Sensors

User Records

Training

The Model

Metrics
- FRR,
- FPR,
- EER,
- ROC.

Authentication Decision

Yes / No

Authentication

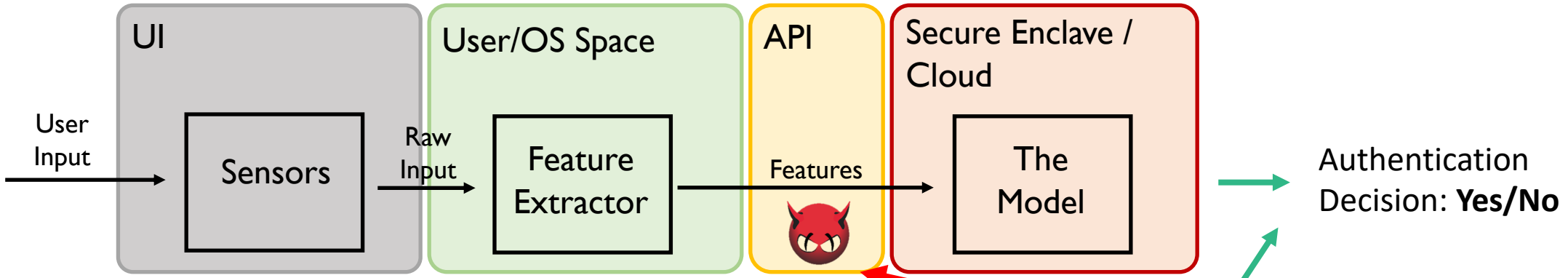
1 Record

Record In Question

Classification

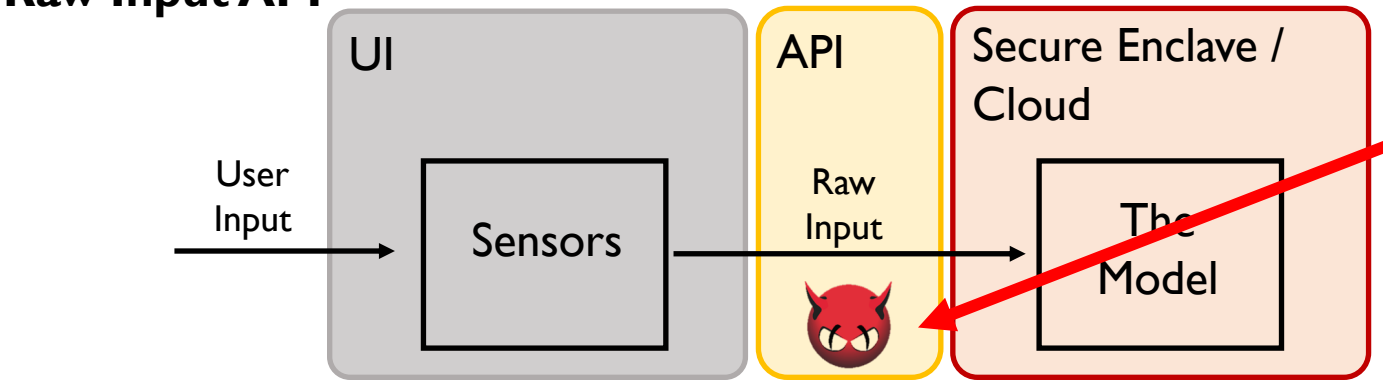
Biometrics as an API

Feature Vector API



- Engineered
- Embeddings

Raw Input API



Attack Surface

What if an attacker had access to these APIs?

What is the success of an attacker?

- Perception: FPR is indicative of success of this attacker.
- Yes, if attacker inputs have the same distribution as biometric data.

- If the API is available, an attacker has more freedom.
- In particular, an attacker can submit random inputs.

Assumptions →

Length of Input
Value Bounds
User Identifier

What is the Security of the biometric system against these Random Inputs?

Contributions

- A notion of Acceptance Region (AR): positively classified region of features.
- Formally and experimentally show AR is larger than positive user's data region.

- Show Random Input attacker with black-box feature API access succeeds more than EER.
- Show Random Input attacker with Raw Input (before feature extraction) API succeeds more than EER

- Demonstrate attack on four real-world biometric schemes, and four ML algorithms.
- Propose mitigation against attackers with either Raw or Feature API access.

- Release our code in our Repo : https://imathatguy.github.io/Acceptance-Region/

Contributions

- A notion of Acceptance Region (AR): positively classified region of features.
- Formally and experimentally show AR is larger than positive user's data region.

- Show Random Input attacker with black-box feature API access succeeds more than EER.
- Show Random Input attacker with Raw Input (before feature extraction) API succeeds more than EER

- Demonstrate attack on four real-world biometric schemes, and four ML algorithms.
- Propose mitigation against attackers with either Raw or Feature API access.

- Release our code in our Repo : https://imathatguy.github.io/Acceptance-Region/

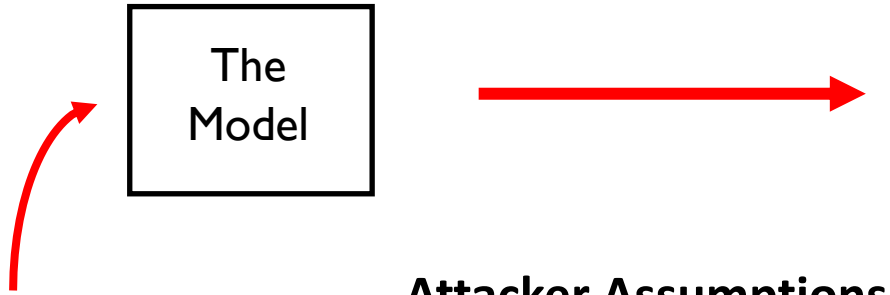
Outline

- What is the Random Input Attacker?
- How we evaluate a Random Input Attacker's Success.
- Are Random Input Attacker successful on real-world datasets?
- Factors that may affect the Success of the Random Input Attacker.
- Evaluation of factors on Synthetic Data.
- Propose a defence mechanism.

