IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A COMPREHENSIVE SURVEY OF DEFENCE MECHANISMS IN ADVERSARIAL MACHINE LEARNING: TAXONOMY, COMPARATIVE ANALYSIS, AND FUTURE DIRECTIONS

Mehul Manani, Prerana Gupta

Lecturer, Department of Computer Engineering, Shri K J Polytechnic, Bharuch, Gujarat, India

Abstract: The increasing deployment of machine learning (ML) systems in critical domains such as healthcare, autonomous vehicles, finance, and cybersecurity has amplified concerns regarding their vulnerability to adversarial attacks. Adversarial Machine Learning (AML) has demonstrated that carefully crafted malicious perturbations can significantly compromise model reliability, security, and trustworthiness. In response, a wide range of defence mechanisms has been proposed to improve model robustness against such threats.

This survey presents a comprehensive review of defence mechanisms in adversarial machine learning, covering major defence paradigms including adversarial training, input transformation and preprocessing, detection-based defences, certified robustness, ensemble methods, and emerging hybrid and adaptive approaches. The paper systematically classifies these techniques, analyses their underlying principles, strengths, limitations, and practical applicability, and provides a comparative evaluation based on robustness, computational cost, scalability, and deployment feasibility. Special attention is given to recent advances, including transformer-specific defences, diffusion-based purification methods, and adaptive defence frameworks.

The study highlights key research gaps such as the absence of universal defence strategies, robustness–accuracy trade-offs, scalability challenges, and limited real-world validation. Finally, future research directions toward unified, explainable, and self-adaptive defence frameworks are discussed. This survey aims to serve as a comprehensive reference for researchers and practitioners working toward secure, trustworthy, and resilient machine learning systems.

Index Terms **-** Adversarial Machine Learning, Defence Mechanisms, Adversarial Training, Certified Robustness, Input Transformation, Hybrid Defences, Robust Machine Learning, Trustworthy AI.

I.INTRODUCTION TO ADVERSERIAL MACHINE LEARNING AND SECURITY

The rapid adoption of machine learning (ML) across critical application domains such as healthcare, autonomous transportation, finance, cybersecurity, and intelligent surveillance has transformed modern computing systems by enabling data-driven decision making at unprecedented scale and accuracy. Despite these remarkable advancements, recent studies have demonstrated that machine learning models—particularly deep neural networks—remain highly vulnerable to carefully crafted adversarial manipulations that can significantly degrade model reliability and trustworthiness [1], [2]. These vulnerabilities have raised serious concerns regarding the secure deployment of ML systems in real-world environments.

Adversarial Machine Learning (AML) has emerged as an important research field focused on understanding and mitigating security vulnerabilities in learning systems. In adversarial settings, malicious actors deliberately exploit weaknesses in machine learning models by manipulating training data, modifying inference inputs, or extracting confidential model information [3], [36]. Among these threats, adversarial examples—small, often imperceptible perturbations added to legitimate inputs—have been

shown to mislead even state-of-the-art deep learning models with high confidence [1], [6]. Furthermore, physical-world adversarial attacks, such as manipulated traffic signs and malicious wearable accessories, demonstrate that these vulnerabilities are not limited to digital environments but extend directly to safety-critical physical systems [4], [35].

To address these threats, a broad range of defence mechanisms has been proposed over the last decade. Early defences focused primarily on heuristic methods such as input preprocessing, feature squeezing, and defensive distillation [8], [10]. However, many of these approaches were later shown to rely on gradient obfuscation and were subsequently bypassed by stronger adaptive attacks [7], [29]. This led to the development of more principled defence strategies, including adversarial training [5], [18], certified robustness techniques [14], [15], randomized smoothing [14], [16], robust optimization [12], and formal verification-based defences [41], which aim to provide stronger and theoretically grounded security guarantees.

Despite significant progress, designing universally robust machine learning systems remains an open challenge. Most existing defences exhibit important limitations, including degradation of clean-data accuracy, high computational complexity, poor scalability to large models, and limited generalization against unseen attacks [13], [64], [81]. Moreover, the rapid emergence of new model paradigms—such as Vision Transformers (ViTs), self-supervised learning systems, and diffusion-based generative models—has introduced new robustness challenges and motivated the development of novel defence strategies [85], [86], [91], [93].

Several survey papers have reviewed adversarial machine learning broadly [36], [37], but comparatively fewer studies have focused exclusively on **defence mechanisms**, especially incorporating the substantial advances made between 2021 and early 2023 [75], [100]. Existing surveys often emphasize attack methodologies while providing only limited discussion of comparative defence performance, certified robustness, and practical deployment considerations. This creates a need for an updated, defence-centric review.

Motivated by this gap, this paper presents a **comprehensive survey of defence mechanisms in adversarial machine learning**, covering classical and emerging strategies proposed up to **April 2023**. Unlike prior surveys, this work focuses specifically on defence methodologies and provides a structured taxonomy, comparative technical analysis, and future research directions. The major contributions of this survey are summarized as follows:

1. **Comprehensive Taxonomy:** A systematic categorization of defence mechanisms including adversarial training, input transformation, detection-based defences, certified robustness, ensemble learning, and emerging adaptive defences.
2. **Technical Analysis:** Detailed examination of defence principles, mathematical foundations, strengths, limitations, and practical applicability.
3. **Comparative Evaluation:** Cross-method comparison based on robustness, computational cost, scalability, interpretability, and deployment readiness.
4. **Emerging Trends:** Coverage of recent advances involving transformer robustness, diffusion-based purification, and adaptive robust learning.
5. **Future Directions:** Identification of open research challenges and promising future directions toward trustworthy and secure machine learning systems.

The remainder of this paper is organized as follows. Section II presents Literature Review. Section III introduces the fundamentals of adversarial machine learning and the underlying threat model. Section IV introduces a taxonomy of defence mechanisms. Sections V through IX discuss major categories of defence strategies in detail. Section X provides comparative analysis and observations, followed by research gaps and future directions in Section XI. Finally, Section XII concludes the paper.

