```
@book{Zimmermann2001,
	Abstract = {Since its inception, the theory of fuzzy sets has advanced in a variety of ways and in many disciplines. Applications of fuzzy technology can be found in artificial intelligence, computer science, control engineering, decision theory, expert systems, logic, management science, operations research, robotics, and others. Theoretical advances have been made in many directions. The primary goal of Fuzzy Set Theory - and its Applications, Fourth Edition is to provide a textbook for courses in fuzzy set theory, and a book that can be used as an introduction. To balance the character of a textbook with the dynamic nature of this research, many useful references have been added to develop a deeper understanding for the interested reader. Fuzzy Set Theory - and its Applications, Fourth Edition updates the research agenda with chapters on possibility theory, fuzzy logic and approximate reasoning, expert systems, fuzzy control, fuzzy data analysis, decision making and fuzzy set models in operations research. Chapters have been updated and extended exercises are included.},
	Author = {Zimmermann, H.-J.},
	Booktitle = {Kluwer, Boston, 2nd ed., 1993.},
	Doi = {10.1007/978-94-010-0646-0},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Zimmermann - Unknown - Fuzzy Set Theory- and Its Applications Fourth Edition Library of Congress Cataloging-in-Publication Data.pdf:pdf},
	Isbn = {978-94-010-3870-6},
	Issn = {19395108},
	Pages = {1--525},
	Pmid = {2943989},
	Title = {{Fuzzy Set Theory---and Its Applications}},
	Url = {http://www.springer.com/us/book/9780792374350{\%}5Cnhttp://link.springer.com/10.1007/978-94-010-0646-0},
	Year = {2001},
	Bdsk-Url-1 = {http://dx.doi.org/10.1007/978-94-010-0646-0}}

@article{Singh2014,
	Abstract = {Recent work has shown deep neural networks (DNNs) to be highly susceptible to well-designed, small perturbations at the input layer, or so-called adversarial examples. Taking images as an example, such distortions are often imperceptible, but can result in 100{\%} mis-classification for a state of the art DNN. We study the structure of adversarial examples and explore network topology, pre-processing and training strategies to improve the robustness of DNNs. We perform various experiments to assess the removability of adversarial examples by corrupting with additional noise and pre-processing with denoising autoencoders (DAEs). We find that DAEs can remove substantial amounts of the adversarial noise. However, when stacking the DAE with the original DNN, the resulting network can again be attacked by new adversarial examples with even smaller distortion. As a solution, we propose Deep Contractive Network, a model with a new end-to-end training procedure that includes a smoothness penalty inspired by the contractive autoencoder (CAE). This increases the network robustness to adversarial exam- ples, without a significant performance penalty.},
	Archiveprefix = {arXiv},
	Arxivid = {arXiv:1412.5068v3},
	Author = {Singh, Khomdram Jolson and Kho, K L Rita and Singh, Sapam Jitu and Devi, Yengkhom Chandrika and Singh, N Basanta and Sarkar, S K},
	Doi = {10.5121/ijcsa.2014.4310},
	Eprint = {arXiv:1412.5068v3},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Gu, Rigazio - 2015 - TOWARDS DEEP NEURAL NETWORK ARCHITECTURES ROBUST TO ADVERSARIAL EXAMPLES.pdf:pdf},
	Issn = {22000011},
	Journal = {Int. J. Comput. Appl.},
	Keywords = {artificial neural,equivalent circuit parameter,generalized
```

```
model,i,matlab,modeling,network,photovoltaic module,simulink,v characteristics},
	Number = {3},
	Pages = {101--116},
	Title = {{Artificial Neural Network Approach for More Accurate Solar Cell Electrical Circuit Model}},
	Url = {http://arxiv.org/abs/1412.5068{\%}5Cnhttp://www.arxiv.org/pdf/1412.5068.pdf},
	Volume = {4},
	Year = {2014},
	Bdsk-Url-1 = {http://dx.doi.org/10.5121/ijcsa.2014.4310}}

@article{AlhusseinFawzi2015,
	Abstract = {The robustness of a classifier to arbitrary small perturbations of the datapoints is a highly
desirable property when the classifier is deployed in real and possibly hostile environments. In this paper, we
propose a theoretical framework for analyzing the robustness of classifiers to adversarial perturbations, and
study two common families of classifiers. In both cases, we show the existence of a fundamental limit on the
robustness to adversarial perturbations, which is expressed in terms of a distinguishability measure between
the classes. Our result implies that in tasks involving small distinguishability, no classifier will be robust to
adversarial perturbations, even if a good accuracy is achieved. Furthermore, we show that robustness to
random noise does not imply, in general, robustness to adversarial perturbations. In fact, in high dimensional
problems, linear classifiers are shown to be much more robust to random noise than to adversarial
perturbations. Our analysis is complemented by experimental results on controlled and real-world data. Up to
our knowledge, this is the first theoretical work that addresses the surprising phenomenon of adversarial
instability recently observed for deep networks (Szegedy et al., 2014). Our work shows that this
phenomenon is not limited to deep networks, and gives a theoretical explanation of the causes underlying
the adversarial instability of classifiers.},
	Archiveprefix = {arXiv},
	Arxivid = {1502.02590v1},
	Author = {{Alhussein Fawzi} and {Omar Fawzi} and {Pascal Frossard}},
	Eprint = {1502.02590v1},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Fawzi, Fawzi,
Frossard - Unknown - Analysis of classifiers' robustness to adversarial perturbations.pdf:pdf},
	Isbn = {1502.02590},
	Number = {2014},
	Pages = {1--14},
	Title = {{Analysis of classifiers' robustness to adversarial perturbations}},
	Url = {http://arxiv.org/abs/1502.02590v1},
	Year = {2015},
	Bdsk-Url-1 = {http://arxiv.org/abs/1502.02590v1}}

@incollection{Warde-Farley,
	Abstract = {This chapter provides a review of a body of recent work on the topic of adversarial
