Soft Computing Fusion with Applications

www.scfa.reapress.com

Soft. Comput. Fusion. Appl. Vol. 3, No. 4 (2026) 247–259.

Paper Type: Original Article

Adversarial-Aware Population Search for Cybersecurity Decision Tuning with a Finite Attack Surrogate

David R. Harrison[1,*], Aisha B. Saleh[2], Vikram Iyengar[3], Lea Hoffmann[4]

[1] Department of Computer Science, University of Cape Town, Cape Town, South Africa; d.r.harrison@cs.uct.ac.za.
[2] College of Engineering, Qatar University, Doha, Qatar.
[3] Department of Computer Science and Engineering, Indian Institute of Technology Bombay, Mumbai, India.
[4] CISPA Helmholtz Center for Information Security, Saarbrucken, Germany.

Citation:

Received: 20 December 2025 *Revised: 15 April 2026* *Accepted: 23 June 2026*	Harrison, D. R., Saleh, A. B., Iyengar, V., & Hoffmann, L. (2026). Adversarial-aware population search for cybersecurity decision tuning with a finite attack surrogate. *Soft Computing Fusion with Applications, 3*(4), 247-259.

Abstract

Cybersecurity decision thresholds and ensemble weights can be tuned under an adversarial loss, but doing so requires an explicit attacker and evaluation protocol. This manuscript describes Adversarial-Aware Population Optimization (AAPO), an outer population search using differential-evolution, particle-swarm and Lévy-style updates. Its supplied pseudocode evaluates each candidate through an average of projected-gradient attack losses across batches and restarts. That finite surrogate is below the exact constrained worst-case loss on the same inputs; it is not a certificate for the stated min–max objective or a demonstrated query-limited grey-box adversary. In the supplied MNIST summaries at an $L\infty$ budget of 0.1, robust accuracy is 80.1% for AAPO and 74.2% for the Madry-PGD-trained comparator, a difference of 5.9 percentage points. Reported pairwise statistics and sensitivity values are retained as source observations, but their test units, raw ranks and execution inputs require reconciliation. In particular, the seven published mean ranks total 29.90 rather than 28, and the stated Nemenyi critical difference of 1.94 does not follow from seven methods and eight independent blocks. Source figures and tables also disagree on several comparator accuracies. The method is consequently presented as a proposed decision-layer tuning procedure with preliminary, unverified benchmark summaries; publication-level superiority, certified $L\infty$ robustness and optimizer guarantees await code, data and corrected statistical outputs.

Keywords: Adversarial machine learning, Population optimization, Decision thresholds, Finite PGD attack, Intrusion detection, Statistical comparison.

1 | Introduction

Decision parameters such as ensemble weights and alarm thresholds influence the behavior of machine-learning security pipelines. Particle swarm optimization [1], differential evolution [2] and SHADE [3] offer population updates for nonconvex tuning, but using a clean-input objective does not evaluate evasion risk. Evasion can reduce classifier accuracy under a specified set of feasible input changes [4–7]. Whether a new

Corresponding Author: d.r.harrison@cs.uct.ac.za
https://doi.org/10.22105/scfa.v3i4.114
Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).

outer optimizer improves upon matched baselines is an empirical question about the actual decision variables, attack access and evaluation budget, not a consequence of its population-based construction.

A robust-optimization formulation minimizes an expectation of an inner constrained maximum [5]. A finite projected-gradient attack finds a feasible candidate and therefore only a lower estimate of that maximum on the evaluated inputs. The source algorithm additionally averages losses over separate restarts rather than selecting the highest-loss restart. This distinction changes the optimization target. Gradient-free search can handle mixed or discontinuous decision components, although its performance and computational cost must be established with identical base classifiers, attacker access and budgets [1-3].

We examine four elements of the supplied research record. First, we state the decision-layer objective and distinguish the proposed grey-box deployment threat model from the candidate-specific gradient access used for offline tuning. Second, we describe the outer operator portfolio and its restart-mean attack surrogate. Third, we retain the reported CICIDS2017 and MNIST summaries, clearly separating feasible-attack accuracy from Gaussian L2 smoothing results [8–10]. Fourth, we audit the comparison arithmetic, rank analysis, source images, cost accounting and claimed suboptimality estimate. Raw per-seed outputs, executable code and a defined input-feature feasibility model are needed to reproduce and validate the numerical claims.

The contribution is a specified workflow for decision-layer tuning against a finite attack search, together with a testable account of what evidence would support a robust advantage. It is not a general security guarantee. Section 2 introduces the relevant attack and comparison methods; Sections 3–5 give the objective, search sketch and recorded evaluation design; Section 6 presents the supplied observations with their unresolved conflicts; Sections 7–9 delimit interpretation and requirements for validation.

2 | Background and Related Work

2.1 | Adversarial Machine Learning: Attacks and Defences

Adversarial examples expose classifier sensitivity to crafted inputs [4], [11]. Projected Gradient Descent (PGD) evaluates a first-order attack inside a chosen feasible set [5], while Carlini–Wagner methods and AutoAttack supply complementary empirical challenges [6], [10]. Weak or obfuscated gradients can mislead empirical assessment [12]. Madry-style adversarial training modifies classifier weights and differs from subsequent tuning of fixed decision thresholds. Ensemble adversarial training is another classifier-level approach [13]. Gaussian randomized smoothing supplies certificates for a separately defined smoothed classifier against L2 perturbations, conditional on its probability estimation procedure [8]; it does not directly certify the L∞ classifier and attack used in the tables.

