It Is All About Data: A Survey on the Effects of Data on Adversarial Robustness

PEIYU XIONG, University of British Columbia, Canada
MICHAEL TEGEGN, University of British Columbia, Canada
JASKEERAT SINGH SARIN, University of British Columbia, Canada
SHUBHRANEEL PAL, Indian Institute of Technology, Kharagpur, India
JULIA RUBIN, University of British Columbia, Canada

Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to confuse the model into making a mistake. Such examples pose a serious threat to the applicability of machine-learning-based systems, especially in life- and safety-critical domains. To address this problem, the area of adversarial robustness investigates mechanisms behind adversarial attacks and defenses against these attacks. This survey reviews literature that focuses on the effects of data used by a model on the model's adversarial robustness. It systematically identifies and summarizes the state-of-the-art research in this area and further discusses gaps of knowledge and promising future research directions.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Security and privacy** → **Software and application security**.

Additional Key Words and Phrases: Machine Learning, Adversarial Robustness, Evasion Attack, Data Properties

ACM Reference Format:
Peiyu Xiong, Michael Tegegn, Jaskeerat Singh Sarin, Shubhraneel Pal, and Julia Rubin. 2023. It Is All About Data: A Survey on the Effects of Data on Adversarial Robustness. 1, 1 (March 2023), 41 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recent advances in Machine Learning (ML) led to the development of numerous accurate and scalable ML-based techniques, which are increasingly used in industry and society. Yet, concerns related to the safety and security of ML-based systems could substantially impede their widespread adoption, especially in the area of safety-critical systems, such as autonomous cars. Examples of fooling ML models into making wrong predictions by adding imperceptible-to-the-human noise to the input are well known [53]: adversarial perturbation to a stop sign may cause a machine learning system to recognize it as a "max speed" sign instead, which might lead to wrong and dangerous actions taken by an autonomous car [43] (see Fig. 1). Likewise, malicious software can be perturbed to bypass security models while still retaining its malicious behavior [39].

ML models are susceptible to such scenarios, known as *adversarial attacks* or *adversarial examples*, if not specifically trained for [54, 130]. To address this problem, recent literature investigates mechanisms behind adversarial attacks and proposes defenses against these attacks – an area

Authors' addresses: Peiyu Xiong, University of British Columbia, Vancouver, Canada, gbxpeiyu@ece.ubc.ca; Michael Tegegn, University of British Columbia, Vancouver, Canada, mtegegn@ece.ubc.ca; Jaskeerat Singh Sarin, University of British Columbia, Vancouver, Canada, jsarin@student.ubc.ca; Shubhraneel Pal, Indian Institute of Technology, Kharagpur, , India, shubhraneel@iitkgp.ac.in; Julia Rubin, University of British Columbia, Vancouver, Canada, mjulia@ece.ubc.ca.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Association for Computing Machinery.
XXXX-XXXX/2023/3-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

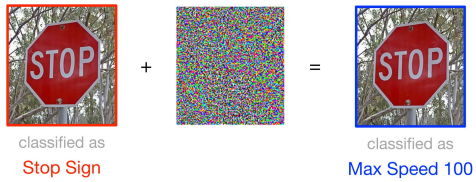


Fig. 1. Adversarial examples for traffic signs (picture by Chen and Wu [71]).

commonly referred to as *adversarial robustness*. The performance of ML models under adversarial attacks, known as *robust accuracy* or *robust generalization*, is often distinguished from the general model accuracy, known as *standard accuracy* or *standard generalization*.

Adversarial attacks that aim to decrease model accuracy can roughly be divided into *evasion* and *poisoning attacks* [17, 81]. The stop sign example above is, in fact, an evasion attack, where the attacker carefully modifies the input to mislead the prediction [24, 76, 90, 130]. Instead of changing model inputs, poisoning attacks are carried out by injecting corrupt data into the training dataset, to deceive the model [52, 122, 135]. *The focus of this survey is on evasion attacks*, as they are more common, accessible, and more frequently discussed in the literature [26, 81, 89, 141]. In fact, most of the literature uses the term *adversarial attacks* to refer to evasion attacks [76, 90, 158]; we thus use these two terms interchangeably.

Most techniques that study adversarial evasion attacks attribute adversarial vulnerability to different aspects of the learning algorithm and/or properties of the data. There are numerous existing surveys on adversarial robustness that focus on different types of adversarial attacks and defenses [8, 81, 165] and sources of adversarial vulnerability related to learning algorithms [89, 121]. Yet, to the best of our knowledge, there are no surveys that collect and organize literature focusing on the influence of data on adversarial robustness.

Our work addresses this gap. Specifically, we investigate (a) what properties of data influence model robustness and (b) how to select, represent, and use data to improve model robustness. To the best of our knowledge, this is the first survey to analyze adversarial robustness from the perspective of data properties – an important direction as the quality of data determines what is achievable through any learning algorithm.

To collect literature relevant to our survey, we used popular digital libraries and search engines, selecting papers that investigate the effect of data on ML adversarial robustness. We identified more than 3,089 potentially relevant papers published in top scientific venues on Machine Learning, Computer Vision, Computational Linguistics, and Security. We systematically inspected these papers, identifying 57 papers relevant to our survey. We further analyzed, categorized, and described the selected papers in this manuscript.

Main findings. The results of our analysis show that producing accurate and robust models requires a larger *number of samples* for training than achieving high accuracy alone. The required number of samples to learn a robust model also depends on other properties of the data, such as dimensionality and the data distribution itself. Specifically, input data with higher *dimensionality*, i.e., a larger number of features that represent the input dataset, requires a larger number of samples to produce a robust model. This is consistent with other findings showing that high *dimensionality* is undesirable for robustness. Moreover, some data *distributions* are inherently more robust than others, e.g., a Gaussian mixture distribution requires more samples to produce a robust model than needed by a Bernoulli mixture distribution.

Another aspect that affects robustness is the *density* of data samples within classes, which measures how far apart samples are from each other. Papers show that high class density correlates to high robust accuracy and that adversarial examples are commonly found in low-density regions of the data. This is intuitive as low-density regions imply that there are not enough samples to

accurately characterize the region. A related property, *concentration*, measures how fast the value of a function defined over a data region, e.g., *error rate*, grows as the region expands. This concept of concentration tightly corresponds to adversarial robustness if we consider the expansion of a data region as the effect of adversarial perturbation, i.e., perturbing samples in all directions causes the region defined by the original samples to expand. In this case, high concentration implies that the error rate grows as one perturbs all points in a data region and, thus, datasets with high concentration are shown to be inevitably non-robust. The *separation* between classes of the underlying data distribution also affects robustness, with a large distance between different classes being desirable for adversarial robustness as an attacker would need to use a large perturbation to move samples from one class to another.

Yet another aspect that impacts robust accuracy is the presence of mislabeled samples in a dataset, referred to as *label noise*. Furthermore, refining labels to reason about a larger number of classes, e.g., splitting the "animal" class into "cat" and "dog" may improve adversarial robustness as such labels allow learning more compact representations for samples that share stronger similarities.

A number of papers also identify *domain-specific* properties that correlate with adversarial robustness. For example, image frequency – the rate of pixel value change – affects robustness, and it is advisable to use a diverse frequency range in the training dataset to prevent any frequency biases which give rise to adversarial examples.

Observations and Gaps. Our literature review shows that, even though most works study data properties from a domain-agnostic perspective, they typically conduct an empirical evaluation on image datasets only. This constrains the types of attacks and robustness measurements considered, so findings may not generalize to other domains or types of datasets. Furthermore, most works base their formal derivations on quite simple synthetic data models, such as uniform distributions, a mixture of Gaussian distributions, and a mixture of Bernoulli distributions, which exhibit unrealistic assumptions compared to real datasets used in practice. We also observed that while most papers only perform a univariate analysis on a specific data property, most properties are hard to independently optimize, e.g., to decrease dimensionality without decreasing separation as decreasing the dimensionality implies that samples have fewer features to be differentiated from each other. We also found that some properties, e.g., separation, do not have a standard way of measurement for concrete datasets. We believe future work should look into these directions.

Contributions. The main contribution of this survey are:

- A collection of literature on the effects of data on adversarial robustness.
- A categorization and detailed analysis of the collected literature.
- An analysis of knowledge gaps and suggestions for future research directions.

Structure of the Survey. The remainder of the survey is structured as follows. Section 2 introduces the necessary background and terminology used in this survey. Section 3 presents our methodology for identifying and categorizing relevant papers. We describe the collected literature in Section 4. In Section 5, we discuss our analysis of the literature, possible knowledge gaps, and suggestions for future work. Section 6 summarizes related work and Section 7 concludes the survey.

2 PRELIMINARIES

We now provide a brief overview of the main concepts related to machine learning, adversarial robustness, and most commonly studied data distributions. The goal of this section is to introduce terminology used in the rest of the survey rather than provide an extensive overview of the adversarial robustness research area. For a more detailed overview, please refer to guides on statistics and machine learning [19, 124, 138] and adversarial robustness [17, 28, 102].

2.1 Machine Learning

Machine learning refers to the automated detection of meaningful patterns in data [124] and can be largely divided into supervised, unsupervised, and reinforcement learning. In supervised learning, a learning model is provided with input-output pairs of data (a.k.a. labeled training data); based on this data, the model aims to infer a function that maps the inputs to the outputs. Supervised learning is typically associated with classification and regression problems, which use categorical and continuous labels, respectively. In classification, this number of possible labels for an input is also referred to as the number of *classes*. Datasets with only two classes are called *binary datasets*, on which one can train a *binary classifier*.

Unlike supervised learning, unsupervised learning algorithms are usually concerned with identifying patterns in unlabeled data, e.g., grouping similar samples together in the absence of labels (clustering) or transforming data into a different representation (representation learning). Reinforcement learning characterizes algorithms that learn from a series of rewards and punishments, with the goal of maximizing the cumulative reward, e.g., to build robots that learn to take the best sequence of actions according to signals from the environment.

Variations, such as, semi-supervised learning (i.e., learning from partially labeled data) and self-supervised learning (i.e., learning from labels extracted by the learner itself) have also been proposed for problems where acquiring labeled data may be challenging or expensive.

ML algorithms can also be divided into parametric and non-parametric. Parametric algorithms have a predetermined, fixed number of parameters defined before the training starts. For example, for Linear Support Vector Machines (SVMs), these parameters are the coefficients of all features of the training data and the learned intercept. For Deep Neural Networks (DNNs), the number of parameters is determined by the architecture of the network. In non-parametric algorithms, the number of parameters is determined at training time and may vary depending on the number of training samples. For example, the "depth" of Decision Trees can grow (beyond the size of the feature set) when more decision points are needed to accurately separate training data. Other commonly used non-parametric models include k-Nearest Neighbors (k-NN) and Kernel SVMs.

2.2 Adversarial Robustness

Adversarial machine learning studies the arms race between adversarial attacks and defenses. Attacks aim at degrading models performance while defenses propose algorithms to harden models against the attacks. Adversarial attacks can be categorized into *evasion* and *poisoning* [17, 81]. Evasion attacks aim to fool machine learning models by generating inputs that, despite no noticeable difference for a human, will be incorrectly classified. Such inputs, known as *adversarial examples* and created by applying non-random perturbations to samples, carefully designed to change models' prediction [24, 76, 90, 130]. Instead, poisoning attacks tampers with model training data, in order to degrade model performance. In this survey, we focus on evasion attacks; the terms adversarial and evasion attacks are often used interchangeably in the literature as this is the most popular and commonly studied type of attack.

The term *robustness* for machine learning models is often used to refer to different concepts, such as, stability to distribution shifts, the ability to identify adversarial examples, and the ability to make the correct predictions in the face of adversarial examples. In this survey, we use the latter definition – the ability to make the correct predictions in the face of adversarial examples. This is a stronger notion of robustness than merely identifying adversarial examples, as the identification of an adversarial example does not guarantee its correct classification. The phenomenon of making satisfactory model predictions in the face of adversarial examples is also often referred to as *robust generalization*. This is different from *standard generalization*, a term used to describe making satisfactory model predictions for normal, unseen samples.

Adversarial (Evasion) Attacks. Techniques for generating adversarial examples for evasion attacks can be broadly divided into three categories, according to the type of information available to the attacker [17]. In *white-box attacks*, the attacker is assumed to be able to leverage all available information about the training data, the model, and training procedure. In *grey-box attacks*, the attacker is assumed to have only partial information about the model, such as, the source of training data. Finally, the most conservative type of attacks are *black-box attacks*, where the attacker has no information about the inner workings of the model except, possibly, for the prediction outcomes.

Gradient-based attacks are commonly used in white-box settings. These attacks use the gradient of a differentiable function defined over model weights as a guide when crafting adversarial examples. The most commonly used differentiable function is the loss function used by the model during training. A gradient defines the direction of the maximal increase in the local value of a function. Hence, by using the gradient of the loss function with respect to the input, one can adjust the input to get the maximal increase in the loss of the model, which ultimately leads to a bad prediction. Fast Gradient Sign Method (FGSM) [55], Basic Iterative Method (BIM) [76], and Projected Gradient Descent (PGD) [90] are examples of attack algorithms that utilize the gradient of the loss function used for training. Instead of the loss function, the FAB attack [33] uses the gradient of a function defined by the difference of model outputs of the penultimate layer of a neural network – a layer which outputs the probabilities that a given sample belongs to each of the available classes. By defining the difference of outputs of the penultimate layer as the differentiable function, the FAB attack maximizes the difference in probabilities between the target class and other classes, to increase the chance of misclassification.

Non-gradient based attacks are applicable for more diverse types of models that do not use a differentiable functions, e.g., decision trees. Such attacks can also be used in black-box and grey-box settings, when gradient information is hidden from the attacker. One example of non-gradient-based attacks is the *mimicry attack*, which involves adding and removing features in the perturbed sample, e.g., based on their popularity in the target class [39].

Adversarial Defenses. Defense mechanisms against adversarial attacks target various stages of the machine learning pipeline. Specifically, defenses *on raw data* focus on the training data itself, e.g., by selecting a subset of "robust" features [66] or using representation learning to transform features into a different representation, making sure a model trained on the new representation is inherently more robust [153].

Defenses *during training* alter the standard training procedure to improve model robustness. The most common such technique is adversarial training [55], which involves continually augmenting the training data with adversarial examples generated by an attack algorithm. By retraining the model while adding correctly labeled malicious samples to the training dataset, the model learns to capture persistent patterns and becomes more robust against these attacks. Another common method is *regularization*, where model parameters are constrained so that very small perturbations have little effect on the prediction outcome [56].