II.LITERATURE REVIEW

Table 1: Literature Review

Defence Category	Key References	Major Contributions	Key Limitations
Adversarial Training	[1], [5], [12], [18], [43], [71], [76]	Introduced robustness through training on adversarial samples; methods such as FGSM, PGD, TRADES, and fast adversarial training significantly improve empirical robustness.	High computational cost; reduced clean accuracy; weak generalization to unseen attacks.
Input Transformation / Preprocessing	[8], [9], [10], [19], [21], [56]	Uses feature squeezing, randomization, denoising, and generative reconstruction to remove adversarial perturbations before classification.	Vulnerable to adaptive attacks; often provides only limited robustness.
Detection-Based Defences	[7], [29], [62]	Detect adversarial samples using statistical anomalies, confidence scores, or activation monitoring before prediction.	High false positives; many methods fail under adaptive attack settings.
Certified Robustness	[14], [15], [16], [17], [41], [42], [68], [73], [98]	Provides mathematically provable guarantees using randomized smoothing, convex relaxation, differential privacy, and formal verification.	High computational complexity; difficult to scale to large models.
Ensemble and Hybrid Defences	[18], [46], [50], [72]	Combines multiple models or multiple defence layers to improve robustness and attack diversity resistance.	Increased memory, inference, and training cost.
Robust Optimization Methods	[5], [12], [39], [52], [54]	Reformulates defence as an optimization problem to directly minimize worst-case adversarial risk.	Optimization is expensive and sensitive to attack assumptions.
Randomization-Based Defences	[10], [16], [47], [56]	Introduces stochasticity through random transformations or noise injection to reduce attack success.	Can degrade model accuracy; limited guarantees.
Generative Model-Based Defences	[21], [22], [91], [93], [94]	Uses GANs and diffusion models for adversarial purification and input reconstruction.	Computationally intensive; slower deployment.
Vision Transformer (ViT) Defences	[85], [86], [87], [89]	Investigates robustness properties and defence strategies specifically for transformer-based vision models.	Still an emerging field; lacks standardized evaluation.
Adaptive / Dynamic Defences	[95], [100]	Dynamically modifies model behaviour or architecture during inference to counter adaptive attackers.	Early-stage research; limited real-world validation.
Survey and Benchmark Studies	[36], [37], [75], [99], [100]	Provide comprehensive reviews, benchmarks, and open challenges in AML defence research.	Mostly descriptive; limited new technical solutions.

The literature indicates that **adversarial training** remains the dominant empirical defence mechanism due to its practical effectiveness, although it suffers from substantial computational overhead and reduced clean-data performance [5], [12]. **Certified robustness methods** provide stronger theoretical guarantees but currently face scalability challenges [14], [41]. Traditional **input transformation** and **detection-based defences** offer lower-cost alternatives but are often vulnerable to adaptive attacks [7], [29]. Recent research has increasingly shifted toward **transformer robustness**, **diffusion-based purification**, and **adaptive**

hybrid frameworks, suggesting a move toward more generalizable and deployment-oriented defence strategies [85], [91], [95].

III.FUNDAMENTALS OF ADVERSARIAL MACHINE LEARNING

Machine learning models have achieved significant success across diverse application domains; however, their vulnerability to intentionally crafted malicious inputs has emerged as a major security concern. Unlike traditional software vulnerabilities, these weaknesses arise from the statistical nature of learning algorithms, high-dimensional feature spaces, and complex nonlinear decision boundaries, which create exploitable regions where small perturbations can cause severe prediction errors without noticeable input changes [1], [2].

Adversarial Machine Learning (AML) studies how attackers exploit such weaknesses and how learning systems can be designed to remain secure and robust under attack [3], [36]. Similar to traditional cybersecurity, adversarial threats target the confidentiality, integrity, and availability of machine learning systems, but through mechanisms unique to data-driven models [37].

A fundamental concept in AML is the **adversarial threat model**, which defines the attacker's goals, knowledge, and capabilities [3], [37]. Attackers may aim to violate **integrity** by causing targeted misclassifications, **availability** by degrading model performance through poisoning, or **privacy** by extracting sensitive training data or model parameters [3], [36], [37].

The effectiveness of an attack depends heavily on the attacker's knowledge of the target model. In **white-box attacks**, the adversary has full access to architecture, parameters, and gradients, representing the strongest threat model [5], [6]. In **gray-box attacks**, only partial system information is available, while **black-box attacks** rely solely on model queries and outputs, yet remain highly practical due to transferability and query-based optimization [31], [33].

Adversarial attacks can occur at multiple stages of the machine learning lifecycle. **Training-time attacks** manipulate training data or labels to corrupt the learned decision boundary, often causing long-term damage [37]. **Inference-time attacks** alter test inputs after deployment and remain the most widely studied adversarial threat [1], [28]. **Model interaction attacks** repeatedly query deployed systems to extract internal logic or infer sensitive information [31], [36].

These attacks are commonly categorized as **evasion attacks**, which manipulate inputs using methods such as FGSM, PGD, and C&W [1], [5], [6]; **poisoning attacks**, which corrupt the training process [37]; **backdoor attacks**, where hidden triggers are embedded during training [37]; and **privacy attacks**, which attempt to recover confidential information through inference or reconstruction [36].

Deep learning models are especially vulnerable due to several inherent characteristics. High-dimensional input spaces create many exploitable directions [2], local linearity enables small perturbations to accumulate effectively [1], overconfidence leads to highly certain incorrect predictions [29], transferability allows attacks to generalize across different models [30], [49], and sensitivity to distribution shifts reduces robustness when inputs deviate slightly from training data [13].

To build secure and trustworthy ML systems, defences must ensure robustness against adversarial perturbations, generalize to unseen and adaptive attacks, scale efficiently to modern architectures, minimize computational overhead, remain interpretable, and ideally provide formal robustness guarantees [14], [41]. These requirements form the foundation for the diverse defence mechanisms discussed in subsequent sections.

IV.TAXONOMY OF DEFENCE MECHANISMS IN ADVERSARIAL MACHINE LEARNING

As adversarial attacks against machine learning systems become increasingly sophisticated, a broad range of defence mechanisms has been developed to improve robustness, reliability, and trustworthiness. These defences differ in their assumptions, deployment stages, computational complexity, and security guarantees, making a structured taxonomy essential for systematic analysis [36], [37]. Broadly, defence strategies can be categorized as proactive defences, which strengthen models before deployment, and reactive defences, which detect or mitigate attacks during inference [75], [100].

Adversarial training is one of the most effective empirical defence strategies, where adversarial examples are incorporated into the training process to learn robust decision boundaries [1], [5]. This is typically formulated as a min-max optimization problem and includes methods such as FGSM adversarial training, PGD-based robust optimization, TRADES, and fast adversarial training [1], [5], [12], [43], [71], [76]. Despite strong robustness, it often incurs high computational cost and reduced clean-data accuracy [13].

Input transformation and preprocessing defences attempt to remove adversarial perturbations before classification by applying techniques such as feature squeezing, JPEG compression, random resizing,

denoising, and generative purification [8], [10], [19], [21], [56]. These methods are efficient and easy to deploy but often fail against adaptive attacks specifically designed to bypass preprocessing [7], [29].