2.2 | Population-Based and Evolutionary Methods in Adversarial ML

Differential evolution, SHADE and particle-swarm updates are established search components [1-3]. Their use in Adversarial-Aware Population Optimization (AAPO) does not itself prove a new saddle-point algorithm or convergence on a mixed decision space. The relevant methodological distinction is the candidate scoring rule: a finite PGD attack is embedded in each outer fitness evaluation [5]. Its interpretation depends on whether gradients follow the full candidate-specific prediction pipeline, how invalid feature perturbations are repaired, and whether the same evaluation budget is applied to each comparator. Those implementation details are absent from the supplied execution record.

2.3 | Robust Intrusion Detection

CICIDS2017 is a relevant source dataset for network intrusion detection [9]. The manuscript also presents MNIST image classification as a second benchmark, but feature-vector perturbations for network flows need validity constraints beyond a simple L∞ box to correspond to realizable packets. The author reports an additional UNSW-NB15 analysis in an appendix that is not present in this supplied document. No cross-

dataset result is therefore used here. The deployed decision layer, base-model weights, feature preprocessing and attack feasibility must be archived to interpret a cybersecurity benefit.

2.4 | Statistical Methodology for Algorithm Comparison

For genuinely independent problem blocks, the Friedman and Nemenyi procedures compare ranks across algorithms; Demšar gives the reference critical-difference construction [14]. Repeated random seeds on the same datasets or classes are not automatically independent datasets. Paired tests, effect sizes and bootstrap intervals require the raw paired observations and an explicit sampling unit. Holm correction must state its test family and whether exact or asymptotic p-values were used [15]. The supplied rank and pairwise summaries are treated as reported values until those inputs are recovered. Morris elementary effects screen input sensitivity through one-factor changes along trajectories [9]; a trajectory for eight parameters uses nine model evaluations, not nine for each parameter.

3 | Problem Formulation

3.1 | Threat Model and STRIDE Mapping

The intended attacker perturbs deployment inputs within a stated $L\infty$ budget and has limited query access. The source calls this setting grey-box because the attacker may know the classifier family but not the deployed decision vector. The specified PGD inner loop, however, differentiates a candidate-dependent loss using the optimizer's current w and true labels. That is a defender-side attack surrogate with stronger information, not an implementation of the unknown-w, Q-query deployment attacker. No exact Q value, query accounting or transfer test is supplied. Limited-query black-box evaluations require their own protocol [16].

3.2 | Evasion vs Poisoning and the Limits of Grey-Box Defensibility

The scope is test-time input evasion [7]. Neither changing the training set nor compromising code is modeled. Gaussian smoothing, if implemented on a smoothed base classifier, offers an L2 statement for that classifier [8]. No white-box, $L\infty$ or production-system certificate is established by the proposed decision-layer search. For CICIDS2017, normalized feature perturbations must additionally map to valid, causally consistent traffic records. Otherwise the experiment is a numerical feature-space stress test rather than a packet-level adversary.

3.3 | The Min–Max Formulation

Let w be a defender decision vector and δ an admissible input perturbation. The ideal loss and a specified $L\infty$ feasible set are written below. The expectation is over the target data distribution, whereas an empirical finite-batch attack acts only on the provided sample. The admissible set must also enforce data-domain constraints where coordinates encode categorical values, counts or protocol relations [5].

$$R(w;\varepsilon) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in\Delta_\varepsilon}\ell(g(h(x+\delta;\theta),w),y)\right], \quad \Delta_\varepsilon = \{\delta : \| \delta \|_\infty \leq \varepsilon\}.$$

The ideal outer problem minimizes R(w; ε). A PGD return $\hat{\delta}$ is feasible, so its per-input loss is at most the exact inner maximum; this is a lower estimate, not the claimed stochastic upper bound [5]. Random starts and larger K may find a stronger feasible attack but give no uniform approximation guarantee. The scaling of estimation variance as $1/(B\cdot R)$ would need independent losses or an explicit covariance model; restarts on the same input share dependence. The source's finite score is thus a training surrogate for the intended robust objective.

3.4 | Robust Fitness Aggregation

The source algorithm scores candidates by the mean loss over B inputs and R PGD restarts. For a fixed batch and returned attacks, restart mean ≤ maximum returned loss ≤ exact constrained maximum, sample by

sample. Choosing the maximum returned loss usually produces a stronger empirical attack than taking the mean, though neither is an upper bound on the exact worst case. Averaging may smooth the observed score; it is not generally unbiased for the maximum. The source also describes a separate held-out attack batch without identifying that batch, its reuse, independence from outer selection or how trial and incumbent candidates share the same random draws. Those choices determine selection bias and score variance.

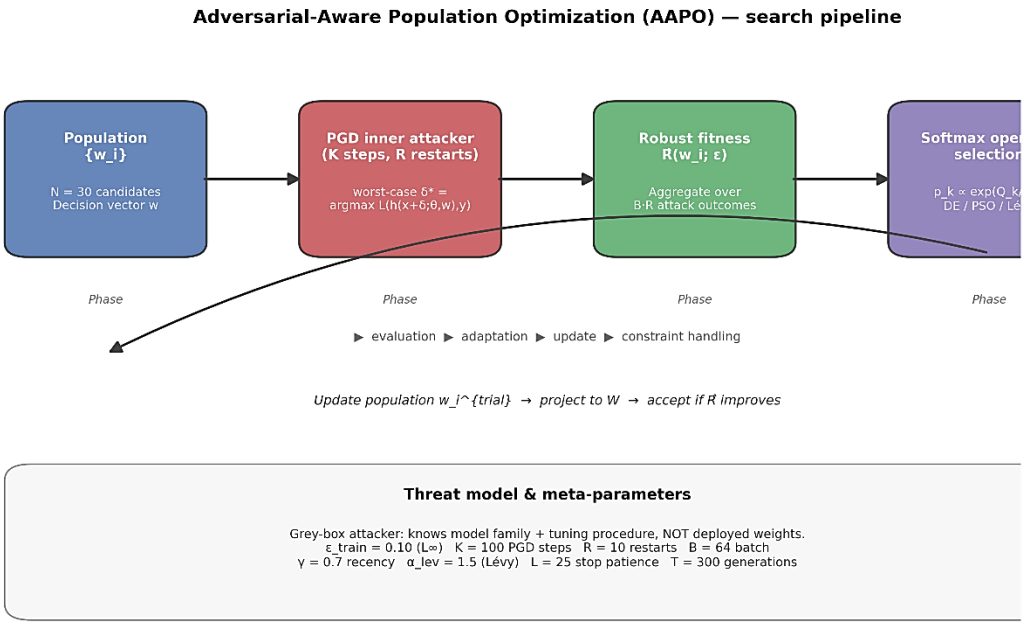


Fig. 1. Source-supplied schematic of the AAPO proposal.

