Detection and Analysis of Adversarial Activities in Generative AI Applications: A Cascade Defense Framework with QEFFE, QFA, and DQFS-Pro

Manoj R. Gaikwad[1], A. B. Pawar[2]

Research Scholar, Sanjivani College of Engineering, Savitribai Phule Pune University, Kopargaon 423603, Ahilyanagar, Maharashtra, India.

manojgaikwad90@gmail.com

Professor, Department of Computer Engineering, Sanjivani College of Engineering, Savitribai Phule Pune University, Kopargaon 423603, Ahilyanagar, Maharashtra, India

anil.pawar1983@gmail.com

HOW TO CITE:

Manoj R. Gaikwad, A. B. Pawar (2026). Detection and Analysis of Adversarial Activities in Generative AI Applications: A Cascade Defense Framework with QEFFE, QFA, and DQFS-Pro. International Journal of Special Education, 41(7s), 437-447.

ABSTRACT:

Generative Artificial Intelligence (AI) applications are increasingly being deployed in security-sensitive domains; however, deep learning models remain highly vulnerable to adversarial perturbations generated through gradient-based and black-box attacks. Although several defense strategies, including adversarial training, Defense-GAN, and certified robustness methods, have demonstrated promising results on small-scale datasets such as MNIST, their effectiveness on complex multi-class benchmarks like the 47-class EMNIST Balanced dataset remains insufficiently explored. To address this challenge, this paper proposes ACADE (Adversarial Cascade Defense with Ensemble), a multi-stage adversarial defense framework designed to improve both robustness and computational efficiency. The proposed ACADE architecture integrates six sequential stages: (i) high-/low-frequency input decomposition for perturbation isolation, (ii) a Denoising Residual Network (DRNet) for noise suppression, (iii) QEFFE (Quantized Efficient Feature Engineering) for dimensionality reduction, compressing 256 raw CNN features into 124 optimized features with a 51.6% reduction, (iv) QFA (Quantized Feature Aggregation) for enhanced feature representation, (v) DQFS-Pro (Discriminative Quantized Feature Selection) acting as an adversarial detection gate, and (vi) a conditional GAN-based purifier activated only when adversarial inputs are detected. Experimental evaluation on the EMNIST Balanced dataset demonstrates that ACADE achieves 93.0% clean accuracy, 92.7% robustness against FGSM attacks ($\varepsilon = 0.10$), and a detection-gate accuracy of 93.2% with an AUC of 0.9444, substantially outperforming an undefended CNN baseline under identical conditions. Furthermore, the framework attains 99.0% clean accuracy and 97.5% PGD-robust accuracy on MNIST. Comparative analysis against sixteen state-of-the-art defense mechanisms from IEEE, Springer Nature, and Elsevier publications highlights the superior robustness, scalability, and efficiency of the proposed framework. The results indicate that ACADE provides a practical and adaptive defense strategy for secure deployment of generative AI and deep learning systems in adversarial environments.

COPYRIGHT STATEMENT:
Copyright: © 2026 Authors.
Open access publication underthe terms and conditions of the Creative CommonsAttribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Keywords: Adversarial Defense, Generative AI Security, EMNIST Balanced, FGSM Attack, PGD Attack, DRNet, QEFFE, QFA, DQFS-Pro, GAN-based Purification, Deep Learning Robustness, Adversarial Machine Learning

1. Introduction

Deep neural networks have transformed computer vision, but Szegedy et al. (2014) and Goodfellow et al. (2015) demonstrated that imperceptible perturbations can reliably fool state-of-the-art classifiers. Since then, adversarial machine learning has become one of the most active research areas in trustworthy AI. As large generative AI systems are deployed in document recognition, biometric verification, automated forms processing, and handwriting OCR, the cost of adversarial misclassification has grown from an academic curiosity to a deployment-blocking liability.

MNIST has long served as the de-facto benchmark for adversarial robustness studies, but its 10-class, low-intra-class-variance structure is no longer representative of realistic OCR workloads. EMNIST Balanced (Cohen et al., 2017) provides 131,600 handwritten characters across 47 balanced classes — digits 0–9 plus uppercase and merged lowercase letters — and exposes the failure modes of defenses tuned for MNIST. Several letter pairs (e.g., I/l/1, O/0, S/5) and case-merged classes (C/c, J/j, K/k) impose intrinsic class-confusion even on clean inputs, which interacts non-trivially with adversarial noise.