examples and generative adversarial networks.},
	Author = {Warde-Farley, David and Goodfellow, Ian},
	Booktitle = {Perturbation, Optim. Stat.},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Hazan, Papandreou,
Tarlow - Unknown - Perturbation, Optimization and Statistics.pdf:pdf},
	Title = {{Adversarial Perturbations of Deep Neural Networks}}}

@article{DiazdeLeon2005,
	Abstract = {Restricted Boltzmann machines (RBMs) are probabilistic graphical models that can be
interpreted as stochastic neural networks. The increase in computational power and the development of
faster learning algorithms have made them applicable to relevant machine learning problems. They attracted
much attention recently after being proposed as building blocks of multi-layer learning systems called deep
belief networks. This tutorial introduces RBMs as undirected graphical models. The basic concepts of
graphical models are introduced first, however, basic knowledge in statistics is presumed. Different learning
```

```
algorithms for RBMs are discussed. As most of them are based on Markov chain Monte Carlo (MCMC)
methods, an introduction to Markov chains and the required MCMC techniques is provided.},
	Author = {{D$\backslash$'iaz de Le{\'{o}}n}, Roc$\backslash$'io and Sucar, Luis Enrique},
	Doi = {10.1007/b101756},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Biggio, Fumera, Roli -
Unknown - Pattern Recognition Systems Under Attack Design Issues and Research Challenges.pdf:pdf},
	Isbn = {978-3-642-41826-6},
	Issn = {1875-7855},
	Journal = {Lect. Notes Comput. Sci. Prog. Pattern Recognition, Image Anal. Comput. Vision, Appl.},
	Keywords = {contour detection,polar signature,region vertebra,selection,template matching,vertebral
mobility analysis,x-ray images},
	Number = {October 2005},
	Pages = {350--357},
	Pmid = {24309266},
	Title = {{Progress in Pattern Recognition, Image Analysis and Applications}},
	Url = {http://link.springer.com/chapter/10.1007/978-3-642-33275-3{\_}2},
	Volume = {3287},
	Year = {2005},
	Bdsk-Url-1 = {http://link.springer.com/chapter/10.1007/978-3-642-33275-3%7B%5C_%7D2},
	Bdsk-Url-2 = {http://dx.doi.org/10.1007/b101756}}

@misc{Tygar2011,
	Abstract = {The author briefly introduces the emerging field of adversarial machine learning, in which
opponents can cause traditional machine learning algorithms to behave poorly in security applications. He
gives a high-level overview and mentions several types of attacks, as well as several types of defenses, and
theoretical limits derived from a study of near-optimal evasion.},
	Address = {New York, New York, USA},
	Annote = {Sparsity is very dangoures for ML in adversary setting !



Should study randomizing classifiers},
	Author = {Tygar, J. D.},
	Booktitle = {IEEE Internet Comput.},
	Doi = {10.1109/MIC.2011.112},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Huang et al. - 2011 -
Adversarial machine learning.pdf:pdf},
	Isbn = {9781450310031},
	Issn = {10897801},
	Keywords = {adversarial machine learning,computer security,intrusion detection,machine learning,spam
email},
	Number = {5},
	Pages = {4--6},
	Pmid = {6015575},
	Publisher = {ACM Press},
	Title = {{Adversarial machine learning}},
	Url = {http://dl.acm.org/citation.cfm?doid=2046684.2046692},
	Volume = {15},
	Year = {2011},
	Bdsk-Url-1 = {http://dl.acm.org/citation.cfm?doid=2046684.2046692},
	Bdsk-Url-2 = {http://dx.doi.org/10.1109/MIC.2011.112}}

@article{Sommer2010,
	Abstract = {In network intrusion detection research, one popular strategy for finding attacks is
monitoring a network's activity for anomalies: deviations from profiles of normality previously learned from
```

```
benign traffic, typically identified using tools borrowed from the machine learning community. However,
despite extensive academic research one finds a striking gap in terms of actual deployments of such
systems: compared with other intrusion detection approaches, machine learning is rarely employed in
operational "real world" settings. We examine the differences between the network intrusion detection
problem and other areas where machine learning regularly finds much more success. Our main claim is that
the task of finding attacks is fundamentally different from these other applications, making it significantly
harder for the intrusion detection community to employ machine learning effectively. We support this claim by
identifying challenges particular to network intrusion detection, and provide a set of guidelines meant to
strengthen future research on anomaly detection.},
	Annote = {Seminal paper about IDS
Argues about failur of ANN in 1980 for DARPA tank use that
Some old folks of the field good survey},
	Archiveprefix = {arXiv},
	Arxivid = {file:///home/spikeh/Dropbox/chp{\%}253A10.1007{\%}252F11553595{\_}10.pdf},
	Author = {Sommer, Robin and Paxson, Vern},
	Doi = {10.1109/SP.2010.25},
	Eprint = {//home/spikeh/Dropbox/chp{\%}253A10.1007{\%}252F11553595{\_}10.pdf},
	Isbn = {978-1-4244-6894-2},
	Issn = {1081-6011},
	Journal = {IEEE Symp. Secur. Priv.},
	Keywords = {-anomaly detection,Computer science,Computer security,Computerized
monitoring,Guidelines,Laboratories,National security,Privacy,Telecommunication traffic,anomaly
detection,detection,intrusion,intrusion detection,machine learning,network security},
	Number = {May},
	Pages = {305--316},
	Pmid = {22057480},
	Primaryclass = {file:},
	Publisher = {IEEE},
	Title = {{Outside the closed world: On using machine learning for network intrusion detection}},
	Url = {http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5504793{\%}5Cnhttp://
ieeexplore.ieee.org/xpls/abs{\_}all.jsp?arnumber=5504793{\%}5Cnhttp://ieeexplore.ieee.org/lpdocs/epic03/
wrapper.htm?arnumber=5504793{\%}5Cnhttp://ieeexplore.ieee.org/articleDetails.jsp?arnumbe},
	Volume = {2},
	Year = {2010},
	Bdsk-Url-1 = {http://dx.doi.org/10.1109/SP.2010.25}}

@article{Lowd2005,
	Abstract = {Many classification tasks, such as spam filtering, intrusion detection, and terrorism
detection, are complicated by an adversary who wishes to avoid detection. Previous work on adversarial
