Table 5. Categorization Table for Papers - Part 1

ID	Paper	Data Property	Problem Setup							Practicality							
			Target Distribution	Model		Robustness Setting				Applicability		Exp.	Type of Evidence				
				Learning Task	Classifier Type	Definition of Robustness	Attacker's Knwl.	Attacker's Tech.	Perturb Bound	Metr.	Tech.		Fml.	Empirical			
														Dataset	Classifier Type	Training Proc.	Attacks
1	Amsaleg et al. [10]	Dimensionality	Any	Any	Any	Radius based	White box	Any	L_2	✓	✗	✗	✓	C-10, IN	k-NN	Standard	N/A
2	Awasthi et al. [12]	Dimensionality	Any	Any	DNNs	Radius based	White box	Any	L_2, L_∞	✓	✗	✗	✗	C-10, C-100	DNNs	Adversarial	PGD
3	Bhagoji et al. [17]	Separation	Any	Binary Classif.	Any	Error-rate based	White box	Gradient based	L_2	✓	✗	✗	✓	C-10, M, FM	DNNs	Adversarial	PGD, FGSM
4	Bhattacharjee et al. [19]	Separation	Any	Binary Classif.	Non-parametric classifiers	Radius based	White box	Distance based	L_2	✓	✓	✓	✓	HM	Histogram, 1-NN	Standard	Distance-based attacks
5	Bhattacharjee et al. [20]	Number of samples, Dimensionality, Separation	Well-separated	Binary Classif.	Linear	Error-rate based	White box	Any	$L_p, p > 2$	✓	✗	✓	✓	N/A	N/A	N/A	N/A
6	Blum et al. [22]	Dimensionality	Any	Any	Randomized smoothed classifier	Radius based	White box	Any	$L_p, p > 2$	✓	✗	✗	✓	C-10	Smoothed DNN	Adversarial	Gaussian noise
7	Bui et al. [26]	Separation	Any	Any	DNNs	Error-rate based	White box	Gradient based	L_p	✗	✓	✗	✗	C-10, M	CNNs	Adversarial	PGD
8	Carbone et al. [27]	Dimensionality	Any	Any	Bayesian neural network	Radius based	White box	Gradient based	L_∞	✗	✗	✓	✓	M, FM, HM	Bayesian neural network	Adversarial	PGD,FGSM
9	Carmon et al. [29]	Number of samples	Gaussian-mixture (theory), Any (application)	Binary Classif.	Any	Radius based	White box	Gradient based	L_2, L_∞	✓	✓	✗	✓	C-10, S	CNNs	Adversarial	PGD
10	Chen et al. [31]	Domain-Specific	Any	Any	CNN	Error-rate based	White box	Any	L_2	✓	✓	✗	✗	C-10, C-100, S, IN, L	CNNs	Standard	PGD, FGSM
11	Chen et al. [33]	Domain-Specific	Image Data	Any	CNNs	Error-rate based	White Box	Gradient based	L_p	✗	✗	✓	✗	C-10, C-100, TI, IN, L	CNNs	Standard, Adversarial	FGSM, BIM PGD, C&W
12	Cheng et al. [35]	Separation	Gaussian Mixture	Any	DNNs	Error-rate based	Any	Any	L_2	✗	✓	✓	✗	C-10, C-100, M	DNNs	Standard, Adversarial	FGSM, PGD, C&W
13	Cullina et al. [43]	Number of samples	Any	Binary Classif.	Any	Error-rate based	White box	Any	L_p	✓	✗	✓	✓	N/A	N/A	N/A	N/A
14	Dan et al. [44]	Number of samples, Dimensionality	Gaussian-mixture	Binary Classif.	Any	Error-rate based	White box	Any	$L_p, p \geq 1$	✓	✗	✓	✓	N/A	N/A	N/A	N/A
15	Daniely et al. [45]	Dimensionality	Any	Any	ReLU networks	Radius-based	White box	Any	L_2	✓	✗	✗	✓	N/A	N/A	N/A	N/A
16	De Palma et al. [47]	Dimensionality	Image Data	Binary Classif.	DNNs	Radius-based	White Box	Any	L_1	✓	✗	✗	✓	M, C-10	DNNs	Standard	Others
17	Deng and Karam [51]	Domain-Specific	Image Data	Any	CNNs	Error-rate based	White Box	GANs-based	L_∞	✓	✓	✓	✗	IN	CNNs	Standard	FTUAP
18	Deng and Karam [52]	Domain-Specific	Image Data	Any	CNNs	Error-rate based	White Box	GANs-based	L_∞	✓	✓	✓	✗	MC, G etc.	CNNs	Standard	FTUAP
19	Ding et al. [53]	Distribution	Any	Any	Any	Error-rate based	White box	Any	Any	✗	✗	✓	✓	C-10, M	DNNs	Adversarial	PGD
20	Diochnos et al. [54]	Dimensionality	Uniform distribu -tion on boolean hypercube	Any	Any	Radius based	White box	Any	L_0	✓	✗	✗	✓	N/A	N/A	N/A	N/A
21	Dohmatob [55]	Concentration	Any	Any	Any	Radius based	White box	Any	L_p, Geodesic	✓	✗	✗	✓	M	DNNs	Adversarial	Not mentioned
22	Dong et al. [56]	Label Quality	Any	Any	DNNs	Error-rate based	White Box	Gradient-based	L_2	✗	✓	✓	✓	C-10, C-100, TI	DNNs	Adversarial	Square RayS