- Code Available in our Repo: https://imathatguy.github.io/Acceptance-Region/

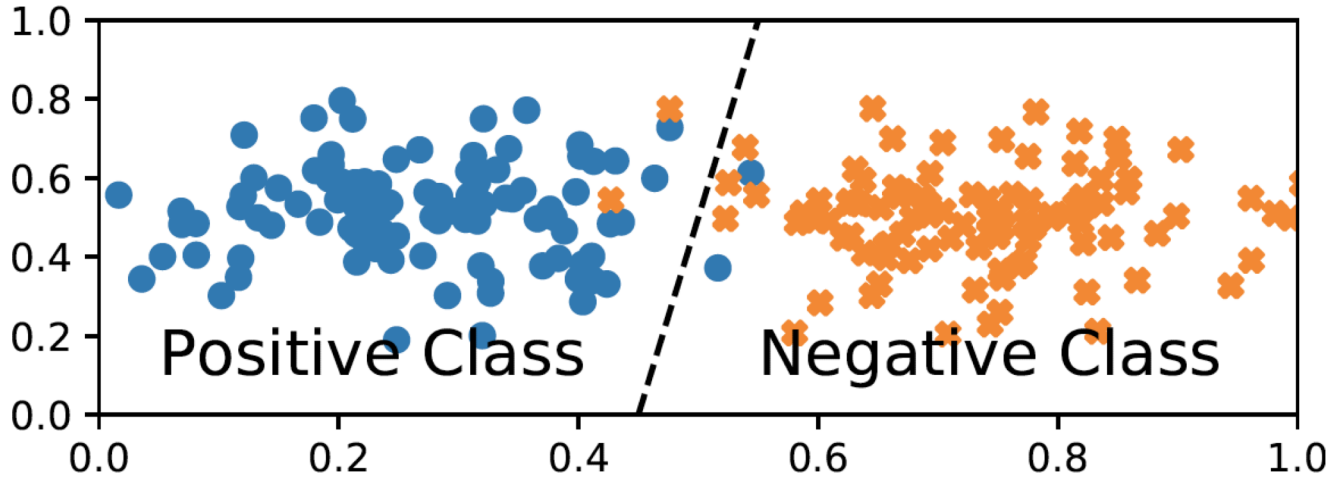
Random Input Attacker

- **How easy can a Random Input Attacker find an accepting sample?**
- The region where biometric samples are labelled as positive, Acceptance Region (**AR**).
- And this is exactly equal to success probability of an attacker submitting random inputs.



Uniformly Sampled
Random Inputs
$\{0-1, 0-1\}$

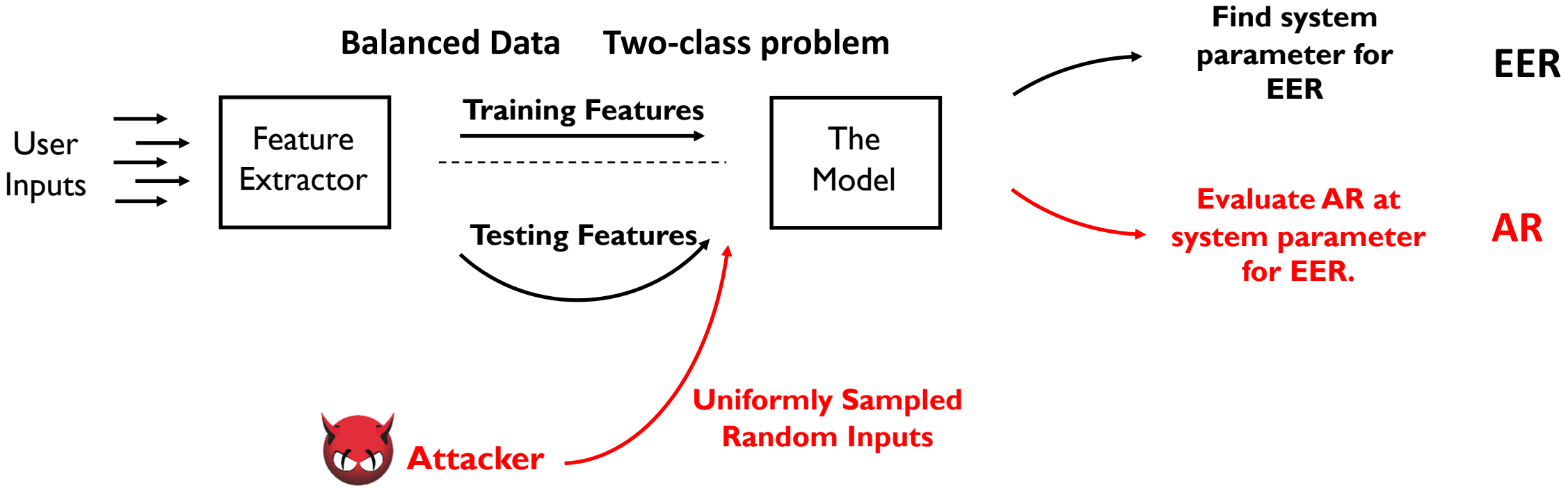
Attacker Assumptions:
Length of Input
Input Bounds
User Identifier