classification has made the unrealistic assump- tion that the attacker has perfect knowledge of the classifier
[2]. In this paper, we introduce the adversarial classifier reverse engineering (ACRE) learning problem, the
task of learning sufficient information about a classifier to construct adversarial attacks. We present efficient
algorithms for re- verse engineering linear classifiers with either continuous or Boolean features and
demonstrate their effectiveness using real data from the domain of spam filtering},
	Author = {Lowd, Daniel and Meek, Christopher},
	Doi = {10.1145/1081870.1081950},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Lowd, Meek -
Unknown - Adversarial Learning.pdf:pdf},
	Isbn = {159593135X},
	Journal = {Proceeding Elev. ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '05},
	Keywords = {adversarial classification,linear classifiers,spam},
	Pages = {641},
	Title = {{Adversarial learning}},
	Url = {http://portal.acm.org/citation.cfm?doid=1081870.1081950},
```

Year = {2005},
Bdsk-Url-1 = {http://portal.acm.org/citation.cfm?doid=1081870.1081950},
Bdsk-Url-2 = {http://dx.doi.org/10.1145/1081870.1081950}}

@article{Xiao2015,
Abstract = {Machine learning algorithms are increasingly being applied in security-related tasks such as spam and malware detection, although their security properties against deliberate attacks have not yet been widely understood. Intelligent and adaptive attackers may indeed exploit specific vulnerabilities exposed by machine learning techniques to violate system security. Being robust to adversarial data manipulation is thus an important, additional requirement for machine learning algorithms to successfully operate in adversarial settings. In this work, we evaluate the security of Support Vector Machines (SVMs) to well-crafted, adversarial label noise attacks. In particular, we consider an attacker that aims to maximize the SVM's classification error by flipping a number of labels in the training data. We formalize a corresponding optimal attack strategy, and solve it by means of heuristic approaches to keep the computational complexity tractable. We report an extensive experimental analysis on the effectiveness of the considered attacks against linear and non-linear SVMs, both on synthetic and real-world datasets. We finally argue that our approach can also provide useful insights for developing more secure SVM learning algorithms, and also novel techniques in a number of related research areas, such as semi-supervised and active learning.},
Author = {Xiao, Huang and Biggio, Battista and Nelson, Blaine and Xiao, Han and Eckert, Claudia and Roli, Fabio},
Doi = {10.1016/j.neucom.2014.08.081},
File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Unknown - 2015 - Support Vector Machines under Adversarial Label Contamination.pdf:pdf;:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Unknown - 2015 - Support Vector Machines under Adversarial Label Contamination(2).pdf:pdf},
Issn = {18728286},
Journal = {Neurocomputing},
Keywords = {Adversarial learning,Label flip attacks,Label noise,Support vector machines},
Pages = {53--62},
Title = {{Support vector machines under adversarial label contamination}},
Volume = {160},
Year = {2015},
Bdsk-Url-1 = {http://dx.doi.org/10.1016/j.neucom.2014.08.081}}

@article{Goodfellow2014,
Abstract = {We propose a new framework for estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions G and D, a unique solution exists, with G recovering the training data distribution and D equal to 1/2 everywhere. In the case where G and D are defined by multilayer perceptrons, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. Experiments demonstrate the potential of the framework through qualitative and quantitative evaluation of the generated samples.},
Archiveprefix = {arXiv},
Arxivid = {arXiv:1406.2661v1},
Author = {Goodfellow, Ij and Pouget-Abadie, J and Mirza, Mehdi},
Eprint = {arXiv:1406.2661v1},
Isbn = {1406.2661},
Issn = {10495258},
Journal = {arXiv Prepr. arXiv {\ldots}},
Pages = {1--9},
Title = {{Generative Adversarial Networks}},

```
	Url = {http://arxiv.org/abs/1406.2661},
	Year = {2014},
	Bdsk-Url-1 = {http://arxiv.org/abs/1406.2661}}

@article{Xue2008,
	Abstract = {We compare discriminative and generative learning as typified by logistic regression and naive Bayes. We show, contrary to a widely- held belief that discriminative classifiers are almost always to be preferred, that there can often be two distinct regimes of per- formance as the training set size is increased, one in which each algorithm does better. This stems from the observation- which is borne out in repeated experiments- that while discriminative learning has lower asymptotic error, a generative classifier may also approach its (higher) asymptotic error much faster.},
	Archiveprefix = {arXiv},
	Arxivid = {http://dx.doi.org/10.1007/s11063-008-9088-7},
	Author = {Xue, Jing Hao and Titterington, D. Michael},
	Doi = {10.1007/s11063-008-9088-7},
	Eprint = {/dx.doi.org/10.1007/s11063-008-9088-7},
	Isbn = {1106300890},
	Issn = {13704621},
	Journal = {Neural Process. Lett.},
	Keywords = {Asymptotic relative efficiency,Discriminative classifiers,Generative classifiers,Logistic regression,Na??ve Bayes classifier,Normal-based discriminant analysis},
	Number = {3},
	Pages = {169--187},
	Pmid = {25246403},
	Primaryclass = {http:},
	Title = {{Comment on "on discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes"}},
	Volume = {28},
	Year = {2008},
	Bdsk-Url-1 = {http://dx.doi.org/10.1007/s11063-008-9088-7}}

@article{Szegedy2013,