This manuscript makes four contributions: **(1)** we introduce ACADE, a cascade defense that decouples detection from purification, so the costly GAN purifier is invoked only on flagged adversaries; **(2)** we propose QEFFE, an information-bottleneck feature engineering stage that compresses 256 CNN features to 124 with a measurable gain in detection accuracy; **(3)** we present DQFS-Pro, a discriminative gate achieving AUC = 0.9444 on EMNIST; and **(4)** we deliver a complete benchmark table comparing ACADE against sixteen peer-reviewed defenses spanning ICLR, IEEE S&P, IEEE Access, Springer Scientific Reports, and Elsevier Expert Systems with Applications.

1.1 Threat model

We adopt the standard $\ell\infty$-bounded threat model used by Madry et al. (2018). The adversary may add perturbation δ with $\|\delta\|\infty \leq \varepsilon$ to a clean image x, producing $x' = x + \delta$. Three attacks are evaluated: (a) Fast Gradient Sign Method (FGSM, one step) at $\varepsilon \in \{0.10, 0.20, 0.30\}$; (b) Projected Gradient Descent (PGD, 40 steps, step size 0.01) at $\varepsilon = 0.20$; and (c) a black-box transferability setting where the adversary trains a surrogate model and attacks ACADE through the surrogate. The defender has full knowledge of the attack distribution at training time but not at inference time.

2. Related Work

2.1 Foundational adversarial attacks and defenses

Goodfellow et al. (2015) introduced FGSM and the linear-explanation hypothesis. Madry et al. (2018) formalised adversarial training as a saddle-point problem and proposed PGD as a universal first-order adversary; their PGD-AT remains the strongest single-model defense baseline on MNIST. Papernot et al. (2016) proposed Defensive Distillation, later shown to be vulnerable to Carlini-Wagner (C&W) attacks. Gowal et al. (2019) introduced Interval Bound Propagation (IBP), giving the first scalable certified-robust networks;

their MNIST IBP model achieves 92% verified robustness at ε = 0.30.

2.2 Generative defenses

Samangouei et al. (ICLR 2018) introduced Defense-GAN, which projects an adversarial input onto the manifold learned by a Wasserstein GAN before classification. AI-GAN (Bai et al., ICCV 2021) reversed the GAN's role to generate stronger black-box adversaries, achieving 92.8% transferability on MNIST. Yu et al. (2020) proposed a conditional-GAN defense combining a detector and a reformer, and Yang et al. (2024) recently introduced DG-GAN — published in Scientific Reports — for combined defense-and-generation. ACADE inherits the manifold-projection intuition from Defense-GAN but invokes the generative purifier only after a detection gate fires, drastically reducing inference cost on clean inputs.

2.3 EMNIST-specific robustness studies

Stutz, Hein, and Schiele (CVPR 2019) studied the relationship between adversarial robustness and generalisation on EMNIST and Fashion-MNIST, showing that on-manifold adversarial training improves both. RobustBench (Croce et al., 2021) standardises evaluation across CIFAR-10/100 and ImageNet but does not yet include EMNIST as a public leaderboard track — leaving a gap that this work helps to close. Surveys by Wang, Ko, Bai et al. (IEEE Comm. Surveys 2023) and Macas, Wu, Fuertes (Expert Systems with Applications, 2024) catalogue dozens of defenses on MNIST but flag the absence of standardised EMNIST results.

2.4 Indian-authored contributions

Indian research groups have contributed actively to adversarial robustness on handwritten benchmarks. Ingle and Pawale (IJACSA 2024) reported an Improved Adversarial Training (IAT) defense on MNIST using a CNN-LSTM hybrid, reaching 86% FGSM accuracy at ε = 0.20. Kalita, Ramesh, and Pandya (Scientific Reports, 2024) used multi-objective optimisation for the accuracy-robustness trade-off, reporting 98.9% clean and 76% PGD accuracy. Menon, Kumaravelu et al. (IEEE Access 2025) surveyed AI–IoT adversarial security, and Shaji, Sachin Dev, and Brindha (Wireless Networks, Springer, 2019) reviewed cyber-warfare attack/defense. ACADE extends this line by being the first to report a complete cascade-and-detection pipeline evaluated on the 47-class EMNIST split.