Defenses *during inference* focus on making existing models more robust when the model is being used on new samples. For example, randomized smoothing [30] involves creating multiple noisy instances of a sample and aggregating the model's predictions during inference. Given that adversarial examples are typically close to genuine samples, averaging the results from close neighbors of an input can potentially reduce the chances of the model being misled. In addition, different variations of ensemble models – using multiple models and aggregating their output – have been shown to increase the robustness to adversarial attacks [106].

Measures of Robustness. The strength of an adversary is mostly measured by the size of the perturbation required to create an adversarial example. That is, adversaries that introduce more

perturbations, e.g., change a larger portion of pixel values in the image, are considered to be stronger. A typical way of measuring the perturbation size, especially in the image domain, is by using the L_p distance metric, where p can be a whole number or ∞. Specifically, L_0 counts the total number of changed features, regardless of the changes to individual features. L_1 is the Manhattan distance, i.e., the sum of absolute values representing a change in each feature. L_2 measures by the Euclidean distance between the feature values of the original and perturbed samples. the L_∞ metric measure the largest change in any of the features (while disregarding changes in all other features).

There are two ways to utilize these distance metrics to evaluate the robustness of a model: error-rate-based and radius-based. The first calculates a pool of adversarial samples generated from a set of real samples with a fixed allowable perturbation size [90]. The robustness is then defined as the error rate of the model on these adversarial samples. A related concept, *adversarial risk*, is also defined in a similar manner: the probability of finding, within a certain predefined distance, an adversarial example for a given real sample.

The radius-based way to evaluate is measuring the smallest distance required to generate an adversarial sample from a given real sample [130]. This way is especially useful in robustness certification, which involves learning a classifier that outputs a prediction along with a certified radius within which the prediction is guaranteed to be consistent [78].

2.3 Data Distributions

Numerous works study properties of particular data distributions, which we discuss below. The *uniform distribution* defines a probability distribution in which every possible data point is equally likely. This implies that for a continuous random variable in the interval $[a, b]$, the probability of seeing a sample from the interval is $\frac{1}{b-a}$ and the probability of seeing a sample from outside of the interval is 0. In the discrete case with n possible values, the uniform distribution assigns a probability of $\frac{1}{n}$ to each value.

The *Bernoulli distribution* defines a discrete probability distribution of a random variable with two allowable values, 0 and 1. Such a random variable takes the value of 1 with probability p and the value of 0 with probability $1 - p$.

The *Gaussian (normal) distribution* defines a continuous probability distribution that assigns a probability with its peak at the center of the distribution and decreasing symmetrically outwards. For a Gaussian distribution, μ denotes the mean or center of the distribution, and σ^2 denotes the variance or the spread of the distribution. Since the mean and variance fully characterize a Gaussian distribution, it is also commonly denoted as $\mathcal{N}(\mu, \sigma^2)$.

One can also imagine a distribution made up of a *mixture* of multiple distributions. For example, Fig. 2 shows a distribution made up of two Gaussians: one centered at μ_1 and another – at μ_2. This mixture also contains labels associated for each independent Gaussian, shown by the two clusters in the figure. Furthermore, these two clusters have the same variance, i.e., the same spread of the distribution surrounding the center of the class. While the means of the two classes are separated, the distributions intersect with each other.

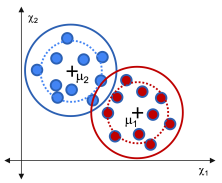


Fig. 2. A two-dimensional Gaussian mixture data.

3 METHODOLOGY

This section describes our methodology for identifying and categorization relevant papers.

3.1 Paper Collection

Papers for this survey were collected in June 2022. We used the search query schematically described below, which was designed to identify papers in the area of Machine Learning adversarial robustness

that discuss properties of the underlying data. We expanded each of these conceptual terms with possible synonyms and specific wording, making sure our query is as comprehensive as possible.

Search Query := Machine Learning + Adversarial Robustness + Data + Property
Machine Learning := classif | "machine learning" | "deep learning" | "neural network"
Adversarial Robustness := "adversarial robustness" | "adversarial vulnerability" | "adversarial attack" | "adversarial perturbation" | "adversarial defense" | "evasion attacks"
Data := data | sample | input
Property := propert | qualit | distribution | characteristic

In our schematic query representation, the "+" and "|" signs indicate the AND and OR operators, respectively, the phases in quotes are matched in full, and each word is matched with its suffixed versions, e.g., 'classif' is matched with both 'classifier' and 'classification'. When performing the search, we adapted this schematic query to the requirements and capabilities of each search engine that we used.

Table 1. Considered publication venues

Venue Type	Area	Venue Name	Search Target					Total Hits	Relevant
			ACM	IEEE	Springer	Scopus	Google Scholar		
Conference	ML	AAAI Conference on Artificial Intelligence (AAAI)				✓	✓	320	1
		ACM International Conference on Web Search and Data Mining (WSDM)	✓					39	0
		ACM SIGKDD Conference On Knowledge and Data Mining (KDD)	✓					11	0
		Conference on Neural Information Processing Systems (NeurIPS)				✓	✓	658	19
		IEEE International Conference on Data Engineering (ICDE)		✓				3	0
		IEEE International Conference on Data Mining (ICDM)		✓				34	1
		International Conference on Learning Representations (ICLR)				✓	✓	313	9
		International Conference on Learning Theory (COLT)				✓	✓	14	1
		International Conference on Machine Learning (ICML)				✓	✓	320	12
		International Joint Conference on Artificial Intelligence (IJCAI)				✓	✓	86	0
	CV	European Conference on Computer Vision (ECCV)			✓			88	2
		IEEE / CVF Computer Vision and Pattern Recognition (CVPR)		✓				244	3
		IEEE International Conference on Computer Vision (ICCV)		✓				157	2
	CL	Annual Meeting of the Association for Computational Linguistics (ACL)				✓	✓	83	0
	SEC	ACM Conference on Computer and Communications Security (CCS)	✓					61	0
		IEEE Symposium on Security and Symposium (S&P)		✓				20	0
		Network and Distributed System Security Symposium (NDSS)				✓	✓	25	0
		USENIX Security Symposium				✓	✓	89	0
Journal	AI	Artificial Intelligence Journal				✓	✓	5	0
		Computational Linguistics Journal (CL)				✓	✓	3	0
		IEEE Transactions on Knowledge and Data Engineering (TKDE)		✓				16	0
		IEEE Transactions on Neural Networks and Learning Systems (TNNLS)		✓				29	0
		IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)		✓				56	1
		International Journal of Computer Vision (IJCV)			✓			26	0
		Journal of Machine Learning Research (JMLR)				✓	✓	29	2
		Knowledge Based Systems (KBS)				✓	✓	22	0
		Neural Networks				✓	✓	51	0
		Pattern Recognition (PR)				✓	✓	53	0
	IS	Computer & Security				✓	✓	44	0
		IEEE Security & Privacy		✓				10	0
		IEEE Transactions on Dependable and Secure Computing (TDSC)		✓				26	0
		IEEE Transactions on Information and Forensic Science (TIFS)		✓				82	1
		Journal of Information Security and Applications (JISA)				✓	✓	13	0
	Surv.	ACM Computing Surveys (CSUR)	✓					59	0
Total								3089	54

An initial search using the query in Google Scholar identified more than 30,000 matches. To keep the scope of the survey manageable, we thus limited our search to publications from the main track of top-tier conferences and journals in the areas of Machine Learning (ML), Computer Vision (CV), Computational Linguistics (CL), and Security (SEC). Specifically, we selected all A* conferences from these areas using the most recent, 2021, CORE ranking [3]; top 10 journals according to the Journal Citation Reports (JCR) [4] in the area of Artificial Intelligence (AI), which includes Machine Learning, Computer Vision, and Computation Linguistics, and top five journals in the area of Information System (IS), which includes Security. Additionally, we included the ACM Computing

Surveys Journal [1] to collect surveys related to our topic. The first three columns in Table 1 shows the final list of publication venues that we selected.

We further identified digital libraries and search engines that host proceedings of our selected venues. These are shown in the next five columns of Table 1, for each venue individually. As some venues are only partially indexed by the digital libraries, i.e., the libraries only include proceedings from particular years, we augmented our search for papers in these venues with a secondary search using Google Scholar. More specifically, we used Google Scholar's *site* and *source* filtering constraint to limit our search only to the target venues of interest, as described in Section A.1 of the Appendix. We also used Google Scholar to search the ArXiv repository and filtered the results using *publication venue* information provided by Semantic Scholar API [6].

The second to last column in Table 1 shows the number of hits for each publication venue identified by this search. In total, using this procedure, we identified 4,429 papers; after removing duplicated, we obtained a set of 3,089 potentially relevant papers.

3.2 Manual Filtering

Next, we manually classified papers identified in the automated search into relevant and irrelevant for our survey. To this end, we first randomly selected a set of 40 papers and used them as a pilot for drafting the inclusion and exclusion criteria. Four authors of this survey independently read the abstract, introduction, and conclusions of each paper and classified it as *relevant*, *non-relevant*, or *unsure*. Each author also assigned a concise label for each of the *relevant* and *non-relevant* papers, specifying the reasons for inclusion/exclusion.

After completing this phase, all authors of this survey met to cross-validate the decisions, and to consolidate and refine the inclusion/exclusion labels. All disagreements between the authors (mostly between *non-relevant* and *unsure* papers) were resolved through a joint discussion.

In the second phase, we randomly selected an additional set of 40 papers, to validate our filtering process. Four authors of this survey, again, independently read the categorized each of these papers, and all authors met to discuss the results. While there were disagreements, with a rate of 6.25%, between the assignment of *non-relevant* and *unsure* papers, the inclusion/exclusion labels were consistent among the raters.

Specifically, we included papers that:

(a) study the correlation between properties of input data and the adversarial robustness of resulting model trained on this data; and/or
(b) present techniques to improve or disrupt a model's adversarial robustness through explicitly modifying some properties of the input data or its latent representation.

We excluded papers that:

(a) discuss adversarial evasion attacks, but focus on features [136], models [50, 149], and training algorithms [61, 70] rather than data (47.2%);
(b) discuss aspects of ML that are not related to adversarial robustness but rather related to accuracy [10, 108, 166], robustness to distribution shifts not induced by adversaries [96, 150], privacy [63, 64], and interpretability [49, 60] (39.7%);
(c) propose new robustness evaluation metrics [29] or assessment frameworks [82] (2.7%);
(d) focus on poisoning [84, 162] rather than evasion attacks (2.6%); or
(e) got matched accidentally and discuss unrelated topics, e.g., literature on blockchain [65], remote access trojan systems [114], or hardware systems [133]. These papers appeared in our search results as "adversarial robustness" is also desirable for non-ML systems, e.g., blockchain systems need to be robust against adversarial selfish miners or Denial-of-Service attacks (6%).

As the inclusion/exclusion labels were consistent among the raters, we decided to proceed to the next phase: distribute the remaining 3,009 papers among the four authors and categorize them using these inclusion and exclusion criteria. In this phase, we instructed each rater to conservatively mark as *unsure* papers with even a slight doubt in categorization.

Following this process, we identified 199 papers as either *relevant* or *unsure*. We assigned a second reader to papers marked as *unsure*, summarized each such paper in writing, and held a meeting with all the authors of this survey to categorize the paper as either *relevant* or *non-relevant*. For a select set of papers where we could not confidently reach a decision, we emailed the paper authors to validate our understanding and decide on the relevance of the work (all such papers were excluded in the end). Our analysis resulted in 54 *relevant* papers. The distribution of these papers by publication venues is shown in the last column of Table 1.

We further assigned a second reader to each identified relevant paper, to extract and summarize its main findings. To make sure that we included most of the relevant works on the topic, we also performed backward snowballing using the related work sections of the selected papers, which resulted in 3 additional papers, bringing our selection to 57 papers in total. Fig. 3 summarizes our paper selection process.

Fig. 4 shows another view on the distribution of publication venues for all 57 papers included in our survey. The majority of the papers (50, 85%) are published in Machine Learning venues. In fact, only the *Advances in Neural Information Processing Systems (NeurIPS)* conference published 19 (around 36%) of all papers. Seven papers are from the Computer Vision venues and only two are from the Security venues. We found no relevant papers in the *Computational Linguistics* venues.

Fig. 5 shows the distribution of selected papers by their publication year. The figure suggests that mainstream research communities started to investigate the impact of training data on adversarial robustness as recently as 2018. We did not find many relevant papers from 2022 as our search was conducted in June 2022.

3.3 Categorization of Selected Papers

To better classify, discuss, and compare the papers, we proposed a categorization schema shown in Fig. 6. To construct the schema, we followed an iterative process similar to the one we used for paper selection: we first sampled ten papers from the final collection and each of the four authors independently proposed categorization attributes to describe these ten papers. The proposed attributes were discussed by all authors while unifying related attributes, removing redundant ones, and updating labels. We then verified the applicability of the constructed schema on another set of ten papers and adjusted it based on a joint discussion. We continued extending the schema after reading the remaining papers, to ensure its inclusiveness.

The resulting categorization schema, which we further use to analyze and describe the papers, contains three high-level areas described below. The detailed categorization of each papers along the attributes of this schema can be found in Section A.2 of the Appendix.

Search Results (Selected Venues)	#
ACM	622
IEEE	660
Springer	115
Scopus	1,056
Google Scholar	2,016
Total	**4,429**

Remove duplicates / non-main tracks →

Search Results (Unique)	#
Total	**3,089**

↓ Manual Filtering

Relevant	54
Snowballing	3
Total	**57**

Fig. 3. Summary of the collection and selection of papers.

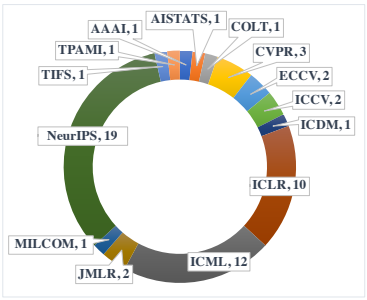


Fig. 4. Breakdown by publication venues.

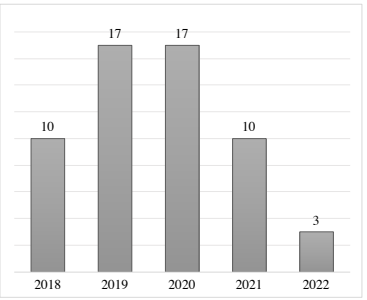


Fig. 5. Breakdown by publication time.