Table 6. Categorization Table for Papers - Part 2

ID	Paper	Data Property	Problem Setup							Practicality							
			Target Distribution	Model		Robustness Setting				Applicability		Exp.	Type of Evidence				
				Learning Task	Classifier Type	Definition of Robustness	Attacker's Knwl.	Attacker's Tech.	Perturb Bound	Metr.	Tech.		Fml.	Empirical: Dataset	Empirical: Classifier Type	Empirical: Training Proc.	Empirical: Attacks
23	Fawzi et al. [58]	Distribution	Distribution generated by smooth generative model	Any	Any	Radius based	White box	Any	Any	✓	✗	✓	✓	C-10, S	DNNs	Adversarial	PGD
24	Garg et al. [62]	Separation	Any	Any	N/A	Error-rate based	White box	Any	Any	✓	✓	✗	✓	M	DNNs	Adversarial	PGD
25	Gilmer et al. [66]	Dimensionality	Concentric n-dimensional spheres	Binary Classif.	DNNs	Radius based	White box	Gradient based	L_2	✓	✗	✓	✓	M	DNNs	Standard	PGD
26	Gourdeau et al. [72]	Number of samples, Dimensionality	Boolean hypercube	Binary Classif.	Monotone Conjunction	Error-rate based	White box	Any	L_0	✓	✗	✗	✓	N/A	N/A	N/A	N/A
27	Gourdeau et al. [73]	Number of samples	Boolean hypercube	Binary Classif.	Monotone Conjunction	Error-rate based	White box	Any	L_0	✓	✗	✗	✓	N/A	N/A	N/A	N/A
28	Gowal et al. [74]	Number of samples	Any	Any	Any	Error-rate based	White box	Any	L_p	✓	✓	✗	✗	C-10, C-100, M, TI	DNNs	Adversarial	AutoAttack
29	Izmailov et al. [85]	Distribution	Any	Binary Classif.	Linear SVM, RBF SVM, NNs	Error-rate based	White box	Gradient based	L_∞	✓	✓	✗	✗	M	Linear SVM, RBF SVM, DNN	Standard	FGSM
30	Javanmard et al. [87]	Number of samples, Dimensionality	Any	Regres.	Linear Regres.	Error-rate based	Black box	Any	L_2	✓	✗	✗	✓	N/A	N/A	N/A	N/A
31	Kumar et al. [94]	Dimensionality	Any	Any	Any	Radius based	Any	Any	$L_p, p > 2$	✓	✗	✗	✓	C-10, IN	DNNs	DNNs	Gaussian noise
32	Lee et al. [98]	Distribution	Any	Any	DNNs	Error-rate based	White box, Black box	Gradient based, Non-gradient based	L_∞	✗	✓	✓	✓	C-10, C-100, S, TI	DNNs	Standard, Adversarial	PGD, FGSM, C&W, Transfer-based attacks
33	Li et al. [100]	Dimensionality	Well separated Balanced distribution	Binary Classif.	ReLU networks	Error-rate based	White Box	Any	L_2, L_∞	✓	✗	✗	✓	N/A	N/A	N/A	N/A
34	Ma et al. [109]	Domain-Specific	Image Data	Any	CNNs	Error-rate based	White Box	Gradient-based	L_2, L_∞	✗	✗	✓	✗	F, CX, D	CNNs	Standard	FGSM, BIM PGD, C&W
35	Mahloujifar et al. [112]	Concentration	Distributions in Lévy families	Any	Any	Error-rate based	White box	Any	L_0	✓	✗	✗	✓	N/A	N/A	N/A	N/A
36	Mahloujifar et al. [113]	Concentration	Any	Any	Any	Error-rate based	White box	Any	L_2, L_∞	✓	✗	✗	✗	C-10, M	DNNs	Adversarial	PGD
37	Mao et al. [115]	Label Quality	Any	Any	DNNs	Error-rate based	White box	Gradient based	L_∞	✓	✗	✓	✓	CS, TO	DNNs	Standard	PGD, FGSM MIM, Houdini
38	Mehrabi et al. [116]	Dimensionality	Gaussian mixture	Regres., Binary Classif.	Linear Regres., Linear classifiers	Error-rate based	White box	Any	$L_p, p \geq 1$	✓	✗	✗	✓	N/A	N/A	N/A	N/A
39	Montasser et al. [119]	Number of samples	Any	Binary Classif.	Any	Error-rate based	White Box	Any	Any	✓	✗	✗	✓	N/A	N/A	N/A	N/A
40	Mustafa et al. [123]	Separation	Any	Any	DNNs	Error-rate based	White box	Gradient based	L_p	✗	✓	✗	✓	C-10, C-100, M, FM, S	CNNs	Standard, Adversarial	PGD, FGSM, BIM, MIM, C&W

Table 7. Categorization Table for Papers - Part 3

ID	Paper	Data Property	Problem Setup							Practicality							
			Target Distribution	Model		Robustness Setting				Applicability		Exp.	Type of Evidence				
				Learning Task	Classifier Type	Definition of Robustness	Attacker's Knwl.	Attacker's Tech.	Perturb Bound	Metr.	Tech.		Fml.	Empirical			
														Dataset	Classifier Type	Training Proc.	Attacks
41	Mygdalis et al. [124]	Separation	Any	Any	DNNs	Error-rate based	Any	Gradient-based	L_2	✗	✓	✓	✗	C-10, C-100, S	DNNs	Standard, Adversarial	FGSM, BIM, MIM
42	Najafi et al. [125]	Number of samples	Any	Any	Any	Error-rate based	White box	Gradient based	L_2, L_∞	✓	✓	✗	✓	C-10, M, S	DNNs	Adversarial	PGD
43	Naseer et al. [126]	Density	Any	Any	DNNs	Radius-based	White Box	Gradient-based	Any	✓	✓	✓	✗	M	DNNs	Standard	FGSM
44	Oritz-Jimenez et al. [130]	Domain-Specific	Any	Any	CNNs	Radius based	White box	Gradient based	L_2	✓	✗	✓	✗	C-10, M, IN	CNNs	Standard, Adversarial	PGD
45	Pang et al. [131]	Distribution, Separation	Any	Any	DNNs	Radius based	White box	Gradient based	L_2	✓	✓	✓	✓	C-10, M, IN	DNNs	Standard	FGSM, BIM ILCM, JSMA
46	Pang et al. [132]	Density, Separation	Any	Any	DNNs	Error-rate based	White box, Black box	Gradient based, Non-gradient based	L_2, L_∞	✓	✓	✗	✓	C-10, C-100, M	DNNs	Standard, Adversarial	PGD, FGSM, Transfer-based attacks
47	Prescott et al. [137]	Concentration	Gaussian (theory), Any (application)	Any	Any	Error-rate based	White box	Any	$L_p, p \geq 2$	✓	✗	✗	✓	C-10, M, FM, S	N/A	N/A	N/A
48	Pydi & Jog [138]	Separation	Any	Binary Classif.	Any	Error-rate based	White box	Gradient based	L_2, L_∞	✓	✗	✗	✓	C-10, M, FM, S	DNNs	Adversarial	N/A
49	Pydi & Jog [139]	Separation	Any	Binary Classif.	Any	Error-rate based	White box	Gradient based	L_2, L_∞	✓	✗	✗	✓	N/A	N/A	N/A	N/A
50	Qaraei et al. [140]	Number of samples	Discrete language data.	Binary Classif.	DNNs	Error-rate based	White Box	Gradient-based	L_0	✓	✓	✗	✗	W, AC	DNNs	Standard	Others
51	Rajput et al. [141]	Dimensionality	Any	Any	Linear classifiers, non-linear classifiers	Radius based	Any	Any	L_2	✓	✗	✗	✓	N/A	N/A	N/A	N/A
52	Richardson & Weiss [144]	Distribution	Gaussian-mixture	Binary Classif.	Bayes optimal, SVM, CNNs	Radius based	White box	Any	L_2	✗	✗	✗	✗	M	Linear SVM, Kernel SVM, CNNs	Standard, Adversarial	C&W
53	Sanyal et al. [147]	Label Quality	Any	Binary Classif.	Any	Error-rate based	White box	Any	Any	✗	✗	✓	✓	C-10, M	DNNs	Standard, Adversarial	PGD
54	Schmidt et al. [148]	Number of samples, Distribution	Gaussian-mixture, Bernoulli-mixture	Binary Classif.	Any	Error-rate based	White box	Any	L_∞	✓	✗	✗	✓	C-10, M, S	DNNs	Adversarial	PGD
55	Shafahi et al. [152]	Dimesionality, Density	N-dimensional hypercube	Any	Any	Radius-based	White box	Any	L_p, Geodesic	✗	✗	✓	✓	C-10, M	CNN	Adversarial	PGD
56	Shamir et al. [154]	Label Quality	Any	Binary Classif.	ReLU networks	Radius-based	White Box	Any	L_0	✓	✗	✗	✓	M	DNNs	Standard	Others
57	Simon-Gabriel et al. [159]	Dimensionality	Any	Any	DNNs	Error-rate based	White box	Any	Any	✓	✗	✗	✓	C-10	DNNs	Adversarial	PGD
58	Song et al. [162]	Density	Any	Any	Any	Error-rate based	Any	Any	Any	✓	✓	✗	✗	C-10, M, FM	CNNs	Adversarial	FGSM, BIM C&W, DeepFool
59	Sun et al. [164]	Domain-Specific	Image Data	Any	CNNs	Radius-based	Any	Corruption	L_2	✗	✓	✓	✗	C-10, C-100, IN	CNNs	Adversarial	Corruption