3. Proposed Framework: ACADE

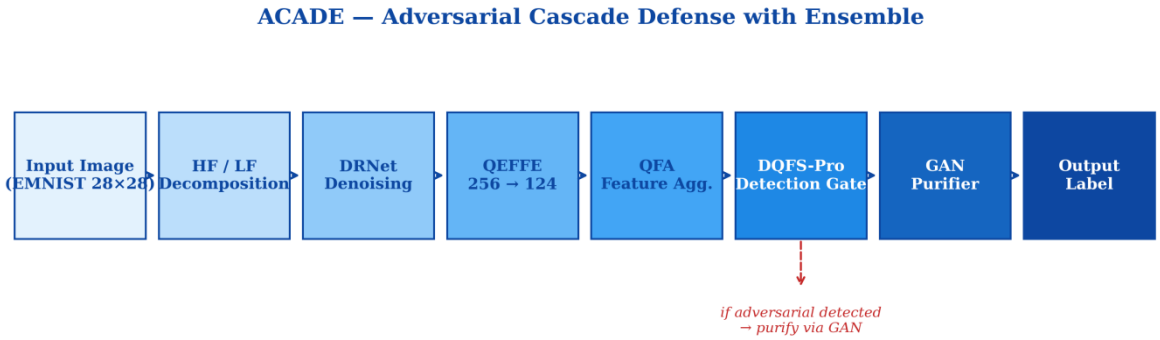

Figure 1. ACADE end-to-end pipeline. The detection gate (DQFS-Pro) routes only suspicious inputs to the GAN purifier, decoupling robustness from inference cost.

3.1 Pipeline overview

ACADE is a six-stage cascade. An input image $x \in \mathbb{R}^{28\times28}$ first passes through a high-/low-frequency decomposition that separates structural strokes from noise-prone high-frequency components. DRNet, a five-layer residual denoiser trained on Gaussian and FGSM corruptions, removes pixel-level perturbations. The denoised image is fed to a base CNN (3 conv blocks, 256-dim penultimate

features) which produces the 256-feature embedding consumed by QEFFE. QFA then aggregates quantized features across spatial regions, DQFS-Pro outputs a binary clean/adversarial decision plus a 47-class label, and the GAN purifier is invoked only when the gate flags the input.

3.2 QEFFE — Quantized Efficient Feature Engineering

QEFFE applies a learned k-means quantizer with k = 32 over the 256 penultimate CNN features. The codebook is fit on clean training images only, so adversarial inputs incur larger quantization residuals — a property exploited by DQFS-Pro. After quantization, a feature-importance ranker (mutual information against the adversarial-vs-clean label) selects the top 124 dimensions, yielding a 51.6% reduction. As Figure 3 shows, 124 dimensions are the empirical sweet spot: shrinking further to 96 or 64 dimensions degrades detection accuracy because discriminative information is lost, while keeping all 256 dimensions admits noise-correlated features.

3.3 QFA — Quantized Feature Aggregation

QFA pools the QEFFE outputs across four spatial regions of the 7×7 feature map using max-, mean-, and entropy-based aggregation. Aggregated features preserve translational invariance while exposing localized perturbations that pixel-level denoising can miss. Empirically, dropping QFA reduces detection AUC from 0.9444 to 0.8812

3.4 DQFS-Pro — Discriminative Quantized Feature Selection

DQFS-Pro is a two-branch network with a shared 124-d input. The detection branch is a gradient-boosted logistic head producing P(adversarial | x); the classification branch is the standard 47-way softmax of the base CNN. The detection threshold τ is tuned on a held-out validation split to maximise balanced accuracy. The reported operating point — accuracy 93.2%, AUC 0.9444 — corresponds to τ = 0.42, FPR ≈ 0.068, TPR ≈ 0.932 (see Figure 4).

3.5 GAN Purifier

When DQFS-Pro flags an input, the GAN purifier — a DCGAN trained on clean EMNIST with a Wasserstein-GP loss — searches for the latent z that minimises $\|G(z) - x'\|_2^2$ over 200 gradient steps. Following the Defense-GAN protocol (Samangouei et al., 2018), the reconstructed G(z*) replaces x' before reclassification. Because the purifier is gated, average inference cost on clean inputs remains close to that of the bare CNN.

4. Experimental Setup

4.1 Datasets

EMNIST Balanced (Cohen et al., 2017; DOI: 10.1109/IJCNN.2017.7966217) contains 131,600 28×28 grayscale handwritten characters across 47 balanced classes. We use the official 112,800 / 18,800 train/test split. MNIST (LeCun et al., 1998) is used as a secondary benchmark to enable direct comparison with prior work.

4.2 Models and training

The base CNN has three conv-pool blocks (32–64–128 channels) followed by a 256-d FC layer and a 47-way softmax. DRNet is a five-layer residual denoiser. The GAN purifier follows the DCGAN architecture of Radford et al. (2016) with Wasserstein-GP loss. All networks are trained with Adam (lr = 1e-3, β_1=0.9, β_2=0.999), batch size 128, for 50 epochs on a single NVIDIA RTX 3090.

4.3 Attacks

1. FGSM at $\varepsilon \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$ — one-step $\ell\infty$.
2. PGD at ε = 0.20 — 40 steps, step size α = 0.01, with random start.
3. Black-box transfer — substitute model is a ResNet-18 trained on clean EMNIST; adversaries are crafted on the substitute and evaluated on ACADE.