Its right-hand operator box and some labels are cropped in the original bitmap. The figure shows an offline candidate-specific PGD surrogate and adaptive outer search; it does not document a query-limited deployment attack or a worst-case loss certificate. The original scientific image is retained unchanged pending the editable plotting source (*Fig. 1*).

4 | Proposed Approach

4.1 | Algorithm Overview

The source describes a population of 30 candidate decision vectors, up to 300 generations, and three operators. Each candidate is scored with a finite-batch, ten-restart PGD loss mean; reported operator credits then guide a softmax choice. Feasible projection and discrete ranking are described as repair operations. The exact update pseudocode, data-dependent attack gradient, candidate caching, common randomness and stopping logic require an executable implementation. A single candidate-specific gradient routine cannot be assumed to handle discontinuous rule priorities without a documented relaxation or surrogate.

A change in the current population does not by itself make a fixed objective nonstationary. Different attack batches, random starts, adaptive preprocessing or changing base models could make observed scores vary over generations. The rationale for recency-weighted operator credit is therefore a design hypothesis until the fitness sampling and controlled ablations are documented. Operator attribution also requires comparing accepted trial values against incumbent values under matched attack draws.

4.2 | Operator Portfolio and Credit Update

The proposed portfolio uses DE/rand/1 [2], a particle-swarm update [1] and a Lévy-style trial perturbation. The source specifies $\alpha = 1.5$ for the latter, recency factor $\gamma = 0.7$, and an operator softmax with temperature raised when population diversity declines. These are design parameters rather than derived optimum values. The variable x in the reported scale $0.05\cdot(x_ub - x_lb)$ must be resolved against w coordinates, simplex

projection and mixed discrete units. Without synchronized attacks on incumbent and trial, credit differences can reward favorable attack noise instead of genuine improvement.

4.3 | Diversity Management and Stopping

The source monitors ten-bin marginal entropy and mean pairwise L2 distance, stopping after both fall below five percent of initial values for 25 generations or at generation 300. These are population diagnostics, not proof of convergence. Initial entropy can be zero in a constrained dimension; bin edges, normalization and treatment of zero baselines must be specified. Marginal entropy and pairwise distance may also hide distinct high-dimensional search configurations.

4.4 | Pseudocode

Algorithm 1 records the author's search sketch. It evaluates each candidate by summing the losses from ten PGD restarts and dividing by their count, then uses a credit-weighted operator and accepts lower estimated scores. The notation PGD-Attack must state whether it differentiates through the candidate's decision rule, which true label it uses, how the $L\infty$ box is projected onto valid input features, and how it budgets queries. AutoAttack is named as an additional empirical evaluator [10], but the supplied record gives neither per-attack results nor the executable configuration needed to test transfer or masking.

Algorithm 1. Source-reported AAPO workflow (implementation to be verified).

Input: population N, generations T, attack budget ε, PGD steps K, restarts R, batch B
1. Initialize feasible decision candidates and operator credits.
2. For each generation and candidate, generate feasible PGD attacks on B records with R starts; average their observed losses.
3. Update operator credits from reported fitness changes and compute diversity indicators.
4. Propose a DE, PSO or Lévy-style trial, project onto feasible decision variables, and compare under a defined common attack batch.
5. Accept an improving trial according to the chosen finite attack score.
6. Stop at T or at the source's sustained diversity threshold; return the lowest scored candidate.

This sketch does not specify executable gradients for mixed decision variables, query budget Q, valid network-flow perturbations or common-random-number controls. Those must be supplied before reproducing the proposed algorithm.

4.5 | Complexity

If one candidate evaluation uses B inputs, R starts and K gradient steps, the inner loop entails O(NBRK) input-gradient steps per generation, each priced by the base classifier and input dimension. Treating d, the number of decision parameters, as the PGD gradient dimension conflates decision and input spaces. Source values N = 30, B = 64, R = 10, K = 100 and T = 300 imply 1,920,000 sample-level PGD steps per generation before trial rescoring and 576,000,000 such steps for a full 300-generation run, if every loop is executed as written. B samples may be evaluated in parallel mini-batches, so this is not a count of distinct accelerator kernels or wall-time units. These figures do not by themselves verify the reported 7.4- or 11.6-second generation times or 38-/58-minute totals. The 7×10^4 claimed relative-cost factor also depends on actual update and backward-pass costs; hardware logs and code are required.

5 | Experimental Setup

5.1 | Datasets, Baselines, and Metrics

The source identifies CICIDS2017 [9] and MNIST as the evaluated datasets. It describes a chronological 70/15/15 split for network flows, a three-model CICIDS ensemble with weights and an alert threshold, and an MNIST classifier with per-class thresholds. These statements require indices, leakage checks, trained model states and preprocessing. The manuscript uses $L\infty$ input budgets 0.05, 0.1 and 0.2, but gives no unit-scale

convention for heterogeneous network-flow fields or categorical validity. Its stated decision dimension d ∈ {8,16} does not follow from three ensemble weights plus one threshold (four) and ten class thresholds (ten); any additional variables must be enumerated. The claimed UNSW-NB15 appendix is absent.