Detection-based defences focus on identifying adversarial inputs using signals such as abnormal confidence scores, feature-space anomalies, gradient inconsistencies, or unusual activation patterns [29], [62]. Their effectiveness depends heavily on maintaining low false-positive rates while resisting adaptive attackers.

Certified robustness methods provide mathematical guarantees that bounded perturbations cannot alter model predictions, making them especially important in safety-critical systems [14], [15]. Major approaches include randomized smoothing, convex relaxation, formal verification, and differential privacy-based certification [14], [15], [16], [17], [41], [42]. However, these methods often suffer from scalability limitations.

Ensemble and hybrid defences combine multiple models or defence mechanisms to improve robustness against diverse attacks, particularly transfer-based attacks [18], [46]. Although these methods generally improve resilience, they also increase computational and deployment complexity.

Randomization-based defences introduce stochasticity into preprocessing or model behaviour through methods such as noise injection, randomized discretization, or stochastic activations, making gradient estimation more difficult for attackers [10], [47]. Nevertheless, adaptive attacks can often bypass these methods using expectation-based optimization [7].

Generative model-based defences use learned data distributions to reconstruct clean inputs from adversarial samples. Approaches such as Defence-GAN and diffusion-based purification methods like DiffPure have demonstrated strong robustness, although they are often computationally expensive during inference [21], [91], [93], [94].

The rise of transformer and foundation model defences reflects the growing importance of securing modern architectures such as Vision Transformers, which exhibit different adversarial behaviours compared to traditional CNNs [85], [86]. Current research focuses on attention regularization, robust tokenization, **transformer**-specific training, and architecture-aware certification, though many challenges remain open [87], [89].

Finally, adaptive and dynamic defences aim to counter evolving attacks through online learning, dynamic model reconfiguration, and self-healing architectures [95]. While promising, these approaches are still largely experimental.

Overall, this taxonomy highlights that no single defence is universally effective [75], [100]. Adversarial training remains dominant for empirical robustness, certified methods provide the strongest theoretical guarantees, and future secure ML systems will likely rely on hybrid, multi-layered defence frameworks combining complementary mechanisms [5], [14].

V.ADVERSARIAL TRAINING-BASED DEFENCE MECHANISMS

Adversarial training is the most widely studied empirical defence in adversarial machine learning and remains the benchmark for evaluating new defence strategies. Its core idea is to include adversarially perturbed samples during training so that the model learns robust decision boundaries and improves resistance to malicious inputs [1], [5]. Unlike preprocessing or detection-based defences, adversarial training directly modifies the optimization objective, making robustness an inherent model property [12], [76].

Mathematically, adversarial training is formulated as a min-max optimization problem, where the inner optimization generates the strongest perturbation within a bounded attack budget, while the outer optimization updates model parameters to minimize worst-case adversarial loss [5]. This robust optimization framework, introduced by Madry et al., has become the standard approach for empirical robustness evaluation.

The earliest implementation was **FGSM-based adversarial training**, proposed by Goodfellow et al. [1], which uses single-step gradient-based perturbations. Although computationally efficient and easy to implement, FGSM suffers from vulnerabilities to stronger iterative attacks and may cause label leaking [34]. To address these limitations, PGD-based adversarial training was introduced by Madry et al. [5], using iterative perturbation generation to achieve significantly stronger robustness. While considered the empirical gold standard, PGD training is computationally expensive and difficult to scale to large architectures such as transformers [64].

A major challenge in adversarial training is balancing robustness and clean-data accuracy. To address this, TRADES, proposed by Zhang et al., introduced a **regularization-based framework** that explicitly optimizes the trade-off between standard accuracy and adversarial robustness [12], [76]. This method remains highly influential in modern robust learning.

Because PGD training is costly, several faster alternatives such as Fast is Better Than Free and Fast Adversarial Training were proposed to reduce computational overhead while maintaining near-PGD robustness on benchmark datasets [43], [71]. Similarly, Ensemble Adversarial Training improves black-box robustness by incorporating adversarial examples generated from multiple surrogate models, increasing attack diversity but requiring additional computational resources [18].

Despite its effectiveness, adversarial training faces important limitations. It is computationally intensive [64], often overfits to specific attack patterns [44], and does not guarantee robustness against all attack types [69]. Its application to large-scale transformer and foundation models also remains challenging [85].

Recent advances from 2021–2023 have focused on improving scalability and adaptability through adaptive training schedules, multi-perturbation robustness, robust distillation, and transformer-aware adversarial learning [59], [63], [72], [77], [85], [89].

Overall, the literature confirms that adversarial training remains the strongest empirical defence currently available [5], [12]. However, the trade-off between robustness and accuracy persists [13], and computational efficiency remains a major barrier to real-world deployment [71]. Future research is therefore moving toward adaptive, scalable, and architecture-aware adversarial training methods [85], [95].

VI. INPUT TRANSFORMATION AND PREPROCESSING DEFENCES

Input transformation and preprocessing defences aim to neutralize adversarial perturbations before inputs reach the classifier. Unlike adversarial training, which modifies the learning process, these methods act as an external protective layer by filtering, compressing, or reconstructing potentially malicious samples [8], [10]. Their underlying assumption is that adversarial perturbations often lie in non-robust or high-frequency regions of the input space and can be removed without significantly affecting legitimate semantic content. Because they are model-independent, computationally efficient, and easy to integrate into existing systems, these defences gained significant popularity [36], [75]. However, their effectiveness against adaptive attacks remains a major concern [7].

One of the earliest methods, **Feature Squeezing**, proposed by Xu et al., reduces input complexity through bit-depth reduction or spatial smoothing, making adversarial perturbations less effective [8]. Although simple and computationally efficient, it often degrades clean image quality and is vulnerable to adaptive attacks [29].

Compression-based defences, particularly **JPEG compression**, remove high-frequency adversarial artifacts through lossy encoding [10]. These methods are inexpensive and easy to deploy but become ineffective when attackers optimize perturbations through the compression pipeline.

Randomization-based preprocessing introduces stochastic operations such as random resizing, padding, cropping, and noise injection to make gradient estimation more difficult for attackers [56]. While effective against standard gradient-based attacks, stronger expectation-over-transformation attacks can often bypass such defences [7].

Image denoising approaches, such as Feature Denoising Networks proposed by Xie et al., remove adversarial noise from intermediate feature representations during inference [19]. These methods outperform simple filtering but increase architectural complexity and typically require retraining.

Autoencoder-based defences reconstruct cleaner inputs by encoding adversarial samples into a latent space and decoding purified outputs. These methods learn task-specific denoising patterns but may introduce reconstruction artifacts and remain vulnerable to adaptive attacks.

GAN-based purification, exemplified by Defence-GAN, projects adversarial inputs onto the learned clean-data manifold before classification [21]. While effective for denoising, GAN-based methods are computationally slow and difficult to scale.

