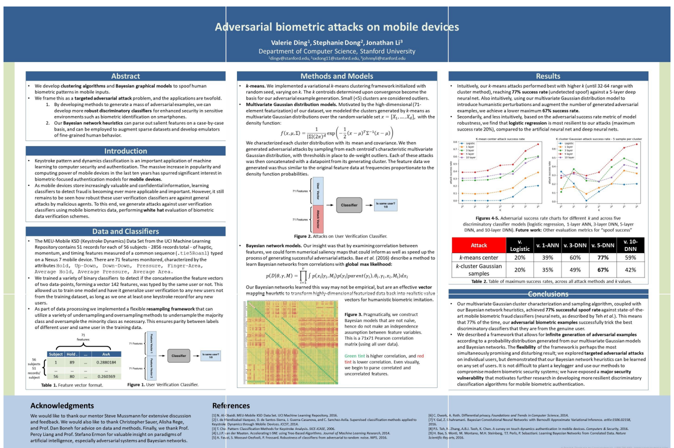
Your settings for this Block Poster are:

Pages Wide	3
Orientation	PORTRAIT
Paper Format	LETTER
Border Setting	Top: 0, Right: 0, Bottom: 0, Left: 0

You can find tips on printing, assembling and putting up your poster at www.blockposters.com

Enjoy your Block Poster!



Ac

Abstract

- We develop **clustering algorithms** and **Bayesian graphical models** to spoof human biometric patterns in mobile inputs.
- We frame this as a **targeted adversarial attack** problem, and the applications are twofold.
    1. By developing methods to generate a mass of adversarial examples, we can develop more **robust discriminatory classifiers** for enhanced security in sensitive environments such as biometric identification on smartphones.
    2. Our **Bayesian network heuristics** can parse out salient features on a case-by-case basis, and can be employed to augment sparse datasets and develop emulators of fine-grained human behavior.

Introduction

- Keystroke pattern and dynamics classification is an important application of machine learning to computer security and authentication. The massive increase in popularity and computing power of mobile devices in the last ten years has spurred significant interest in biometric-focused authentication models for **mobile devices**.
- As mobile devices store increasingly valuable and confidential information, learning classifiers to detect fraud is becoming ever more applicable and important. However, it still remains to be seen how robust these user verification classifiers are against general attacks by malicious agents. To this end, we generate attacks against user verification classifiers using mobile biometrics data, performing **white hat** evaluation of biometric data verification schemes.

Data and Classifiers

- The MEU-Mobile KSD (Keystroke Dynamics) Data Set from the UCI Machine Learning Repository contains 51 records for each of 56 subjects - 2856 records total - of haptic, momentum, and timing features measured of a common sequence (`.tie5Roanl`) typed on a Nexus 7 mobile device. There are 71 features monitored, characterized by the attributes `Hold, Up-Down, Down-Down, Pressure, Finger-Area, Average Hold, Average Pressure, Average Area.`
- We trained a variety of binary classifiers to detect if the concatenation the feature vectors of two data-points, forming a vector 142 features, was typed by the same user or not. This allowed us to train one model and have it generalize user verification to any new users not

dversarial biometric attacks on mobile devic

Valerie Ding[1], Stephanie Dong[2], Jonathan Li[3]
Department of Computer Science, Stanford University
[1]dingv@stanford.edu, [2]sxdong11@stanford.edu, [3]johnnyli@stanford.edu

Methods and Models

- ***k*-means.** We implemented a variational *k*-means clustering framework initialized with random seed, varying on *k*. The *k* centroids determined upon convergence become the basis for our adversarial example generation. Small (<5) clusters are considered outliers.
- **Multivariate Gaussian distribution models.** Motivated by the high-dimensional (71-element featurization) of our dataset, we modeled the clusters generated by *k*-means as multivariate Gaussian distributions over the random variable set $x = [X_1, \dots, X_d]$, with the density function:

$$f(x, \mu, \Sigma) = \frac{1}{|\Sigma|(2\pi)^d} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

We characterized each cluster distribution with its mean and covariance. We then generated adversarial attacks by sampling from each centroid's characteristic multivariate Gaussian distribution, with thresholds in place to de-weight outliers. Each of these attacks was then concatenated with a datapoint from its generating cluster. The feature data we generated was thus similar to the original feature data at frequencies proportionate to the density function probabilities.