Table 8. Categorization Table for Papers - Part 4

ID	Paper	Data Property	Problem Setup							Practicality							
			Target Distribution	Model		Robustness Setting				Applicability		Exp.	Type of Evidence				
				Learning Task	Classifier Type	Definition of Robustness	Attacker's Knwl.	Attacker's Tech.	Perturb Bound	Metr.	Tech.		Fml.	Empirical			
														Dataset	Classifier Type	Training Proc.	Attacks
60	Uesato et al. [175]	Number of samples	Gaussian-mixture (theory), Any (application)	Binary Classif.	Any	Error-rate based	White box	Any	L_∞	✓	✓	✗	✓	C-10, S	DNNs	Adversarial	PGD, FGSM
61	Wan et al. [177]	Distribution, Separation	Any	Any	Any	Error-rate based	White box	Any	L_∞	✗	✓	✗	✓	C-10, M, IN	DNNs	Standard	FGSM, BIM, ILCM, C&W
62	Wang et al. [179]	Domain-Specific	Any	Any	CNNs	Error-rate based	White box	Gradient based	L_2	✓	✗	✓	✗	C-10	CNNs	Standard, Adversarial	PGD, FGSM
63	Wang et al. [184]	Dimensionality, Separation	Any	Binary Classif.	kNN	Radius based	White box	Any	L_2	✓	✓	✗	✗	M, M1V7, HM	k-NN	Adversarial	Direct attack, Transfer-based attacks
64	Wang et al. [182]	Number of samples	Gaussian-mixture (theory), Any (application)	Binary Classif. (theory), Any (appl.)	DNNs	Error-rate based	White Box	Gradient-based	L_p	✓	✓	✓	✓	C-10, S	DNNs	Standard, Adversarial	PGD
65	Weber et al. [185]	Dimensionality	Hierarchial data	Any	Any	Error-rate based	White box	Any	Check	✓	✓	✓	✗	IN	Hyperbolic perceptron	Adversarial	Gradient based
66	Wu et al. [186]	Number of samples	Any	Any	DNNs	Error-rate based	White box	Gradient based	L_∞	✓	✓	✗	✗	C-10, C-100	DNNs	Standard, Adversarial	PGD C&W Transfer-based attacks
67	Xing et al. [188]	Number of samples	Sub-Gaussian (theory), Any (application)	Binary Classif.	Linear classifers (theory), Any (application)	Error-rate based	White Box	Any	L_2, L_∞	✓	✓	✓	✓	C-10, C-100, S	DNNs	Standard, Adversarial	PGD
68	Xu & Liu [190]	Number of samples	Any	Multi-Class Classif.	Any	Error-rate based	White Box	Any	$L_p, p \geq 1$	✓	✗	✗	✓	N/A	N/A	N/A	N/A
69	Yang et al. [195]	Separation	Any	Any	DNNs	Error-rate based	White box	Gradient based	L_2	✓	✓	✗	✓	C-10, C-100, M, TI	DNNs	Adversarial	PGD
70	Yang et al. [197]	Separation	Any	Binary Classif.	Non-parametric classifiers	Radius based	White box	Distance based	L_2	✓	✓	✗	✓	HM	Histogram, 1-NN	Standard	Distance based
71	Yin et al. [199]	Domain-Specific	Any	Any	Any	Error-rate based	White box	Any	L_2	✓	✗	✗	✗	C-10, IN	DNNs	Adversarial	Corruptions, PGD
72	Yin et al. [198]	Dimensionality	Any	Any	Linear classifiers, DNNs	Error-rate based	White box	Any	L_∞	✓	✗	✗	✓	M	Linear classifiers, ReLU networks	Adversarial	PGD
73	Zhang et al. [204]	Domain-Specific	Image Data	Any	CNNs	Error-rate based	White Box	Gradient-based	L_1, L_∞	✗	✗	✓	✗	IN	CNNs	Standard	UAP
74	Zhang et al. [206]	Density	Any	Any	Any	Error-rate based	White box	Any	L_2, L_∞	✓	✗	✗	✗	C-10, M, FM	DNNs	Adversarial	C&W
75	Zhang & Evans [210]	Concentration	Gaussian (theory), Any (application)	Any	Any	Error-rate based	White box	Any	L_2, L_∞	✓	✗	✗	✓	C-10	DNNs	Standard, Adversarial	AutoAttack
76	Zhang et al. [205]	Label Quality	Any	Any	DNNs	Error-rate based	White Box	Gradient-based	L_∞	✗	✓	✓	✗	C-100, IN	DNNs	Standard	FGSM, PGD
77	Zhu et al. [213]	Density	Any	Any	DNNs	Error-rate based	Any	Any	L_∞	✓	✗	✗	✓	IN	CNNs	Adversarial	PGD Transfer-based attacks

Tables 5-8 include the detailed categorization of papers collected in this survey. We used the following abbreviation to denote the datasets discussed in the papers: $\boxed{\text{M}}$ for MNIST, $\boxed{\text{FM}}$ for Fashion-MNIST [187], $\boxed{\text{S}}$ for SVHN, $\boxed{\text{C-10}}$ for CIFAR-10, $\boxed{\text{C-100}}$ for CIFAR-100, $\boxed{\text{IN}}$ for ImageNet [93], $\boxed{\text{TI}}$ for Tiny Images Dataset, $\boxed{\text{CA}}$ for CelebA [106], $\boxed{\text{HM}}$ for Halfmoon, $\boxed{\text{M1V7}}$ for MNIST 1v7, $\boxed{\text{A}}$ for abalone [75], $\boxed{\text{L}}$ for LSUN [201], $\boxed{\text{CS}}$ for Cityscapes [40], $\boxed{\text{TO}}$ for Taskonomy [203], $\boxed{\text{W}}$ for Wikipedia-31K [18], $\boxed{\text{AC}}$ for AmazonCat-13K [18] $\boxed{\text{MC}}$ for MINC [15], $\boxed{\text{G}}$ for GTOS [193], $\boxed{\text{F}}$ for Fundoscopy [91], $\boxed{\text{CX}}$ for Chest X-Ray [183], $\boxed{\text{D}}$ for Dermoscopy [39].