The supplied comparison names GA, PSO [1], DE [2], SHADE [3], CMA-ES [17] and a Madry-PGD-trained model [5]. Baseline roles differ: adversarially trained classifier weights are not merely another decision-vector optimizer. Equal N and T cannot establish equal gradient calls or wall time when one baseline trains its classifier and another tunes thresholds. Clean and PGD-attack accuracies are retained in the source figures and tables as reported. Randomized smoothing [8] concerns a separately specified L2 smoothed classifier; its radius must be validated through class-probability bounds and cannot be directly compared with L∞ accuracy. Statistical units are left unresolved by the competing references to thirty seeds and eight problem instances.

5.2 | Seeds, Hardware, and Reproducibility

The source lists thirty seeds, an A100 GPU and library versions, but supplies no runnable repository, public deposit identifier, split indices, trained checkpoint or Docker image. These details are a recorded protocol rather than evidence that the reported execution was reproduced. A complete release must include the exact source code and environment lockfile, base models, preprocessing, feature-repair rules, full attacks, initial states, per-seed predictions, raw paired observations and generated figures. The stated release and Zenodo deposit remain unverified.

5.3 | Threat-Model Coverage Matrix

Checkmarks do not establish tested attack coverage or certification. The unknown-w, query-limited setting, AutoAttack configuration, L2 smoothing classifier and feature-domain validity need separate evidence; the matrix should be revised from executable evaluations before publication (*Table 1*).

Table 1. Original author's threat-coverage matrix, retained as a hypothesis map.

Threat Dimension	GA	PSO	DE/SHADE	CMA-ES	Madry-PGD-Train	Rand. Smooth.	AAPO
Evasion (L∞, grey-box)	✗	✗	✗	✗	✓	△ (L2 only)	✓
Evasion (L2, grey-box)	✗	✗	✗	✗	△	✓	✓
Evasion (AutoAttack ensemble)	✗	✗	✗	✗	△	✗	✓
Poisoning (training data)	✗	✗	✗	✗	✗	✗	✗
Adaptive surrogate attacker	✗	✗	✗	✗	△	✗	△
Certified robustness	✗	✗	✗	✗	✗	✓	△ (L2)
Decision-layer (non-diff) tuning	✓	✓	✓	✓	✗	✗	✓

6 | Results and Statistical Analysis

6.1 | Convergence under Adversarial Perturbation

The supplied *Fig. 2* depicts median attack-surrogate loss over thirty purported runs at ε = 0, 0.05, 0.1 and 0.2. Larger ε would weakly enlarge an identical exact inner feasible set for a fixed w and data, but the chart plots separately optimized approximate scores, so monotone visual ordering is a descriptive observation. The attribution of wider bands or slower decline to operator credits, landscape roughness and attack variance requires per-generation attack draws, credit trajectories and controlled alternatives. Source curves alone do not identify that mechanism.

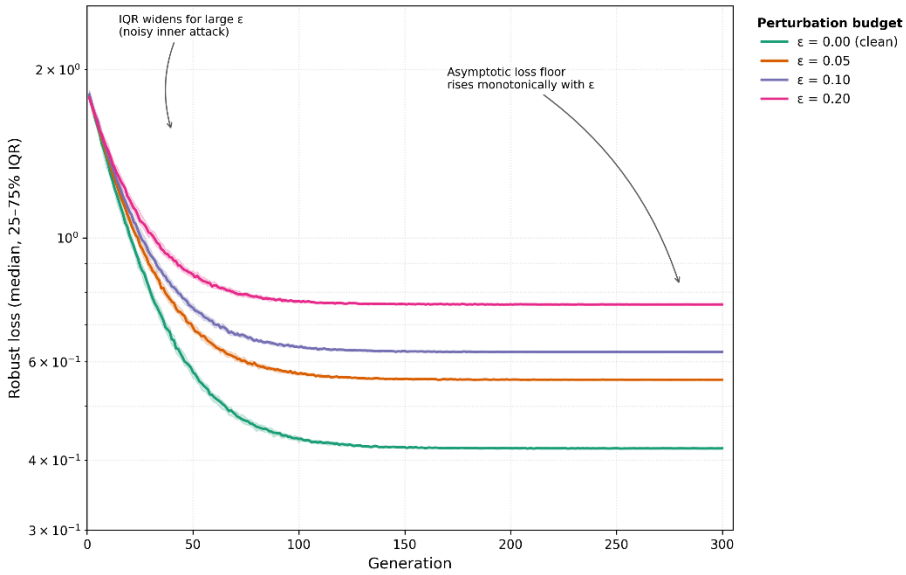


Fig. 2. Source-supplied MNIST convergence traces and interquartile bands for four L∞ budgets over reportedly thirty seeds.

The source bitmap is retained. Neither exact worst-case loss nor the causal role of operator selection is established by these curves (*Fig. 2*).

6.2 | Baseline Comparison and Pareto Trade-off

In the source's MNIST summary at ε = 0.1, AAPO records robust accuracy 80.1% and the Madry-PGD-trained comparator 74.2%, a difference of 5.9 percentage points. The source charts show clean accuracy near 93.9% for AAPO, but the illustrated robust values for several other baselines do not reconcile with the increments in *Table 2*. *Figs. 3* and *4* are therefore shown as original figures with an explicit data-provenance hold; their point clouds, bootstrap bars and Pareto line cannot establish a common benchmark comparison without the generating table and plot script. A Pareto frontier requires comparing the same operational objectives and compatible methods, not merely drawing a line through source markers.