	Abstract = {Deep neural networks are highly expressive models that have recently achieved state of the art performance on speech and visual recognition tasks. While their expressiveness is the reason they succeed, it also causes them to learn uninter- pretable solutions that could have counter-intuitive properties. In this paper we report two such properties. First, we find that there is no distinction between individual high level units and random linear combinations of high level units, according to various methods of unit analysis. It suggests that it is the space, rather than the individual units, that contains the semantic information in the high layers of neural networks. Second, we find that deep neural networks learn input-output mappings that are fairly discontinuous to a significant extent. We can cause the network to misclas- sify an image by applying a certain hardly perceptible perturbation, which is found by maximizing the network's prediction error. In addition, the specific nature of these perturbations is not a random artifact of learning: the same perturbation can cause a different network, that was trained on a different subset of the dataset, to misclassify the same input.},
	Archiveprefix = {arXiv},
	Arxivid = {arXiv:1312.6199v4},
	Author = {Szegedy, Christian and Zaremba, W and Sutskever, I},
	Doi = {10.1021/ct2009208},
	Eprint = {arXiv:1312.6199v4},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Szegedy et al. - 2013 - Intriguing properties of neural networks.pdf:pdf},
	Isbn = {1549-9618},
	Issn = {15499618},
	Journal = {arXiv Prepr. arXiv {\ldots}},
	Month = {dec},
	Pages = {1--10},
	Pmid = {22545027},
	Title = {{Intriguing properties of neural networks}},
	Url = {http://arxiv.org/abs/1312.6199},
	Year = {2013},
	Bdsk-Url-1 = {http://arxiv.org/abs/1312.6199},
	Bdsk-Url-2 = {http://dx.doi.org/10.1021/ct2009208}}

@article{Kearns1993,
	Abstract = {In this paper w e study an extension of the distribution?free model of learning in troduced b yV alian t ??? ? ?also kno wn as the pr ob ably appr oximately c orr ct or P C model? that allo e A ws the presence of malicious errors in the examples giv en to a learning algorithm? Suc h errors are generated b y an adv ersary with un bounded computational po er and access to the en w tire history of the learning algorithm?s computation? Th us? w e study a w orst?case model of errors? Our results include general methods for bounding the rate of error tolerable b yan y learning algorithm? e?cien t algorithms tolerating non trivial rates of malicious errors? and equiv alences bet een problems of learning with errors and standard com w binatorial optimization problems? ?},
	Author = {Kearns, Michael and Li, Ming},
	Doi = {10.1137/0222052},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Kearns, Li - 1993 - LEARNING IN THE PRESENCE OF MALICIOUS ERRORS.pdf:pdf},
	Isbn = {0-89791-264-0},
	Issn = {0097-5397},
	Journal = {SIAM J. Comput.},
	Number = {4},
	Pages = {807--837},
	Title = {{Learning in the Presence of Malicious Errors}},
	Url = {http://epubs.siam.org/doi/abs/10.1137/0222052},
	Volume = {22},
	Year = {1993},
	Bdsk-Url-1 = {http://epubs.siam.org/doi/abs/10.1137/0222052},
	Bdsk-Url-2 = {http://dx.doi.org/10.1137/0222052}}

@book{Scholz2011,
	Abstract = {This paper focuses on resource-aware and cost-effective indoor-localization at room-level using RFID technology. In addition to the tracking information of people wearing active RFID tags, we also include information about their proximity contacts. We present an evaluation using real-world data collected during a conference: We complement state-of-the-art machine learning approaches with strategies utilizing the proximity data in order to improve a core localization technique further.},
	Address = {Berlin, Heidelberg},
	Archiveprefix = {arXiv},
	Arxivid = {arXiv:1207.6324},
	Author = {Scholz, Christoph and Doerfel, Stephan and Atzmueller, Martin and Hotho, Andreas and Stumme, Gerd and Gunopulos, Dimitrios and Hofmann, Thomas and Malerba, Donato and Vazirgiannis, Michalis},
	Booktitle = {Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)},
	Doi = {10.1007/978-3-642-23808-6},
	Editor = {Blockeel, Hendrik and Kersting, Kristian and Nijssen, Siegfried and {\v{Z}}elezn{\'{y}}, Filip},
	Eprint = {arXiv:1207.6324},
	Isbn = {978-3-642-23807-9},
	Issn = {03029743},
	Keywords = {Computer Science},
	Number = {PART 3},
	Pages = {129--144},
	Pmid = {22183238},
	Publisher = {Springer Berlin Heidelberg},
	Series = {Lecture Notes in Computer Science},
	Title = {{Machine Learning and Knowledge Discovery in Databases}},
	Url = {http://www.springerlink.com/content/a270055474n727j2/},
	Volume = {6913},
	Year = {2011},
	Bdsk-Url-1 = {http://www.springerlink.com/content/a270055474n727j2/},
	Bdsk-Url-2 = {http://dx.doi.org/10.1007/978-3-642-23808-6}}

@article{Kantchelian2015,
	Abstract = {Recent work has successfully constructed adversarial "evading" instances for differentiable
prediction models. However generating adversarial instances for tree ensembles, a piecewise constant class
of models, has remained an open problem. In this paper, we construct both exact and approximate evasion
algorithms for tree ensembles: for a given instance x we find the "nearest" instance x' such that the classifier
predictions of x and x' are different. First, we show that finding such instances is practically possible despite
tree ensemble models being non-differentiable and the optimal evasion problem being NP-hard. In addition,
we quantify the susceptibility of such models applied to the task of recognizing handwritten digits by
measuring the distance between the original instance and the modified instance under the L0, L1, L2 and L-
infinity norms. We also analyze a wide variety of classifiers including linear and RBF-kernel models, max-
ensemble of linear models, and neural networks for comparison purposes. Our analysis shows that tree
ensembles produced by a state-of-the-art gradient boosting method are consistently the least robust models
notwithstanding their competitive accuracy. Finally, we show that a sufficient number of retraining rounds with
L0-adversarial instances makes the hardened model three times harder to evade. This retraining set also
marginally improves classification accuracy, but simultaneously makes the model more susceptible to L1, L2
and L-infinity evasions.},
	Annote = {Great paper but couldn
t underestand the main idea of the linear programming.