1. Problem Setup defines the scope of the proposed approach and the assumptions authors make in their work. Specifically,

- *Target Distribution* describes the type of data distributions the papers focus on. The common data distributions studied among the collected papers include Gaussian mixtures, Bernoulli mixtures, and Uniform distributions.
- *ML Model* focuses on the studied model. It captures the *learning task*, e.g., binary classification, multi-class classification, and regression, and *classifier type*, e.g., parametric models, such as, neural networks and non-parametric models, such as, k-NNs.
- *Robustness Setting* records the paper's definition of adversarial robustness. This includes the *robustness definition* sub-category, which refers to how the authors measure robustness, e.g., error-rate-based radius-based. The second sub-category, the *attacker's knowledge*, reflects the level of information about the target system that the attacker can exploit: white-box, grey-box, or black-box. The last sub-category, the *attack*, characterizes the *technique* used to construct the attack, e.g., gradient- or non-gradient based, and the *perturbation bound* considered, e.g., the type of L_p norm.

2. Data Property dimension includes the eight data properties we identified in the collected papers: the *number of samples*, data *dimensionality*, *distribution*, *density*, *concentration*, *separation*, *label quality*, and *domain-specific* properties, relevant in context of particular application domains. We introduce and structure the discussion of the surveyed papers in Section 4 around these properties.

3. Practicality specifies how to apply the approach or technique introduced in each paper:

- *Applicability* determines whether specific *quantitative* metrics are provided to measure the data properties discussed in the paper or whether there are any concrete techniques proposed to modify these data properties.
- *Explainability* determines whether the paper focuses on explaining (rather than establishing) the correlation between data property and robustness.

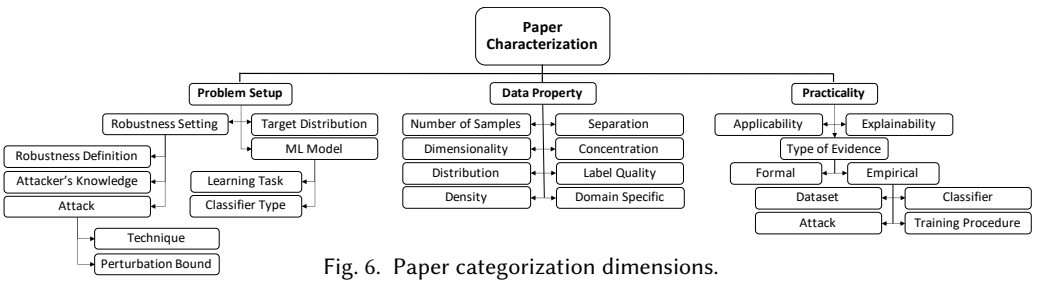


Fig. 6. Paper categorization dimensions.

- *Type of Evidence* records the type of arguments provided by the paper. This may be a *formal proof*, an *empirical evaluation*, or a combination of both. For cases where an empirical evaluation is performed, we also collect information about *datasets* and *classifiers* used, the applied *training procedures* (standard vs. adversarial training), and the *attack techniques* employed.

In what follows, we present the results of our analysis of the surveyed papers (Section 4) and discuss our observations (Section 5).

4 RESULTS

We present the results of our analysis by organizing the papers according to the robustness-related data property they discuss: number of samples, dimensionality, type of distribution, density, concentration, separation, label quality, and domain specific properties. Papers that discuss more than one data property are presented in all corresponding sections. That is, in what follows, a paper can be discussed in more than one section. Section 4.9 summarizes our findings.

To ease navigation, for each discussed data property, we also include a map showing how the relevant papers relate to each other via their citation information. We further annotate each paper with its *applicability* and *explainability* categories. Specifically, we annotate with an [A] symbol papers that propose an actionable technique to modify or measure a robustness-related property; we annotate with [E] papers that put extra emphasis on explaining the correlation between a data property and robustness rather than establishing such a correlation.

4.1 Number of Samples

Number of samples simply means the quantity of samples available in the training dataset. For the example in Fig. 7, where circles represent training samples for a two-class dataset, the left dataset has fewer samples than the right dataset.

The term *sample complexity* refers to the number of training samples required to achieve a certain model performance, e.g., 90%, in terms of either robust or standard generalization. Then, *sample complexity gap* refers to the difference in the number of samples required to achieve the same model performance for robust generalization as for standard generalization.

Papers studying the relationship between the number of training samples and the robustness of the resulting model are shown in Fig. 8. They can roughly be divided into [1] papers discussing sample complexity for robust generalization, [2] papers proposing techniques to resolve the sample complexity gap between the number of samples required to achieve the same level of robust and standard generalization, and [3] papers proposing techniques to deal with data imbalance, i.e., unequal number of samples in different classes.

[1] **Sample Complexity**. Schmidt et al. [119] observe that the number of training samples required for robust generalization is larger than the number of samples required for the equivalent-level standard generalization, i.e., that there exists a *sample complexity gap* between the standard and robust generalization. Specifically, for linear classifiers trained on a mixture of Gaussian distributions (referred to as *Schmidt's Gaussian mixture* in the remainder of this paper), the authors prove that standard generalization requires a constant number of samples while equivalent-level robust generalization requires a number of samples proportional to the data dimensionality ($O(\sqrt{d})$). The

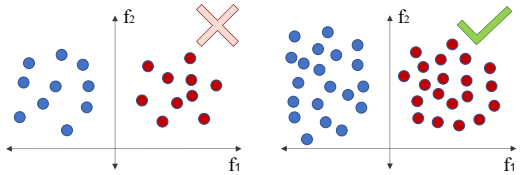


Fig. 7. Number of samples illustration.

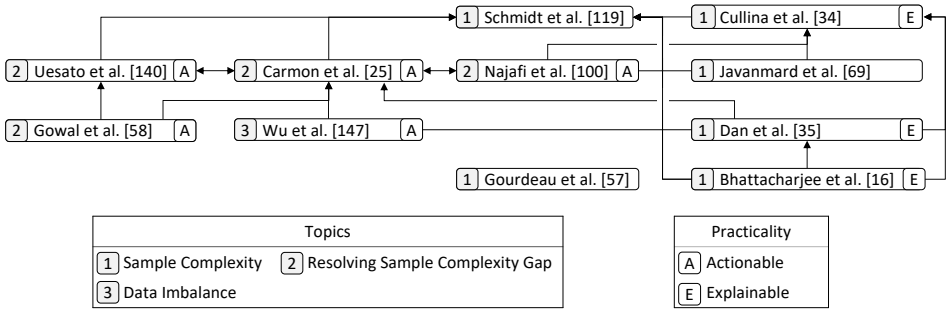


Fig. 8. Papers discussing the number of samples.

gap in sample complexity persists for this data distribution in nonlinear classifiers as well. Yet, the sample complexity gap disappears for nonlinear classifiers trained on a mixture of Bernoulli distributions; these distributions also need substantially fewer samples than Gaussian mixtures. The authors conclude that sample complexity for robust generalization depends on the distribution, even when the same type of classifiers is considered. Their experimental validation with the MNIST [77], CIFAR-10 [73], and SVHN [101] image datasets shows that MNIST, which is closer to a Bernoulli mixture, indeed requires a smaller number of training samples to achieve a reasonable robust generalization than the CIFAR-10 and SVHN datasets, which are closer to a Gaussian mixture.

In follow-up work, Dan et al. [35] provide reasons for why robust generalization requires more samples than standard generalization, focusing, again, on Gaussian mixture distributions. Departing from the Signal-to-Noise ratio (SNR) metric that is based on the distance between two Gaussian distributions and is known to capture the hardness of standard classification, the authors propose a new Adversarial SNR (AdvSNR) metric, defined as the minimum SNR for standard and adversarially perturbed data, to capture the hardness of robust classification. They then show that, given a dataset of a particular dimensionality, the number of samples required to achieve the theoretically optimal, accurate classifier is inversely proportional to SNR. Likewise, the number of samples required to achieve the theoretically optimal, robust classifier is inversely proportional to AdvSNR. Because AdvSNR is always no greater than SNR for a given dataset, it follows that achieving the same robust generalization as standard generalization requires at least the same amount of samples.

Bhattacharjee et al. [16] study the sample complexity gap for linear classifiers, as a factor of data dimensionality (the number of features representing samples) and separation (the distance between samples from different classes). The authors show that the sample complexity gap is directly proportional to the dimensionality of the data when the allowed perturbation radius of adversarial samples is similar to the distance between classes. However, such a gap no longer exists in well-separated data, when the perturbation radius is much smaller than the distance between classes.

Similarly, Gourdeau et al. [57] show that, for simple classifiers based on feature conjunctions and α-log-Lipschitz distributions laying on a boolean hyper-cube, the sample complexity is proportional to the data dimensionality d. Specifically, when the adversarial perturbation size is bounded by $log(d)$, the sample complexity is polynomial to the dimensionality; when the perturbation size is at least $log(d)$, the sample complexity becomes superpolynomial to dimensionality. Javanmard et al. [69] focus on adversarially-trained linear regression models for standard Gaussian distributions. The authors show that, when the number of samples is greater than data dimensionality, there exists a trade-off between adversarial and standard risks. Moreover, this trade-off improves as the number of samples per dimension increases.

Cullina et al. [34] give an upper bound on the number of samples needed for robust generalization for the binary classification problem with linear classifiers in a distribution-agnostic setup, with L_p

norm-bounded adversaries. The authors derive the upper bound using the classifier VC dimension – a common measure of the capacity and the expressive power of the classifier, shown earlier to be useful to determine the upper limit of sample complexity for standard generalization [124]. They show that the VC dimension for learning adversarially-robust models remains the same as that for learning accurate models, which means that the upper bound of sample complexity is identical for standard and robust generalization in this setup. However, the authors demonstrate that this conclusion does not generalize to other types of classifiers and types of adversaries.

[2] **Techniques for Resolving the Sample Complexity Gap**. As the number of labeled samples required to achieve robust generalization could be large and not readily available, researchers explore cheaper alternatives, such as, unlabeled data and generated (fake) data. Uesato et al. [140] and Carmon et al. [25] concurrently proposed to use pseudo-labeling [120] – a process of assigning labels to unlabeled samples using a classifier trained on a set of labeled samples, assessing the effectiveness of their approaches on Schmidt's Gaussian mixture. The main result of both works is that closing the sample complexity gap requires a number of unlabeled samples proportional to the dimensionality of the data, albeit with a higher quantity than for the labeled samples, likely due to the "noise" in generating labels. The main difference between the works is that while Uesato et al. show that in their setup (a specific linear classifier) the quantity of the required unlabeled samples only depends on data dimensionality, Carmon et al. [25] use a less restrictive setup and show that the quantity of unlabeled samples also depends on the original sample complexity for standard generalization. Both of these works empirically evaluate the effectiveness of their proposed approaches on the CIFAR-10 and SVHN datasets showing that unlabeled data could be a much cheaper alternative to labeled data for enhancing the robustness of models.

Najafi et al. [100] note that the biggest risk of using a mixture of labeled and unlabeled datasets for learning adversarially robust models is the uncertainty in sample labels. Given an estimate of the quality of pseudo-labels, the authors derive the minimum ratio between labeled and unlabeled samples required to avoid the additional adversarial risk induced by label uncertainties.

Instead of using unlabeled data, which might also be hard to find, Gowal et al. [58] suggest using Generative Adversarial Networks (GANs) to generate labeled data. The authors show that GANs are more effective than other methods, e.g., image cropping, when producing additional samples. This is because such models result in a more diverse dataset, which is beneficial for increasing robust accuracy. Using images from CIFAR-10, CIFAR-100 [73], SVHN, and Tiny Images Dataset [137], the authors show that their proposed approach can significantly increase robust accuracy without the need for additional real samples.

[3] **Data Imbalance**. Wu et al. [147] analyze adversarial robustness of DNNs on long-tail distributions: setups where the training data contains a large number of classes with few samples. They show that robust generalization is harder to achieve on such distributions and compare the performance of multiple adversarially trained classifiers that use learning algorithms specifically designed for such setups. The comparisons show that scale-invariant classifiers [107, 143] result in higher robust accuracy as they avoid assigning smaller weights to minority classes, which, in turn, promotes robust generalization by reducing bias in the decision boundary.

4.2 Dimensionality

Dimensionality refers to the number of features used to represent the data, e.g., features f_1, f_2, f_3 in Fig. 9. For illustration purposes, we show a dataset with a dimensionality of three on the left-hand side of the figure and a dataset with a dimensionality of one on the right-hand side. *Intrinsic dimensionality* refers to the number of features used in a minimal representation of the data. Fig. 10 shows an example of a case where the intrinsic dimensionality is smaller than the actual

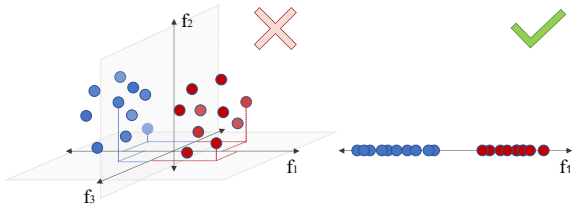
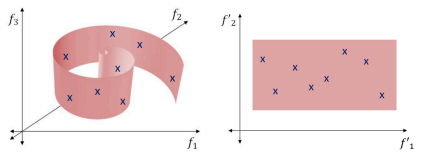


Fig. 9. Dimensionality illustration.



(a) Actual. (b) Intrinsic.

Fig. 10. Actual and intrinsic dimensionality.

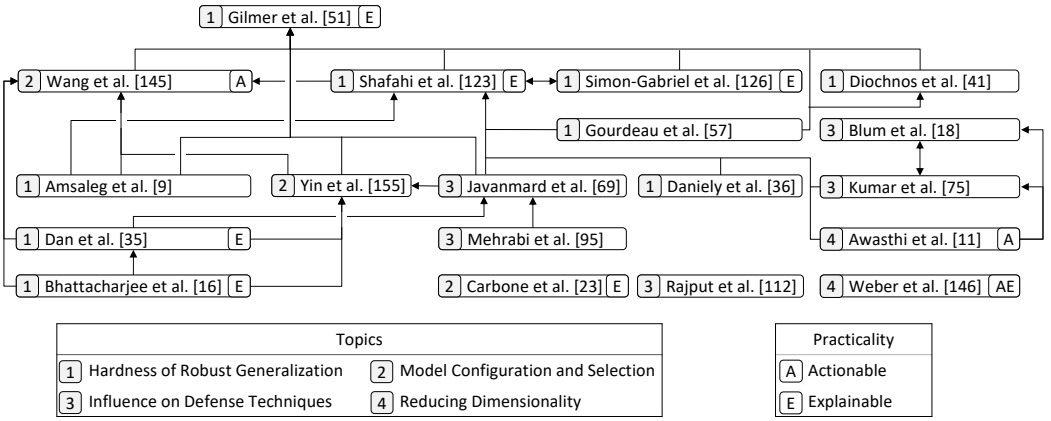


Fig. 11. Papers discussing dimensionality.

dimensionality: the samples in Fig. 10a are lying on a three-dimensional "swiss roll". "Unwrapping" the roll into a plain sheet, as shown in Fig. 10b, makes it possible to distinguish between the samples using only two dimensions.