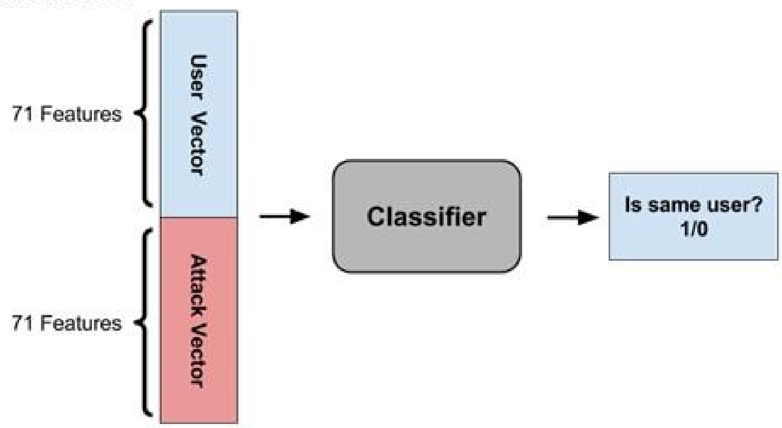


Figure 2. Attacks on User Verification Classifier.

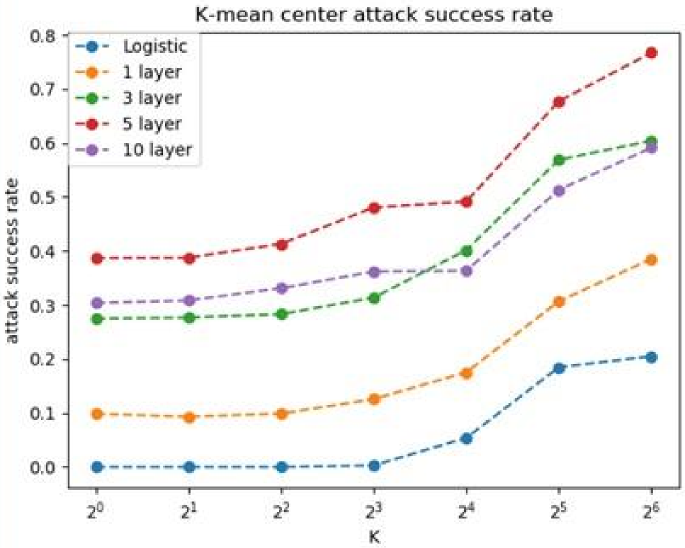
- **Bayesian network models.** Our insight was that by examining correlation between features, we could form numerical saliency maps that could inform as well as speed up the process of generating successful adversarial attacks. Bae *et al.* (2016) describe a method to learn Bayesian networks from correlations with **global max likelihood:**

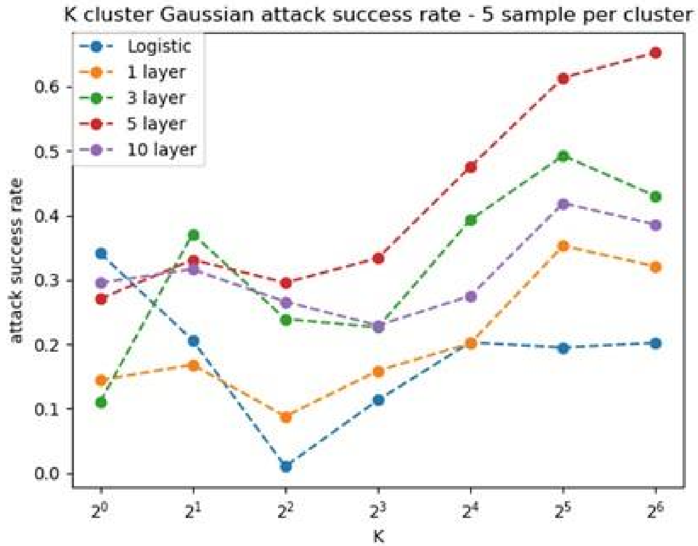
$$p(D|\theta, \gamma, M) = \prod_{i=1}^{v} \int p(x_i|\gamma_i, M_i) p(\gamma_i | parent(\gamma_i), \theta_i, \gamma_i, x_i, M_i) dx_i$$