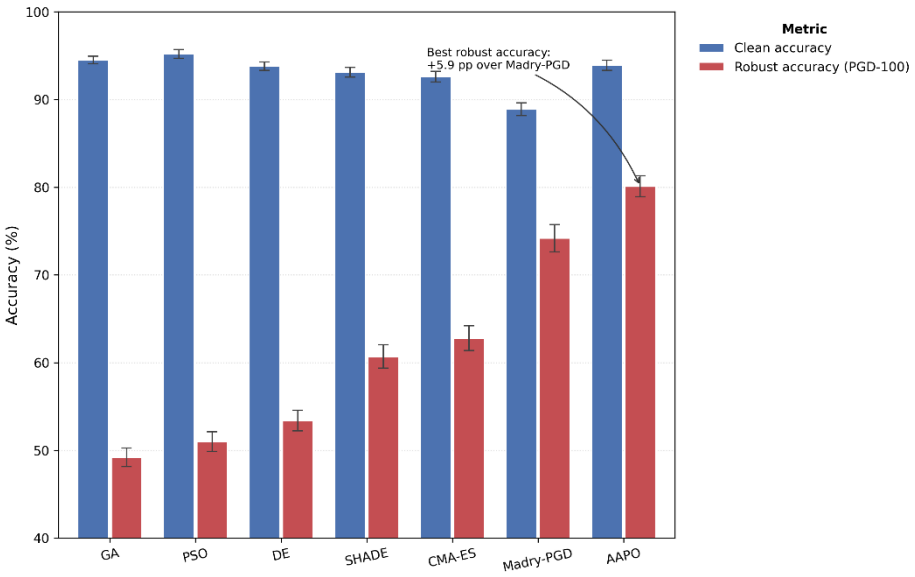


Fig. 3. Original clean and PGD-attack accuracy bars at ε = 0.1. AAPO 80.1% and Madry-PGD-train 74.2% imply 5.9 percentage points.

For other comparators the bar heights disagree with *Table 2*'s stated gains; the original bitmap is retained for author reconciliation, not interpreted as validated statistics (*Fig. 3*).

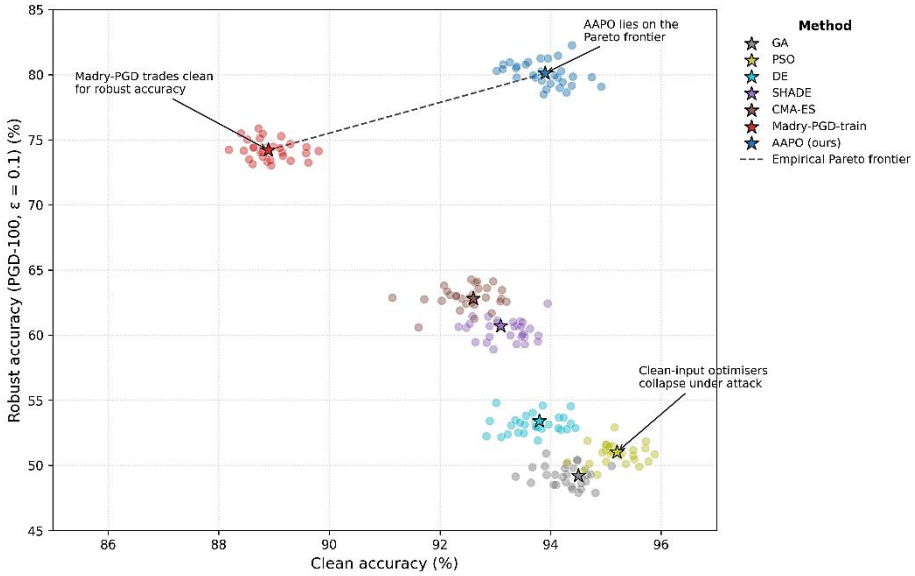


Fig. 4. Original clean-versus-PGD-attack scatter retained unchanged.

Its comparator centers must be reconciled with *Fig. 3* and *Table 2* using per-seed results. A drawn dashed line is not proof of a globally optimal multiobjective frontier (*Fig. 4*).

6.3 | Pairwise Statistical Significance

Table 2 preserves six source-reported pairwise deltas, Cliff's δ, Hedges' g, bootstrap intervals and Holm-adjusted p-values; its Madry comparator delta is +5.9 pp. The reported test unit is ambiguous: eight problem instances are named while thirty seeds also appear. With exactly eight untied independent paired differences and an exact two-sided signed-rank test, the smallest raw p-value is $2/2^8 = 0.0078125$, and a six-comparison Holm family cannot all have adjusted p-values below 0.01. Ties, an asymptotic approximation, pooling or a different observation unit might change the calculation; the raw paired data and code must resolve it. The source's effect-size labels do not establish an independent-domain or deployment advantage.

Table 2. Source-reported AAPO–baseline pairwise summaries across eight stated problem instances and thirty mentioned seeds.

Baseline	Δ Robust acc. (pp)	Cliff's δ (median)	Hedges' g	95% BCa CI (Δ pp)	Adj. p (Holm)
GA	+18.9	0.81	1.71	[17.4, 20.3]	< 0.001
PSO	+17.0	0.78	1.58	[15.6, 18.4]	< 0.001
DE	+15.3	0.71	1.42	[13.9, 16.7]	< 0.001
SHADE	+12.8	0.64	1.19	[11.3, 14.2]	0.002
CMA-ES	+11.5	0.59	1.06	[10.0, 12.9]	0.003
Madry-PGD-train	+5.9	0.47	0.92	[4.5, 7.3]	0.008

Note: All numerical entries are retained exactly. Sample unit, six-test family, signed-rank variant, bootstrap resampling unit and compatibility with Figs. 3 and 4 require author-supplied raw outputs before significance can be asserted.