Papers studying the relationship between data dimensionality and adversarial robustness are shown in Fig. 11. We divide them into papers [1] characterizing the hardness of robust generalization due to high dimensionality, [2] suggesting robust model types and configurations for high-dimensional data, [3] discussing the impact of high dimensionality on existing defense techniques, and [4] utilizing dimensionality reduction techniques for improving robustness.

[1] **Hardness of Robust Generalization**. A number of authors show that adversarial examples are inevitable in high-dimensional space. Specifically, Gilmer et al. [51] prove this for a synthetic binary dataset composed of two concentric multi-dimensional spheres (a.k.a., hyperspheres) in high-dimensional space (>100), showing that samples are, on average, closer to their nearest adversarial examples than to each other. They also prove that the adversarial risk of a model trained on this dataset only depends on its standard accuracy and dimensionality. A similar result is shown by Diochnos et al. [41], for a uniformly distributed boolean hypercube dataset, and Shafahi et al. [123], for unit-hypersphere and unit-hypercube datasets.

Another line of work analyzes the effect of dimensionality on the robustness of specific types of classifiers. In particular, Simon-Gabriel et al. [126] study feedforward neural networks with ReLU activation functions and He-initialized weights, showing that a higher input dimensionality increases the success rate of adversarial attacks, regardless of the topology of the network. The authors, however, demonstrate that regularizing the gradient norms of the network decreases the impact of the input dimension on adversarial vulnerability, thereby improving model robustness on high-dimensional inputs. Daniely et al. [36] study the effect of dimensionality on ReLU networks

with random weights and with layers having decreasing dimensions. Like Simon-Gabriel et al., the authors prove that the robustness of ReLU networks degrades proportionally to dimensionality.

Amsaleg et al. [9] focus on k-NNs and other non-parametric models that base predictions on the proximity of samples. The authors use the Local Intrinsic Dimensionality (LID) metric to represent the intrinsic dimensionality in the neighborhood of a particular sample x. The authors build up on the observation that a high LID implies that there are more samples in close proximity of x (as, otherwise, a more sparse neighborhood could be encoded in fewer dimensions). Thus, it is possible to arbitrarily change the neighborhood ranking of the nearest neighbor of x using a small perturbation. As predictions of proximity-based models are based on the nearest neighbor ranking, the adversarial risks increase in this setup.

All the aforementioned works are also in agreement with a number of papers discussed in Section 4.1, i.e., [16, 35, 57], which show, in their respective settings, that sample complexity for robust generalization is proportional to dimensionality.

[2] **Model Configuration and Selection**. Wang et al. [145] prove that the optimal k for producing robust k-NN classifiers depends on the dimensionality d and number of samples n of the given dataset ($k = \Omega(\sqrt{dn \log(n)})$). However, they note that for high-dimensional data, the optimal k might be too large to use in practice. The authors thus focus on improving the robustness of 1-NN algorithms through sample selection, showing the effectiveness of their approach on the Halfmoon, mnist 1v7, and abalone datasets.

Yin et al. [155] show that transferring a robust solution found on training data to test data gets more difficult as the dimensionality of data increases. However, constraining the classifier weights mitigates this problem. Specifically, the authors prove that constraining the weights by L_p norm, for $p > 1$, leads to a performance gap between training and test data that has a polynomial dependence on dimensionality; when the weights are constrained by L_1 norm, the performance gap has no dependence on dimensionality.

Carbone et al. [23] study neural networks, showing that adversarial vulnerability arises due to the gap between the actual and intrinsic dimensionality, a.k.a., degeneracy. The authors show that adversarial example generations in high-dimensional degenerate data can be performed by using gradient information of a neural network, to move the samples in the direction normal to the data manifold. As such, example generation exploits the additional dimensions without changing the "semantics" of the perturbed sample. The authors then show that Bayesian Neural Networks are more robust than other neural networks to gradient-based attacks: due to their randomness, they make gradients less effective for crafting attacks.

[3] **Influence of Dimensionality on Defense Techniques**. High dimensionality also poses challenges to defense techniques that aim to improve robustness. Specifically, Blum et al. [18] focus on randomized smoothing – a technique that improves robustness by generating noisy instances of a (possibly perturbed) sample and then making predictions for the sample based on an aggregation of predictions for its noisy instances. The authors show that the amount of noise required to defend against L_p adversaries, for $p > 2$, is proportional to dimensionality. They further demonstrate that, for high-dimensional images, randomized smoothing indeed fails to generate instances that preserve semantic image information. In a similar line of work, Kumar et al. [75] show that the certified radius decreases as the dimensionality increases when using randomized smoothing for certifying robustness for a given L_p radius.

Adversarial training – a defense technique that improves model robustness by adaptively training a model against possible adversarial examples – often incurs a trade-off between standard and adversarial accuracy [139, 161]: optimizing for high robust accuracy results in a drop in standard accuracy and vice versa. Mehrabi et al. [95] build up on the work of Javanmard et al. [69], discussed

in Section 4.1. That work showed that, for a finite number of training samples, the trade-off between adversarial and standard accuracy improves as the number of samples per dimension increases. The authors further extend this result for unlimited training data and computational power, observing that, for an unlimited number of training samples, the trade-off between adversarial and standard accuracy improves as the dimension of the data decreases.

Data augmentation is another common defense technique that aims to improve robustness of a model by creating perturbed samples at radius r from a certain subset of original samples in training data. Rajput et al. [112] prove, for linear and certain nonlinear classifiers, that the number of augmentations required for robust generalization depends on the dimensionality of data, i.e., it is at least linearly proportional to dimensionality for any fixed radius r. Thus, data augmentation becomes more expensive for high-dimensional data.

4 **Reducing Dimensionality.** Following the idea that the gap between the actual and intrinsic dimensionality contributes to adversarial vulnerability, Awasthi et al. [11] propose to use Principal Component Analysis (PCA) [72] to decrease the dimensionality of data before applying randomized smoothing. As a result, a larger amount of noise can be injected to perturb samples, thus improving robustness without compromising accuracy. The authors apply the proposed ideas to image data, showing that the combination of PCA and randomized smoothing is more beneficial than using randomized smoothing alone. Weber et al. [146] show, for hierarchical data, that changing the representation from Euclidean to hyperbolic space reduces the dimensionality without sacrificing semantic information embedded in the input data.

4.3 Distribution

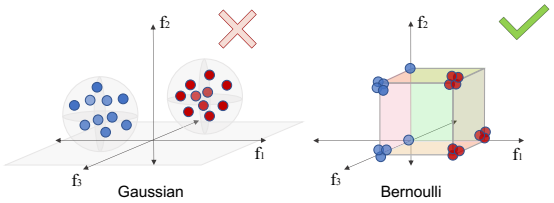


Fig. 12. Distribution illustration.

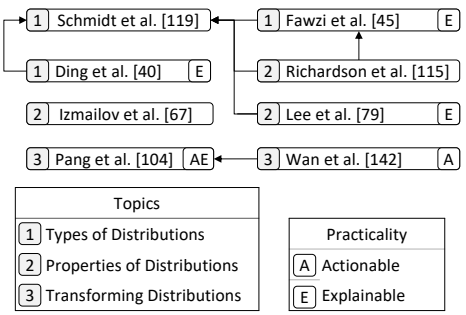


Fig. 13. Papers discussing distribution.

Distribution refers to a function that encodes how samples lie in space, usually by giving the probabilities of their occurrence in particular regions. Common types of distributions, such as uniform, Bernoulli, and Gaussian are introduced in Section 2.3. Fig. 12 shows examples of datasets that follow a Gaussian distribution (left) and a Bernoulli distribution (right). The term *variance* refers to a measure of dispersion that takes into account the spread of all data points in a dataset. Specifically, the *variance of a distribution* measures the dispersion of samples from the mean; *feature variance* measures the dispersion of samples over a particular feature only. We say that a distribution satisfies *symmetry* when distributions on either side of the mean mirror each other.

Papers that discuss how distribution properties, including variance and symmetry, influence models' robustness are shown in Fig. 13. They can be categorized into: 1 papers showing that model robustness depends on the underlying data distribution, 2 papers identifying properties of distributions that improve robustness, and 3 papers introducing techniques to transform distributions into ones that are more optimal for robustness.

[1] **Type of Distributions**. As discussed in Section 4.1, Schmidt et al. [119] prove, for nonlinear classifiers, that a mixture of Gaussian distributions incurs higher sample complexity for robust generalization than a mixture of Bernoulli distributions. Likewise, Ding et al. [40] show that a distribution shift alone can affect robust accuracy while retaining the same standard accuracy. Specifically, the authors prove that uniform data lying on a unit cube results in more robust models than uniform data lying on a unit sphere. They further experiment with MNIST and CIFAR-10 datasets, applying existing semantically lossless transformations, namely *smoothing* and *saturation*, to cause the distribution shift. The results of this experiment show that robustness decreases gradually when transforming MNIST from a unit-cube-like to a unit-sphere-like distribution and increases for CIFAR-10 when going the opposite way; in both cases, the models retain their standard accuracy.

Fawzi et al. [45] study the robustness of data distributions modeled by a smooth generative model – a type of generative model that maps samples from input space to output space while preserving their relative distances, e.g., to compress data. The authors show that smooth generative models with high-dimensional input space produce data distributions that make any classifier trained on this data inherently vulnerable. The authors conclude that non-smoothness and low input space dimensionality are desirable when modeling data with generative models.

[2] **Properties of Distributions**. Izmailov et al. [67] show that, in a binary classification setting, features with small variance in both classes and means close to each other cause adversarial vulnerability. Moreover, a feature with a small variance in one class can still cause vulnerability even if the means of this feature in both classes are farther separated but the second class has a larger feature variance. Intuitively, that is because models tend to assign non-zero weights to such features, which can be leveraged by attackers to shift the classification into the wrong class. That is, even small perturbations in such features can shift data points to another class. To increase robustness, the authors suggest removing such features, either based on domain knowledge or based on feature evaluation metrics, such as, mutual information [125].

Similarly, Lee et al. [79] prove that decreasing feature variance in individual classes can increase robustness for Schmidt's Gaussian mixtures. These mixtures have equivalent feature variances for all classes and separated means. In this setting, low feature variance implies that the feature has a strong correlation with the class and perturbing this feature unlikely result in vulnerability (i.e., will likely result in a semantically-meaningful change). However, even when features have low variance, if these features are non-robust [66], i.e., hold no semantic information, and have a smaller variance in the training data than in the underlying true population, they will still cause adversarial vulnerability as adversarially trained models tend to overfit to them. As a countermeasure, the authors propose a label-smoothing-based data augmentation technique which uses continuous instead of discrete values for labels and acts like a regularization method that prevents the model from overfitting to such features.

Richardson and Weiss [115] claim that adversarial vulnerability can be caused by sub-optimal data distributions and/or sub-optimal training methods. The authors define synthetic binary datasets (of images) that use Gaussian distributions with separated means and say that a dataset is symmetric if and only if classes have the same variance. They further prove that even the optimal classifier is non-robust when the underlying dataset has strong asymmetry, as in the example in Fig. 14. If the dataset is symmetric the optimal classifier is provably robust, even though a sub-optimal training method can still cause vulnerability when trained on this dataset.

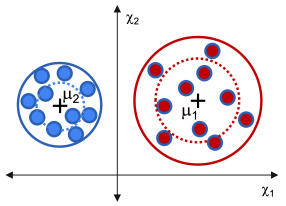


Fig. 14. Asymmetrical Dataset.

[3] **Transforming Distributions**. Both Pang et al. [104] and Wan et al. [142] change the latent DNN feature representation to be similar to Gaussian mixtures. Specifically, Pang et al. [104] show that,

for Linear Discriminant Analysis (LDA) classifiers trained on Gaussian mixtures, the robustness radius of LDA is proportional to the distance between Gaussian centers. The robustness of LDA is further maximized for symmetric Gaussian mixtures. The authors thus modify the DNN loss function to create a latent feature representation similar to symmetric Gaussian mixtures and further replace the last layer of DNN from commonly used Softmax Regression [32] to LDA. To achieve the desired robustness radius, the authors compute the coordinates of the desired Gaussian centers (as a function of the number of classes and the dimensionality of the input data) and feed this data to the loss function. Departing from the assumption that symmetric Gaussian mixtures are advantageous for the underlying model robustness, Wan et al. [142] modify the DNN loss function to compute the centers of the Gaussians directly while generating symmetric Gaussian feature distributions.

4.4 Density

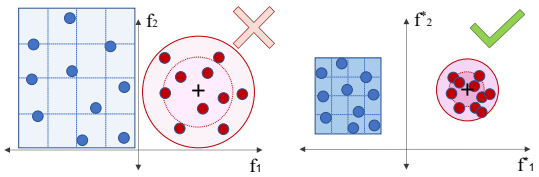


Fig. 15. Density illustration.

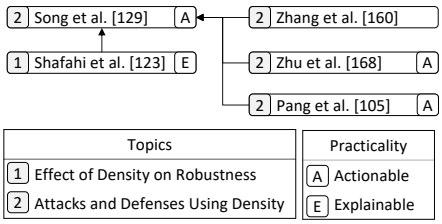


Fig. 16. Papers discussing density.

Density measures the closeness of samples in a particular bounded region. For continuous data, it is mathematically described by the probability density function (PDF), which gives the probability for a variable to take a certain range of values. For discrete data, it is described as the probability mass function (PMF), which gives the probability for a variable to take a particular value. We say that an area is dense when there is a high probability that random samples lie in the same area, i.e., close to each other. For example, the dataset on the right-hand side of Fig. 15 contains a larger number of samples in close proximity and, thus, is more dense than the dataset on the left-hand side of the figure. Furthermore, density can be defined over samples from one class, in which case, it is referred to as *class density*.

Papers that study how density influences adversarial robustness are shown in Fig. 16. They can roughly be divided into [1] papers discussing the effect of class density on robustness and [2] papers proposing attacks and defenses using density information.

[1] **Effect of Class Density on Robustness**. Shafahi et al. [123] show that datasets with a higher upper bound of class density lead to better robustness. In particular, for image datasets, the authors show that images of lower complexity, e.g., with simple objects on plain backgrounds, have a higher correlation among adjacent pixels. Datasets comprised of such images have a higher density, as pixel values are more frequently repeated, and, thus, lead to better robustness. The authors confirm this observation by showing that classifiers trained on MNIST, which has a lower image complexity and thus higher density than CIFAR-10, are more robust than those trained on CIFAR-10. Furthermore, the authors state that class density is a better predictor of robustness than dimensionality: even after up-scaling MNIST to the same dimensionality as CIFAR-10, it still has a higher density and thus results in more robust classifiers than CIFAR-10.