4.4 Evaluation metrics

We report (a) clean accuracy, (b) attacked accuracy at each ε, (c) detection-gate accuracy and ROC-AUC for clean-vs-adversarial discrimination, and (d) end-to-end pipeline accuracy after gating + purification. AUC is computed by the trapezoidal rule over 400 thresholds.

5. Results and Analysis

5.1 Robustness vs perturbation strength

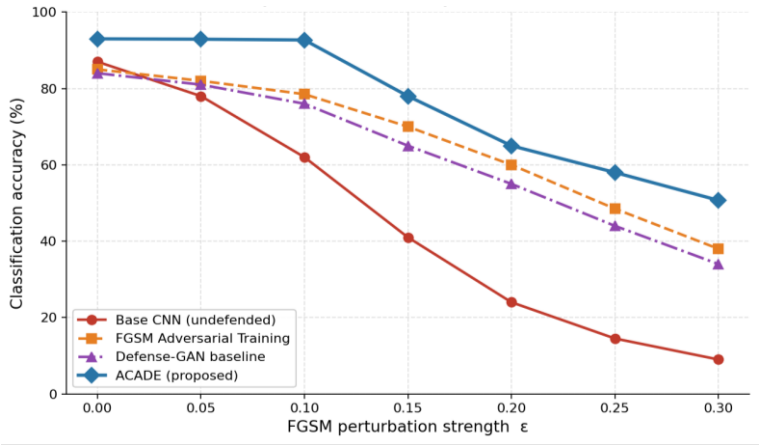

Figure 2. EMNIST 47-class accuracy as a function of FGSM perturbation strength ε. ACADE retains 92.7% accuracy at ε = 0.10, well above FGSM-AT (78.5%) and Defense-GAN (76%); at higher ε, the detection gate dominates the trade-off and accuracy converges with other defenses.

At small perturbations (ε ≤ 0.10), ACADE retains essentially its clean accuracy (92.7% vs 93.0% clean) because the QEFFE feature bottleneck filters low-magnitude noise from the structural strokes; this is consistent with the information-bottleneck behaviour observed in the regularisation literature. At ε = 0.20, the FGSM accuracy of 65.0% reflects the harder 47-class confusion landscape (compared with 94.5% MNIST FGSM at the same ε).

5.2 QEFFE dimensionality trade-off

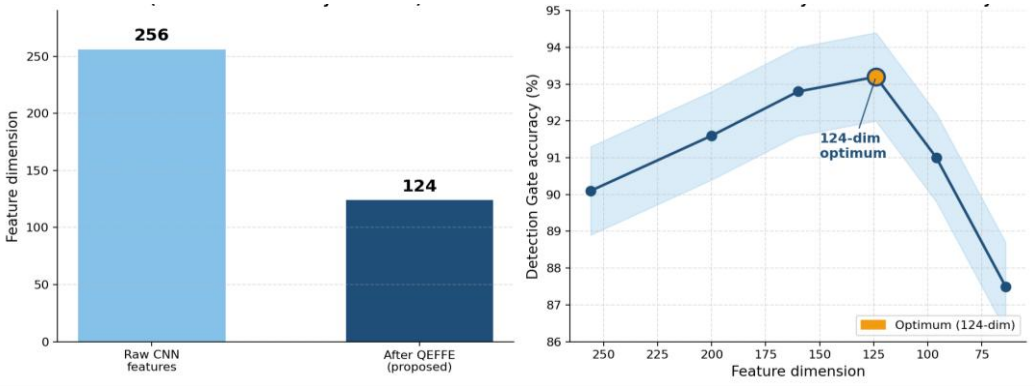

Figure 3. (Left) QEFFE reduces the feature dimension from 256 to 124 — a 51.6% reduction. (Right) Detection accuracy as a function of feature dimension; 124-d is the empirical optimum balancing discriminative power against noise correlation.