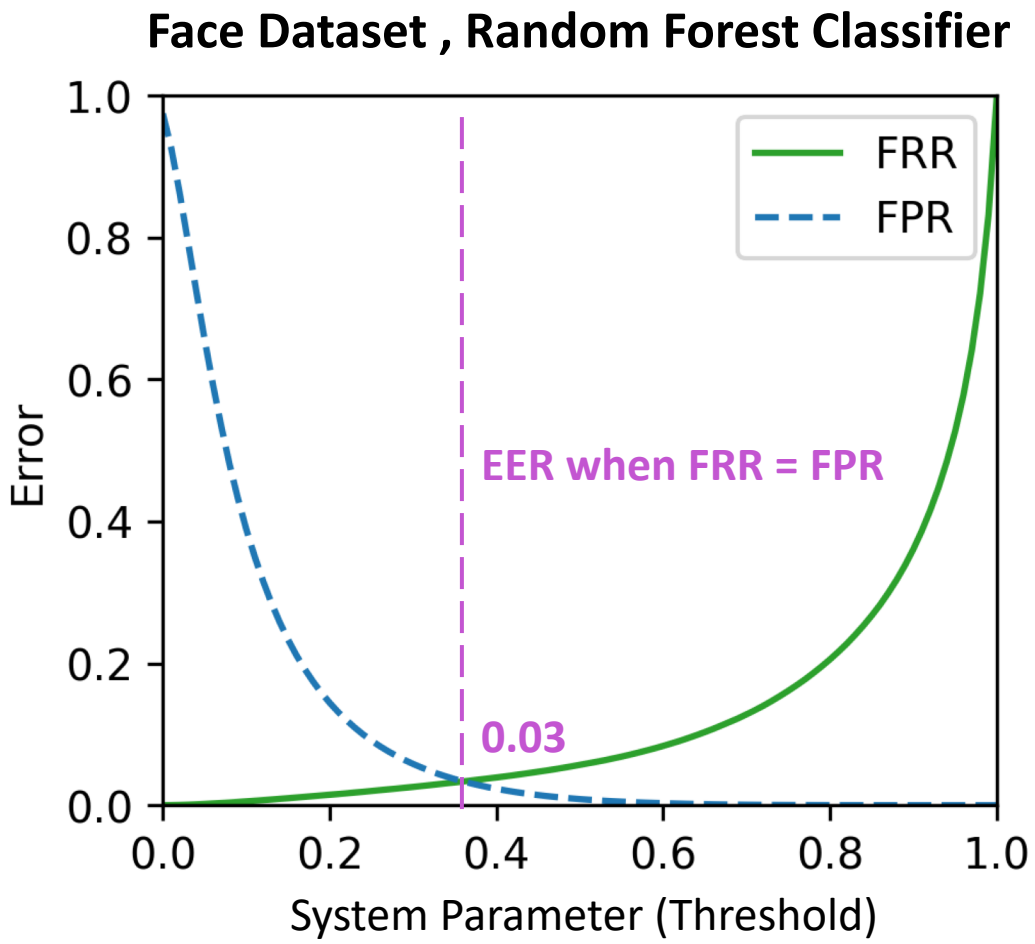
Evaluation Methodology

Gait Touch
Face Voice
Synthetic

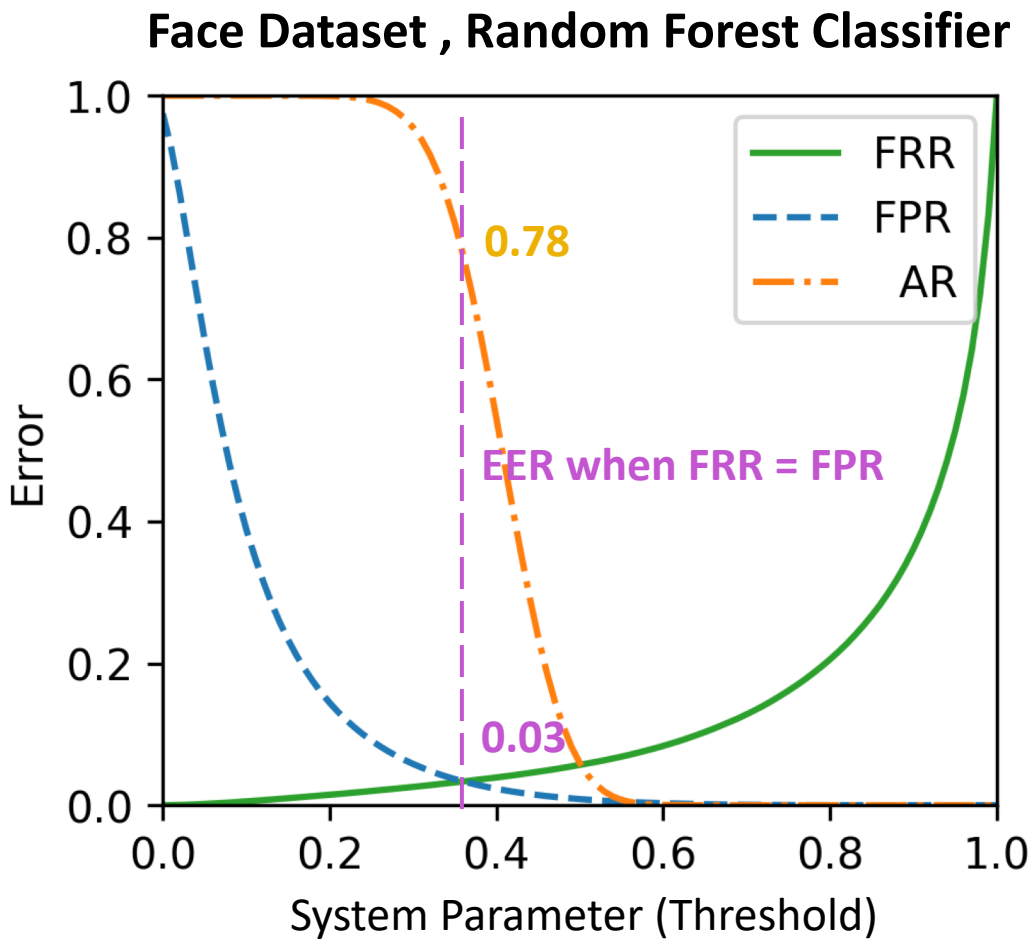
Linear SVM Radial SVM
Random Forest DNN



Real-world Data Evaluation