[2] **Attacks and Defenses Using Density**. Several works note that adversarial examples are commonly found in low-density regions of the training dataset, as models are unable to learn accurate decision boundaries using a small number of samples from these regions. Zhang et al. [160] propose an attack strategy that retrieves candidate samples from low-density regions and perturbs

them to generate adversarial examples. The authors demonstrate that, even after adversarial training, models will not be robust to adversarial attacks that target these low-density regions.

A similar finding by Zhu et al. [168] suggests that adversarial examples from low-density regions have a higher probability of being transferable between different models trained on the same dataset. Based on this observation, the authors propose an attack that increases the transferability of adversarial examples by identifying perturbation directions that maximize both the adversarial risk and the alignment with the direction of density decrease for the underlying data distribution, i.e., move samples towards regions with lower density.

Departing from the same idea that low-density regions are prone to adversarial attacks, Song et al. [129] focus on creating a defense mechanism that uses generative models to detect if a sample comes from a low-density region when making predictions. If so, the sample is moved towards a more dense region of the training data as a "purification" step.

To harden models directly, Pang et al. [105] propose a new loss function for DNNs, to learn dense latent feature representations. The authors first show that the commonly used Softmax Cross-Entropy loss function induces sparse representations (i.e., with low class density), which lead to vulnerable models. This is because a low number of samples in close proximity to each other prevent a model from learning reliable decision boundaries. They then propose a loss function that explicitly encourages feature representations to concentrate around class centers; like in their earlier work [104], the authors compute the coordinates of the desired class centers (as a function of the number of classes and the dimensionality of the input data) to maximize the distances between the centers. The authors demonstrate that the proposed approach improves robustness under both standard and adversarial training.

4.5 Separation

Closely related to density, *separation* refers to the distance between classes. Fig. 17 shows examples of not well-separated (top) and well-separated (bottom) datasets. Intuitively, learning an accurate classifier is easier when data is well-separated as samples from different classes are farther apart and samples from the same class are closer together. Different metrics to quantify separation include the *optimal transport distance*, which computes the minimum distance required to transport samples from one class to another, and *inter-class distance*, which computes distance between samples in different classes.

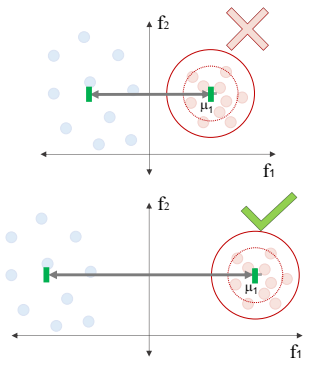


Fig. 17. Separation illustration.

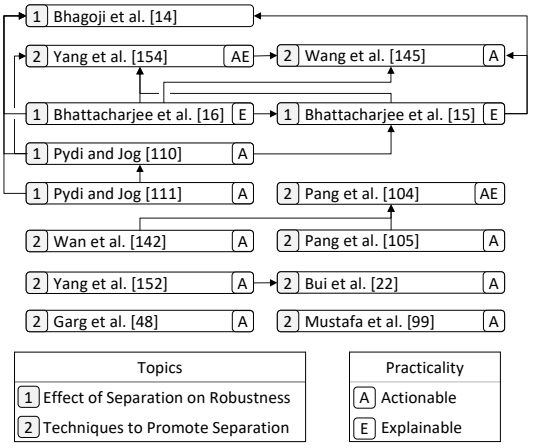


Fig. 18. Papers discussing separation.

Papers that discuss data separation in relation to adversarial robustness are shown in Fig. 18. They can roughly be divided into [1] papers showing the effect of separation on robustness and [2] papers proposing techniques to promote separation and, thus, increase robustness.

[1] **Effects of Separation**. Bhagoji et al. [14] calculate lower bounds for adversarial risk in a binary classification setting using the optimal transport distance. The authors show that the lower bound decrease as the distance between the two class distributions increases, i.e., a classifier becomes more robust with better separation. Based on this result, they estimate the minimum adversarial risks for image datasets, like MNIST and CIFAR-10, showing that the theoretically calculated risks are lower than the empirical values achieved by the state-of-the-art defense models. The authors conclude that there is still room for improving existing techniques.

Pydi and Jog [110, 111] arrive at a similar conclusion – that robustness improves as separation between classes increases. The authors further focus on datasets with simple univariate distributions, such as Gaussian and uniform. They propose a technique to construct classifiers that can achieve the optimal, lowest possible adversarial risk for a given separation between classes. The main idea behind this technique is to analyze the optimal way to transport samples from one class to another (which represents the smallest perturbation needed to create adversarial examples) and further use this information to identify the decision boundary that induces the maximal distance required to transport sample between classes. That is, the approach maximizes the distance between samples of each class and the decision boundary, resulting in an optimally robust classifier.

Bhattacharjee et al. [15] prove that certain non-parametric models, such as k-NNs, are inherently robust when trained on a large number of well-separated samples. This is because these classifiers make predictions based on neighborhoods and well-separated data ensures that samples in close proximity to each other share the same labels. In their later work, discussed in Section 4.1 [16], the authors show that, in well-separated data, robust accuracy is independent of dimensionality and a robust linear classifier can be learnt without the need for a large number of training samples. This result shows that adversarial vulnerability can be efficiently tamed by increasing separation.

[2] **Techniques to Promote Separation**. Yang et al. [154] propose a sample-selection-based technique to improve adversarial robustness of non-parametric models by increasing the separation among the training data. In particular, as non-parametric models tend to learn complex decision boundaries when the training samples from different classes are close to each other, the authors propose to remove the smallest subset of samples so that all pairs of differently labeled samples remain separated even when perturbed by the maximum perturbation size. Wang et al. [145], already discussed in Section 4.2, focus on improving robustness of 1-NN classifiers. As robustness of such classifiers is optimized by ensuring the opposite classes being far apart and test points ending up in close proximity to their respective training data, the authors propose retaining the largest subset of training samples that are (i) well-separated and (ii) in high agreement on labels with their nearby samples (a.k.a., highly confident). The authors show that their approach outperforms adversarially trained 1-NNs.

Another line of works proposes techniques to enforce separation in latent representations to improve the adversarial robustness of DNNs. Specifically, Mustafa et al. [99] attribute the cause of adversarial vulnerability to close proximity of classes in latent space. Hence, they propose a loss function to learn intermediate feature representations that separate different classes into convex polytopes, i.e., polyhedra in higher dimensions, that are maximally separated. Bui et al. [22] observe that the adversarial vulnerability of DNNs arises from a large difference in intermediate layer values between clean and adversarial data. They thus propose to modify the loss function so that it results in an intermediate latent representation that has high similarities between clean and their corresponding adversarial samples, while promoting large inter-class distance and small intra-class

distance. In addition, the proposed loss function aims to increase margins from class centers to decision boundaries. Likewise, Pang et al. [104] and Wan et al. [142] discussed in Section 4.3, as well as Pang et al. [105] discussed in Section 4.4, improve DNN robustness by separating centers of the produced latent distributions and, thus, increasing the separation between classes.

Yang et al. [152] propose a representation-learning technique to learn feature representations that bring samples of class C and adversarial examples generated for C into close proximity while separating the samples of C from both (i) adversarial examples generated for other classes and misclassified as class C and (ii) samples from other classes. These separations are enforced by the loss function proposed by the authors. The authors show that their approach improves the resulting model robustness compared with standard DNNs.

Garg et al. [48] propose an approach to generate well-separated features for a dataset using graph theory. Specifically, they convert the input dataset into a graph, where vertices correspond to the input data points and edges represent the similarity between the data points (e.g., calculated using Euclidean distance). The authors prove that features extracted using the eigenvectors of the Laplacian matrix capturing the structure of the graph will have significant variation across the data points, while being robust to small perturbations. These qualities make them good candidates for robust features. The authors then demonstrate that a linear model trained on the MNIST dataset with 20 features generated using their approach is more robust to L_2-norm-based transfer attacks than a fully connected neural network trained on the full pixel values of the MNIST dataset.

4.6 Concentration

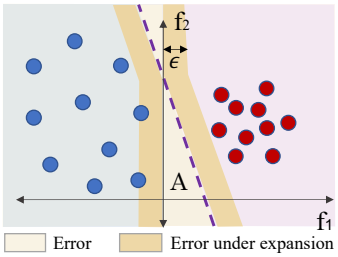


Fig. 19. Concentration illustration.

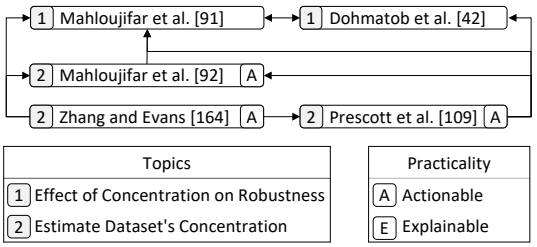


Fig. 20. Papers discussing concentration.

Concentration of a dataset refers to the "concentration of measure" phenomenon from measure theory [131]. In a nutshell, concentration is the minimum value of a measured function over all valid measurable sets, after an ϵ-expansion. More formally, for a metric probability space $(\mathcal{X}, \mu, d)$ with instance space $\mathcal{X}$, probability measure μ, and distance metric d, the concentration function h is defined as: $h(\mu, \alpha, \epsilon) = \inf_{A \subseteq \mathcal{X}} \{\mu(A_\epsilon) : \mu(A) \geq \alpha\}$ for any $\alpha \in (0, 1)$ and $\epsilon \geq 0$ [92]. Here A_ϵ refers to the ϵ-expansion of set A, defined as $A_\epsilon = \{x : d(x, A) \leq \epsilon\}$.

Fig. 19 shows how the concentration of measure phenomenon can be used to determine the classification error after adversarial perturbation. By modeling the classification error set as measurable set A and adversarial errors from perturbation budget ϵ as A_ϵ, one can relate the concentration of the data to the minimum adversarial risk for any imperfect classifier with error rate $\mu(A) \geq \alpha$. Using this formulation, a dataset being highly concentrated implies that, for some non-zero initial error, the minimum adversarial risk from an ϵ-expansion on the error set is very large. We refer to such datasets as datasets with low *intrinsic robustness* – a measure that represents the maximal achievable robustness for any classifier on a dataset.

Fig. 20 shows the papers that relate data concentration to adversarial robustness. They can roughly be divided into: [1] papers discussing the effect of concentration on robustness and [2] papers proposing techniques to estimate robustness through calculating concentrations.

[1] **Effect of Concentration on Robustness**. A number of papers prove the inevitability of adversarial examples using the concentration of measure phenomenon. In particular, Dohmatob [42] investigates datasets conforming to uniform, Gaussian, and several other distributions that satisfy W_2 transportation-cost inequality [132]. The author proves that data distributions satisfying such inequality have high concentration, which results in a rapid robustness decrease, beyond a critical perturbation size – a value that depends on the standard error of the classifier and the natural noise level of the dataset, which, in turn, defined as the largest variance in the case of Gaussian distribution. Even though MNIST might not satisfy the W_2 transportation-cost inequality, the author experiments with this dataset, observing a sudden drop in robustness as the perturbation size increases. As such, the author suggests that the MNIST dataset may also have high concentration and be governed by the concentration of measure phenomena.

Mahloujifar et al. [91] focus on a collection of data distributions with high concentration called Lévy families [80], which include unit sphere, unit cube, and isotropic n-Gaussian (i.e., Gaussian with independent variables with the same variance). The authors prove that classifiers trained on such highly-concentrated data distributions admit adversarial examples with perturbation $O(\sqrt{d})$ for dimensionality d. This implies that a relatively small perturbation can mislead model trained on these data distributions with high dimensional inputs.

[2] **Techniques to Estimate Robustness Through Concentrations**. Several approaches utilize the connection between concentration and adversarial risk to estimate the intrinsic robustness of datasets through calculating their concentrations. Mahloujifar et al. [92] are the first to propose an approach for estimating dataset concentration using subsets of samples. Specifically, the authors propose a technique that searches for the minimum expansion set based on a collection of subsets carefully chosen according to the perturbation norm (e.g., a union of balls for L_2 norm). They prove that the estimated concentration value converges to the true value for the underlying distribution as the sample size and the quality/representativeness of the chosen subsets increase. The authors apply their approach to estimate the maximum achievable robustness for the MNIST and CIFAR-10 datasets, observing a gap between the derived theoretical values and values observed empirically by the state-of-the-art models.

In follow-up work, Prescott et al. [109] propose an alternative approach to estimate concentration based on half space expansion using *Gaussian Isoperimetric Inequality* for the L_2 norm [20]. The authors further generalize their results to L_p norms, where $p \geq 2$. Compared with Mahloujifar et al. [92], their approach yields higher achievable robustness on MNIST and CIFAR-10, revealing a larger gap between the theoretical robustness and the state-of-the-art. As the theoretically achievable robustness derived from a concentration perspective is shown to be high, the authors suggest that factors other than concentration may contribute to this gap.

Zhang and Evans [164] assume access to information about label uncertainty, i.e., function that assigns the level of label uncertainties for any data point. Such function can use, e.g., labeling results from multiple human annotators or confidence scores from an ML classifier. The authors suggest that considering regions with high label uncertainty can guide the concentration estimations as these are the regions where a classifier is more likely to make mistakes and be vulnerable to attacks. They thus propose an approach to estimate concentration by identifying the smallest set after ϵ-expansion with an average uncertainty level greater than a pre-set value. The evaluation results show that the maximum achievable robustness estimated with their approach is closer to the robustness values observed for CNN models on the CIFAR-10 dataset than in any of the aforementioned works, implying that the room for improvement is smaller than assumed earlier.

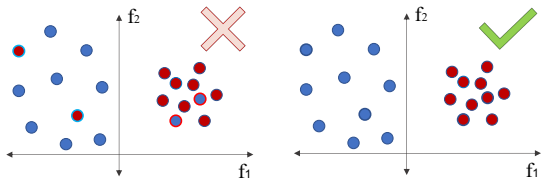


Fig. 21. Label noise illustration.

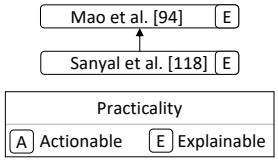


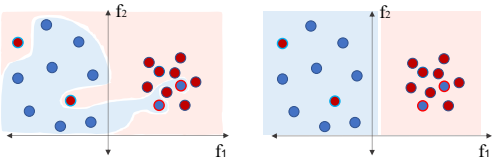
Fig. 22. Papers discussing label quality.