Our Bayesian networks learned this way may not be empirical, but are an effective **vector mapping heuristic** to transform highly-dimensional featurized data back into realistic value es

Results

- Intuitively, our *k*-means attacks performed best with higher *k* (until 32-64 range with cluster method), reaching **77% success rate** (undetected spoof) against a 5-layer deep neural net. Also intuitively, using our multivariate Gaussian distribution model to introduce humanistic perturbations and augment the number of generated adversarial examples, we achieve a lower maximum **67% success rate**.
- Secondarily, and less intuitively, based on the adversarial success rate metric of model robustness, we find that **logistic regression** is most resilient to our attacks (maximum success rate 20%), compared to the artificial neural net and deep neural nets.





Figures 4-5. Adversarial success rate charts for different *k* and across five discriminatory classifier models (logistic regression, 1-layer ANN, 3-layer DNN, 5-layer DNN, and 10-layer DNN). **Future work:** Other evaluation metrics for "spoof success"

Attack	v. Logistic	v. 1-ANN	v. 3-DNN	v. 5-DNN	v. 10-DNN
k-means center	20%	39%	60%	**77%**	59%
k-cluster Gaussian samples	20%	35%	49%	**67%**	42%

Table 2. Table of maximum success rates, across all attack methods and *k* values.

Conclusions

from the training dataset, as long as we one at least one keystroke record for any new users.

- As part of data processing we implemented a flexible **resampling framework** that can utilize a variety of undersampling and oversampling methods to undersample the majority class and oversample the minority class as necessary. This ensures parity between labels of different user and same user in the training data.

71 features

56 subjects 51 records/ subject

Subject	Hold .	...	AvA
1	89	...	0.2880184
...	...	...	...
56	80	...	0.260369

Table 1. Feature vector format.

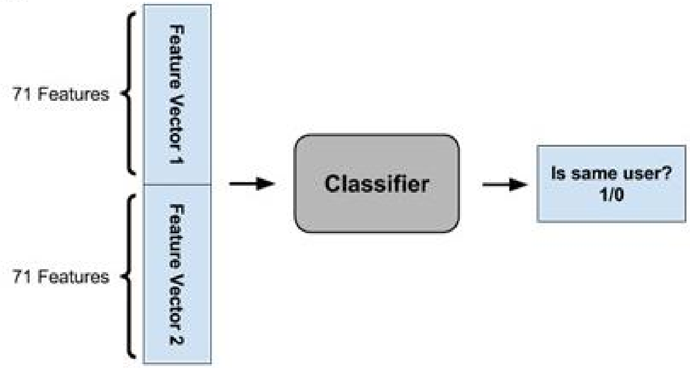


Figure 1. User Verification Classifier.

Acknowledgments

We would like to thank our mentor Steve Mussmann for extensive discussion and feedback. We would also like to thank Christopher Sauer, Alisha Rege, and Prof. Dan Boneh for advice on data and methods. Finally, we thank Prof. Percy Liang and Prof. Stefano Ermon for valuable insight on paradigms of artificial intelligence, especially adversarial systems and Bayesian networks.

Refe

[1] N. Al-(
[2] I. de M
Keystroke
[3] T. Cho
[4] L.J.P. v
[5] A. Fav

mapping heuristic to transform highly-dimensional featurized data back into realistic value vectors for humanistic biometric imitation.