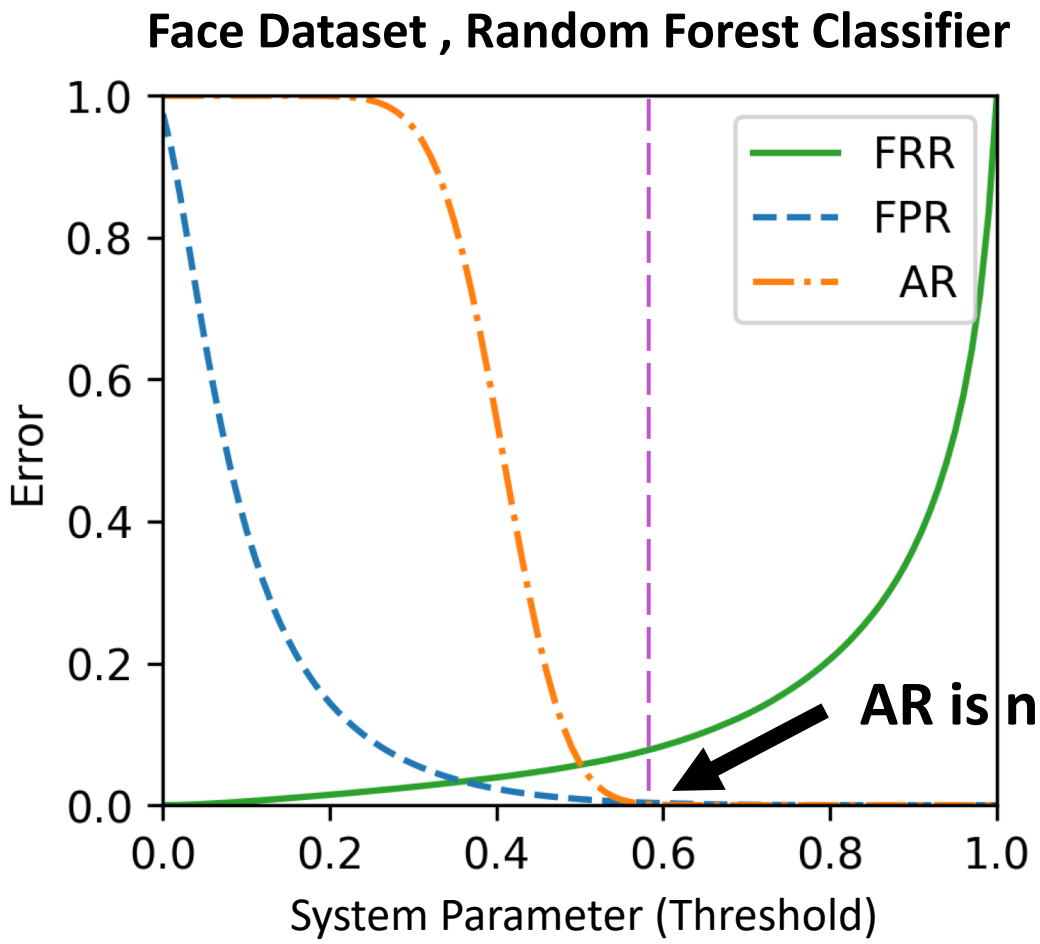
Real-world Data Evaluation



In many cases AR exceeds the EER

Hides a vulnerability not revealed by EER

Real-world Data Evaluation

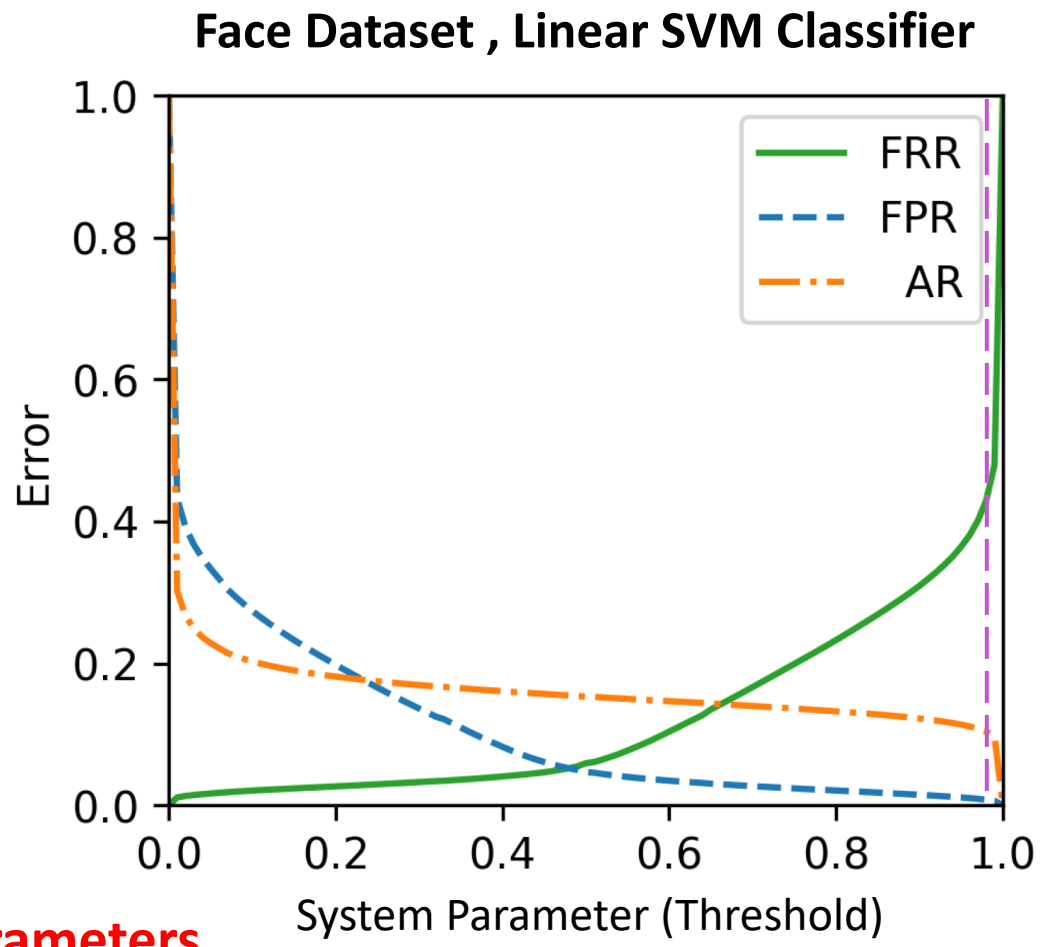


AR is now zero, Problem Solved?