REFERENCES

[1] [n. d.]. ACM Computing Surveys Journal. https://dl.acm.org/journal/csur.
[2] [n. d.]. Advances in Neural Information Processing Systems (NeurIPS). https://proceedings.neurips.cc.
[3] [n. d.]. CORE ranking (Conference Portal). http://portal.core.edu.au/conf-ranks/.
[4] [n. d.]. Journal Citation Reports (JCR). https://jcr.clarivate.com/jcr/home.
[5] [n. d.]. Proceedings of Machine Learning Research. https://proceedings.mlr.press.
[6] [n. d.]. Semantic Scholar Academic APIs. https://www.semanticscholar.org/product/api.
[7] Albert Ahumada and Heidi Peterson. 1992. Luminance-model-based DCT quantization for color image compression. *Human Vision, Visual Process Display III* 1666 (02 1992).
[8] Naveed Akhtar and Ajmal Mian. 2018. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access* 6 (2018), 14410–14430.
[9] Naveed Akhtar, Ajmal S. Mian, Navid Kardan, and Mubarak Shah. 2021. Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey. *IEEE Access* 9 (2021), 155161–155196.
[10] Laurent Amsaleg, James Bailey, Amélie Barbe, Sarah M. Erfani, Teddy Furon, Michael E. Houle, Miloš Radovanović, and Xuan Vinh Nguyen. 2021. High Intrinsic Dimensionality Facilitates Adversarial Attack: Theoretical Evidence. *IEEE Transactions on Information Forensics and Security* 16 (2021), 854–865.
[11] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. 2019. Intrinsic Dimension of Data Representations in Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*. 6111–6122.
[12] Pranjal Awasthi, Himanshu Jain, Ankit Singh Rawat, and Aravindan Vijayaraghavan. 2020. Adversarial Robustness via Robust Low Rank Representations. In *Advances in Neural Information Processing Systems (NeurIPS)*. 11391–11403.
[13] Ms. Aayushi Bansal, Dr. Rewa Sharma, and Dr. Mamta Kathuria. 2021. A Systematic Review on Data Scarcity Problem in Deep Learning: Solution and Applications. *Comput. Surveys* 54, 208 (2021), 1–29.
[14] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A Survey on Data Augmentation for Text Classification. *Comput. Surveys* (2022).
[15] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. 2015. Material Recognition in the Wild with the Materials in Context Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[16] Shai Ben-David, Nicolò Cesa-Bianchi, David Haussler, and Philip.M. Long. 1995. Characterizations of Learnability for Classes of (0, ..., n)-Valued Functions. *J. Comput. System Sci.* 50, 1 (1995), 74–86.
[17] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. 2019. Lower Bounds on Adversarial Robustness from Optimal Transport. In *Advances in Neural Information Processing Systems (NeurIPS)*.
[18] Kush Bhatia, Kunal Dahiya, Himanshu Jain, Purushottam Kar, Anshul Mittal, Yasgiteja Prabhu, and Manik Varma. 2016. The extreme classification repository: Multi-label datasets and code. http://manikvarma.org/downloads/XC/XMLRepository.html
[19] Robi Bhattacharjee and Kamalika Chaudhuri. 2020. When Are Non-Parametric Methods Robust?. In *International Conference on Machine Learning (ICML)*. 832–841.
[20] Robi Bhattacharjee, Somesh Jha, and Kamalika Chaudhuri. 2021. Sample Complexity of Robust Linear Classification on Separated Data. In *Conference on Learning Theory (COLT)*. 884–893.
[21] Battista Biggio and Fabio Roli. 2018. Wild Patterns: Ten Years after The Rise of Adversarial Machine Learning. *Pattern Recognition* 84 (2018), 317–331.
[22] Avrim Blum, Travis Dick, Naren Manoj, and Hongyang Zhang. 2020. Random Smoothing Might Be Unable to Certify L_∞ Robustness for High-Dimensional Images. 21, 211 (2020), 8726–8746.
[23] Giuseppe Bonaccorso. 2017. *Machine Learning Algorithms: A Reference Guide to Popular Algorithms for Data Science and Machine Learning*. Packt Publishing.
[24] Christer Borell. 1975. The Brunn-Minkowski Inequality in Gauss Space. *Inventiones mathematicae* 30 (1975), 207–216.
[25] Sebastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. 2019. Adversarial Examples from Computational Constraints. In *International Conference on Machine Learning (ICML)*. 831–840.
[26] Anh Bui, Trung Le, He Zhao, Paul Montague, Oliver deVel, Tamas Abraham, and Dinh Phung. 2020. Improving Adversarial Robustness by Enforcing Local and Global Compactness. In *European Conference on Computer Vision (ECCV)*. 209–223.
[27] Ginevra Carbone, Matthew Wicker, Luca Laurenti, Andrea Patane, Luca Bortolussi, and Guido Sanguinetti. 2020. Robustness of Bayesian Neural Networks to Gradient-Based Attacks. In *International Conference on Neural Information Processing Systems (NeurIPS)*. 15602–15613.
[28] Nicholas Carlini and David A. Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *Symposium on Security and Privacy (SP)*. 39–57.
[29] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. 2019. Unlabeled Data Improves Adversarial Robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[30] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. Adversarial Attacks and Defences: A Survey. *ArXiv* (2018).
[31] Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. 2021. Amplitude-Phase Recombination: Rethinking Robustness of Convolutional Neural Networks in Frequency Domain. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 458–467.
[32] Pin-Yu Chen and Cho-Jui Hsieh. 2023. *Adversarial Robustness for Machine Learning*. Academic Press.
[33] Yiting Chen, Qibing Ren, and Junchi Yan. 2022. Rethinking and Improving Robustness of Convolutional Neural Networks: a Shapley Value-based Approach in Frequency Domain. In *Advances in Neural Information Processing Systems*.
[34] Wuxinlin Cheng, Chenhui Deng, Zhiqiang Zhao, Yaohui Cai, Zhiru Zhang, and Zhuo Feng. 2021. SPADE: A Spectral Method for Black-Box Adversarial Robustness Evaluation. In *International Conference on Machine Learning (ICML)*. 1814–1824.
[35] Zhen Cheng, Fei Zhu, Xu-Yao Zhang, and Cheng-Lin Liu. 2023. Adversarial Training with Distribution Normalization and Margin Balance. *Pattern Recognition* (2023).
[36] Fan. R. K. Chung. 1997. *Spectral Graph Theory*. American Mathemetical Society.
[37] Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A. Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. 2023. Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning. *Comput. Surveys* (2023).
[38] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified Adversarial Robustness via Randomized Smoothing. In *International Conference on Machine Learning (ICML)*. 1310–1320.
[39] The International Skin Imaging Collaboration. 2019. https://www.isic-archive.com.
[40] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3213–3223.
[41] J.S. Cramer. 2002. *The Origins of Logistic Regression*. Technical Report 2002-119/4. Tinbergen Institute.
[42] Francesco Croce and Matthias Hein. 2020. Minimally Distorted Adversarial Examples with a Fast Adaptive Boundary Attack. In *International Conference on Machine Learning (ICML)*. 2196–2205.
[43] Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. 2018. PAC-Learning in the Presence of Evasion Adversaries. In *Advances in Neural Information Processing Systems (NeurIPS)*. 228–239.
[44] Chen Dan, Yuting Wei, and Pradeep Ravikumar. 2020. Sharp Statistical Guarantees for Adversarially Robust Gaussian Classification. In *International Conference on Machine Learning (ICML)*. 2345–2355.
[45] Amit Daniely and Hadas Schacham. 2020. Most ReLU Networks Suffer from L2 Adversarial Perturbations. In *Advances in Neural Information Processing Systems (NeurIPS)*. 6629–6636.
[46] David.H.Haussler, Nick. Littlestone, and Manfred.K. Warmuth. 1994. Predicting 0, 1-Functions on Randomly Drawn Points. *Information and Computation* 115, 2 (1994), 248–292.
[47] Giacomo De Palma, Bobak Kiani, and Seth Lloyd. 2021. Adversarial Robustness Guarantees for Random Deep Neural Networks. In *International Conference on Machine Learning (ICML)*. 2522–2534.
[48] Akshay Degwekar, Preetum Nakkiran, and Vinod Vaikuntanathan. 2019. Computational Limitations in Robust Classification and Win-Win Results. In *Conference on Learning Theory (COLT)*. 994–1028.
[49] Luca Demetrio, Scott E. Coull, Battista Biggio, Giovanni Lagorio, Alessandro Armando, and Fabio Roli. 2021. Adversarial EXEmples: A Survey and Experimental Evaluation of Practical Attacks on Machine Learning for Windows Malware Detection. *ACM Transactions on Privacy and Security* 24, 4 (2021), 1–31.
[50] Ambra Demontis, Marco Melis, Battista Biggio, Davide Maiorca, Dan Arp, Konrad Rieck, Igino Corona, Giorgio Giacinto, and Fabio Roli. 2019. Yes, Machine Learning Can Be More Secure! A Case Study on Android Malware Detection. *IEEE Transactions on Dependable and Secure Computing (TDSC)* 16, 4 (2019), 711–724.
[51] Yingpeng Deng and Lina J. Karam. 2020. Frequency-Tuned Universal Adversarial Perturbations. In *Computer Vision – ECCV 2020 Workshops*. 494–510.
[52] Yingpeng Deng and Lina J. Karam. 2022. Frequency-Tuned Universal Adversarial Attacks on Texture Recognition. *IEEE Transactions on Image Processing (TIP)* 31 (2022), 5856–5868.
[53] Gavin Weiguang Ding, Kry Yik Chau Lui, Xiaomeng Jin, Luyu Wang, and Ruitong Huang. 2019. On the Sensitivity of Adversarial Robustness to Input Data Distributions. In *International Conference on Learning Representations (ICLR)*.
[54] Dimitrios I. Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. 2018. Adversarial Risk and Robustness: General Definitions and Implications for the Uniform Distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*. 10380–10389.
[55] Elvis Dohmatob. 2019. Generalized No Free Lunch Theorem for Adversarial Robustness. In *International Conference on Machine Learning (ICML)*. 1646–1654.