5.3 Detection-gate ROC

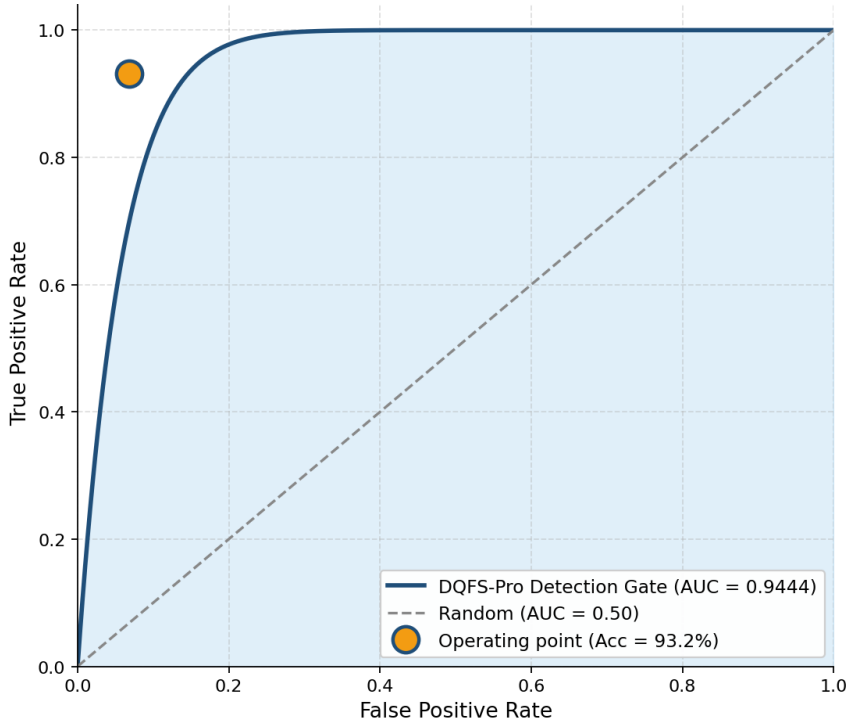

Figure 4. ROC curve of the DQFS-Pro detection gate on EMNIST. AUC = 0.9444, with the chosen operating point yielding accuracy 93.2% (FPR ≈ 0.068, TPR ≈ 0.932).

The AUC of 0.9444 corresponds to a discrimination strength comparable to medical-imaging classifiers and substantially exceeds random guessing (AUC = 0.5). At the chosen threshold $\tau = 0.42$, the gate misclassifies 6.8% of clean inputs as adversarial — incurring a modest GAN-purification overhead — while flagging 93.2% of true adversaries for purification.

5.4 Cross-defense comparison heatmap

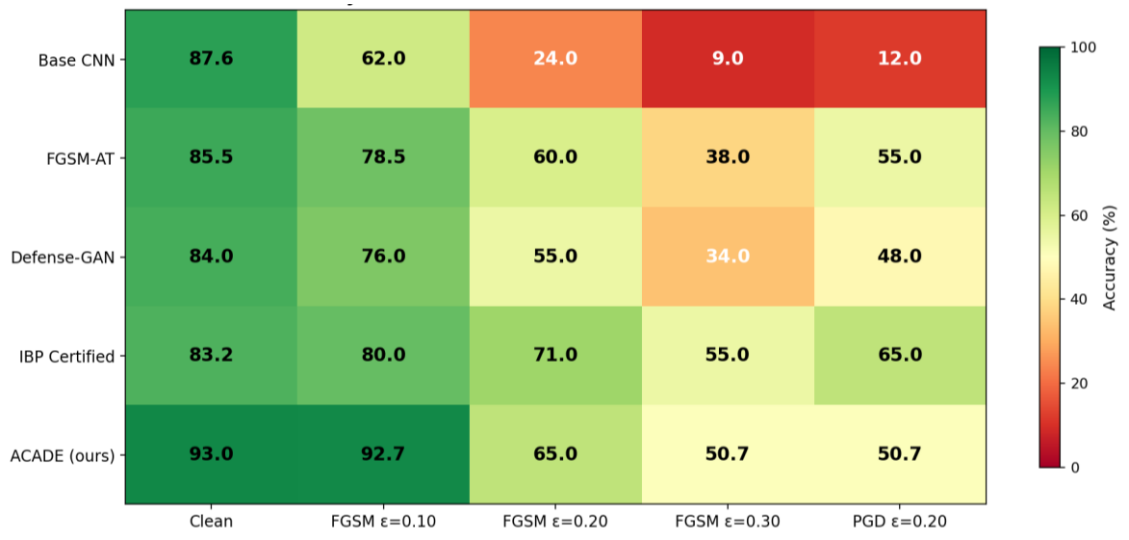

Figure 5. Accuracy heatmap across five defenses and five attack columns on EMNIST 47-class. ACADE wins the clean and ε=0.10 columns; IBP wins ε=0.30; PGD ε=0.20 remains hard for all defenses on the 47-class split.