In many cases AR exceeds the EER

Hides a vulnerability not revealed by EER

Real-world Data Evaluation

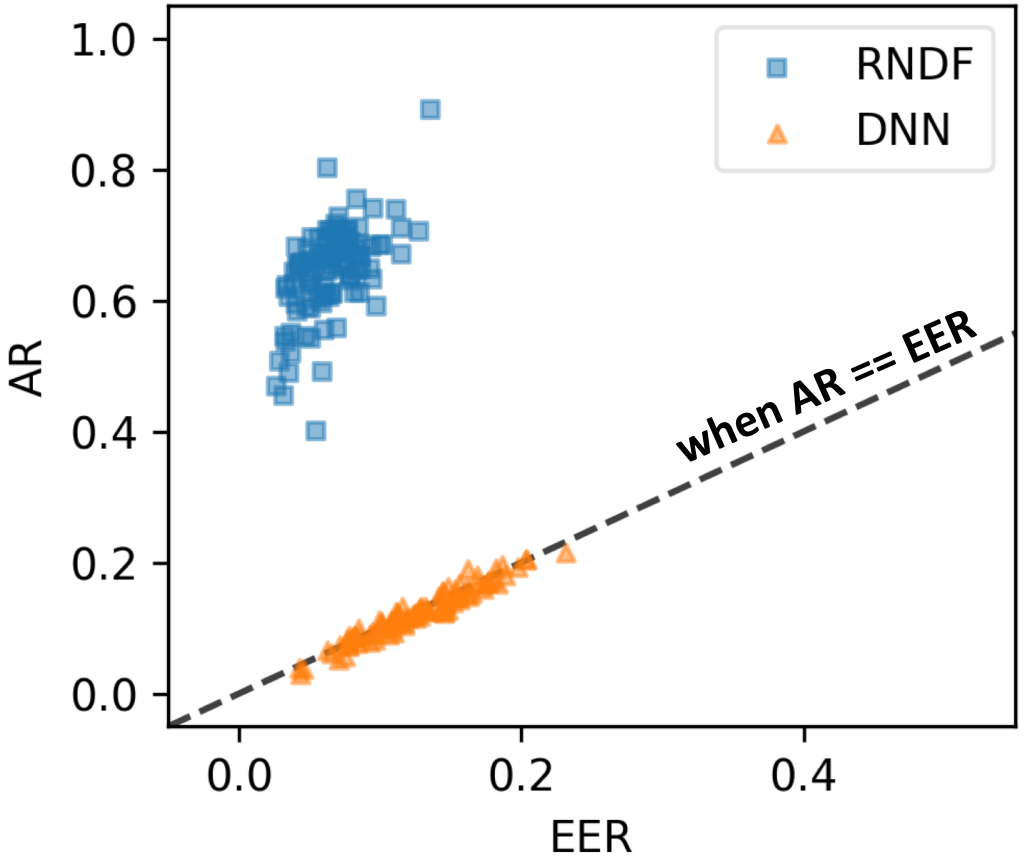


A flat AR response can be observed in many configurations

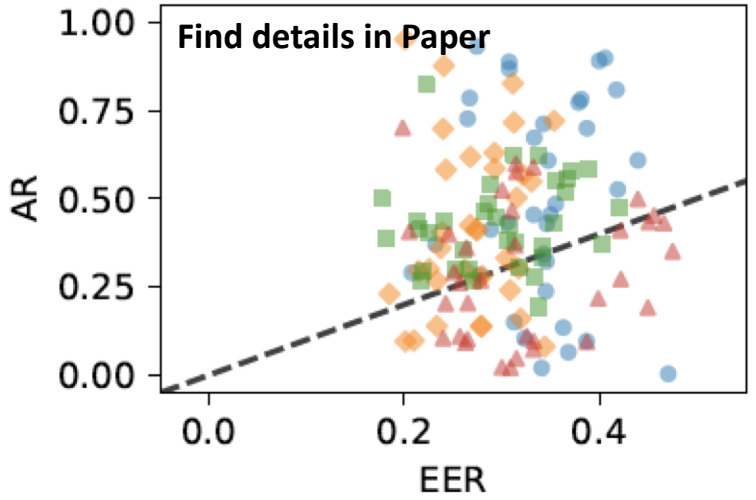
Simply adjusting system parameters is ineffective in mitigating the Random Input Attacker.

Real-world Data Evaluation - Individuals

Face Dataset, Random Forests & DNN Classifiers



Touch, All Classifiers



Relationship between a user's AR and EER not always guaranteed.

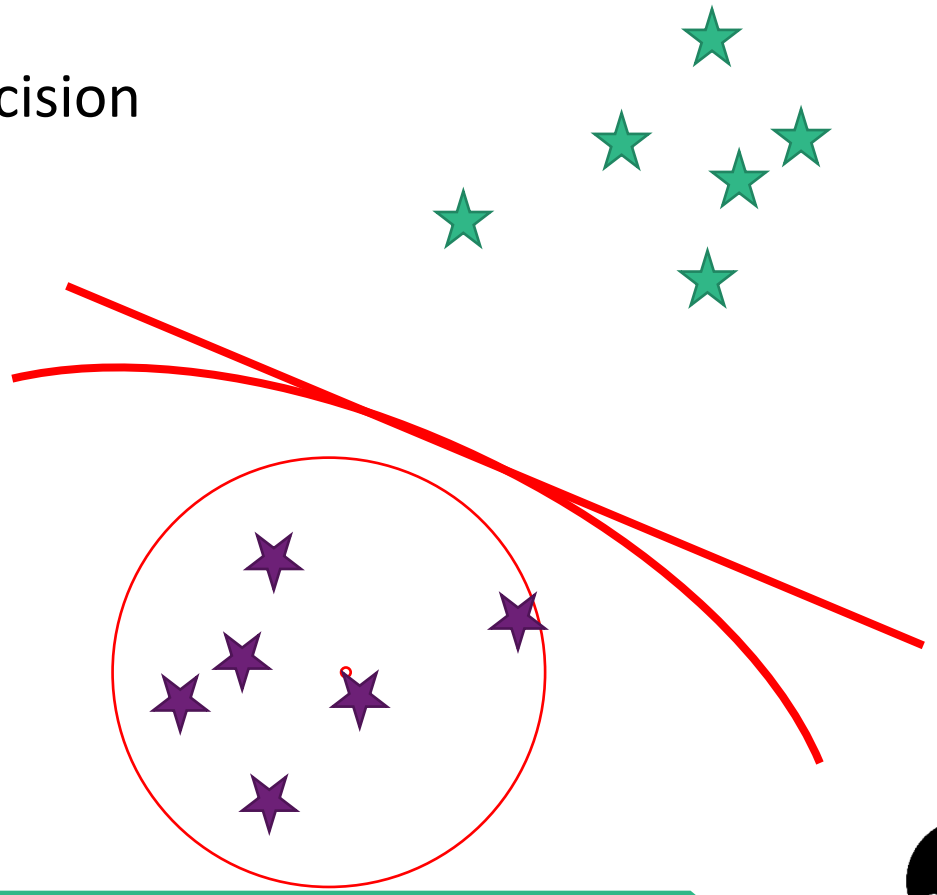
Recap – Real World Data

- ✓ Random Input Attacker, Leverages an exposed Feature Vector Input API to submit crafted inputs
- ✓ The Acceptance Region an approximate measure of success of a Random Input Attacker
- ✓ The Random Input attacker has success comparable to EER in user averages.
- ✓ An individual's EER is not a reliable indicator of Random Input Attacker success

- Outline factors that may affect the Success of the Random Input Attacker.
- Evaluation of factors on Synthetic Data.
- Propose a defence mechanism.