[56] Chengyu Dong, Liyuan Liu, and Jingbo Shang. 2022. Label Noise in Adversarial Training: A Novel Perspective to Study Robust Overfitting. In *Advances in Neural Information Processing Systems (NeurIPS)*.
[57] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust Physical-World Attacks on Deep Learning Visual Classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1625–1634.
[58] Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. 2018. Adversarial Vulnerability for Any Classifier. In *Advances in Neural Information Processing Systems (NeurIPS)*. 1186–1195.
[59] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2015. Fundamental Limits on Adversarial Robustness. In *ICML Workshop on Deep Learning*.
[60] Sid Ahmed Fezza, Yassine Bakhti, Wassim Hamidouche, and Olivier Déforges. 2019. Perceptual Evaluation of Adversarial Attacks for CNN-based Image Classification. In *International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6.
[61] Benoit Frenay and Michel Verleysen. 2014. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* 25, 5 (2014), 845–869.
[62] Shivam Garg, Vatsal Sharan, Brian Hu Zhang, and Gregory Valiant. 2018. A Spectral View of Adversarially Robust Features. In *Advances in Neural Information Processing Systems (NeurIPS)*. 10159–10169.
[63] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of Neural Networks is Fragile. In *AAAI Conference on Artificial Intelligence (AAAI)*. 3681–3688.
[64] Partha Ghosh, Arpan Losalka, and Micheal. J. Black. 2019. Resisting Adversarial Attacks using Gaussian Mixture Variational Autoencoders. In *AAAI Conference on Artificial Intelligence (AAAI)*. 541–548.
[65] Justin Gilmer, Ryan P. Adams, Ian J. Goodfellow, David G. Andersen, and George E. Dahl. 2018. Motivating the Rules of the Game for Adversarial Example Research. *ArXiv* (2018).
[66] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian J. Goodfellow. 2018. Adversarial Spheres. In *International Conference on Learning Representations (ICLR)*.
[67] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. 2022. Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
[68] Ian Goodfellow, Nicolas Papernot, Sandy Huang, Rocky Duan, Pieter Abbeel, and Jack Clark. 2017. Attacking Machine Learning with Adversarial Examples. https://openai.com/blog/adversarial-example-research/.
[69] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR)*.
[70] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR)*.
[71] Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J. Cree. 2021. Regularization of neural networks by enforcing Lipschitz continuity. *Machine Learning* 110 (2021), 393–416.
[72] Pascale Gourdeau, Varun Kanade, Marta Kwiatkowska, and James Worrell. 2021. On the Hardness of Robust Classification. *Journal of Machine Learning Research* 22, 273 (2021), 12521–12549.
[73] Pascale Gourdeau, Varun Kanade, Marta Kwiatkowska, and James Worrell. 2022. Sample Complexity Bounds for Robustly Learning Decision Lists against Evasion Attacks. In *International Joint Conference on Artificial Intelligence, (IJCAI)*. 3022–3028.
[74] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy Mann. 2021. Improving Robustness using Generated Data. In *Advances in Neural Information Processing Systems (NeurIPS)*. 4218–4233.
[75] UCI Machine Learning Group. 1995. Abalone Dataset. https://archive.ics.uci.edu/ml/datasets/abalone.
[76] Wenbo Guo, Dongliang Mu, Jun Xu, Purui Su, Gang Wang, and Xinyu Xing. 2018. LEMNA: Explaining Deep Learning Based Security Applications. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 364–379.
[77] Yiwen Guo, Long Chen, Yurong Chen, and Changshui Zhang. 2021. On Connections between Regularizations for Improving DNN Robustness. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 12 (2021), 4469–4476.
[78] Sicong Han, Chenhao Lin, Chao Shen, Qian Wang, and Xiaohong Guan. 2023. Interpreting Adversarial Examples in Deep Learning: A Review. *Comput. Surveys* (2023).
[79] Haibo He and Edwardo A. Garcia. 2009. Learning From Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 21, 9 (2009), 1263–1284.
[80] Xinlei He and Yang Zhang. 2021. Quantifying and Mitigating Privacy Risks of Contrastive Learning. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 845–863.
[81] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022. Membership Inference Attacks on Machine Learning: A Survey. *Comput. Surveys* 54, 235 (2022), 1–37.