But get very same idea as mine -{\textgreater} finding adverserial examples and use them to harden the
classifier!},
	Archiveprefix = {arXiv},
	Arxivid = {1509.07892},
	Author = {Kantchelian, Alex and Tygar, J. D. and Joseph, Anthony D.},
	Eprint = {1509.07892},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Kantchelian, Tygar,
Joseph - 2015 - Evasion and Hardening of Tree Ensemble Classifiers.pdf:pdf},
	Isbn = {9781510829008},
	Month = {sep},
	Pages = {1--9},
	Title = {{Evasion and Hardening of Tree Ensemble Classifiers}},
	Url = {http://arxiv.org/abs/1509.07892},
	Volume = {48},
	Year = {2015},
	Bdsk-Url-1 = {http://arxiv.org/abs/1509.07892}}

@article{Corona2013,
	Abstract = {Intrusion Detection Systems (IDSs) are one of the key components for securing computing
infrastructures. Their objective is to protect against attempts to violate defense mechanisms. Indeed, IDSs
themselves are part of the computing infrastructure, and thus they may be attacked by the same adversaries
they are designed to detect. This is a relevant aspect, especially in safety-critical environments, such as
hospitals, aircrafts, nuclear power plants, etc. To the best of our knowledge, this survey is the first work to
present an overview on adversarial attacks against IDSs. In particular, this paper will provide the following
```

```
original contributions: (a) a general taxonomy of attack tactics against IDSs; (b) an extensive description of
how such attacks can be implemented by exploiting IDS weaknesses at different abstraction levels; (c) for
each attack implementation, a critical investigation of proposed solutions and open points. Finally, this paper
will highlight the most promising research directions for the design of adversary-aware, harder-to-defeat IDS
solutions. To this end, we leverage on our research experience in the field of intrusion detection, as well as
on a thorough investigation of the relevant related works published so far. ?? 2013 Elsevier Inc. All rights
reserved.},
        Annote = {This paper has the list of softwares to attack IDS
and generate anomolous traffic

Poisoning: 24, 27, 7 102, 8,81, 142
mimicary attack: 49 76 75
Poisoning 116, 105, 27, 131,81, "102" "8"
Robust Statistic and boosting: 131, 135},
        Author = {Corona, Igino and Giacinto, Giorgio and Roli, Fabio},
        Doi = {10.1016/j.ins.2013.03.022},
        File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Corona, Giacinto, Roli
- Unknown - Adversarial Attacks against Intrusion Detection Systems Taxonomy, Solutions and Open
Issues.pdf:pdf},
        Isbn = {0020-0255},
        Issn = {00200255},
        Journal = {Inf. Sci. (Ny).},
        Keywords = {Adversarial environment,Computer security,Intrusion detection system},
        Pages = {201--225},
        Title = {{Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open
issues}},
        Volume = {239},
        Year = {2013},
        Bdsk-Url-1 = {http://dx.doi.org/10.1016/j.ins.2013.03.022}}

@inproceedings{Papernot2016,
        Abstract = {Deep learning algorithms have been shown to perform extremely well on many classical
machine learning problems. However, recent studies have shown that deep learning is vulnerable to
adversarial samples: inputs crafted to force a deep neural network (DNN) to provide adversary-selected
outputs. Such attacks can seriously undermine the security of the system supported by the DNN, sometimes
with devastating consequences. For example, autonomous vehicles can be crashed, illicit or illegal content
can bypass content filters, or biometric authentication systems can be manipulated to allow improper access.
In this work, we introduce a defensive mechanism called defensive distillation to reduce the effectiveness of
adversarial samples on DNNs. We analytically investigate the generalizability and robustness properties
granted by the use of defensive distillation when training DNNs. We also empirically study the effectiveness
of our defense mechanisms on two DNNs placed in adversarial settings. The study shows that defensive
distillation can reduce effectiveness of sample creation from 95{\%} to less than 0.5{\%} on a studied DNN.