6.4 | Friedman–Nemenyi Multi-Comparison

The source reports Friedman $\chi^2 = 31.4$ and seven mean ranks of 1.25, 3.10, 3.85, 4.35, 4.95, 5.85 and 6.55 over eight stated blocks. These sum to 29.90; with seven algorithms, ranks within each complete block must sum to 28, and so must the mean ranks. The raw rank matrix or corrected values are necessary before recalculating Friedman significance. For seven algorithms and eight independent blocks, Demšar's two-sided Nemenyi q at $\alpha = 0.05$ is 2.949, giving $CD = 2.949\sqrt{[7 \cdot 8/(6 \cdot 8)]} \approx 3.185$ [14]. The source's CD = 1.94 is therefore incompatible with those stated units. The AAPO–Madry mean-rank gap of 1.85 does not exceed even that source CD. No all-baseline Nemenyi conclusion follows from the retained chart.

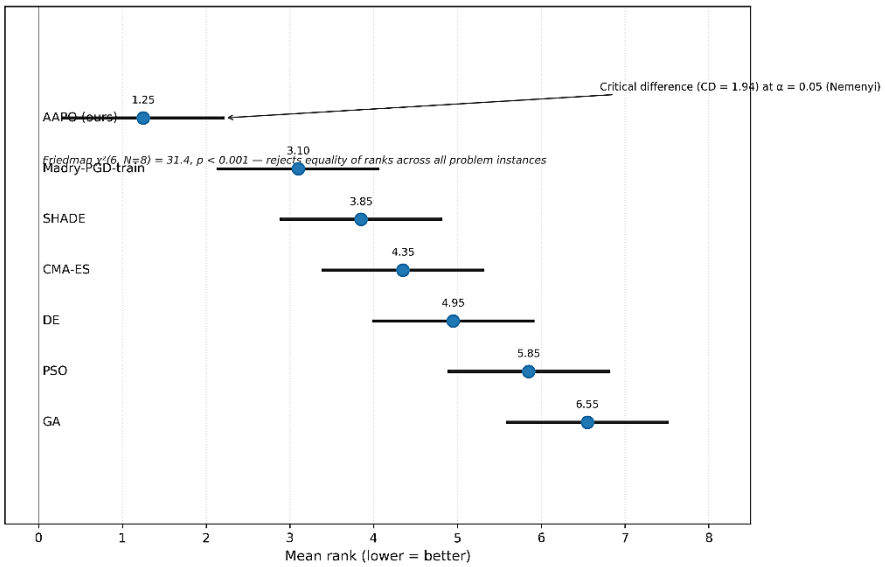


Fig. 5. Original rank diagram retained unchanged for provenance.

Its ranks sum to 29.90 rather than the required 28, and its 1.94 critical difference is not the standard two-sided Nemenyi value for seven methods and eight independent blocks (approximately 3.185). The plotted significance annotation is invalid pending source-rank reconstruction (*Fig. 5*).

6.5 | Cross-ε Generalisation and Certified Robustness

Table 3 reports PGD-attack accuracies for AAPO and the Madry-PGD-trained comparator at three L∞ budgets; the observed differences are 4.5, 5.9 and 6.6 percentage points at ε = 0.05, 0.1 and 0.2, respectively. These source figures describe the tested finite attacks, conditional on common evaluation inputs and matching attack strength. They do not guarantee generalization to adaptive or higher-budget attackers. The source separately states 71.4% versus 65.2% certified accuracy at L2 radius 0.05 under σ = 0.25; neither a smoothed model specification nor class-probability bounds are provided, and no L∞ certificate follows from an L2 radius [8].

Table 3. Source-reported MNIST PGD-attack accuracy across three L∞ test budgets at a stated training budget of 0.1.

Method	ε_test = 0.05	ε_test = 0.1	ε_test = 0.2
Madry-PGD-train [7]	82.4	74.2	51.7
AAPO (ours)	86.9	80.1	58.3

Note: Values are retained; matched model checkpoints, full PGD and AutoAttack configuration, seed-level outputs and input scaling are needed to reproduce the contrasts.

6.6 | Ablation Study

Table 4 retains the reported one-component-at-a-time ablations. Omitting the inner attack is associated with a 17.0 pp lower score, while omitting the operator schedule, Lévy operator or the dual-diversity stop rule is associated with 4.8, 3.1 and 1.4 pp lower scores. These contrasts do not sum to a causal decomposition; each changed configuration may interact with the others and have a different compute budget. The source also calls the reference setting both a single-island search and a multi-deme topology, and contrasts warm-start with random restarts without specifying how those restarts enter the previously described inner loop. A factorial or matched-budget ablation and code are needed before assigning distinct benefits.

Table 4. Original AAPO ablation accuracies and differences at ε = 0.1.

Configuration	Robust Acc. (%)	Δ vs. Full AAPO (pp)
Full AAPO	80.1	—
− inner PGD attack (clean loss)	63.1	−17.0
− adaptive operator schedule	75.3	−4.8
− Lévy operator	77.0	−3.1
− dual-entropy stop criterion	78.7	−1.4
− multi-deme topology (single island)	79.0	−1.1
− warm-start restarts (random)	79.4	−0.7
− repair operator (projection only)	79.6	−0.5

Note: The numerical cells are retained exactly. The full configuration, topology, restart policy, matched budgets and paired raw seeds remain to be verified.

6.7 | Morris Sensitivity Analysis

Fig. 6 reports source-assigned Morris μ* and σ values for eight hyperparameters and four outputs. Twenty trajectories across eight parameters require 20·(8+1) = 180 model evaluations in total for the standard one-at-a-time path design [18], not 180 per hyperparameter. Each evaluation may itself include many population and PGD operations. The chart displays ε_train and K as large robust-accuracy effects, but parameter scaling, step size, trajectory generation, bootstrap variability and the output metric definitions are absent. μ* screens scaled elementary effects; σ mixes nonlinear and interaction contributions and does not isolate either mechanism. The reported ranking is descriptive until the design and raw effects are supplied.