More recently, **diffusion-based purification methods**, including DiffPure and related frameworks, have achieved state-of-the-art robustness by iteratively denoising corrupted inputs through reverse diffusion processes [91], [93], [94]. Although highly effective, their extreme inference cost limits real-time applicability.

Overall, preprocessing defences provide a lightweight and practical first layer of protection [8], [10]. However, deterministic transformations alone are rarely sufficient against adaptive adversaries [7], [29]. Learned purification methods generally offer stronger robustness but at significantly higher computational cost [21], [93]. As a result, modern research increasingly treats input transformation as a complementary component within **multi-layered defence architectures**, often combined with adversarial training or detection mechanisms [94], [100].

VII. DETECTION-BASED DEFENCE MECHANISMS

Detection-based defences aim to identify adversarial inputs before they reach the final prediction stage. Unlike adversarial training, which improves inherent model robustness, or preprocessing methods, which attempt to remove perturbations, detection-based approaches treat adversarial samples as anomalous or out-of-distribution inputs that should be flagged, rejected, or sent for further verification [29], [36]. Their key assumption is that adversarial examples, although visually similar to legitimate samples, often exhibit abnormal statistical, geometric, or internal feature-space behaviour that can be distinguished from clean data [62]. This makes detection particularly valuable in real-world systems where rejecting suspicious inputs may be safer than producing incorrect predictions.

A typical detection framework consists of two components: the primary classifier, which predicts the class label, and a detector module, which determines whether the input is clean or adversarial. If detected as malicious, the input may be rejected or redirected for manual review.

Statistical detection methods identify adversarial samples by analysing deviations in input distributions, prediction confidence, entropy, or likelihood under learned data models. These methods are simple to implement and do not require architectural changes, but they often fail against carefully optimized adaptive attacks.

Feature-space detection methods monitor hidden-layer representations and identify abnormal internal behaviour using techniques such as Mahala Nobis distance, latent-space clustering, and nearest-neighbour anomaly detection. These methods are often more reliable than input-level detection but may suffer from representation drift.

Confidence-based detection relies on abnormal prediction probabilities or inconsistent confidence scores to identify attacks. These approaches are computationally lightweight and suitable for real-time systems, but attackers can often manipulate confidence values directly to evade detection [29].

Auxiliary classifier detectors use separate neural networks trained specifically to distinguish clean and adversarial inputs. While highly effective against known attack types, they often generalize poorly to unseen attacks and require attack-specific retraining.

Gradient and sensitivity-based methods detect attacks by analysing abnormal gradient responses or model sensitivity patterns. Although theoretically well-motivated, these approaches are computationally expensive for large modern architectures.

Activation signature detection examines neuron activation patterns to identify unusual internal responses. This approach is especially useful for detecting backdoor attacks, where hidden triggers often activate unique neuron pathways [40].

Out-of-distribution (OOD) detection treats adversarial inputs as abnormal samples outside the training distribution. Common methods include energy-based detection, density estimation, and uncertainty-based approaches. This area has become increasingly important for transformer and foundation-model security.

Despite promising results, detection-based defences face major challenges. Adaptive attackers can explicitly optimize to evade detectors [7], [29], false positives may reject legitimate unusual inputs, and many methods overfit to benchmark datasets, limiting real-world generalization. These limitations have prevented detection from becoming a standalone universal defence.

Recent advances focus on uncertainty-aware detection, transformer attention anomaly monitoring, ensemble detectors, self-supervised anomaly detection, and multimodal defence systems [85], [95]. Overall, the literature suggests that detection-based defences are most effective as a **secondary security layer** rather than a primary defence mechanism [36]. Future robust ML systems will likely combine detection with adversarial training and preprocessing to build stronger **defence-in-depth architectures**.

VIII. CERTIFIED ROBUSTNESS AND PROVABLE DEFENCE MECHANISMS

While empirical defences such as adversarial training improve robustness against known attacks, they cannot guarantee protection against stronger or previously unseen adversaries. This limitation led to the development of **certified robustness**, which aims to provide mathematical guarantees that a model's prediction will remain unchanged within a predefined perturbation range [14], [15]. Unlike empirical robustness, which depends on specific attack methods, certified robustness offers **attack-independent guarantees**, making it particularly important in safety-critical applications such as autonomous vehicles, medical diagnosis, and industrial automation.

The core idea of certified robustness is to determine whether a model can maintain consistent predictions for all possible perturbations within a specified threat boundary. The larger the certified radius, the stronger the guaranteed robustness.

One of the most widely used certified defence methods is **Randomized Smoothing**, introduced by Cohen et al. [14]. This technique improves robustness by adding random noise to inputs and making predictions based on aggregated outputs. Its major advantages are model independence, scalability to deep neural networks, and formal robustness guarantees under common threat models. However, it may reduce clean-data accuracy and increase inference time due to repeated sampling.

Another important approach is **Convex Relaxation**, which approximates nonlinear neural network behaviour using linear bounds to obtain provable robustness guarantees [15], [42]. These methods often provide tighter guarantees than randomized smoothing but are more difficult to scale to large modern architectures.

Formal verification methods use symbolic reasoning and optimization techniques such as mixed-integer programming and SMT solvers to prove exact robustness properties [41]. These methods provide the strongest guarantees available but suffer from extremely high computational complexity, making them practical only for relatively small networks.

Some studies have explored **Differential Privacy-based certification**, where privacy-preserving training mechanisms also provide robustness guarantees under certain assumptions [17]. Although this offers the dual advantage of privacy and security, it often reduces model accuracy because of added training noise.

Lipschitz-based certification improves robustness by constraining how much a model's output can change in response to small input changes. Techniques such as spectral normalization and gradient regularization are commonly used [52]. While theoretically elegant, these methods often produce conservative robustness estimates.

Another scalable approach is **Interval Bound Propagation (IBP)**, which estimates robustness by propagating uncertainty intervals through network layers [73]. IBP is computationally faster than exact verification and scales well to larger networks, although its guarantees may become loose for very deep models.

Recent advances (2021–2023) have focused on scalable certification for transformer architectures, tighter smoothing methods, hybrid adversarial-certified training, certification under semantic perturbations, and robust large-model safety [85], [98]. These developments show a growing effort to move certified robustness beyond small benchmark datasets toward practical deployment.

Overall, the literature shows that **certified robustness provides the strongest theoretical guarantees in adversarial machine learning** [14], [41]. However, stronger guarantees usually come with higher computational cost. Randomized smoothing remains the most practical large-scale certification method, while formal verification offers exact but less scalable guarantees. Therefore, certified robustness is essential for **high-assurance machine learning systems**, especially where empirical defences alone are insufficient. Table 2 presents the strengths and weaknesses comparison of various defence methods.