Such dramatic gains can be explained by the fact that distillation leads gradients used in adversarial sample
creation to be reduced by a factor of 10{\^{}}30. We also find that distillation increases the average minimum
number of features that need to be modified to create adversarial samples by about 800{\%} on one of the
DNNs we tested.},
        Archiveprefix = {arXiv},
        Arxivid = {1511.04508},
        Author = {Papernot, Nicolas and McDaniel, Patrick and Wu, Xi and Jha, Somesh and Swami,
Ananthram},
        Booktitle = {Proc. - 2016 IEEE Symp. Secur. Privacy, SP 2016},
        Doi = {10.1109/SP.2016.41},
        Eprint = {1511.04508},
        File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Papernot et al. - 2015
```

```
- Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks.pdf:pdf},
	Isbn = {9781509008247},
	Month = {nov},
	Pages = {582--597},
	Pmid = {7546524},
	Title = {{Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks}},
	Url = {http://arxiv.org/abs/1511.04508},
	Year = {2016},
	Bdsk-Url-1 = {http://arxiv.org/abs/1511.04508},
	Bdsk-Url-2 = {http://dx.doi.org/10.1109/SP.2016.41}}

@article{Milenkoski2015,
	Abstract = {The evaluation of computer intrusion detection systems (which we refer to as intrusion detection systems) is an active research area. In this article, we survey and systematize common practices in the area of evaluation of such systems. For this purpose, we define a design space structured into three parts: workload, metrics, and measurement methodology. We then provide an overview of the common practices in evaluation of intrusion detection systems by surveying evaluation approaches and methods related to each part of the design space. Finally, we discuss open issues and challenges focusing on evaluation methodologies for novel intrusion detection systems.},
	Author = {Milenkoski, Aleksandar and Vieira, Marco and Kounev, Samuel and Avritzer, Alberto and Payne, Bryan D},
	Doi = {10.1145/2808691},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Milenkoski et al. - 2015 - Evaluating Computer Intrusion Detection Systems.pdf:pdf},
	Isbn = {0360-0300},
	Issn = {03600300},
	Journal = {ACM Comput. Surv.},
	Keywords = {Computer intrusion detection systems,measurement methodology,metrics,workload generation},
	Month = {sep},
	Number = {1},
	Pages = {1--41},
	Publisher = {ACM},
	Title = {{Evaluating Computer Intrusion Detection Systems: A Survey of Common Practices}},
	Url = {http://dl.acm.org/citation.cfm?doid=2808687.2808691},
	Volume = {48},
	Year = {2015},
	Bdsk-Url-1 = {http://dl.acm.org/citation.cfm?doid=2808687.2808691},
	Bdsk-Url-2 = {http://dx.doi.org/10.1145/2808691}}

@article{Gornitz2009,
	Abstract = {Anomaly detection for network intrusion detection is usually considered an unsupervised task. Prominent techniques, such as one-class support vector machines, learn a hyper-sphere enclosing network data, mapped to a vector space, such that points outside of the ball are considered anomalous. However, this setup ignores relevant information such as expert and background knowledge. In this paper, we rephrase anomaly detection as an active learning task. We propose an effective active learning strategy to query low- confidence observations and to expand the data basis with minimal labeling effort. Our empirical evaluation on network intrusion detection shows that our approach consistently outperforms existing methods in relevant scenarios.},
	Address = {New York, New York, USA},
	Author = {G{\"{o}}rnitz, Nico and Kloft, Marius and Rieck, Konrad and Brefeld, Ulf},
	Doi = {10.1145/1654988.1655002},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/G{\"{o}}rnitz et al. - 2009 - Active learning for network intrusion detection.pdf:pdf},
	Isbn = {9781605587813},
	Issn = {15437221},
	Journal = {Proc. 2nd ACM Work. Secur. Artif. Intell. AISec 09},
	Keywords = {information retrieval {\&} textual information access,learning,statistics {\&} optimisation,theory {\&} algorithms},
	Pages = {47},
	Publisher = {ACM Press},
	Title = {{Active Learning for Network Intrusion Detection}},
	Url = {http://eprints.pascal-network.org/archive/00005488/},
	Year = {2009},
	Bdsk-Url-1 = {http://eprints.pascal-network.org/archive/00005488/},
	Bdsk-Url-2 = {http://dx.doi.org/10.1145/1654988.1655002}}

@misc{Yang2011,
	Abstract = {We study the problem of active learning in a stream-based setting, allowing the distribution of the examples to change over time. We prove upper bounds on the number of prediction mistakes and number of label requests for established disagreement-based active learning algorithms, both in the realizable case and under Tsybakov noise. We further prove minimax lower bounds for this problem.},
	Author = {Yang, Liu},
	Booktitle = {Adv. Neural Inf. Process. Syst.},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Yang - 2011 - Active Learning with a Drifting Distribution.pdf:pdf},
	Isbn = {9781618395993},
	Pages = {1--14},
	Title = {{Active Learning with a Drifting Distribution}},
	Url = {http://papers.nips.cc/paper/4190-active-learning-with-a-drifting-distribution},
	Volume = {1},
	Year = {2011},
	Bdsk-Url-1 = {http://papers.nips.cc/paper/4190-active-learning-with-a-drifting-distribution}}

@article{Barreno2006,
	Abstract = {Machine learning systems offer unparalled flexibility in deal- ing with evolving input in a variety of applications, such as intrusion detection systems and spam e-mail filtering. How- ever, machine learning algorithms themselves can be a target of attack by a malicious adversary. This paper provides a framework for answering the question, ``Can machine learn- ing be secure?'' Novel contributions of this paper include a taxonomy of different types of attacks on machine learn- ing techniques and systems, a variety of defenses against those attacks, a discussion of ideas that are important to security for machine learning, an analytical model giving a lower bound on attacker's work function, and a list of open problems.},
	Annote = {Great paper proposing an attack on hyperspheir anomoly detection algorithms very great theoratical analysis},
	Author = {Barreno, Marco and Nelson, Blaine and Sears, Russell and Joseph, Anthony D. and Tygar, J. D.},