[82] Huawei Huang, Wei Kong, Sicong Zhou, Zibin Zheng, and Song Guo. 2021. A Survey of State-of-the-Art on Blockchains: Theories, Modelings, and Tools. *Comput. Surveys* 54, 44 (2021), 1–42.
[83] Ingo. Höntsch and Lina.J. Karam. 2002. Adaptive Image Coding with Perceptual Distortion Control. *IEEE Transactions on Image Processing* 11, 3 (2002), 213–222.
[84] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Mądry. 2019. Adversarial Examples Are Not Bugs, They Are Features. In *Advances in Neural Information Processing Systems (NeurIPS)*. 125–136.
[85] Rauf Izmailov, Shridatt Sugrim, Ritu Chadha, Patrick McDaniel, and Ananthram Swami. 2018. Enablers of Adversarial Attacks in Machine Learning. In *IEEE Military Communications Conference (MILCOM)*. 425–430.
[86] Joern-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. 2019. Excessive Invariance Causes Adversarial Vulnerability. In *International Conference on Learning Representations (ICLR)*.
[87] Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. 2020. Precise Tradeoffs in Adversarial Training for Linear Regression. In *International Conference on Learning Theory (COLT)*. 2034–2078.
[88] Jongheon. Jeong and Jinwoo. Shin. 2020. Consistency Regularization for Certified Robustness of Smoothed Classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*. 10558–10570.
[89] Xi Wu Jiefeng Chen. 2019. Robust Attribution Regularization. https://www.altacognita.com/robust-attribution/.
[90] Ian. T. Jolliffe. 2002. *Principal Component Analysis*. Springer.
[91] Kaggle. 2015. Kaggle Diabetic Retinopathy Detection Challenge. https://www.kaggle.com/c/diabetic-retinopathy-detection.
[92] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. 2009. CIFAR-10 and CIFAR-100 Datasets. https://www.cs.toronto.edu/~kriz/cifar.html.
[93] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* 60 (2012), 84 – 90.
[94] Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. 2020. Curse of Dimensionality on Randomized Smoothing for Certifiable Robustness. In *International Conference on Machine Learning (ICML)*. 5458–5467.
[95] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2017. Adversarial Machine Learning at Scale. In *International Conference on Learning Representations (ICLR)*.
[96] Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. 1998. The MNIST Database of Handwritten Digits. http://yann.lecun.com/exdb/mnist/.
[97] Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. 2019. Tight Certificates of Adversarial Robustness for Randomly Smoothed Classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*.
[98] Saehyung Lee, Hyungyu Lee, and Sungroh Yoon. 2020. Adversarial Vertex Mixup: Toward Better Adversarially Robust Generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 269–278.
[99] Paul Lévy. 1951. *Problèmes concrets d'analyse fonctionnelle*. Gauthier-Villers.
[100] Binghui Li, Jikai Jin, Han Zhong, John E. Hopcroft, and Liwei Wang. 2022. Why Robust Generalization in Deep Learning is Difficult: Perspective of Expressive Power. In *Advances in Neural Information Processing Systems (NeurIPS)*.
[101] Deqiang Li, Qianmu Li, Yanfang (Fanny) Ye, and Shouhuai Xu. 2021. Arms Race in Adversarial Malware Detection: A Survey. *Comput. Surveys* 55, 1 (2021), 1–35.
[102] Xiang Ling, Shouling Ji, Jiaxu Zou, Jiannan Wang, Chunming Wu, Bo Li, and Ting Wang. 2019. DEEPSEC: A Uniform Platform for Security Analysis of Deep Learning Model. In *IEEE Symposium on Security and Privacy (SP)*. 673–690.
[103] Jinxin Liu, Michele Nogueira, Johan Fernandes, and Burak Kantarci. 2022. Adversarial Machine Learning: A Multilayer Review of the State-of-the-Art and Challenges for Wireless and Mobile Systems. *IEEE Communications Surveys & Tutorials* 24, 1 (2022), 123–159.
[104] Xuanqing Liu, Si Si, Xiaojin Zhu, Yang Li, and Cho-Jui Hsieh. 2019. A Unified Framework for Data Poisoning Attack to Graph-Based Semi-Supervised Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*. 9780–9790.
[105] Yue Liu, Chakkrit Tantithamthavorn, Li Li, and Yepang Liu. 2022. Deep Learning for Android Malware Defenses: A Systematic Literature Review. *Comput. Surveys* (2022).
[106] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*. 3730–3738.
[107] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. 2013. An Insight Into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics. *Information Science (inf.Sci)* 250 (2013), 113–141.
[108] Ana C. Lorena, Luís P. F. Garcia, Jens Lehmann, Marcilio C. P. Souto, and Tin Kam Ho. 2020. How Complex Is Your Classification Problem? A Survey on Measuring Classification Complexity. *Comput. Surveys* 52, 107 (2020), 1–34.
[109] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. 2021. Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems. *Pattern Recognition* 110 (2021), 107332.

[110] Gabriel Resende Machado, Eugênio Silva, and Ronaldo Ribeiro Goldschmidt. 2021. Adversarial Machine Learning in Image Classification: A Survey Toward the Defender's Perspective. *Comput. Surveys* 55, 8 (2021), 1–38.

[111] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations (ICLR)*.

[112] Saeed Mahloujifar, Dimitrios I. Diochnos, and Mohammad Mahmoody. 2019. The Curse of Concentration in Robust Learning: Evasion and Poisoning Attacks from Concentration of Measure. In *AAAI Conference on Artificial Intelligence (AAAI)*. 4536–4543.

[113] Saeed Mahloujifar, Xiao Zhang, Mohammad Mahmoody, and David Evans. 2019. Empirically Measuring Concentration: Fundamental Limits on Intrinsic Robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*. 5209–5220.

[114] Davide Maiorca, Battista Biggio, and Giorgio Giacinto. 2019. Towards Adversarial Malware Detection: Lessons Learned from PDF-based Attacks. *Comput. Surveys* 52, 78 (2019), 1–36.

[115] Chengzhi Mao, Amogh Gupta, Vikram Nitin, Baishakhi Ray, Shuran Song, Junfeng Yang, and Carl Vondrick. 2020. Multitask Learning Strengthens Adversarial Robustness. In *European Conference on Computer Vision (ECCV)*. 158–174.

[116] Mohammad Mehrabi, Adel Javanmard, Ryan A. Rossi, Anup Rao, and Tung Mai. 2021. Fundamental Tradeoffs in Distributionally Adversarial Training. In *International Conference on Machine Learning (ICML)*. 7544–7554.

[117] Ali H. Mezher, Yingpeng Deng, and Lina J. Karam. 2022. Visual Quality Assessment of Adversarially Attacked Images. In *European Workshop on Visual Information Processing (EUVIP)*. 1–5.

[118] Eric Mintun, Alexander Kirillov, and Saining Xie. 2021. On Interaction Between Augmentations and Corruptions in Natural Corruption Robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*. 3571–3583.

[119] Omar Montasser, Steve Hanneke, and Nathan Srebro. 2022. Adversarially Robust Learning: A Generic Minimax Optimal Learner and Characterization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[120] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, and Stefano Soatto. 2018. Robustness of Classifiers to Universal Perturbations: A Geometric Perspective. In *International Conference on Learning Representations (ICLR)*.

[121] Jose Garcia Moreno-Torres, Troy Raeder, Rocío Alaíz-Rodríguez, N. Chawla, and Francisco Herrera. 2012. A Unifying View on Dataset Shift in Classification. *Pattern Recognition* 45 (2012), 521–530.

[122] Aamir Mustafa, Salman Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. 2019. Adversarial Defense by Restricting the Hidden Space of Deep Neural Networks. In *IEEE International Conference on Computer Vision (ICCV)*. 3384–3393.

[123] Aamir Mustafa, Salman H Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. 2020. Deeply supervised discriminative learning for adversarial defense. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 9 (2020), 3154–3166.

[124] Vasileios Mygdalis and Ioannis Pitas. 2022. Hyperspherical Class Prototypes for Adversarial Robustness. *Pattern Recognition* (2022).

[125] Amir Najafi, Shin ichi Maeda, Masanori Koyama, and Takeru Miyato. 2019. Robustness to Adversarial Perturbations in Learning from Incomplete Data. In *Advances in Neural Information Processing Systems (NeurIPS)*. 5541–5551.

[126] Mahum Naseer, Bharath Srinivas Prabakaran, Osman Hasan, and Muhammad Shafique. 2023. UnbiasedNets: a dataset diversification framework for robustness bias alleviation in neural networks. *Machine Learning* (2023), 1–28.

[127] Balas K. Natarajan. 2004. On learning sets and functions. *Machine Learning* 4 (2004), 67–97.

[128] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

[129] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. 2018. Adversarial Robustness Toolbox v1.2.0. *ArXiv* (2018).

[130] Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. 2020. Hold Me Tight! Influence of Discriminative Features on Deep Network Boundaries. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2935–2946.