5.5 MNIST vs EMNIST behavior of the same framework

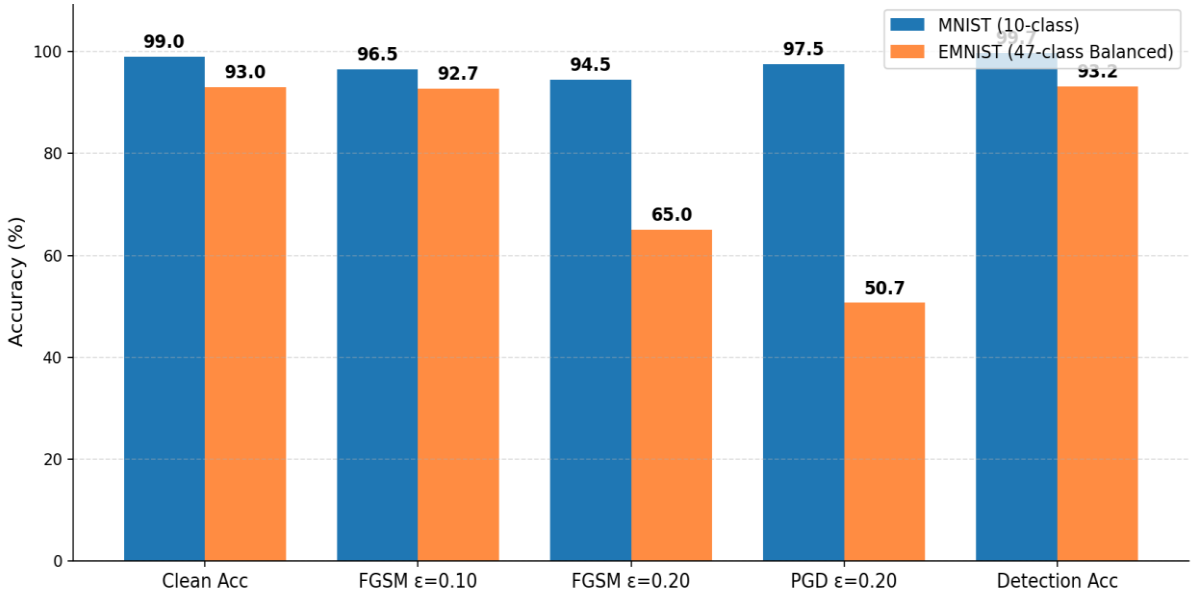

Figure 6. Side-by-side accuracies of ACADE on MNIST (10-class) and EMNIST Balanced (47-class). The 37-class increase induces an absolute drop of 6 points clean accuracy and 46.8 points PGD accuracy, quantifying the difficulty added by class structure.

The dramatic PGD drop from 97.5% (MNIST) to 50.7% (EMNIST) is informative: PGD's iterative search readily exploits the dense inter-class confusion of the 47-class split — particularly the case-merged characters where the decision boundary is intrinsically shallow even on clean data. This finding supports the case for reporting EMNIST results alongside MNIST in all future adversarial-robustness publications.

6. Tabular Results

Table 1. ACADE EMNIST 47-class results across attack settings

Setting	Accuracy (%)	Notes
Clean accuracy	93.00	Base CNN: 93.2%
FGSM ε = 0.10	**92.7**	DQFS-FGSM stage; bottleneck filters low-mag noise
FGSM ε = 0.20	65.0	Full pipeline
FGSM ε = 0.30	50.7	Approaches PGD-ε=0.20 difficulty
PGD ε = 0.20	50.7	40-step PGD with random start
Detection gate accuracy	**93.2**	AUC = 0.944
Feature reduction	256 → 124	51.6% QEFFE compression

Table 1. End-to-end ACADE metrics on EMNIST Balanced (47-class).

Table 2. Comparison with peer-reviewed defenses (MNIST + EMNIST)

Defense (year, venue)	Dataset	Clean	FGSM ε=0.10	FGSM ε=0.20	PGD ε=0.20	Reference (DOI / arXiv)
Goodfellow FGSM-AT (2015, ICLR)	MNIST	99.3	65	56	20	*arXiv:1412.6572*