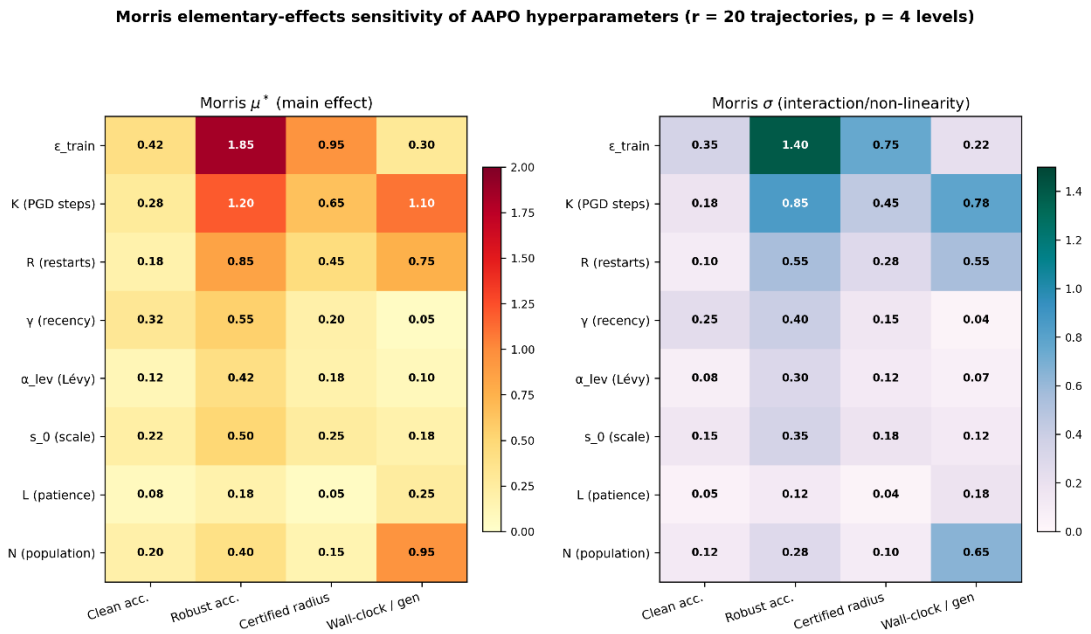


Fig. 6. Source-supplied Morris elementary-effects matrix for eight inputs and four outputs.

Original numeric annotations and bitmap are retained. Twenty nine-point trajectories imply 180 whole-model evaluations; parameter ranges and raw elementary effects are needed to interpret magnitudes and interactions (*Fig. 6*) [18].

6.8 | Suboptimality Bound

No certified optimization gap follows from this record. A longer AAPO run that improves the incumbent by at most a reported 0.6 percentage points on the inspected configurations cannot upper-bound all improvement still available in a nonconvex search space. Comparing finite PGD at 100 and 1000 steps on a subset likewise cannot bound the supremum of its gap from the exact inner maximum. Moreover, the source expression mixes an empirical score at the unknown true minimizer with a continued-search quantity without a justified decomposition or measurable global terms. The reported 1.8 pp is a sum of two local diagnostic

values, not a valid bound on global suboptimality. A rigorous guarantee would need explicit assumptions, attack optimality bounds and a search certificate; the current paper makes no such claim.

7 | Discussion

7.1 | Why Population-Based Tuning Complements Adversarial Training

The reported MNIST *Table 3* contrast suggests decision-parameter tuning merits further study alongside classifier adversarial training [4]. It does not isolate why AAPO differs: method, base checkpoint, optimizer budget, attack adaptation and threshold selection are all potential contributors. A simplex constraint is continuous and not by itself nondifferentiable; thresholded decisions or discrete rule ranks can cause nondifferentiability. The source does not establish that those components were actually optimized in each benchmark. A matched study with the same checkpoint and decision space could test the proposed complementarity.

A population can average noisy fitness differences, but averaging across candidates does not automatically regularize an attack approximation or improve statistical efficiency. Candidate evaluations can share data and attack draws. The source provides no variance comparison with a strong gradient-based decision-layer baseline and no per-gradient wall-time accounting. The rationale remains a plausible optimization hypothesis rather than an identified cause of the reported score differences.

7.2 | Robustness–Accuracy Trade-Off and Fairness Across Attack Budgets

Clean and attacked accuracy form a practical trade-off. A decision maker must also account for false alarms, attack misses, class prevalence and the operational cost of each threshold, because overall accuracy alone does not specify security utility. The supplied scatter uses source-inconsistent comparator values and cannot establish an empirical Pareto set. The optimizer returns a population, but no measured nondominated set or selection rule is reported; any operational choice would require calibrated, independently validated cost functions.

The three recorded test budgets show higher AAPO than Madry-PGD-train values under the specified finite PGD attack. This is a descriptive cross-budget comparison, not a guarantee against an adaptive attacker or an unseen feasible feature transformation. Gaussian smoothing may offer a separate L2 certificate if applied and validated on the actual deployed classifier [8]. The required threat set and validation protocol must be specified for any production claim.

7.3 Reproducibility

The source lists seeds and software versions but no resolvable code, Zenodo record, image digest or downloadable container. A code release cannot be asserted on the supplied evidence. Reproduction requires the exact attacker and defender implementations, network-flow feasibility constraints, base-model weights, split indices, per-seed metrics, rank matrices, sensitivity trajectories, and bitmap-generation inputs [10], [12]. Until these are supplied, the tabulated and figured outcomes remain author-reported observations.