Table 2: Comparative Strengths and Weaknesses

Method	Guarantee Strength	Scalability	Practicality
Randomized Smoothing	Medium	High	High
Convex Relaxation	High	Medium	Medium
Formal Verification	Very High	Low	Low
Differential Privacy	Medium	Medium	Medium
Lipschitz Methods	Moderate	High	Medium
IBP	Medium	High	High

IX.ENSEMBLE, HYBRID, AND EMERGING DEFENCE MECHANISMS

As adversarial attacks continue to evolve, single defence mechanisms often fail to provide comprehensive protection across diverse threat models. This has led to the development of **ensemble, hybrid, and emerging adaptive defences**, which combine multiple complementary strategies to improve robustness, generalization, and real-world deploy ability [18], [72], [95]. Similar to the traditional cybersecurity principle of **defence-in-depth**, these methods rely on multiple protective layers rather than a single security mechanism.

Ensemble defence mechanisms improve robustness by combining predictions from multiple independently trained models. Since adversarial examples generated for one model may not transfer equally across others, ensembles can reduce attack success rates and improve black-box robustness [18]. Common approaches include model diversity, architecture diversity, and attack-diverse adversarial training. However, these methods increase memory usage and inference time.

Hybrid defence architectures combine multiple defence families within a single pipeline, such as preprocessing, detection, adversarial training, and certified robustness. This layered approach provides broader attack coverage and stronger overall security, but it also increases system complexity and may introduce error propagation between defence layers.

Robust distillation adapts knowledge distillation for adversarial defence by transferring robustness from a stronger teacher model to a smaller student model [72]. This improves deployment efficiency and reduces computational cost, although full robustness may not always transfer effectively.

Recent work has also explored **generative and diffusion-based hybrid defences**, where adversarial inputs are purified using generative models before classification. Methods such as DiffPure have demonstrated strong resistance against multiple attacks [91], [93], [94], but their high inference latency limits real-time deployment.

The rise of **transformer architectures** has introduced new defence challenges and opportunities. Current research focuses on attention regularization, token-level masking, robust patch embeddings, and transformer-specific adversarial training to address vulnerabilities unique to self-attention mechanisms [85], [86].

Another emerging direction is **dynamic and adaptive defence**, where models modify their behaviour during inference based on observed inputs. Examples include runtime model switching, adaptive thresholds, and online weight updates [95]. These approaches are promising against evolving attackers but remain largely experimental.

Meta-learning-based defences aim to train models that can rapidly adapt to unseen attack strategies, improving generalization and reducing dependence on predefined attack knowledge. Similarly, **self-supervised and foundation model defences** leverage large-scale pretraining to improve natural robustness and reduce reliance on explicit adversarial examples.

Comparative analysis shows that stronger robustness usually comes with greater computational cost and deployment complexity. Ensemble and robust distillation methods are relatively mature, while diffusion-based, transformer-specific, and adaptive defences remain active research areas.

Overall, the literature highlights that **multi-layered defence mechanisms consistently outperform isolated methods** [72], [95]. Hybrid pipelines are becoming the dominant practical strategy, while diffusion models, transformer defences, and adaptive learning are emerging as major future directions. These trends indicate a clear shift from isolated defence methods toward **integrated trustworthy AI systems** capable of handling evolving adversarial threats. Table 3 represents the comparative analysis of the emerging defence mechanism.

Table 3: Comparative Analysis of Emerging Defence Mechanism

Defence Type	Robustness	Cost	Scalability	Maturity
Ensemble	High	High	Medium	Mature
Hybrid	Very High	Very High	Medium	Growing
Robust Distillation	Medium	Low	High	Mature
Diffusion-Based	Very High	Very High	Low	Emerging
Transformer Defence	Medium–High	High	Medium	Emerging
Adaptive Defence	Potentially High	Medium	Medium	Experimental

X.COMPARATIVE ANALYSIS AND DISCUSSION

Table 4 summarizes the major defence categories discussed in this survey.

Table 4: Comparative Analysis of Defence Mechanisms in Adversarial Machine Learning

Defence Category	Robustness Strength	Computational Cost	Scalability	Theoretical Guarantee	Practical Deployment
Adversarial Training	Very High	Very High	Medium	No	High
Input Transformation	Medium	Low	High	No	Very High
Detection-Based	Medium	Medium	High	No	High
Certified Robustness	High	Very High	Low–Medium	Yes	Medium
Ensemble Methods	High	High	Medium	Partial	Medium
Hybrid Defences	Very High	Very High	Medium	Partial	High
Diffusion-Based	Very High	Extremely High	Low	No	Low
Transformer Defences	Medium–High	High	Medium	Limited	Emerging
Adaptive Defences	Potentially Very High	Medium	Medium	No	Experimental

After reviewing the major defence categories in adversarial machine learning, it is clear that no single method provides universal protection against all attack types. Each defence mechanism offers unique strengths and limitations in terms of robustness, computational cost, scalability, interpretability, and deployment feasibility [75], [99], [100]. Therefore, comparative analysis is essential to identify the most suitable defence strategy for different application scenarios.

A broad comparison of existing defences shows that adversarial training remains the strongest empirical defence but requires very high computational resources. Input transformation and preprocessing methods are computationally efficient and easy to deploy but provide relatively weaker robustness. Detection-based defences offer an effective secondary security layer but often struggle against adaptive attacks. Certified robustness methods provide formal security guarantees, making them highly trustworthy, but their scalability remains limited for large modern models. Ensemble and hybrid defences generally improve robustness by combining multiple complementary mechanisms, although this increases system complexity and resource requirements. Emerging methods such as diffusion-based, transformer-specific, and adaptive defences show strong potential but remain computationally demanding or experimentally immature.

One of the most important findings is the robustness–computational cost trade-off. Stronger defences typically require greater computational resources. For example, PGD-based adversarial training offers high robustness but significantly increases training time [5], while certified defences provide strong theoretical guarantees but become computationally expensive for large networks [41]. In contrast, preprocessing methods are inexpensive but offer limited protection [8].

Another well-known challenge is the accuracy–robustness trade-off, where increasing adversarial robustness often reduces clean-data accuracy and sometimes affects model calibration and interpretability [12]. This remains one of the central unresolved problems in adversarial machine learning.

Scalability is also becoming increasingly important. While many defences perform well on small and medium-sized neural networks, their effectiveness and feasibility decline as models grow larger. For transformer-based and foundation models, lightweight hybrid methods are currently more practical than computationally intensive certified or adversarial-training approaches [85], [86].