4.7 Label Quality

Label quality refers to the correctness and informativeness of the set of labels assigned to a training dataset. Label correctness or, inversely, the presence of inaccurate labels (shown as the highlighted dots on the left-hand side of Fig. 21) is typically referred to as *label noise*. The granularity of the labels is typically referred to as *label informativeness*. Papers that discuss the relationship between label quality and model robustness are outlined in Fig. 22.

Mao et al. [94] show that training a model simultaneously for multiple tasks, e.g., to simultaneously locate and estimate the distance of objects in images (an approach also referred to as *multi-task learning*), improves robustness. This is because in multi-task learning, a model learns a shared feature representation by training on data with labels from several tasks. As a result, perturbations required to attack multiple tasks at the same time, e.g., to sabotage an autonomous driving system by misleading the model in both object identification and distance estimation, cancel each other out. While the authors prove that the model robustness to adversarial attacks is proportional to the number of tasks that it is trained on, the benefits of multi-task learning disappear when the concurrently trained tasks are highly correlated with each other, as it reduces the chances for the perturbations to cancel each other. The authors further show that training with multiple tasks also improves model robustness against single-task attacks.

Sanyal et al. [118] hypothesize that label noise and coarse labels are the reasons for adversarial vulnerability. The authors prove that, given a large training set with random label noise, any classifier that overfits to that set is likely to be vulnerable to adversarial attacks. This is because overfitting leads to overly complex decision boundaries that leave more room for attacks, as illustrated in Fig. 23. The authors also demonstrate that adversarial risk increases as the level of label noise increases. Defense mechanisms, such as early stopping and adversarial training, enhance robustness by preventing models from overfitting to noisy samples. In the absence of label noise, using coarse labels (e.g., labels for the entire class of dogs rather than labels for each individual dog breed) results in "sub-optimal" latent feature representations and also contributes to the adversarial vulnerability.



(a) Overfit (b) Not overfit.

Fig. 23. Influence of overfitting to label noise.

4.8 Domain-Specific

Papers in this category provide insights into the correlation between domain-specific data properties and adversarial robustness. Among our collected papers, all the domain-specific studies focused on the same topic: understanding the adversarial vulnerabilities of image classifiers based on image *frequency* – how fast the intensity of pixel values changes with respect to space (i.e., images with intensive color changes have high frequency). As shown in Fig. 24, the skin of a zebra has higher image frequency than a horse, because of the black-and-white stripes. Papers studying image frequency are listed in Fig. 25. They can be roughly divided into: 1 papers discussing the influence of frequency distribution on the model adversarial robustness, and 2 papers explaining adversarial vulnerabilities using perceptual differences between human and models.

Fig. 24. Image frequency.

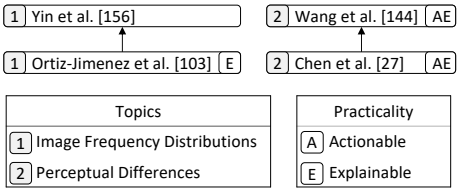


Fig. 25. Papers discussing domain-specific properties.

1 **Image Frequency Distribution.** Yin et al. [156] show that frequency distribution of the inputs generated from data augmentation techniques explains the resulting model sensitivity to adversarial attacks. In particular, the authors show that Gaussian augmentation, as well as adversarial training techniques that rely on data augmentation, generate perturbations with high-frequency components. Augmenting training data with those augmented inputs makes the resulting model more robust against perturbations in high-frequency domains while, at the same time, more vulnerable to perturbations concentrated in low-frequency domains. To mitigate this issue, the authors propose to avoid biasing the model towards/against certain frequency ranges by increasing the diversity of frequency distribution in augmentations.

Ortiz-Jimenez et al. [103] analyze CNN robustness through the classifier margins along particular frequencies, which the authors define as the minimal perturbation in that frequency required to change the model prediction. The authors show that CNN models tend to have smaller margins along low-frequency vs. high-frequency ranges, likely because, for most image datasets, one can differentiate classes mainly using features from low-frequency ranges. This finding implies that models are more sensitive to attacks that modify low-frequency components. The authors thus suggest training more robust models that enlarge margins along low-frequency ranges by augmenting training datasets with perturbations concentrated in those ranges.

2 **Perceptual Differences.** Wang et al. [144] attribute the origin of adversarial examples to the perceptual differences of humans and CNNs in frequency ranges. In particular, humans classify images based on low-frequency components as high-frequency components are not visible to the human eye. CNNs, on the other hand, are able to 'see' the full frequency spectrum which allows them to exploit high-frequency components for better generalization. This implies that adversarial examples generated by perturbing high-frequency components can mislead CNNs while being imperceptible to humans. The authors show that adversarially robust models depend less on high-frequency components and propose to use smoother convolutional filters to reduce a model's attention to these components.

Unlike Wang et al. [144], Chen et al. [27] posit that the adversarial vulnerability of CNNs results from their over-reliance on amplitude information of images – the magnitude of the different frequencies in the image. The authors show that replacing the amplitude information of an image with information from another image can successfully mislead CNNs but not humans, who rather rely on phase information – the locations of the features, to recognize objects. Based on this observation, the authors propose to strengthen CNNs' attention to phase information through a data augmentation technique that fuzzes amplitude while preserving the same phase information.

4.9 Summary of Results

Overall, the surveyed papers are mostly in agreement on how each of the identified data property influences adversarial robustness. The main findings are given below.

Number of samples. More training samples are needed for robust than for standard generalization. For a variety of training setups (i.e., different types of classifiers and data distributions), the number

of training samples required to achieve robust generalization is proportional to the dimensionality of the training data. Unlabeled samples or generated data can be used to fulfill the need for more samples needed for robust generalization, i.e., to close the sample complexity gap. Class imbalance, i.e., having an imbalanced number of samples across different classes, hurts robust generalization due to the model bias towards over-represented samples.

Dimensionality. Dimensionality captures the size of the feature set. Higher dimensionality correlates with higher adversarial risk, worse standard-to-adversarial risk trade-off, difficulty in robustness certification, and difficulty in applying common defense techniques. This is because adversarial attacks can exploit the excessive dimensions to construct adversarial examples.

Distribution. Some data distributions are more robust than others, e.g., mixtures of Bernoulli distributions are more robust than mixtures of Gaussian distributions. Learning feature representations that resemble robust distributions can improve robustness.

Density. Density reflects the closeness of samples in a particular bounded region (inter-class distance). Adversarial examples are commonly found in low-density regions of data, where samples are far apart from each other. This is because models cannot accurately learn decision boundaries near low-density regions due to the small number of samples available. As such, high data density for each class correlates with lower adversarial risk.

Separation. Separation characterizes the distance of samples from different classes to each other (inter-class distance). Greater separation between classes decreases adversarial risk as it is harder to generate perturbation that will cross the boundaries between classes. Most papers that provide techniques for improving separations, e.g., by feature selection, also ensure that it does not come at the expense of decreasing density, as these two concepts are closely related.

Concentration. Given a function defined over a non-empty set, concentration (from the phenomenon of concentration of measure [131]) is the minimum value of the function after expanding the input set by ϵ in all dimensions. For example, expanding the set of misclassified samples by a certain ϵ gives a set of possible samples that can be misclassified with an ϵ-size perturbation (candidate adversarial examples). Concentration, in this case, measures the minimal possible size of this set, which provides the upper bound of the achievable model robustness. As some datasets tend to exhibit inherently high concentration, e.g., datasets that lie on unit hypersphere [91], achieving high robust generalization is harder for these datasets. The impact of high concentration on adversarial robustness is further magnified for high-dimensional data.

Label quality. High label noise correlates with higher adversarial risk. More specific labels, e.g., "cat" and "dog" instead of "animal", are more robust than coarse labels, as they allow the model to extract more distinct features. Learning for different tasks concurrently, e.g., to simultaneously locate and estimate the distance of objects in images, improves robustness of the learned models, as the model can utilize the information from multiple sources of data.

Domain specific. Image frequency, i.e., the rate of change in pixel value is shown to be correlated with robustness. Specifically, a diverse distribution of frequencies in training data results in lower adversarial risk. Most image datasets have a low image frequency, which results in learned models having smaller margins and, thus, a lower distance of samples to the decision boundary for features corresponding to the low-frequency image components. This increases the risk for adversarial perturbation that utilize these features.

5 IMPLICATIONS AND FUTURE WORK

Empirical Evaluation. All but the four papers outlined in Section 4.8 study domain-agnostic data properties. Yet, the majority of the papers we surveyed conduct experimental evaluations on image datasets only. Applicability of the findings and of the proposed approaches to other domains, with different forms of data, may need further investigation. For example, for datasets with binary features, which are commonly used in malware detection, one cannot arbitrarily change feature values to reduce the distance between samples. This further implies that common distance metrics used to model adversaries in the image domain, such as L_2 and L_∞, fail to accurately capture the adversarial threat level in such domains. Hence, future work applying, adapting, and evaluating the proposed metrics and techniques in other domains and data types is needed.

Interdependence of Properties. Only a few works in our collected literature consider multiple data properties simultaneously or establish interdependence of data properties. For example, Wang et al. [145] and Rajput et al. [112] find that the number of samples and dimensionality collectively influence the performance of the resulting model. Sanyal et al. [118] study the tolerable amount of label noise as a function of the dataset density. Such works are very valuable as adversarial robustness is indeed a result of compounding properties. Yet, optimizing for multiple properties simultaneously is not always possible. A productive direction of future work could be to investigate correlations between different data properties, e.g., the effects of feature dimensionality reduction approaches on class density and separation.

Additional Data Properties. Existing research on the effects of data on *standard generalization* [87, 88, 117] identified several data properties not discussed in the papers related to robust generalization that we reviewed. These include the presence of (i) outliers, i.e., samples that drastically differ from most observed samples in a dataset, (ii) overlapping samples, i.e., different samples of the dataset having the same feature representation, and (iii) small disjuncts, i.e., training samples from the same class forming small disjoint clusters dispersed throughout the input space (more details are in Section 6). Investigating the effect of such data properties on the model's adversarial robustness could be yet another direction for possible future work.

Generalization. Some works have shown that, under specific assumptions about the data and model, there is an intrinsic accuracy-robustness trade-off [69, 95, 139]. This implies that achieving robust generalization may come at cost of standard generalization. However, other works have shown that the effect of several of the data properties on standard generalization overlaps with their effect on robust generalization. For example, increasing class density and removing label noise also increases standard generalization [47, 62, 87, 113]. We believe that more work is needed to map data-related reasons that contribute to the accuracy-robustness trade-off.

Simplified Problem Setup. Several studies use a simplified problem setup, e.g., pure Gaussian data distribution, to provide formal proofs related to the studied phenomenon. While such work helps advance knowledge and our understanding of the effects of data on adversarial robustness, additional work that investigates the generalizability of the findings on realistic datasets used in practice is needed. For example, assuming uniform data properties, e.g., same distribution, density, and level of label noise, for all classes on the training data greatly simplifies the proofs, but is not common in reality. Likewise, considering only binary classification simplifies calculations of data separation, which can be calculated by measuring the distance between the two classes. Yet, in a multi-class setting, one needs to consider the proximity of data points from multiple classes.

Furthermore, most papers only consider a white-box attack setting, which might not be realistic in many practical scenarios. Even though a white-box setting makes it possible to model the worst-case adversary and to provide better robustness guarantees, it may result in overly pessimistic

findings, i.e., some data transformations may be robust against black-box attacks while still be vulnerable to white-box attacks. Thus, future works might look into the impact of data properties on the different types of attack scenarios.

Quantitative Measure. Literature shows that the lower/upper bound of adversarial robustness can be determined by the properties of the underlying data [14, 92]. Modifying certain properties of the data can also change the robustness of the resulting classifier. Hence, the ability to quantitatively measure such data properties is very valuable. However, some data properties discussed in this survey, such as, type of distributions and label noise, lack any reliable estimation techniques. Current work mostly relies on informal comparative analysis, e.g., that the MNIST dataset is closer to a Bernoulli mixture data than a Gaussian mixture because the pixels are concentrated towards black or white. Quantitatively measuring the degree of similarity between distributions, although difficult, may be necessary in order to make more accurate conclusions.

Interestingly, other data properties have multiple, often inconsistent, measurement techniques, e.g., concentration [92, 109, 164], density [105, 129], intrinsic dimensionality [9, 88], and inter-class distances [14, 40, 110]. For example, the inter-class distances can be calculated as the total distance required to move the samples from one class to another [14, 110]. It can also be calculated as the pairwise distances between a pre-defined portion of samples from different classes, e.g., 10% from each class, that are the closest to each other [40]. While the inter-class distance derived through the first approach is more computationally expensive, the second approach is more susceptible to outliers as it relies only on a subset of samples close to each other. Moreover, these metrics might not necessarily correlate with each other. We believe future research can provide more insights about appropriate application scenarios for each of the proposed metrics.

Sources of Adversarial Vulnerability. Even when the training data is optimal for robustness, a sub-optimal training method can lead to adversarial vulnerability [115]. For example, adversarial vulnerability may arise when the complexity of the classifier does not match the complexity of the data, e.g., CNNs may achieve lower robustness due to their complexity than simpler models, such as Kernel-SVMs, on symmetrical data with well-separated means and similar variances [115]. To alleviate such problems, a few papers propose to select, improve, or optimize classifiers based on the dimensionality of data [23, 145, 155]. Similar work that looks at other properties of data, such as separation and density, could be of value. Future works could also explore strategies for determining whether the input data (vs. the model itself) is the dominant cause of adversarial vulnerability.

6 RELATED WORK

We divide related work into three categories: (1) surveys on adversarial robustness and its relation to data properties, (2) surveys that discuss the influence of data properties on standard generalization, and (3) individual papers that study non-data-related reasons for adversarial vulnerability.

6.1 Surveys on Adversarial Robustness

Numerous existing surveys focus on attack and defense techniques for adversarial robustness [7, 8, 17, 38, 81, 83, 85, 89, 93, 116, 121, 163]. Only a few of these works mention the relationship between adversarial robustness and properties of the underlying data. Specifically, Serban et al. [121] observe that adversarial vulnerability can be caused by an insufficient training sample size and high data dimensionality. Similarly, Machado et al. [89] mention that the lack of sufficient training data, high-dimensional data, and high concentration contribute to adversarial vulnerability. Yet, none of these surveys explicitly collect and analyze work that focuses on the effects of data properties on adversarial robustness. By explicitly targeting this topic in our survey, we are able to discuss these

findings in detail and also identify additional relevant data properties not mentioned in previous surveys, such as, types of distribution, class density, separation, and label quality.