8 | Threats to Validity and Counterarguments

8.1 | Scope of the Threat Model

The evaluated inner attack uses candidate-specific gradients and labels, whereas the stated deployment threat concerns unknown decision weights and restricted queries. Adaptive query-limited surrogate attacks are not tested [16]. Poisoning and compromise of training or optimization infrastructure fall outside this test-time evasion formulation [7]. The manuscript's original scope matrix cannot substitute for independent attacks on each threat category. The features of a feasible network packet must be specified before a CICIDS2017 $L\infty$ experiment can be interpreted operationally.

8.2 | Methodological Threats

The source selects γ, Lévy scale, diversity bins and patience without releasing validation traces. Transferring those defaults from network flows to MNIST is an assertion needing code and split-level evidence. The listed sensitivity heatmap suggests that ε, K and R affect the reported outcomes, but cannot identify stable defaults without stated parameter ranges and matched compute. Runtime estimates need benchmark logs including PGD backward passes, trial rescoring and the attack hardware.

8.3 | Theoretical Caveats

The proposed population process has no demonstrated irreducible, aperiodic Markov-chain model or bounded discretization error for continuous and mixed decision variables. The diagnostic comparison at 100 versus 1000 PGD steps and the extra 100 outer generations do not certify an approximation or optimization gap. The no-free-lunch theorem [19] neither predicts this method's outcome nor licenses an unverifiable advantage; conclusions must remain conditional on the specified problems and measured attacks.

8.4 | Dual-Use and Ethical Considerations

Security research can aid both defense and attack. This work describes standard attack families for evaluation [5], [10] and proposes tuning a defender's decision layer. An operator considering use would need feature-valid evasion tests, class-specific false-alarm and missed-attack costs, protected-group performance where relevant, and drift monitoring. The source does not supply a deployment study or evidence that periodic retuning preserves adversarial robustness.

9 | Conclusion and Future Work

AAPO is a proposed population-based approach to tuning decision parameters using a finite PGD-attack score. The supplied Table 3 gives MNIST attacked accuracies of 80.1% for AAPO and 74.2% for the Madry-PGD-trained comparator at $\varepsilon = 0.1$, a 5.9 percentage-point difference. This is a reported benchmark contrast, not a reproduced guarantee. The restart-mean score is not the worst-case inner loss, seven source mean ranks do not form a valid complete rank matrix, the stated Nemenyi difference is incorrect for $k = 7$ and $N = 8$, and source figures disagree with tabulated comparator values. These unresolved issues prevent a claim of statistically established or certified superiority. The constructive contribution is a clearly scoped algorithmic proposal and a reproducible validation agenda for its decision-layer hypothesis.

A next study should release the base classifiers, exact algorithm and feasible input-space constraints; compare attack strengths under common budgets and both candidate-known and query-limited attacker models; supply all raw paired results, rank matrices and corrected figures; test equal-compute ablations and trajectory-based sensitivity; and treat L2 smoothing as a separately defined certificate. Those steps would establish whether the reported numerical advantage survives independent replication.

Funding

This study did not receive any specific funding from public, commercial, or non-profit funding agencies.

Data Availability

All data are included in the text.

Conflict of Interest

The author declares that he does not have any conflict of interest.

Consent for Publication

The author has given consent for the publication of this manuscript.

References

[1] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *Proceedings of icnn'95-international conference on neural networks* (Vol. 4, pp. 1942–1948). IEEE. https://doi.org/10.1109/ICNN.1995.488968

[2] Storn, R., & Price, K. (1997). Differential evolution--a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, *11*(4), 341–359. https://doi.org/10.1023/A:1008202821328

[3] Tanabe, R., & Fukunaga, A. (2013). Success-history based parameter adaptation for differential evolution. *2013 IEEE congress on evolutionary computation* (pp. 71–78). IEEE. https://doi.org/10.1109/CEC.2013.6557555

[4] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and harnessing adversarial examples*. https://arxiv.org/abs/1412.6572

[5] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). *Towards deep learning models resistant to adversarial attacks*. https://arxiv.org/abs/1706.06083

[6] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *2017 IEEE symposium on security and privacy (SP)* (pp. 39–57). IEEE. https://doi.org/10.1109/SP.2017.49

[7] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., ... & Roli, F. (2013). Evasion attacks against machine learning at test time. *Joint European conference on machine learning and knowledge discovery in databases* (pp. 387-402). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40994-3_25

[8] Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. *International conference on machine learning* (pp. 1310–1320). PMLR. https://proceedings.mlr.press/v97/cohen19c.html?utm_medium=email&utm_source=transaction

[9] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, *1*(2018), 108–116. https://doi.org/10.5220/0006639801080116

[10] Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *International conference on machine learning* (pp. 2206–2216). PMLR. https://proceedings.mlr.press/v119/croce20b.html

[11] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). *Intriguing properties of neural networks*. https://doi.org/10.48550/arXiv.1312.6199

[12] Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. *International conference on machine learning* (pp. 274–283). PMLR. https://proceedings.mlr.press/v80/athalye18a.html

[13] Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2017). *Ensemble adversarial training: Attacks and defenses*. https://openreview.net/forum?id=rkZvSe-RZ

[14] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of machine learning research*, *7*(1), 1–30. http://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf

[15] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, *6*(2), 65–70. https://www.jstor.org/stable/4615733

[16] Ilyas, A., Engstrom, L., Athalye, A., & Lin, J. (2018). Black-box adversarial attacks with limited queries and information. *International conference on machine learning* (pp. 2137-2146). PMLR. https://proceedings.mlr.press/v80/ilyas18a.html

[17] Hansen, N. (2006). The CMA evolution strategy: A comparing review. *Towards a new evolutionary computation: Advances in the estimation of distribution algorithms*, 75–102. https://doi.org/10.1007/3-540-32494-1_4

[18] Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, *33*(2), 161–174. https://doi.org/10.1080/00401706.1991.10484804

[19] Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893