A major weakness across many defences is their limited generalization against adaptive attacks. Defences that perform well against one attack often fail against stronger or modified attack strategies, demonstrating the ongoing attacker–defender arms race [7], [29], [44].

From a trustworthiness perspective, certified defences offer the strongest interpretability because their guarantees are mathematically provable [14]. In contrast, ensemble and hybrid systems often behave as complex black boxes, making them harder to interpret and validate.

Defence selection must also be application-specific. Safety-critical domains such as autonomous driving and healthcare often require certified or hybrid defences, while cloud services benefit more from detection and adversarial training. Lightweight preprocessing remains suitable for edge devices, whereas advance models will likely require new adaptive defence frameworks.

Despite significant progress, several open challenges remain, including the absence of universal defences, continuously evolving attackers, inconsistent evaluation benchmarks, limited real-world deployment, and insufficient study of modern architectures [75], [100].

Overall, the field is clearly moving toward unified multi-layered defence frameworks that combine adversarial training, preprocessing, detection, certification, and adaptive response mechanisms. The literature suggests that adversarial training remains the strongest empirical defence, certified robustness remains the gold standard for theoretical assurance, and hybrid defences currently offer the best practical balance. The future of adversarial machine learning security will likely depend on integrated, adaptive, and scalable defence ecosystems.

XI.RESEARCH GAPS AND FUTURE DIRECTIONS

Despite significant progress in adversarial machine learning defence research, existing solutions are still far from achieving universal, scalable, and trustworthy robustness. Most current defence mechanisms perform effectively only under specific assumptions, attack models, or benchmark datasets, revealing several unresolved challenges that must be addressed to build truly secure machine learning systems [75], [99], [100].

One of the most critical research gaps is the **lack of universal defence mechanisms**. Existing methods are typically specialized for specific attack categories—for example, adversarial training mainly targets evasion attacks [5], certified robustness focuses on bounded perturbations [14], and detection methods concentrate on anomaly identification [29]. However, real-world attackers often combine multiple strategies, making isolated defences insufficient. Future work must therefore focus on **unified multi-layered defence architectures** capable of protecting against diverse attack surfaces simultaneously.

Another major challenge is the **limited generalization of defences against adaptive attacks**. Many methods perform well against known threats but fail when attackers modify their strategies. Examples include adaptive attacks bypassing preprocessing defences [7] and stronger optimization methods defeating detectors [29]. This highlights the need for **dynamic and continuously learning defences** rather than static security mechanisms.

The **robustness–accuracy trade-off** also remains a fundamental problem. Improving adversarial robustness often reduces clean-data accuracy, calibration quality, and computational efficiency [12], [76]. Future research should focus on developing defence strategies that achieve a better balance between security and predictive performance.

Computational scalability is another important limitation. Many of the strongest defences, such as PGD-based adversarial training, formal verification, and diffusion-based purification, are computationally expensive and difficult to deploy in large-scale or resource-constrained environments [5], [41], [93]. Future work should prioritize lightweight defences, hardware-aware optimization, and efficient training methods, particularly for edge and mobile AI systems.

A growing research need is **defence for modern deep learning architectures**, including Vision Transformers and multimodal systems. Most existing work focuses on CNN-based image models, while newer architectures introduce unique vulnerabilities such as attention manipulation and token-level attacks [85], [86]. Designing defences specifically for these advanced architectures remains an important open problem.

The field also suffers from a **lack of standardized benchmarks**. Current studies often use different datasets, threat assumptions, attack strengths, and evaluation metrics, making fair comparison difficult [75]. Establishing standardized benchmarks, unified protocols, and open reproducible leaderboards would significantly accelerate progress.

Another emerging direction is **explainability and trustworthy defence**. Most current defences focus only on robustness while ignoring interpretability. Future systems should be able to explain why an input was rejected or why a prediction is considered secure, making the integration of explainable AI and adversarial defence increasingly important.

Privacy-aware robustness is also gaining attention, as real-world systems require both security and privacy simultaneously. Promising directions include combining differential privacy with robustness, secure federated adversarial training, and privacy-preserving certification [17], particularly in sensitive domains such as healthcare and finance.

A major concern is the **real-world deployment gap**. Many defences demonstrate strong results only on academic datasets such as MNIST and CIFAR-10, while real-world systems involve noisy inputs, distribution shifts, physical constraints, and human interaction. Future work should emphasize deployment-scale testing, industrial case studies, and field validation.

Looking ahead, the long-term vision is **self-adaptive secure AI**—machine learning systems capable of detecting attacks, adapting behaviour, and updating defences autonomously. This represents a shift from static robustness toward intelligent and self-defending systems.

Overall, the next decade of adversarial machine learning research is expected to focus on **unified hybrid defences, scalable certified robustness, modern architecture security, explainable robustness, adaptive defence systems, privacy-integrated learning, and standardized benchmarking**. Together, these directions define the future roadmap toward **secure and trustworthy AI**.

XII.CONCLUSION

This survey presented a comprehensive review of defence mechanisms in adversarial machine learning, covering major defence categories such as adversarial training, input transformation, detection-based methods, certified robustness, and emerging hybrid and adaptive defences. The study highlights that although substantial progress has been made in improving model robustness, no single defence mechanism currently offers complete protection against all adversarial threats. Among existing techniques, adversarial training remains the most effective empirical defence, while certified robustness provides the strongest theoretical guarantees. However, both approaches face important challenges related to computational overhead, scalability, and the trade-off between model robustness and predictive accuracy.

The comparative analysis further indicates that the future of secure machine learning lies in multi-layered defence architectures that combine robustness, detection, certification, and adaptability rather than relying on isolated methods. As machine learning systems continue to expand into real-world and safety-critical applications, defence mechanisms must evolve toward more scalable, adaptive, and explainable solutions. Ultimately, achieving trustworthy machine learning will require a shift from static protection methods to intelligent, self-adaptive security frameworks capable of maintaining long-term robustness, reliability, and user trust in dynamic threat environments.

REFERENCES:

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy,
"Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.
[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus,
"Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2014.
[3] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami,
"The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Security Privacy*, pp. 372–387, 2016.
[4] A. Kurakin, I. Goodfellow, and S. Bengio,
"Adversarial examples in the physical world," in *Artificial Intelligence Safety and Security*, Chapman and Hall/CRC, pp. 99–112, 2018.
[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu,
"Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2018.
[6] N. Carlini and D. Wagner,
"Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Security Privacy*, pp. 39–57, 2017.
[7] A. Athalye, N. Carlini, and D. Wagner,
"Obfuscated gradients give a false sense of security: Circumventing defences to adversarial examples," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 274–283, 2018.
[8] W. Xu, D. Evans, and Y. Qi,
"Feature squeezing: Detecting adversarial examples in deep neural networks," in *Proc. NDSS*, 2018.
[9] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow,
"Thermometer encoding: One hot way to resist adversarial examples," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2018.