Factors Effecting the Acceptance Region

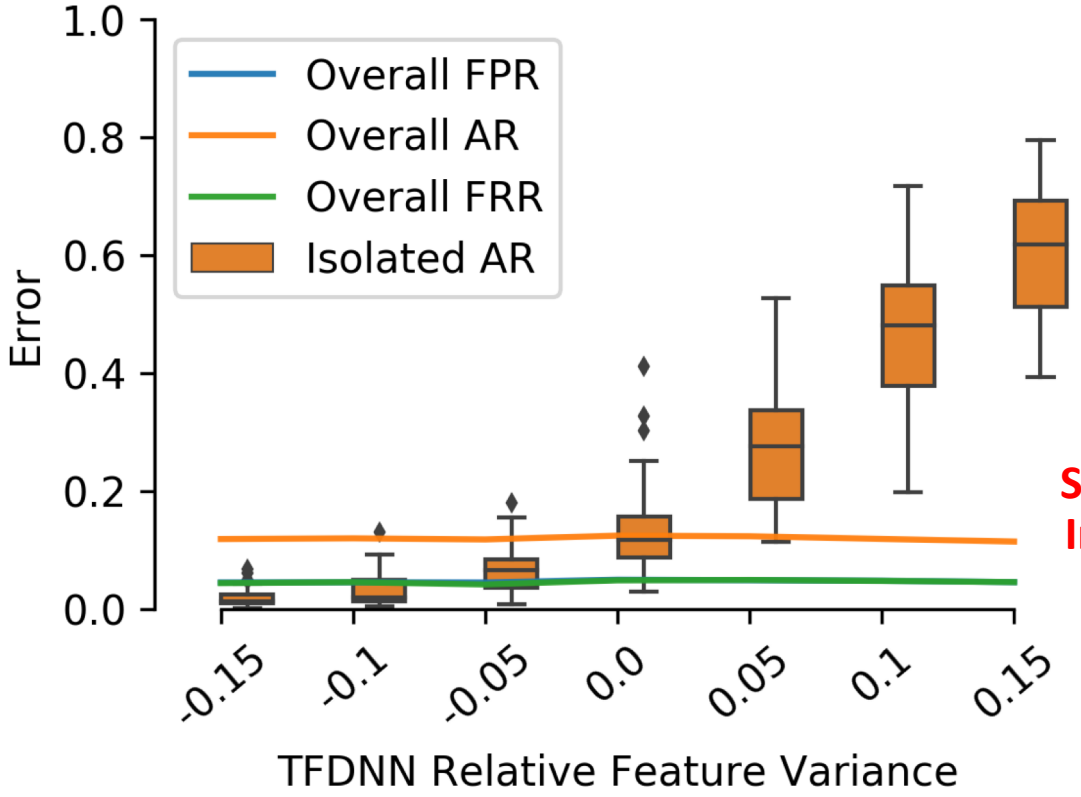
- Both the positive and negative examples are expected to be highly concentrated.
- It is desirable for models to bound it's decision boundary around this region
- However model-based classifiers do not penalize empty space.



- Variability of the Positive class.
- Variability of the Negative class.

Synthetic Data Evaluation – Positive User Variance

Synthetic Data, DNN Classifier

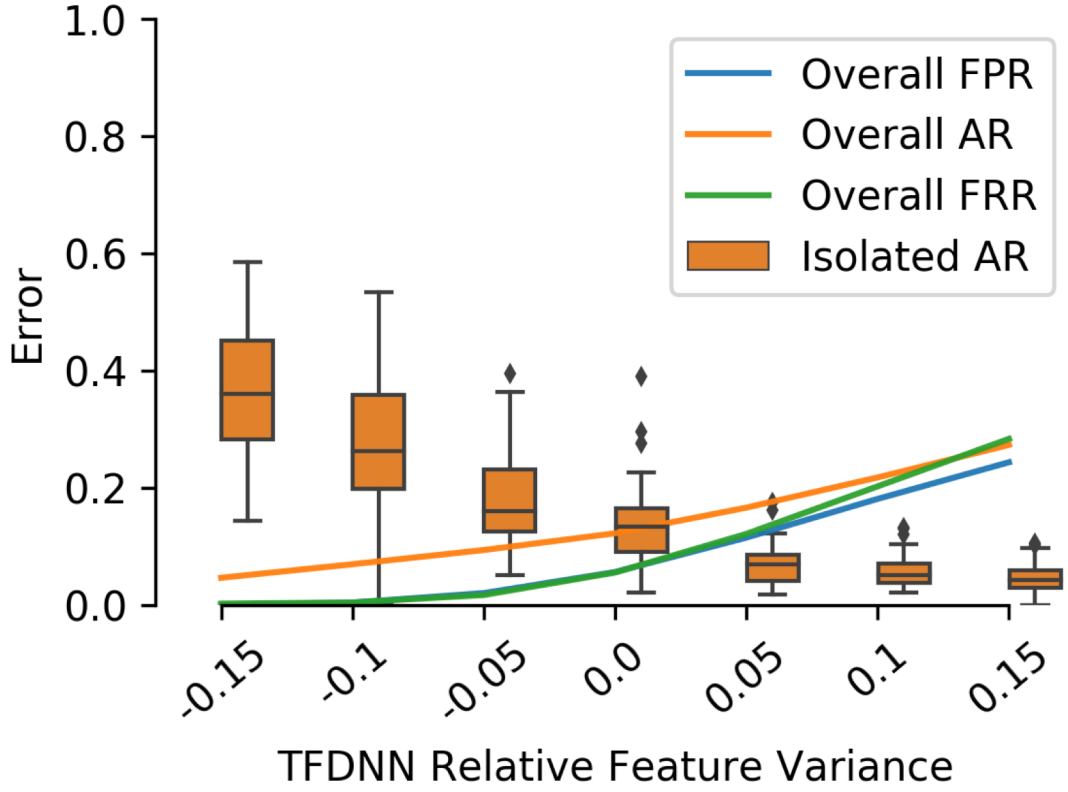


System-wide success of the Random Input Attacker may not capture the large vulnerability of a few users.

A user with high feature variance, will be more susceptible to a Random Input Attacker

Synthetic Data Evaluation – Negative User Variance

Synthetic Data, DNN Classifier



A User's vulnerability to the Random Input Attacker can be decreased by only increasing the variance of the Negative Class.