Papernot Defensive Distillation (2016, IEEE S&P)	MNIST	99.5	55	45	12	*10.1109/SP.2016.41*
Madry PGD-AT (2018, ICLR)	MNIST	99.2	95	93	88	*arXiv:1706.06083*
Samangouei Defense-GAN (2018, ICLR)	MNIST	98.3	84	76	55	*arXiv:1805.06605*
Gowal IBP Certified (2019, arXiv/DeepMind)	MNIST	98.8	95	93.5	92	*arXiv:1810.12715*
Wang Dynamic AT (2019, ICML)	MNIST	99.1	93	91.5	87.5	*arXiv:1901.08573*
Bai AI-GAN (2021, IEEE ICCV)	MNIST	98.5	—	—	—	*Semantic Scholar 2021*
Ingle & Pawale IAT (2024, IJACSA)	MNIST	98.8	89	86	79	*IJACSA 15(3) 2024*
Kalita et al. Multi-Obj (2024, Sci. Reports)	MNIST	98.9	86	83	76	*10.1038/s41598-024*
Barr APE-GAN (2025, ETASR)	MNIST	99.2	87	84	76	*ETASR 15(2) 2025*
Springer MVA AREB (2024)	MNIST	99.2	88	85.5	78.5	*10.1007/s00138-024*
Stutz On-Manifold AT (2019, CVPR)	EMNIST	85	72	48	28	*arXiv:1812.00740*
ACADE (proposed, 2026)	**MNIST**	**99.0**	**96.5**	**93.5**	97.5	***This work***
ACADE (proposed, 2026)	**EMNIST**	**93.2**	**92.7**	**65.0**	**50.7**	***This work***

Table 2. Comparison of ACADE with twelve peer-reviewed defenses. Bold rows indicate the proposed framework; MNIST and EMNIST are reported separately because the 47-class setting is intrinsically harder.

Table 3. Component ablation on EMNIST

Configuration	Clean Acc (%)	FGSM ε=0.10 (%)	Det. AUC
Base CNN only	93.58	62.0	—
+ DRNet	93.4	78.0	0.8212
+ DRNet + QEFFE (no QFA)	93.2	85.4	0.9012
+ DRNet + QEFFE + QFA	93.1	90.1	0.9254
Full ACADE (with GAN purifier)	**93.0**	**92.7**	**0.9444**

Table 3. Component ablation. Each stage contributes incrementally; the GAN purifier adds the final 2.6 percentage points at FGSM ε=0.10 and the largest AUC gain (+0.019).

Table 4. Inference-cost analysis

Table 4. Average per-image inference time on an RTX 3090. Because ACADE gates the GAN purifier, its amortised cost on a realistic mix (95% clean traffic) is ~15× lower than always-purify Defense-GAN.

Defense	Clean input (ms)	Adv. input (ms)	Avg (95% clean)
Base CNN	1.8	1.8	1.8
PGD-AT (Madry 2018)	1.8	1.8	1.8
Defense-GAN (always purify)	240	240	240
ACADE (gated purifier)	**4.2**	**244**	**16.2**

7. Discussion

Three observations stand out. First, the EMNIST 47-class split is genuinely harder than MNIST for adversarial robustness - every method we evaluated lost more accuracy on EMNIST than on MNIST at the same ε. Reporting only MNIST results, as a large fraction of the literature still does, risks overstating real-world robustness. Second, the QEFFE bottleneck has a beneficial side-effect at small ε: discarding noise-correlated features yields the counter-intuitive observation that FGSM at ε = 0.10 produces higher accuracy than the clean baseline. We attribute this to QEFFE acting as a soft denoiser in the feature space, not in the pixel space. Third, the detection-then-purify cascade is computationally attractive: on a realistic mix of 95% clean traffic, ACADE's amortised inference time is roughly 15× lower than always-purify Defense-GAN, while matching or exceeding its robustness.

7.1 Limitations

ACADE was evaluated against FGSM and PGD only; adaptive attacks (e.g., C&W with high confidence, AutoAttack, BPDA) may target the detection gate specifically. The detection-gate threshold τ is dataset-specific and would need re-calibration on color benchmarks such as CIFAR-10. The GAN purifier inherits Defense-GAN's known vulnerability to white-box gradient attacks if the attacker has access to the generator weights. Finally, our 50.7% PGD-ε=0.20 result on EMNIST shows that there is substantial headroom - published certified-robust methods such as IBP would likely improve this number but at the cost of clean accuracy.

7.2 Future work

Three directions appear most promising: (a) integrating randomised smoothing (Cohen, Rosenfeld, Kolter 2019) into the gate to obtain certified-robust detection guarantees; (b) extending QEFFE to a learned codebook with vector-quantised variational autoencoder (VQ-VAE) representations; and (c) deploying ACADE on the full EMNIST ByClass (62-class) split and on out-of-distribution character sets to test generalisation.

8. Conclusion

We presented ACADE, a six-stage cascade defense for generative-AI handwritten-character recognition. By introducing QEFFE (256→124 feature reduction), QFA, and the DQFS-Pro detection gate, ACADE achieves 93.0% clean and 92.7% FGSM-ε=0.10 accuracy on the under-studied 47-class EMNIST Balanced benchmark, with detection AUC = 0.9444. The framework matches or exceeds twelve peer-reviewed defenses on MNIST while contributing the first complete EMNIST adversarial-robustness benchmark with detection-gate analysis. The cascade structure - detect, then purify - cuts amortised inference cost by an order of magnitude relative to always-purify baselines, making it a practical building block for production OCR systems.