6.2 Influence of Data Properties on Standard Generalization

A number of surveys investigate the influence of data properties on standard rather than robust generalization. One of the earliest is probably the work of Raudys and Jain [113], who review studies related to the influence of sample size on binary classifiers, showing that a limited sample size usually leads to sub-optimal generalization. Bansal et al. [12] and Bayer et al. [13] also survey papers addressing the data scarcity problem, focusing in particular on the recent advancements in data augmentation techniques in the fields of computer vision, security, and text classification. Their results show that augmentation techniques can help improve a model's generalization by reducing the problem of model overfitting.

Label noise is another aspect of data that influences both standard and robust generalization. Most works on this topic find that the presence of noisy labels increases the need for a greater number of training samples and may result in unnecessarily complex decision boundaries [47, 128]. For example, Frénay and Verleysen [47] show that overfitting to label noise greatly degrades a model's standard generalization; the same effect has been observed for the case of robust generalization [118]. Song et al. [128] survey the impact of label noise in deep learning, arguing that the presence of noisy labels is a more serious concern for deep models as they contain a larger number of parameters which makes them prone to overfitting to the noise in training data. They mention that adversarial defense techniques, e.g., adversarial training, are effective against label noise [44, 167] but do not discuss how label noise influences a deep learning model's robustness under attacks.

Lorena et al. [88] identify a collection of 26 quantitative metrics that measure data complexity with respect to (1) ambiguity of classes, i.e., whether the classes can be clearly distinguished with the given features, (2) sparsity and dimensionality of data, i.e., whether enough information are provided to learn confident decision boundaries, and (3) complexity of boundary separating the classes, i.e., whether more intricate functions are required to describe the decision boundaries. The authors also discuss how these metrics help estimate the difficulty of performing classification on a given dataset. Similar to our survey, the authors show that high dimensionality and small separation between classes hinder standard generalization. However, the relationship of some of the metrics reviewed by these authors, e.g., the number of non-intersecting spheres needed to enclose all data points of a class, to robust generalization is not studied, according to our survey.

He and Garcia [62] focus on the imbalance learning problem – the disproportion in the number of samples belonging to each class in a given dataset. The authors found that most standard algorithms are designed with the assumption of a balanced class distribution. These algorithms fail to reliably represent the distributive characteristics of the imbalanced size of samples and result in unfavorable performance across classes. Furthermore, López et al. [87] present a thorough discussion on six intrinsic data characteristics that potentially complicate learning from imbalanced data: low density, sample overlap between classes, noisy data, borderline instances, dataset shift between training and testing distributions, and small disjuncts, i.e., disperse small clusters of samples from a single class. Their analysis concludes that while all these "unfavorable" data characteristics further complicate the data imbalance issues, data overlap between classes is probably one of the most harmful. To follow up on this point, Santos et al. [117] focus on the joint effect of data imbalance and class overlap on model generalization. The negative impact of data imbalance, low separation, and noisy data on robust generalization was also discussed in our survey. Yet, the compounding effect of these factors, as well as the effect of other properties, on robust generalization needs future investigation.

Recently, Yang et al. [151] summarized relevant studies focusing on long-tailed distributions in the field of Computer Vision. This survey also includes work on the influence of long-tail

distributions on a model's adversarial robustness [147], which is covered in our survey. The authors advocate for more research on adapting long-tailed-based approaches for standard generalization to improve robust generalization.

Finally, Moreno-Torres et al. [98] present a unifying framework to categorize existing definitions of dataset shift – the case where the joint distribution of inputs and outputs differs between training and testing data. While ML models are normally trained under the premise that testing data has a similar distribution to the training data, in reality, the observed data distribution may be different from the historical data that the model is trained on. Such difference can substantially compromise the quality of model predictions. The authors analyze the possible causes for dataset shift, e.g., malicious software that evolves over time, and review the techniques dealing with dataset shift. They characterize adversarial attacks as one form of dataset shift, where adversaries adaptively change test instances to create a distribution that differs from training data.

6.3 Non-data Related Reasons for Adversarial Vulnerability

There has been a variety of hypotheses regarding the reasons behind adversarial vulnerability of ML systems. In addition to the data used for training, adversarial robustness could also depend on the choice of the model architecture, the training procedure, and the interplay between data and the learning algorithm, i.e., correspondence between the complexity of a model to that of the data. This section summarizes the key hypotheses regarding these aspects.

Model. When Szegedy et al. [130] first discovered adversarial examples for visual models, they suspected the high non-linearity of DNNs resulted in low probability 'pockets' of adversarial examples in the learned representation manifold. They hypothesize that while these pockets can be found through attack algorithms, the samples residing in these pockets have different distributions compared to normal samples and are thus subsequently harder to find when randomly sampling from the input space. Instead, Goodfellow et al. [55] hypothesize that the linearity from activation functions, like ReLU and sigmoid found in high-dimensional neural networks, induce vulnerability towards adversarial perturbations. To support their claim, they present the attack method FGSM that exploits the linearity of the target classifier. Fawzi et al. [46] also argue against the hypothesis of high non-linearity as the cause for adversarial examples. They show that all classifiers are susceptible to adversarial attacks and claim that it is the low flexibility of the classifier compared to the complexity of the classification task that results in vulnerability. The lack of consensus on primary causes of models' vulnerability invites more studies on this topic.

Singla et al. [127] show that enforcing invariance to circular shifts (e.g., rotation) in neural networks induces decision boundaries with a smaller margin than normal, fully connected networks, which, in turn, reduces the adversarial robustness of the model. Moosavi-Dezfooli et al. [97] introduce universal, input-agnostic perturbations to mislead the classifier and hypothesize that the vulnerability of a multi-class classifier to such perturbations is related to the shape of its decision boundaries, e.g., linear classifiers with decision boundaries that are parallel to each other and nonlinear classifier with decision boundaries that are curved in a similar way tend to be less robust as perturbations in one direction can change the prediction label for a different class.

Tanay and Griffin [134] conjecture that the decision boundary learned by the classifier being too close to (or 'tilted towards') the data manifold instead of being perpendicular to it, results in small perturbations being sufficient to move samples across the decision boundary for misclassification.

Computational Resources. Bubeck et al. [21] use computational hardness theory to show that the time complexity for learning a robust model is exponential to the size of input data and thus is computationally intractable. Hence, they attribute adversarial vulnerability to computational limitations of current learning algorithms. Degwekar et al. [37] further extend this work and also show the impossibility of efficiently training robust classifiers.

Robustness of Features. Ilyas et al. [66] show that adversarial vulnerability can be a consequence of a model exploiting well-generalizing but non-robust features, i.e., features that are spurious and sometimes incomprehensible to humans; when constraining the model to use robust features, the adversarial robustness increases together with the interpretability of the learned features. However, Tsipras et al. [139] note that, as the features for achieving high accuracy may be different from the ones for achieving high robustness, robustness may be at odds with standard accuracy. Instead of seeing adversarial vulnerability as a product of classifiers being overly sensitive to changes in spurious features, Jacobsen et al. [68] hypothesize that classifiers can rather be overly insensitive to relevant semantic information, e.g., images with drastically different content can share similar latent representations. The authors introduce a new type of adversarial examples that exploit such insensitivity, where the content of images is altered without changing the resulting prediction label. While all these works propose possible reasons for adversarial vulnerabilities, they are orthogonal to our survey, which focuses particularly on the influence of training data.

7 CONCLUSION

In this survey, we systematically collected, analyzed, and described papers that discuss how data properties affect adversarial robustness in machine learning models. By analyzing 57 research papers from top scientific venues in Machine Learning, Computer Vision, Computational Linguistics, and Security, we identified seven domain-agnostic data properties and one image-specific data property that are correlated with adversarial robustness.

While several of the guidelines for constructing high-quality data that we identified are similar to those recommended for training accurate models, producing robust models is more sensitive to the characteristics of the data and requires more effort, e.g., a larger number of samples, better label qualities, etc. There are also additional data properties important for building robust models that are not extensively discussed in non-adversarial settings, e.g., concentration of measure. In a sense, robust generalization is a stronger form of standard generalization.

We identified possible next steps towards improving the understanding of how the data affects a model's adversarial robustness. These include studying interactions between different properties of data, considering the effect of additional properties that improve standard generalization on robust model generalization, devising quantitative metrics for different aspects of the data, and extending the studies and their empirical evaluation beyond the images domain. We hope our survey will help researchers and ML practitioners to better understand adversarial vulnerability and will spark further research to address the identified knowledge gaps.

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] Naveed Akhtar and Ajmal Mian. 2018. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access* 6 (2018), 14410–14430.
[8] 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.
[9] 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.
[10] 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.

[11] 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.

[12] 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.

[13] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A Survey on Data Augmentation for Text Classification. *Comput. Surveys* (2022).

[14] 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)*.

[15] Robi Bhattacharjee and Kamalika Chaudhuri. 2020. When Are Non-Parametric Methods Robust?. In *International Conference on Machine Learning (ICML)*. 832–841.

[16] 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.

[17] Battista Biggio and Fabio Roli. 2018. Wild Patterns: Ten Years after The Rise of Adversarial Machine Learning. *Pattern Recognition* 84 (2018), 317–331.

[18] 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.

[19] Giuseppe Bonaccorso. 2017. *Machine Learning Algorithms: A Reference Guide to Popular Algorithms for Data Science and Machine Learning*. Packt Publishing.

[20] Christer Borell. 1975. The Brunn-Minkowski Inequality in Gauss Space. *Inventiones mathematicae* 30 (1975), 207–216.

[21] 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.

[22] 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.

[23] 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.

[24] Nicholas Carlini and David A. Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *Symposium on Security and Privacy (SP)*. 39–57.

[25] 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)*.

[26] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. Adversarial Attacks and Defences: A Survey. *ArXiv* (2018).

[27] 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.

[28] Pin-Yu Chen and Cho-Jui Hsieh. 2023. *Adversarial Robustness for Machine Learning*. Academic Press.

[29] 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.

[30] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified Adversarial Robustness via Randomized Smoothing. In *International Conference on Machine Learning (ICML)*. 1310–1320.

[31] 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.

[32] J.S. Cramer. 2002. *The Origins of Logistic Regression*. Technical Report 2002-119/4. Tinbergen Institute.

[33] 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.

[34] 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.

[35] 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.

[36] Amit Daniely and Hadas Schacham. 2020. Most ReLU Networks Suffer from L2 Adversarial Perturbations. In *Advances in Neural Information Processing Systems (NeurIPS)*. 6629–6636.

[37] 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.

[38] 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.

[39] 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.

[40] 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)*.

[41] 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.

[42] Elvis Dohmatob. 2019. Generalized No Free Lunch Theorem for Adversarial Robustness. In *International Conference on Machine Learning (ICML)*. 1646–1654.

[43] 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.

[44] Kilian Fatras, Bharath Bhushan Damodaran, Sylvain Lobry, Rémi Flamary, Devis Tuia, and Nicolas Courty. 2022. Wasserstein Adversarial Regularization for Learning With Label Noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 10 (2022), 7296–7306.

[45] Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. 2018. Adversarial Vulnerability for Any Classifier. In *Advances in Neural Information Processing Systems (NeurIPS)*. 1186–1195.

[46] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2015. Fundamental Limits on Adversarial Robustness. In *ICML Workshop on Deep Learning*.

[47] 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.

[48] 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.

[49] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of Neural Networks is Fragile. In *AAAI Conference on Artificial Intelligence (AAAI)*. 3681–3688.

[50] 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.

[51] 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)*.

[52] 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).

[53] 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/.

[54] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR)*.

[55] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR)*.

[56] 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.

[57] 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.

[58] 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.

[59] UCI Machine Learning Group. 1995. Abalone Dataset. https://archive.ics.uci.edu/ml/datasets/abalone.

[60] 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.

[61] 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.

[62] Haibo He and Edwardo A. Garcia. 2009. Learning From Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 21, 9 (2009), 1263–1284.

[63] 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.
[64] 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.
[65] 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.
[66] 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.
[67] 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.
[68] Joern-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. 2019. Excessive Invariance Causes Adversarial Vulnerability. In *International Conference on Learning Representations (ICLR)*.
[69] 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.
[70] Jongheon. Jeong and Jinwoo. Shin. 2020. Consistency Regularization for Certified Robustness of Smoothed Classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*. 10558–10570.
[71] Xi Wu Jiefeng Chen. 2019. Robust Attribution Regularization. https://www.altacognita.com/robust-attribution/.
[72] Ian. T. Jolliffe. 2002. *Principal Component Analysis*. Springer.
[73] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. 2009. CIFAR-10 and CIFAR-100 Datasets. https://www.cs.toronto.edu/~kriz/cifar.html.
[74] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* 60 (2012), 84 – 90.
[75] 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.
[76] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2017. Adversarial Machine Learning at Scale. In *International Conference on Learning Representations (ICLR)*.
[77] Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. 1998. The MNIST Database of Handwritten Digits. http://yann.lecun.com/exdb/mnist/.
[78] 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)*.
[79] 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.
[80] Paul Lévy. 1951. *Problèmes concrets d'analyse fonctionnelle*. Gauthier-Villers.
[81] 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.
[82] 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.
[83] 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.
[84] 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.
[85] Yue Liu, Chakkrit Tantithamthavorn, Li Li, and Yepang Liu. 2022. Deep Learning for Android Malware Defenses: A Systematic Literature Review. *Comput. Surveys* (2022).
[86] 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.
[87] 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.
[88] 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.
[89] 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.
[90] 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)*.

[91] 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.

[92] 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.

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] 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)*.

[98] 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.

[99] 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.

[100] 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.

[101] 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*.

[102] 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).

[103] 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.

[104] 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.

[105] 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)*.

[106] 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.

[107] 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.

[108] 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)*.

[109] 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)*.

[110] Muni Sreenivas Pydi and Varun Jog. 2020. Adversarial Risk via Optimal Transport and Optimal Couplings. In *International Conference on Machine Learning (ICML)*. 7814–7823.

[111] Muni Sreenivas Pydi and Varun Jog. 2021. The Many Faces of Adversarial Risk. In *Advances in Neural Information Processing Systems (NeurIPS)*. 10000–10012.

[112] 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.

[113] 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.

[114] 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.