	Doi = {10.1145/1128817.1128824},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Barreno et al. - Unknown - Can Machine Learning Be Secure.pdf:pdf},
	Isbn = {1595932720},
	Journal = {Proc. 2006 Symp. Information, Comput. Commun. Secur.},
	Keywords = {adversarial learning,computer networks,computer secu-,game theory,intrusion detection,machine learning,rity,security metrics,spam filters,statistical learning},
	Number = {March},
	Pages = {16--25},
	Title = {{Can machine learning be secure?}},
	Url = {http://dl.acm.org/citation.cfm?id=1128824},
	Year = {2006},
	Bdsk-Url-1 = {http://dl.acm.org/citation.cfm?id=1128824},
	Bdsk-Url-2 = {http://dx.doi.org/10.1145/1128817.1128824}}

@inproceedings{Zhao2012,
	Abstract = {Active learning has played an important role in many areas because it can reduce human efforts by just selecting most informative instances for training. Nevertheless, active learning is vulnerable in adversarial environments, including intrusion detection or spam filtering. The purpose of this paper was to reveal how active learning can be attacked in such environments. In this paper, three contributions were made: first, we analyzed the sampling vulnerability of active learning; second, we presented a game framework of attack against active learning; third, two sampling attack methods were proposed, including the adding attack and the deleting attack. Experimental results showed that the two proposed sampling attacks degraded sampling efficiency of naive-bayes active learner.},
	Author = {Zhao, Wentao and Long, Jun and Yin, Jianping and Cai, Zhiping and Xia, Geming},
	Booktitle = {Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)},
	Doi = {10.1007/978-3-642-34620-0_21},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Zhao et al. - 2012 - Sampling Attack against Active Learning in Adversarial Environment.pdf:pdf},
	Isbn = {9783642346194},
	Issn = {03029743},
	Keywords = {Active Learning,Adversarial Environment},
	Pages = {222--223},
	Publisher = {Springer, Berlin, Heidelberg},
	Title = {{Sampling attack against active learning in adversarial environment}},
	Url = {http://link.springer.com/10.1007/978-3-642-34620-0{\_}21},
	Volume = {7647 LNAI},
	Year = {2012},
	Bdsk-Url-1 = {http://link.springer.com/10.1007/978-3-642-34620-0%7B%5C_%7D21},
	Bdsk-Url-2 = {http://dx.doi.org/10.1007/978-3-642-34620-0_21}}

@inproceedings{??rndi??2014,
	Abstract = {Learning-based classifiers are increasingly used for detection of various forms of malicious data. However, if they are deployed online, an attacker may attempt to evade them by manipulating the data. Examples of such attacks have been previously studied under the assumption that an attacker has full knowledge about the deployed classifier. In practice, such assumptions rarely hold, especially for systems deployed online. A significant amount of information about a deployed classifier system can be obtained from various sources. In this paper, we experimentally investigate the effectiveness of classifier evasion using a real, deployed system, PDFrate, as a test case. We develop a taxonomy for practical evasion strategies and adapt known evasion algorithms to implement specific scenarios in our taxonomy. Our experimental results reveal a substantial drop of PDFrate's classification scores and detection accuracy after it is exposed even to simple attacks. We further study potential defense mechanisms against classifier evasion. Our experiments reveal that the original technique proposed for PDFrate is only effective if the executed attack exactly matches the anticipated one. In the discussion of the findings of our study, we analyze some potential techniques for increasing robustness of learning-based systems against adversarial manipulation of data.},
	Author = {??rndi??, Nedim and Laskov, Pavel},
	Booktitle = {Proc. - IEEE Symp. Secur. Priv.},
	Doi = {10.1109/SP.2014.20},
	Isbn = {9781479946860},
	Issn = {10816011},
	Month = {may},
	Pages = {197--211},
	Publisher = {IEEE},
	Title = {{Practical evasion of a learning-based classifier: A case study}},
	Url = {http://ieeexplore.ieee.org/document/6956565/},
	Year = {2014},
	Bdsk-Url-1 = {http://ieeexplore.ieee.org/document/6956565/},
	Bdsk-Url-2 = {http://dx.doi.org/10.1109/SP.2014.20}}

@article{Cheng2012,
	Abstract = {Detecting attacks disguised by evasion techniques is a challenge for signature-based
Intrusion Detection Systems (IDSs) and Intrusion Prevention Systems (IPSs). This study examines five
common evasion techniques to determine their ability to evade recent systems. The denial-of-service (DoS)
attack attempts to disable a system by exhausting its resources. Packet splitting triestochop dataintosmall
packets, so that a system may not completely reassemble the packets for signature matching. Duplicate
insertion can mislead a system if the system and the target host discard different TCP/IP packets with a
duplicate offset or sequence. Payload mutation fools a system with a mutative payload. Shellcode mutation
transforms an attacker's shellcode to escape signature detection. This study assesses the effectiveness of
these techniques on three recent signature-based systems, and among them, explains why Snort can be
evaded. The results indicate that duplicate insertion becomes less effective on recent systems, but packet
splitting, payload mutation and shellcode mutation can be still effective against them.},
	Annote = {Packet Splitting is a form of attack against signiture matching IDS},
	Author = {Cheng, Tsung Huan and Lin, Ying Dar and Lai, Yuan Cheng and Lin, Po Ching},
	Doi = {10.1109/SURV.2011.092311.00082},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Cheng et al. - 2012 -
Evasion techniques Sneaking through your intrusion detectionprevention systems.pdf:pdf},
	Isbn = {1553-877X},
	Issn = {1553877X},
	Journal = {IEEE Commun. Surv. Tutorials},
	Keywords = {Attacks,Evasion,IDS/IPS,Signature},
	Number = {4},
	Pages = {1011--1020},
	Pmid = {315392500005},
	Title = {{Evasion techniques: Sneaking through your intrusion detection/prevention systems}},
	Volume = {14},
	Year = {2012},
	Bdsk-Url-1 = {http://dx.doi.org/10.1109/SURV.2011.092311.00082}}

@article{Biggio2013,
	Abstract = {In security-sensitive applications, the success of machine learning depends on a thorough