[10] C. Guo, M. Rana, M. Cisse, and L. van der Maaten,
"Countering adversarial images using input transformations," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2018.
[11] H. Kannan, A. Kurakin, and I. Goodfellow,
"Adversarial logit pairing," arXiv:1803.06373, 2018.
[12] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan,
"Theoretically principled trade-off between robustness and accuracy," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 7472–7482, 2019.
[13] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry,
"Robustness may be at odds with accuracy," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2019.
[14] J. Cohen, E. Rosenfeld, and J. Z. Kolter,
"Certified adversarial robustness via randomized smoothing," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 1310–1320, 2019.
[15] E. Wong and J. Z. Kolter,
"Provable defences against adversarial examples via the convex outer adversarial polytope," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 5286–5295, 2018.
[16] G. Yang, T. Duan, E. Hu, H. Salman, I. Razenshteyn, and J. Li,
"Randomized smoothing of all shapes and sizes," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020.
[17] Y. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana,
"Certified robustness to adversarial examples with differential privacy," in *Proc. IEEE Symp. Security Privacy*, pp. 656–672, 2019.
[18] E. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart,
"Ensemble adversarial training: Attacks and defences," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2018.
[19] C. Xie, Y. Wu, L. van der Maaten, A. Yuille, and K. He,
"Feature denoising for improving adversarial robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 501–509, 2019.
[20] A. Mustafa, S. Khan, M. Hayat, J. Shen, and L. Shao,
"Image super-resolution as a defence against adversarial attacks," *IEEE Trans. Image Process.*, vol. 29, pp. 1711–1724, 2020.
[21] R. Samangouei, M. Kabkab, and R. Chellappa,
"Defence-GAN: Protecting classifiers against adversarial attacks using generative models," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2018.
[22] P. Samangouei, M. Kabkab, and R. Chellappa,
"Defence-GAN: Protecting classifiers against adversarial attacks using generative models," *arXiv preprint* arXiv:1805.06605, 2018.
[23] S. Gu and L. Rigazio,
"Towards deep neural network architectures robust to adversarial examples," *arXiv preprint* arXiv:1412.5068, 2014.
[24] A. Raghunathan, J. Steinhardt, and P. Liang,
"Certified defences against adversarial examples," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2018.
[25] N. Papernot and P. McDaniel,
"Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning," *arXiv preprint* arXiv:1803.04765, 2018.
[26] K. He, X. Zhang, S. Ren, and J. Sun,
"Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778, 2016.
[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton,
"ImageNet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 25, pp. 1097–1105, 2012.
[28] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard,
"DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2574–2582, 2016.
[29] N. Carlini and D. Wagner,
"Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proc. ACM Workshop Artif. Intell. Security (AISec)*, pp. 3–14, 2017.

[30] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li,
"Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 9185–9193, 2018.
[31] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin,
"Black-box adversarial attacks with limited queries and information," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 2137–2146, 2018.
[32] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh,
"Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proc. ACM Workshop Artif. Intell. Security (AISec)*, pp. 15–26, 2017.
[33] J. Brendel, J. Rauber, and M. Bethge,
"Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2018.
[34] A. Kurakin, I. Goodfellow, and S. Bengio,
"Adversarial machine learning at scale," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2017.
[35] A. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song,
"Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1625–1634, 2018.
[36] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. M. Leung,
"A survey on adversarial machine learning," *IEEE Access*, vol. 6, pp. 12103–12136, 2018.
[37] B. Biggio and F. Roli,
"Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognit.*, vol. 84, pp. 317–331, 2018.
[38] H. Salman, G. Yang, H. Zhang, C.-J. Hsieh, and P. Zhang,
"A convex relaxation barrier to tight robustness verification of neural networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.
[39] M. Balunovic and M. Vechev,
"Adversarial training and provable defences: Bridging the gap," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.
[40] A. Singh, H. Gehr, M. Püschel, and M. Vechev,
"Boosting robustness certification of neural networks," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2019.
[41] V. Tjeng, K. Xiao, and R. Tedrake,
"Evaluating robustness of neural networks with mixed integer programming," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2019.
[42] E. Wong, F. Schmidt, J. H. Metzen, and J. Z. Kolter,
"Scaling provable adversarial defences," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018.
[43] J. Andriushchenko and N. Flammarion,
"Understanding and improving fast adversarial training," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 16048–16059, 2020.
[44] R. Rice, E. Wong, and J. Z. Kolter,
"Overfitting in adversarially robust deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 8093–8104, 2020.
[45] Y. Carmon, A. Raghunathan, L. Schmidt, J. Duchi, and P. Liang,
"Unlabeled data improves adversarial robustness," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.
[46] T. Pang, K. Xu, C. Du, N. Chen, and J. Zhu,
"Improving adversarial robustness via promoting ensemble diversity," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 4970–4979, 2019.
[47] Y. Zhang and P. Liang,
"Defending against whitebox adversarial attacks via randomized discretization," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, pp. 684–693, 2019.
[48] J. Hayes and G. Danezis,
"Learning universal adversarial perturbations with generative models," in *Proc. IEEE Symp. Security Privacy Workshops*, pp. 43–49, 2018.