[131] Tianyu Pang, Chao Du, and Jun Zhu. 2018. Max-Mahalanobis Linear Discriminant Analysis Networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 4013–4022.

[132] Tianyu Pang, Kun Xu, Yinpeng Dong, Chao Du, Ning Chen, and Jun Zhu. 2020. Rethinking Softmax Cross-Entropy Loss for Adversarial Robustness. In *International Conference on Learning Representations (ICLR)*.

[133] Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. 2019. Improving Adversarial Robustness via Promoting Ensemble Diversity. In *International Conference on Machine Learning (ICML)*. 4970–4979.

[134] Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Jun Zhu, and Hang Su. 2020. Boosting Adversarial Training with Hypersphere Embedding. In *Advances in Neural Information Processing Systems (NeurIPS)*. 7779–7792.

[135] Fabio Pierazzi, Feargus Pendlebury, Jacopo Cortellazzi, and Lorenzo Cavallaro. 2020. Intriguing Properties of Adversarial ML Attacks in the Problem Space. In *Symposium on Security and Privacy (SP)*.
[136] Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. 2021. The Intrinsic Dimension of Images and Its Impact on Learning. In *International Conference on Learning Representations (ICLR)*.
[137] Jack Prescott, Xiao Zhang, and David Evans. 2021. Improved Estimation of Concentration Under Lp-Norm Distance Metrics Using Half Spaces. In *International Conference on Learning Representations (ICLR)*.
[138] Muni Sreenivas Pydi and Varun Jog. 2020. Adversarial Risk via Optimal Transport and Optimal Couplings. In *International Conference on Machine Learning (ICML)*. 7814–7823.
[139] Muni Sreenivas Pydi and Varun Jog. 2021. The Many Faces of Adversarial Risk. In *Advances in Neural Information Processing Systems (NeurIPS)*. 10000–10012.
[140] Mohammadreza Qaraei and Rohit Babbar. 2022. Adversarial examples for extreme multilabel text classification. *Machine Learning* (2022), 1–25.
[141] Shashank Rajput, Zhili Feng, Zachary Charles, Po-Ling Loh, and Dimitris Papailiopoulos. 2019. Does Data Augmentation Lead to Positive Margin?. In *International Conference on Machine Learning (ICML)*. 5321–5330.
[142] S.J. Raudys and A.K. Jain. 1991. Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13, 3 (1991), 252–264.
[143] Mohammad Rezaeirad, Brown Farinholt, Hitesh Dharmdasani, Paul Pearce, Kirill Levchenko, and Damon McCoy. 2018. Schrödinger's RAT: Profiling the Stakeholders in the Remote Access Trojan Ecosystem. In *USENIX Security Symposium*. 1043–1060.
[144] Eitan Richardson and Yair Weiss. 2021. A Bayes-Optimal View on Adversarial Examples. *Journal of Machine Learning Research (JMLR)* 22, 221 (2021), 10076–10103.
[145] Ishai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach. 2021. Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain. *Comput. Surveys* 54, 5 (2021), 1–36.
[146] Miriam Santos, Pedro Henriques Abreu, Nathalie Japkowicz, Alberto Fernández, Carlos Soares, Szymon Wilk, and Joao Santos. 2022. On the Joint-Effect of Class Imbalance and Overlap: A Critical Review. *Artificial Intelligence Review* (2022), 1–69.
[147] Amartya Sanyal, Puneet K. Dokania, Varun Kanade, and Philip Torr. 2021. How Benign is Benign Overfitting?. In *International Conference on Learning Representations (ICLR)*.
[148] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. 2018. Adversarially Robust Generalization Requires More Data. In *Advances in Neural Information Processing Systems (NeurIPS)*. 5019–5031.
[149] H. Scudder. 1965. Probability of Error of Some Adaptive Pattern-Recognition Machines. *IEEE Transactions on Information Theory* 11, 3 (1965), 363–371.
[150] Alex Serban, Erik Poll, and Joost Visser. 2020. Adversarial Examples on Object Recognition: A Comprehensive Survey. *Comput. Surveys* 53, 3 (2020), 1–38.
[151] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. 2018. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*. 6106–6116.
[152] Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. 2019. Are Adversarial Examples Inevitable?. In *International Conference on Learning Representations (ICLR)*.
[153] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.
[154] Adi Shamir, Itay Safran, Eyal Ronen, and Orr Dunkelman. 2019. A Simple Explanation for the Existence of Adversarial Examples with Small Hamming Distance. *ArXiv* abs/1901.10861 (2019).
[155] Claude E. Shannon. 1949. *The Mathematical Theory of Communication*. University of Illinois Press.
[156] Mahmood. Sharif, Lujo. Bauer, and Michael. K. Reiter. 2018. On the Suitability of Lp-Norms for Creating and Preventing Adversarial Examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1686–16868.
[157] Yucheng Shi, Yahong Han, Yu-an Tan, and Xiaohui Kuang. 2022. Decision-based Black-box Attack Against Vision Transformers via Patch-wise Adversarial Removal. In *Advances in Neural Information Processing Systems (NeurIPS)*. 12921–12933.
[158] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *IEEE Symposium on Security and Privacy (SP)*. 3–18.
[159] Carl-Johann Simon-Gabriel, Yann Ollivier, Bernhard Schölkopf, Léon Bottou, and David Lopez-Paz. 2019. First-order Adversarial Vulnerability of Neural Networks and Input Dimension. In *International Conference on Machine Learning (ICML)*. 5809–5817.
[160] Vasu Singla, Songwei Ge, Ronen Basri, and David Jacobs. 2021. Shift Invariance Can Reduce Adversarial Robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*. 1858–1871.