vetting of their resistance to adversarial data. In one pertinent, well-motivated attack scenario, an adversary
may attempt to evade a deployed system at test time by carefully manipulating attack samples. In this work,
we present a simple but effective gradient-based approach that can be exploited to systematically assess the
security of several, widely-used classification algorithms against evasion attacks. Following a recently
proposed framework for security evaluation, we simulate attack scenarios that exhibit different risk levels for
the classifier by increasing the attacker's knowledge of the system and her ability to manipulate attack
samples. This gives the classifier designer a better picture of the classifier performance under evasion
attacks, and allows him to perform a more informed model selection (or parameter setting). We evaluate our
approach on the relevant security task of malware detection in PDF files, and show that such systems can
be easily evaded. We also sketch some countermeasures suggested by our analysis.},
	Annote = {read 11},
	Author = {Biggio, Battista and Corona, Igino and Maiorca, Davide and Nelson, Blaine and {\v{S}}
rndi{\'{c}}, Nedim and Laskov, Pavel and Giacinto, Giorgio and Roli, Fabio},
	Doi = {10.1007/978-3-642-40994-3_25},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Biggio et al. -
Unknown - Evasion Attacks Against Machine Learning at Test Time.pdf:pdf},
	Isbn = {978-3-642-40993-6},
	Issn = {03029743},
	Journal = {Mach. Learn. Knowl. Discov. Databases},
	Pages = {387--402},
	Title = {{Evasion Attacks against Machine Learning at Test Time}},
	Url = {http://dx.doi.org/10.1007/978-3-642-40994-3{\_}25},
	Volume = {8190},
	Year = {2013},
	Bdsk-Url-1 = {http://dx.doi.org/10.1007/978-3-642-40994-3%7B%5C_%7D25},
	Bdsk-Url-2 = {http://dx.doi.org/10.1007/978-3-642-40994-3_25}}

@article{Joseph2009,
	Annote = {[2,8,22] {\textless}{\~{}}{\~{}}{\~{}} some defensive techniques against adverserial machine
learning
and 4

[31-26] {\textless}{\~{}}{\~{}}{\~{}} red herring

[28] {\textless}{\~{}}{\~{}}{\~{}} PCA sensitivity},
	Author = {Joseph, Anthony D and Taft, Nina},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Joseph, Taft - 2009 -
ANTIDOTE Understanding and Defending against Poisoning of Anomaly Detectors.pdf:pdf},
	Isbn = {9781605587707},
	Journal = {Traffic},
	Keywords = {adversarial learning,network traffic analysis,principal components analysis,robust
statistics},
	Number = {November},
	Pages = {1--14},
	Title = {{ANTIDOTE : Understanding and Defending against}},
	Year = {2009}}

@article{Kantarcioglu,
	Author = {Kantarcioglu, Murat and Xi, Bowei},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Kantarcioglu, Xi -
Unknown - Adversarial Data Mining for Cyber Security.pdf:pdf},
	Title = {{Adversarial Data Mining for Cyber Security}}}

@article{Rubinstein2009,
	Abstract = {We consider systems that use PCA-based detectors obtained from a comprehensive view
of the network's traffic to identify anomalies in backbone networks. To assess these detectors' susceptibility
to adversaries wishing to evade detection, we present and evaluate short-term and long-term data poisoning
schemes that trade-off between poisoning duration and the volume of traffic injected for poisoning. Stealthy
Boiling Frog attacks significantly reduce chaff volume,while only moderately increasing poisoning duration.
ROC curves provide a comprehensive analysis of PCA-based detection on contaminated data, and show
that even small attacks can undermine this otherwise successful anomaly detector.},
	Annote = {read [5,6] for traffic generation},
	Author = {Rubinstein, Benjamin I.P. and Nelson, Blaine and Huang, Ling and Joseph, Anthony D. and
Lau, Shing-hon and Rao, Satish and Taft, Nina and Tygar, J. D.},
	Doi = {10.1145/1639562.1639592},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Nelson, Joseph, Taft -
Unknown - Stealthy Poisoning Attacks on PCA-based Anomaly Detectors.pdf:pdf},
	Issn = {01635999},
	Journal = {ACM SIGMETRICS Perform. Eval. Rev.},
	Number = {October},
	Pages = {73},
	Title = {{Stealthy poisoning attacks on PCA-based anomaly detectors}},
	Volume = {37},
	Year = {2009},
	Bdsk-Url-1 = {http://dx.doi.org/10.1145/1639562.1639592}}

@article{Kantchelian2013,
	Abstract = {In this position paper, we argue that to be of practical interest, a machine-learning based
security system must engage with the human operators beyond feature engineering and instance labeling to
address the challenge of drift in adversarial environments. We propose that designers of such systems
broaden the classification goal into an explanatory goal, which would deepen the interaction with system's
operators.$\backslash$r$\backslash$nTo provide guidance, we advocate for an approach based on
maintaining one classifier for each class of unwanted activity to be filtered. We also emphasize the necessity
for the system to be responsive to the operators constant curation of the training set. We show how this
paradigm provides a property we call isolation and how it relates to classical causative attacks.$
\backslash$r$\backslash$nIn order to demonstrate the effects of drift on a binary classification task, we also
report on two experiments using a previously unpublished malware data set where each instance is
timestamped$\backslash$r$\backslash$naccording to when it was seen.},
	Annote = {45 IDS semantic gap
Read 8: it is a robust anomoly model against "poisining attack"

Read 37,21: both for detecting drift},
	Author = {Kantchelian, Alex and Afroz, Sadia and Huang, Ling and Islam, Aylin Caliskan and Miller,
Brad and Tschantz, Michael Carl and Greenstadt, Rachel and Joseph, Anthony D and Tygar, J.D.},
	Doi = {10.1145/2517312.2517320},
	File = {:Users/pooria/Library/Application Support/Mendeley Desktop/Downloaded/Kantchelian et al. -
2013 - Approaches to Adversarial Drift.pdf:pdf},
	Isbn = {9781450324885},
	Issn = {15437221},
	Journal = {AISec},
	Keywords = {D46 [Security and Pro-tection],H12 [User/Machine Systems],Invasive
software,Learning,concept drift,malware classification},
	Pages = {99--109},
	Title = {{Approaches to Adversarial Drift}},
	Year = {2013},
	Bdsk-Url-1 = {http://dx.doi.org/10.1145/2517312.2517320}}
```