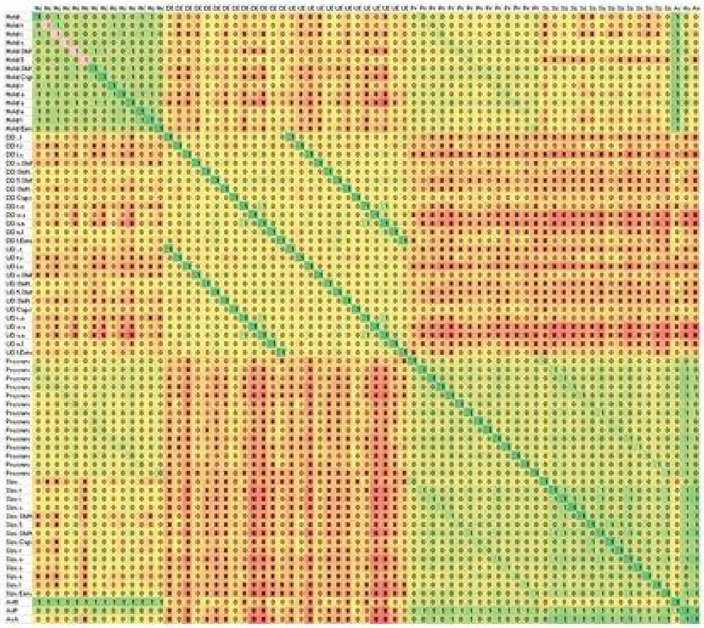


Figure 3. Pragmatically, we construct Bayesian models that are not naïve, hence do not make an independence assumption between feature variables. This is a 71x71 Pearson correlation matrix (using all user data).

Green tint is higher correlation, and red tint is lower correlation. Even visually, we begin to parse correlated and uncorrelated features.

erences

Obaidi. MEU-Mobile KSD Data Set. UCI Machine Learning Repository, 2016.

Mendizabal-Vazquez, D. de Santos-Sierra, J. Guerra-Casanova, and C. Sanchez-Avila. Supervised classification methods applied to e Dynamics through Mobile Devices. *ICCST*, 2014.

. Pattern Classification Methods for Keystroke Analysis. *SICE-ICASE*, 2006.

van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*, 2014.

wzi, S. Moosavi-Dezfooli, P. Frossard. Robustness of classifiers: from adversarial to random noise. *NIPS*, 2016.

[6] C. Dwork,

[7] Y. Gal, Z. G 2016.

[8] P.S. Teh, N

[9] H. Bae, S. *Scientific Rep*

Conclusions

- Our multivariate Gaussian cluster characterization and sampling algorithm, coupled with our Bayesian network heuristics, achieved **77% successful spoof rate** against state-of-the-art mobile biometric fraud classifiers (neural nets, as described by Teh *et al.*). This means that 77% of the time, our **adversarial biometric examples** successfully trick the best discriminatory classifiers that they are from the genuine user.
- We described a framework that allows for **infinite generation of adversarial examples** according to a probability distribution generated from our multivariate Gaussian models and Bayesian networks. The **flexibility** of the framework is perhaps the most simultaneously promising and disturbing result; we explored **targeted adversarial attacks** on individual users, but demonstrated that our Bayesian network heuristics can be learned on any set of users. It is not difficult to plant a keylogger and use our methods to compromise modern biometric security systems; we have exposed a **major security vulnerability** that motivates further research in developing more resilient discriminatory classification algorithms for mobile biometric authentication.

A. Roth. Differential privacy. *Foundations and Trends in Computer Science*, 2014.
Ghahramani. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. *arXiv:1506.02158*,

I. Zhang, A.B.J. Teoh, K. Chen. A survey on touch dynamics authentication in mobile devices. *Computers & Security*, 2016.
Monti, M. Montano, M.H. Steinberg, T.T. Perls, P. Sebastiani. Learning Bayesian Networks from Correlated Data. *Nature
ports*, 2016.