[161] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning From Noisy Labels With Deep Neural Networks: A Survey. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)* (2022), 1–19.
[162] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. 2018. PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples. In *International Conference on Learning Representations (ICLR)*.
[163] Lichao Sun, Yingtong Dou, Carl Yang, Kai Zhang, Ji Wang, S Yu Philip, Lifang He, and Bo Li. 2022. Adversarial Attack and Defense on Graph Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* (2022).
[164] Sun, Jiachen and Mehra, Akshay and Kailkhura, Bhavya and Chen, Pin-Yu and Hendrycks, Dan and Hamm, Jihun and Mao, Z. Morley. 2022. A Spectral View of Randomized Smoothing Under Common Corruptions: Benchmarking and Improving Certified Robustness. In *European Conference on Computer Vision(ECCV)*. 654–671.
[165] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing Properties of Neural Networks. In *International Conference on Learning Representations (ICLR)*.
[166] Michel Talagrand. 1996. A New Look at Independence. *The Annals of Probability* (1996), 1–34.
[167] Michel Talagrand. 1996. Transportation Cost for Gaussian and Other Product Measures. *Geometric and Functional Analysis* 6 (1996), 587–600.
[168] Mingtian Tan, Junpeng Wan, Zhe Zhou, and Zhou Li. 2021. Invisible Probe: Timing Attacks with PCIe Congestion Side-channel. In *IEEE Symposium on Security and Privacy (SP)*. 322–338.
[169] Thomas Tanay and Lewis D. Griffin. 2016. A Boundary Tilting Perspective on the Phenomenon of Adversarial Examples. *ArXiv* (2016).
[170] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. 2022. A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning. *Comput. Surveys* (2022).
[171] Liang Tong, Bo Li, Chen Hajaj, Chaowei Xiao, Ning Zhang, and Yevgeniy Vorobeychik. 2019. Improving Robustness of ML Classifiers against Realizable Evasion Attacks Using Conserved Features. In *USENIX Conference on Security Symposium (USENIX Security)*. 285–302.
[172] Antonio Torralba, Rob Fergus, and William T. Freeman. 2008. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 11 (2008), 1958–1970.
[173] Jerome Friedman Trevor Hastie, Robert Tibshirani. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer New York, NY.
[174] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness May Be at Odds with Accuracy. In *International Conference on Learning Representations (ICLR)*.
[175] Jonathan Uesato, Jean-Baptiste Alayrac, Po-Sen Huang, Robert Stanforth, Alhussein Fawzi, and Pushmeet Kohli. 2019. Are Labels Required for Improving Adversarial Robustness?. In *Advances in Neural Information Processing Systems (NeurIPS)*. 12214–12223.
[176] viso.ai. 2022. What is Adversarial Machine Learning? Attack Methods in 2022. https://viso.ai/deep-learning/adversarial-machine-learning/.
[177] Weitao Wan, Jiansheng Chen, Cheng Yu, Tong Wu, Yuanyi Zhong, and Ming-Hsuan Yang. 2022. Shaping Deep Feature Space towards Gaussian Mixture for Visual Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
[178] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5265–5274.
[179] H. Wang, X. Wu, Z. Huang, and E. P. Xing. 2020. High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8681–8691.
[180] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, Binxin Jiao, Yue Zhang, and Xing Xie. 2023. On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective. *ArXiv* (2023).
[181] Tianfeng Wang, Zhisong Pan, Guyu Hu, Yexin Duan, and Yu Pan. 2022. Understanding Universal Adversarial Attack and Defense on Graph. *International Journal on Semantic Web Information Systems* 18, 1 (2022), 1–21.
[182] Wentao Wang, Han Xu, Xiaorui Liu, Yaxin Li, Bhavani Thuraisingham, and Jiliang Tang. 2021. Imbalanced Adversarial Training with Reweighting. *ArXiv* (2021).
[183] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3462–3471.
[184] Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. 2018. Analyzing the Robustness of Nearest Neighbors to Adversarial Examples. In *International Conference on Machine Learning (ICML)*. 5133–5142.

[185] Melanie Weber, Manzil Zaheer, Ankit Singh Rawat, Aditya K Menon, and Sanjiv Kumar. 2020. Robust Large-margin Learning in Hyperbolic Space. In *Advances in Neural Information Processing Systems (NeurIPS)*. 17863–17873.
[186] Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. 2021. Adversarial Robustness under Long-Tailed Distribution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8655–8664.
[187] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *ArXiv* (2017).
[188] Yue Xing, Qifan Song, and Guang Cheng. 2022. Why Do Artificially Generated Data Help Adversarial Robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 954–966.
[189] Peiyu Xiong, Michael Tegegn, Jaskeerat Singh Sarin, and Shubhraneel Pal. 2023. *Supplementary Materials*. *https://resess.github.io/artifacts/DataForMLRobustness*.
[190] Jingyuan Xu and Weiwei Liu. 2022. On Robust Multiclass Learnability. In *Advances in Neural Information Processing Systems (NeurIPS)*. 32412–32423.
[191] Kaidi Xu, Hongge Chen, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Mingyi Hong, and Xue Lin. 2019. Topology Attack and Defense for Graph Neural Networks: An Optimization Perspective. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 3961–3967.
[192] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. 2021. A Fourier-based Framework for Domain Generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14378–14387.
[193] Jia Xue, Hang Zhang, Kristin J. Dana, and Ko Nishino. [n. d.]. Differential Angular Imaging for Material Recognition.
[194] Lu Yang, He Jiang, Qing Song, and Jun Guo. 2022. A Survey on Long-Tailed Visual Recognition. *International Journal of Computer Vision* 130, 7 (2022), 1837–1872.
[195] Shuo Yang, Zeyu Feng, Pei Du, Bo Du, and Chang Xu. 2021. Structure-Aware Stabilization of Adversarial Robustness with Massive Contrastive Adversaries. In *IEEE International Conference on Data Mining (ICDM)*. 807–816.
[196] Shuo Yang, Tianyu Guo, Yunhe Wang, and Chang Xu. 2021. Adversarial Robustness through Disentangled Representations. In *AAAI Conference on Artificial Intelligence (AAAI)*. 3145–3153.
[197] Yao-Yuan Yang, Cyrus Rashtchian, Yizhen Wang, and Kamalika Chaudhuri. 2020. Robustness for Non-Parametric Classification: A Generic Attack and Defense. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 941–951.
[198] Dong Yin, Ramchandran Kannan, and Peter Bartlett. 2019. Rademacher Complexity for Adversarially Robust Generalization. In *International Conference on Machine Learning (ICML)*. 7085–7094.
[199] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer. 2019. A Fourier Perspective on Model Robustness in Computer Vision. In *Advances in Neural Information Processing Systems (NeurIPS)*. 13276–13286.
[200] Xiaoyan Yin, Wanyu Lin, Kexin Sun, Chun Wei, and Yanjiao Chen. 2023. A2S2-GNN: Rigging GNN-Based Social Status by Adversarial Attacks in Signed Social Networks. *IEEE Transactions on Information Forensics and Security (TIFS)* 18 (2023), 206–220.
[201] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. 2015. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *ArXiv* (2015).
[202] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Transactions Neural Networks and Learning Systems* (2019).
[203] Amir R. Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling Task Transfer Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3712–3722.
[204] Chaoning Zhang, Philipp Benz, Adil Karjauv, and In So Kweon. 2021. Universal Adversarial Perturbations Through the Lens of Deep Steganography: Towards a Fourier Perspective. In *AAAI Conference on Artificial Intelligence (AAAI)*. 3296–3304.
[205] Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. 2021. Delving Deep Into Label Smoothing. *IEEE Transactions on Image Processing (TIP)* 30 (2021), 5984–5996.
[206] Huan Zhang, Hongge Chen, Zhao Song, Duane Boning, Inderjit Dhillon, and Cho-Jui Hsieh. 2019. The Limitations of Adversarial Training and the Blind-Spot Attack. In *International Conference on Learning Representations (ICLR)*.
[207] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *International Conference on Machine Learning (ICML)*. 7472–7482.
[208] Hengtong Zhang, Tianhang Zheng, Jing Gao, Chenglin Miao, Lu Su, Yaliang Li, and Kui Ren. 2019. Data Poisoning Attack against Knowledge Graph Embedding. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 4853–4859.
[209] Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial Attacks on Deep-Learning Models in Natural Language Processing: A Survey. *ACM Transactions on Intelligent Systems and Technology* 11, 3 (2020), 1–41.

[210] Xiao Zhang and David Evans. 2022. Incorporating Label Uncertainty in Understanding Adversarial Robustness. In *International Conference on Learning Representations (ICLR)*.
[211] Shuai Zhou, Chi Liu, Dayong Ye, Tianqing Zhu, Wanlei Zhou, and Philip S. Yu. 2022. Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity. *Comput. Surveys* (2022).
[212] Hangyu Zhu and Yaochu Jin. 2020. Multi-Objective Evolutionary Federated Learning. *IEEE Transactions on Neural Networks and Learning Systems* 31, 4 (2020), 1310–1322.
[213] Yao Zhu, Jiacheng Sun, and Zhenguo Li. 2022. Rethinking Adversarial Transferability from a Data Distribution Perspective. In *International Conference on Learning Representations (ICLR)*.