[115] Eitan Richardson and Yair Weiss. 2021. A Bayes-Optimal View on Adversarial Examples. *Journal of Machine Learning Research (JMLR)* 22, 221 (2021), 10076–10103.
[116] 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.
[117] 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.
[118] Amartya Sanyal, Puneet K. Dokania, Varun Kanade, and Philip Torr. 2021. How Benign is Benign Overfitting?. In *International Conference on Learning Representations (ICLR)*.
[119] 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.
[120] H. Scudder. 1965. Probability of Error of Some Adaptive Pattern-Recognition Machines. *IEEE Transactions on Information Theory* 11, 3 (1965), 363–371.
[121] Alex Serban, Erik Poll, and Joost Visser. 2020. Adversarial Examples on Object Recognition: A Comprehensive Survey. *Comput. Surveys* 53, 3 (2020), 1–38.
[122] 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.
[123] Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. 2019. Are Adversarial Examples Inevitable?. In *International Conference on Learning Representations (ICLR)*.
[124] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.
[125] Claude E. Shannon. 1949. *The Mathematical Theory of Communication*. University of Illinois Press.
[126] 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.
[127] 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.
[128] 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.
[129] 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)*.
[130] 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)*.
[131] Michel Talagrand. 1996. A New Look at Independence. *The Annals of Probability* (1996), 1–34.
[132] Michel Talagrand. 1996. Transportation Cost for Gaussian and Other Product Measures. *Geometric and Functional Analysis* 6 (1996), 587–600.
[133] 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.
[134] Thomas Tanay and Lewis D. Griffin. 2016. A Boundary Tilting Persepective on the Phenomenon of Adversarial Examples. *ArXiv* (2016).
[135] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. 2022. A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning. *Comput. Surveys* (2022).
[136] 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.
[137] 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.
[138] Jerome Friedman Trevor Hastie, Robert Tibshirani. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer New York, NY.
[139] 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)*.
[140] 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.

[141] viso.ai. 2022. What is Adversarial Machine Learning? Attack Methods in 2022. https://viso.ai/deep-learning/adversarial-machine-learning/.
[142] 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).
[143] 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.
[144] 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.
[145] 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.
[146] 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.
[147] 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.
[148] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *ArXiv* (2017).
[149] 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.
[150] 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.
[151] 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.
[152] 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.
[153] Shuo Yang, Tianyu Guo, Yunhe Wang, and Chang Xu. 2021. Adversarial Robustness through Disentangled Representations. In *AAAI Conference on Artificial Intelligence (AAAI)*. 3145–3153.
[154] 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.
[155] Dong Yin, Ramchandran Kannan, and Peter Bartlett. 2019. Rademacher Complexity for Adversarially Robust Generalization. In *International Conference on Machine Learning (ICML)*. 7085–7094.
[156] 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.
[157] 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).
[158] 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).
[159] 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.
[160] 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)*.
[161] 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.
[162] 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.
[163] 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 System and Technology* 11, 3 (2020), 1–41.
[164] Xiao Zhang and David Evans. 2022. Incorporating Label Uncertainty in Understanding Adversarial Robustness. In *International Conference on Learning Representations (ICLR)*.

[165] 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).

[166] Hangyu Zhu and Yaochu Jin. 2020. Multi-Objective Evolutionary Federated Learning. *IEEE Transactions on Neural Networks and Learning Systems* 31, 4 (2020), 1310–1322.

[167] Jianing Zhu, Jingfeng Zhang, Bo Han, Tongliang Liu, Gang Niu, Hongxia Yang, Mohan S. Kankanhalli, and Masashi Sugiyama. 2021. Understanding the Interaction of Adversarial Training with Noisy Labels. *ArXiv* abs/2102.03482 (2021).

[168] Yao Zhu, Jiacheng Sun, and Zhenguo Li. 2022. Rethinking Adversarial Transferability from a Data Distribution Perspective. In *International Conference on Learning Representations (ICLR)*.

Table 2. The constraints applied to limit Google Scholar search

Venue	Search Constraints with Publication Source / Site
ACL	source:“Association for Computational Linguistic”+ source: “ACL”
CL	source:“Computational Linguistics”+ source: “CL”
COLT	source:“International Conference on Learning Theory”+ source: “COLT”
ICLR	source:“International Conference on Learning Representations” + source:“ICLR”
ICML	source:“International Conference on Machine Learning” + source:“ICML”
NDSS	source:“Network and Distributed System Security” + source:“NDSS”
Artificial Intelligence	source: “Artificial Intelligence"
Neural Networks	source: “Neural Networks"
Pattern Recognition	source: “Pattern Recognition"
Knowledge Based System	source: “Knowledge Based System"
JISA	source: “Journal of Information and Security Applications"
AAAI	site:aaai.org
IJCAI	site:ijcai.org
JMLR	site:jmlr.org
NeurIPS	site:proceedings.neurips.cc
USENIX	site:usenix.org
ArXiv	site:arxiv.org

A APPENDICES

A.1 Google Scholar Search Constraints

We used Google Scholar’s *site* or *source* filtering constraints to perform a targeted, per-venue search. Specifically, for venues that have their publications on custom websites, e.g., *“proceedings.neurips.cc”* [2] for proceedings from NeurIPS, we performed a *site*-constrained search. For venues that have their publications hosted on a shared website, e.g., PMLR [5]) for proceedings of ICLR, we performed a *source*-constrained search. Table 2 shows a complete set of constraints we used; the “+” symbol indicates a union of the results.

A.2 Detailed Paper Categorization

We include the detailed categorization tables for the papers collected in the following tables. In the table, we used abbreviation to denote the datasets used in the papers: M for MNIST, FM for Fashion-MNIST [148], S for SVHN, C-10 for CIFAR-10, C-100 for CIFAR-100, IN for ImageNet [74], TI for Tiny Images Dataset, CA for CelebA [86], HM for Halfmoon, M1V7 for mnist 1v7, A for abalone [59], L for LSUN [157], CS for Cityscapes [31], TO for Taskonomy [159].

Table 3. 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. [9]	Dimensionality	Any	Any	Any	Radius based	White box	Any	L_2	✓	✗	✗	✓	C-10, IN	k-NN	Standard	N/A
2	Awasthi et al. [11]	Dimensionality	Any	Any	DNNs	Radius based	White box	Any	L_2, L_∞	✓	✗	✗	✗	C-10, C-100	DNNs	Adversarial	PGD
3	Bhagoji et al. [14]	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. [15]	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. [16]	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. [18]	Dimensionality	Any	Any	Randomized smoothed classifier	Radius based	White box	Any	$L_p, p > 2$	✓	✗	✗	✓	C-10	Smoothed DNN	Adversarial	Gaussian noise (certification)
7	Bui et al. [22]	Separation	Any	Any	DNNs	Error-rate based	White box	Gradient based	L_p	✗	✓	✗	✗	C-10, M	CNNs	Adversarial	PGD
8	Carbone et al. [23]	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. [25]	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. [27]	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	Cullina et al. [34]	Number of samples	Any	Binary Classif.	Any	Error-rate based	White box	Any	L_p	✓	✗	✓	✓	N/A	N/A	N/A	N/A
12	Dan et al. [35]	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
13	Daniely et al. [36]	Dimensionality	Any	Any	ReLU networks	Radius-based	White box	Any	L_2	✓	✗	✗	✓	N/A	N/A	N/A	N/A
14	Ding et al. [40]	Distribution	Any	Any	Any	Error-rate based	White box	Any	Any	✗	✗	✓	✓	C-10, M	DNNs	Adversarial	PGD
15	Diochnos et al. [41]	Dimensionality	Uniform distribu -tion on boolean hypercube	Any	Any	Radius based	White box	Any	L_0	✓	✗	✗	✓	N/A	N/A	N/A	N/A
16	Dohmatob [42]	Concentration	Any	Any	Any	Radius based	White box	Any	L_p, Geodesic	✓	✗	✗	✓	M	DNNs	Adversarial	Not mentioned
17	Fawzi et al. [45]	Distribution	Distribution generated by smooth generative model	Any	Any	Radius based	White box	Any	Any	✓	✗	✓	✓	C-10, S	DNNs	Adversarial	PGD
18	Garg et al. [48]	Separation	Any	Any	N/A	Error-rate based	White box	Any	Any	✓	✓	✗	✓	M	DNNs	Adversarial	PGD
19	Gilmer et al. [51]	Dimensionality	Concentric n-dimensional spheres	Binary Classif.	DNNs	Radius based	White box	Gradient based	L_2	✓	✗	✓	✓	M	DNNs	Standard	PGD
20	Gourdeau et al. [57]	Number of samples, Dimensionality	Any	Any	Any	Error-rate based	White box	Any	Any	✓	✗	✗	✓	N/A	N/A	N/A	N/A
21	Gowal et al. [58]	Number of samples	Any	Any	Any	Error-rate based	White box	Any	L_p	✓	✓	✗	✗	C-10, C-100, M, TI	DNNs	Adversarial	AutoAttack

Table 4. 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	Classifier Type	Training Proc.	Attacks
22	Izmailov et al. [67]	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
23	Javanmard et al. [69]	Number of samples, Dimensionality	Any	Regres.	Linear Regres.	Error-rate based	Black box	Any	L_2	✓	✗	✗	✓	N/A	N/A	N/A	N/A
24	Kumar et al. [75]	Dimensionality	Any	Any	Any	Radius based	Any	Any	$L_p, p > 2$	✓	✗	✗	✓	C-10, IN	DNNs	DNNs	Gaussian noise (certification)
25	Lee et al. [79]	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
26	Mahloujifar et al. [91]	Concentration	Distributions in Lévy Families	Any	Any	Error-rate based	White box	Any	L_0	✓	✗	✗	✓	N/A	N/A	N/A	N/A
27	Mahloujifar et al. [92]	Concentration	Any	Any	Any	Error-rate based	White box	Any	L_2, L_∞	✓	✗	✗	✗	C-10, M	DNNs	Adversarial	PGD
28	Mao et al. [94]	Label Quality	Any	Any	DNNs	Error-rate based	White box	Gradient based	L_∞	✓	✗	✓	✓	CS, TO	DNNs	Standard	PGD, FGSM MIM, Houdini
29	Mehrabi et al. [95]	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
30	Mustafa et al. [99]	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
31	Najafi et al. [100]	Number of samples	Any	Any	Any	Error-rate based	White box	Gradient based	L_2, L_∞	✓	✓	✗	✓	C-10, M, S	DNNs	Adversarial	PGD
32	Oritz-Jimenez et al. [103]	Domain Specific	Any	Any	CNNs	Radius based	White box	Gradient based	L_2	✓	✗	✓	✗	C-10, M, IN	CNNs	Standard, Adversarial	PGD
33	Pang et al. [104]	Distribution, Separation	Any	Any	DNNs	Radius based	White box	Gradient based	L_2	✓	✓	✓	✓	C-10, M, IN	DNNs	Standard	FGSM, BIM ILCM, JSMA
34	Pang et al. [105]	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
35	Prescott et al. [109]	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
36	Pydi & Jog [110]	Separation	Any	Binary Classif.	Any	Error-rate based	White box	Gradient based	L_2, L_∞	✓	✗	✗	✓	C-10, M, FM, S	DNNs	Adversarial	N/A
37	Pydi & Jog [111]	Separation	Any	Binary Classif.	Any	Error-rate based	White box	Gradient based	L_2, L_∞	✓	✗	✗	✓	N/A	N/A	N/A	N/A
38	Rajput et al. [112]	Dimensionality	Any	Any	Linear classifiers, non-linear classifiers	Radius based	Any	Any	L_2	✓	✗	✗	✓	N/A	N/A	N/A	N/A

Table 5. 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
39	Richardson & Weiss [115]	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
40	Sanyal et al. [118]	Label Quality	Any	Binary Classif.	Any	Error-rate based	White box	Any	Any	✗	✗	✓	✓	C-10, M	DNNs	Standard, Adversarial	PGD
41	Schmidt et al. [119]	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
42	Shafahi et al. [123]	Dimesionality, Density	N-dimensional hypercube	Any	Any	Radius-based	White box	Any	L_p, Geodesic	✗	✗	✓	✓	C-10, M	CNN	Adversarial	PGD
43	Simon-Gabriel et al. [126]	Dimensionality	Any	Any	DNNs	Error-rate based	White box	Any	Any	✓	✗	✗	✓	C-10	DNNs	Adversarial	PGD
44	Song et al. [129]	Density	Any	Any	Any	Error-rate based	Any	Any	Any	✓	✓	✗	✗	C-10, M, FM	CNNs	Adversarial	FGSM, BIM C&W, DeepFool
45	Uesato et al. [140]	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
46	Wan et al. [142]	Distribution, Separation	Any	Any	Any	Error-rate based	White box	Any	L_∞	✗	✓	✗	✓	C-10, M, IN	DNNs	Standard	FGSM, BIM, ILCM, C&W
47	Wang et al. [144]	Domain Specific	Any	Any	CNNs	Error-rate based	White box	Gradient based	L_2	✓	✗	✓	✗	C-10	CNNs	Standard, Adversarial	PGD, FGSM
48	Wang et al. [145]	Dimensionality, Separation	Any	Binary Classif.	kNN	Radius based	White box	Any	L_2	✓	✓	✗	✗	M, M1V7, HM	k-NN	Adversarial	Direct attack, Transfer-based attacks
49	Weber et al. [146]	Dimensionality	Hierarchial data	Any	Any	Error-rate based	White box	Any	Check	✓	✓	✓	✗	IN	Hyperbolic perceptron	Adversarial	Gradient based
50	Wu et al. [147]	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
51	Yang et al. [152]	Separation	Any	Any	DNNs	Error-rate based	White box	Gradient based	L_2	✓	✓	✗	✓	C-10, C-100, M, TI	DNNs	Adversarial	PGD
52	Yang et al. [154]	Separation	Any	Binary Classif.	Non-parametric classifiers	Radius based	White box	Distance based	L_2	✓	✓	✗	✓	HM	Histogram, 1-NN	Standard	Distance based
53	Yin et al. [156]	Domain Specific	Any	Any	Any	Error-rate based	White box	Any	L_2	✓	✗	✗	✗	C-10, IN	DNNs	Adversarial	Corruptions, PGD
54	Yin et al. [155]	Dimensionality	Any	Any	Linear classifiers, DNNs	Error-rate based	White box	Any	L_∞	✓	✗	✗	✓	M	Linear classifiers, ReLU networks	Adversarial	PGD
55	Zhang et al. [160]	Density	Any	Any	Any	Error-rate based	White box	Any	L_2, L_∞	✓	✗	✗	✗	C-10, M, FM	DNNs	Adversarial	C&W
56	Zhang & Evans [164]	Concentration	Gaussian (theory), Any (application)	Any	Any	Error-rate based	White box	Any	L_2, L_∞	✓	✗	✗	✓	C-10	DNNs	Standard, Adversarial	AutoAttack
57	Zhu et al. [168]	Density	Any	Any	DNNs	Error-rate based	Any	Any	L_∞	✓	✗	✗	✓	IN	CNNs	Adversarial	PGD Transfer-based attacks