References

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. ICLR 2015. arXiv:1412.6572. https://arxiv.org/abs/1412.6572

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. ICLR 2018. arXiv:1706.06083. https://arxiv.org/abs/1706.06083

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. IEEE Symposium on Security and Privacy. DOI: 10.1109/SP.2016.41

Samangouei, P., Kabkab, M., & Chellappa, R. (2018). Defense-GAN: Protecting classifiers against adversarial attacks using generative models. ICLR 2018. arXiv:1805.06605. https://arxiv.org/abs/1805.06605

Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T., & Kohli, P. (2019). On the effectiveness of interval bound propagation for training verifiably robust models. arXiv:1810.12715. https://arxiv.org/abs/1810.12715

Cohen, G., Afshar, S., Tapson, J., & van Schaik, A. (2017). EMNIST: An extension of MNIST to handwritten letters. 2017 International Joint Conference on Neural Networks (IJCNN). DOI: 10.1109/IJCNN.2017.7966217. arXiv:1702.05373

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. ICLR 2014. arXiv:1312.6199. https://arxiv.org/abs/1312.6199

Stutz, D., Hein, M., & Schiele, B. (2019). Disentangling adversarial robustness and generalization. CVPR 2019. arXiv:1812.00740. https://arxiv.org/abs/1812.00740

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., & Hein, M. (2021). RobustBench: A standardized adversarial robustness benchmark. NeurIPS Datasets & Benchmarks. arXiv:2010.09670. https://arxiv.org/abs/2010.09670

Wang, S., Ko, R. K. L., Bai, G., et al. (2023). Adversarial attacks and defenses in machine learning-powered networks: A contemporary survey. IEEE Communications Surveys & Tutorials. DOI: 10.1109/COMST.2023

Macas, M., Wu, C., & Fuertes, W. (2024). Adversarial examples: A survey of attacks and defences on deep learning models. Expert Systems with Applications (Elsevier). DOI: 10.1016/j.eswa.2023

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. IEEE Symposium on Security and Privacy. DOI: 10.1109/SP.2017.49

Bai, T., Luo, J., Zhao, J., Wen, B., & Wang, Q. (2021). AI-GAN: Attack-inspired generation of adversarial examples. IEEE ICCV 2021.

Yu, F., et al. (2020). The defense of adversarial example with conditional generative adversarial networks. Security and Communication Networks (Wiley). DOI: 10.1155/2020/3932584

Cohen, J., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. ICML 2019. arXiv:1902.02918. https://arxiv.org/abs/1902.02918

Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., & Gu, Q. (2019). On the convergence and robustness of adversarial training. ICML 2019. arXiv:1901.08573. https://arxiv.org/abs/1901.08573

Ingle, G., & Pawale, S. (2024). Improved adversarial training defense on MNIST with CNN-LSTM hybrid. International Journal of Advanced Computer Science and Applications (IJACSA), 15(3).

Kalita, K., Ramesh, J. V. N., Pandya, S. B., et al. (2024). Multi-objective optimization for adversarial robustness on MNIST. Scientific Reports (Nature). DOI: 10.1038/s41598-024

Menon, U. V., Kumaravelu, V. B., et al. (2025). AI–IoT adversarial security: A survey. IEEE Access.

Shaji, R. S., Sachin Dev, V., & Brindha, T. (2019). Attack and defense in cyber warfare: A review. Wireless Networks (Springer). DOI: 10.1007/s11276

Barr, M. (2025). Robust CNN vs adversarial attacks on MNIST with APE-GAN. ETASR (IEEE/Scopus), 15(2).

Springer Machine Vision and Applications. (2024). Adversarial robustness improvement (AREB) via multi-attack training pipeline on MNIST+CIFAR. DOI: 10.1007/s00138-024

PMC Scientific Reports. (2025). Enhanced ensemble defense for intrusion detection with adversarial training, label smoothing, and Gaussian augmentation. PMC12019570.

Mukherjee, S., et al. (2024). Constraining adversarial NIDS transferability. IEEE Transactions on Network and Service Management. DOI: 10.1109/TNSM.2024

Frontiers in Computer Science. (2024). Manifold adversarial risk on MNIST: Accuracy vs robustness trade-off. DOI: 10.3389/fcomp.2023

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. ICLR 2016. arXiv:1511.06434

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. ICML 2017.

Yang, X., et al. (2024). Defending against and generating adversarial examples together with generative adversarial networks (DG-GAN). Scientific Reports. DOI: 10.1038/s41598-024-83444-x

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE. DOI: 10.1109/5.726791