[49] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard,
"Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1765–1773, 2017.
[50] F. Tramèr and D. Boneh,
"Adversarial training and robustness for multiple perturbations," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.
[51] H. Zhang, H. Chen, C. Xiao, B. Li, D. Boning, and C.-J. Hsieh,
"Towards stable and efficient training of verifiably robust neural networks," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.
[52] M. Balunovic and M. Vechev,
"Adversarial training and provable defences: Bridging the gap," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.
[53] J. Z. Kolter and E. Wong,
"Provable defences in the adversarial polytope framework," *Foundations and Trends in Machine Learning*, vol. 13, no. 3, pp. 313–398, 2020.
[54] G. Zhang, Y. Yu, and M. Jordan,
"Adversarial robustness through local linearization," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.
[55] J. Ding, X. Wang, and Q. Gu,
"Improving adversarial robustness via channel-wise activation suppressing," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.
[56] C. Xie, K. Wang, Z. Zhang, Y. Zhou, S. Xie, and A. Yuille,
"Mitigating adversarial effects through randomization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2018.
[57] J. Buckman, A. Roy, and I. Goodfellow,
"Input transformations and defences against adversarial examples," *arXiv preprint* arXiv:1711.00117, 2017.
[58] T. Pang, K. Xu, and J. Zhu,
"Max-Margin adversarial training," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 1517–1524, 2019.
[59] Y. Balaji, T. Goldstein, and J. Hoffman,
"Instance adaptive adversarial training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1665–1674, 2020.
[60] C. Cui, K. Ren, and Y. Wang,
"Understanding and improving adversarial training under label noise," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, pp. 6645–6653, 2021.
[61] M. Andriushchenko and N. Flammarion,
"Understanding and improving fast adversarial training," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 16048–16059, 2020.
[62] J. Wang, H. Zhang, and E. Xing,
"Should adversarial training be cast as anomaly detection?" in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.
[63] P. Maini, E. Wong, and J. Z. Kolter,
"Adversarial robustness against the union of multiple perturbation models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 6640–6650, 2020.
[64] M. Gowal, C. Qin, J. Uesato, T. Mann, and P. Kohli,
"Uncovering the limits of adversarial training against norm-bounded adversarial examples," *arXiv preprint* arXiv:2010.03593, 2020.
[65] S. Croce and M. Hein,
"Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 2206–2216, 2020.
[66] F. Croce and M. Hein,
"Mind the box: l1-APGD for sparse adversarial attacks on image classifiers," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 2201–2211, 2021.
[67] J. Uesato, B. O'Donoghue, P. Kohli, and A. Oord,
"Adversarial risk and the dangers of evaluating against weak attacks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 5025–5034, 2018.

[68] Y. Salman, G. Yang, and P. Zhang,
"Provably robust deep learning via adversarially trained smoothed classifiers," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.
[69] N. Carlini, F. Tramer, K. D. Dvijotham, L. G. Nguyen, and N. Papernot,
"On evaluating adversarial robustness," *arXiv preprint* arXiv:1902.06705, 2019.
[70] A. Raghunathan, J. Steinhardt, and P. Liang,
"Semidefinite relaxations for certifying robustness to adversarial examples," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018.
[71] E. Wong, L. Rice, and J. Z. Kolter,
"Fast is better than free: Revisiting adversarial training," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.
[72] S. Goldblum, R. Schwarzschild, A. Patel, and T. Goldstein,
"Adversarially robust distillation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, pp. 3996–4003, 2020.
[73] C. Liu, J. Yang, and P. Liang,
"Enhancing certified robustness with noise injection," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, pp. 103–112, 2021.
[74] H. Salman, A. Ilyas, L. Engstrom, and A. Madry,
"Do adversarially robust ImageNet models transfer better?" in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 3533–3545, 2020.
[75] B. Wu, S. Chen, and W. Wang,
"Recent advances in adversarial machine learning defence mechanisms," *IEEE Access*, vol. 10, pp. 45890–45912, 2022.
[76] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan,
"TRADES: A trade-off-inspired defence against adversarial examples," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 7472–7482, 2019.
[77] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu,
"Improving adversarial robustness requires revisiting misclassified examples," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.
[78] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang,
"Recent advances in adversarial training for adversarial robustness," *Neurocomputing*, vol. 488, pp. 241–258, 2022.
[79] Y. Dong, T. Pang, H. Su, and J. Zhu,
"Evading defences to transferable adversarial examples by translation-invariant attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4312–4321, 2019.
[80] F. Croce and M. Hein,
"Minimally distorted adversarial examples with a fast adaptive boundary attack," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 2196–2205, 2020.
[81] F. Croce and M. Hein,
"Reliable evaluation of adversarial robustness with AutoAttack," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 2206–2216, 2020.
[82] J. Ding, X. Wang, and Q. Gu,
"A robust benchmark for evaluating adversarial defence methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5220–5235, 2022.
[83] M. Andriushchenko, M. Hein, and N. Flammarion,
"Square attack: A query-efficient black-box adversarial attack via random search," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 484–501, 2020.
[84] S. Rebuffi, X. Bortolotto, and A. Vedaldi,
"Cross-domain adversarial robustness," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.
[85] M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. Khan, and M.-H. Yang,
"On improving adversarial transferability of vision transformers," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2022.
[86] R. Shao, Z. Shi, J. Yi, P.-Y. Chen, and C.-J. Hsieh,
"On the adversarial robustness of vision transformers," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2022.
[87] K. Mahmood, R. Mahmood, and M. van Dijk,
"On the robustness of vision transformers to adversarial examples," *arXiv preprint* arXiv:2104.02610, 2021.

[88] A. Benz, S. Hamdi, and T. Brox,
"Adversarial robustness of self-supervised learning representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 7462–7471, 2021.
[89] Y. Salman, P. Jain, and A. Madry,
"Do vision transformers improve adversarial robustness?" in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022.
[90] H. Shi, M. Xu, and J. Liu,
"Certified robustness of transformer architectures: A survey," *ACM Comput. Surveys*, vol. 55, no. 8, pp. 1–32, 2023.
[91] Y. Yoon, J. Kim, and S. Oh,
"Diffusion models for adversarial purification and robust inference," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022.
[92] M. Lee and D. Kim,
"Robust evaluation of diffusion-based adversarial purification," *arXiv preprint* arXiv:2303.09051, 2023.
[93] C. Nie, Z. Wang, and Y. Chen,
"DiffPure: Diffusion models for robust defence against adversarial attacks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 18187–18198, 2022.
[94] M. Manani and P. Gupta,
"Enhancing Machine Learning Security: A Comprehensive Survey Of Threats And Attacks On Machine Learning Systems", IJCRT, Vol. 9, Issue 2. Pg. 5585-5602, February-2021.
[95] J. Li, H. Zhang, and Y. Liu,
"Adaptive adversarial defence via dynamic model reconfiguration," *Pattern Recognit.*, vol. 135, Art. no. 109154, 2023.
[96] B. Wu, S. Wei, M. Zhu, and Q. Liu,
"Defences in adversarial machine learning: A survey," *arXiv preprint* arXiv:2312.08890, 2023.
[97] Y. Zhao, X. Wang, and J. Wang,
"Benchmarking certified robustness methods for deep neural networks," *IEEE Access*, vol. 10, pp. 104221–104239, 2022.
[98] H. Yang, C. Qin, and J. Uesato,
"Towards practical certified defences for large-scale deep learning," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2022.
[99] T. Bai and Q. Wang,
"A survey of robust adversarial training in pattern recognition: Fundamental, theory, and methodologies," *Pattern Recognit.*, vol. 131, Art. no. 108889, 2022.
[100] P. Bountakas, A. Zarras, A. Lekidis, and C. Xenakis,
"Defence strategies for adversarial machine learning: A survey," *Comput. Sci. Rev.*, vol. 49, Art. no. 100573, 2023.