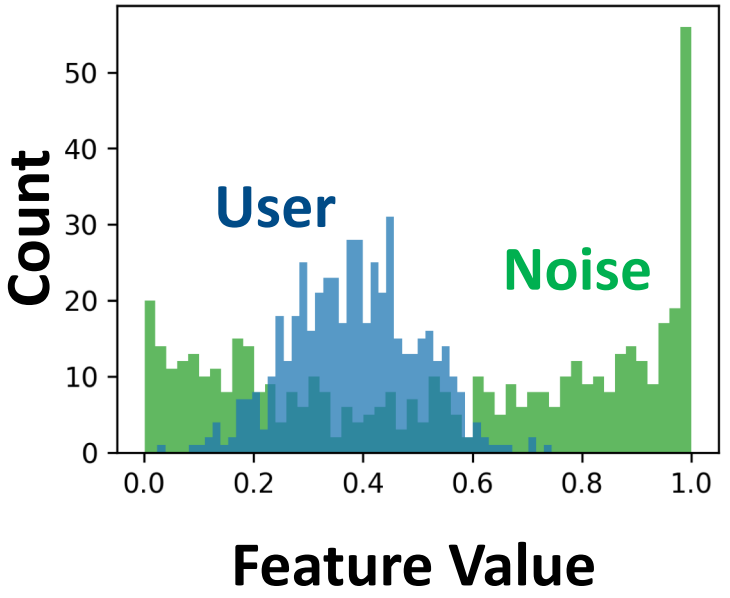
Recap – Synthetic Data

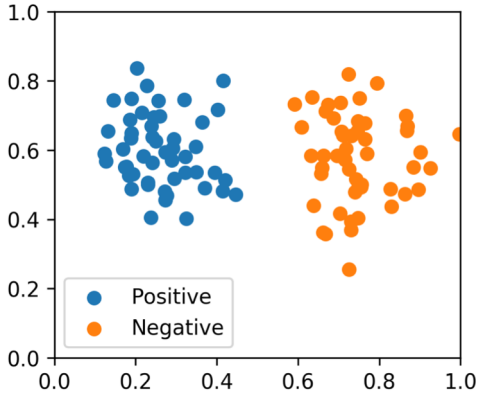
- ✓ A user with high feature variance, will be more susceptible to a Random Input Attacker
- ✓ A User's vulnerability to the Random Input Attacker can be decreased by only increasing the variance of the Negative Class.

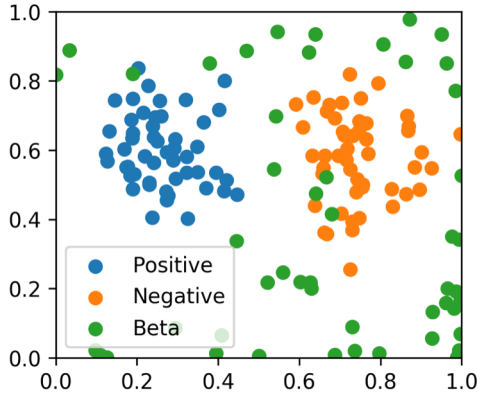
- Propose a defence mechanism.

Proposed Defence Mechanism

- If we can increase the variance of the Negative class, we can reduce the success of the Random Input Attacker.
- We can increase negative class variation with noise.
- Conveniently, Beta-distributed noise, will sample values distant from a user's values.
  - Far away from user values, minimize impact on existing EER
  - Data manipulation is algorithm Independent
- Train user model with noise vectors sampled from beta-distributions defined from the user's training samples.







Proposed Defence Mechanism – Beta Noise

- Maintain balanced dataset.
  - 1/3 positive, 1/3 negative, 1/3 beta noise.

Before Defence | After Defence

	EER	AR	EER	AR
Gait	0.09	0.03	0.09	0.00
Touch	0.21	0.23	0.21	0.00
Face	0.03	0.78	0.03	0.00
Voice	0.04	0.01	0.04	0.00

Random Forests

Before Defence | After Defence

EER	AR	EER	AR
0.215	0.20	0.170	0.00
0.325	0.30	0.375	0.00
0.095	0.10	0.065	0.04
0.115	0.08	0.090	0.02

DNN

The AR has been substantially reduced below EER

Conclusions

- Proposal of the Random Input Attacker
- Probability of success by the Random Input Attacker is comparable to EER.
  - Tuning system parameters may not necessarily mitigate the Random Attacker.
- EER is not a consistent indicator of the Random Input Attacker's success
- Class variance tied to the success of the Random Input Attacker.
- Mitigation the Random Input Attacker with beta-distributed noise at training.

More in the Paper

- Formal Treatment of Random Input Attacker and Acceptance Region
- Success of the Raw Input API Random Attacker
  - More biometric modalities, and ML algorithms
- More Factors affecting Acceptance Region
  - Distance-based classifier
  - Number of Users.
- Defending against Raw input Random Attacks.
  - Beta noise not completely sufficient
  - Additional protections proposed.

What else?

- Is the Random Input attacker equally effective against one-class and multi-class approaches to authentication?
- The effects of a non-balanced dataset on the success of the Random Input Attacker.
- Is the vulnerability of the Random Input Attacker as measured by the Acceptance Region prevalent in other ML applications?

Thank You

Question(s)?

https://imathatguy.github.io/Acceptance-Region/

For details and further info:

Benjamin Zhao
(benjamin.zhao@unsw.edu.au)

More details in Paper + Repo

Repo QR Code