Uncertainty-Aware Classifier with Physics-Based Rejection (UA-PBR): A Proof-of-Concept Under Computational Constraints

Mohsen Mostafa *

Posted Date: 12 May 2026

doi: 10.20944/preprints202603.0748.v3

Keywords: physics-informed machine learning; bayesian deep learning; reject option classification; out-of-distribution detection; scientific machine learning; partial differential equations (PDEs); darcy flow; uncertainty quantification; robustness

 doi:10.20944/preprints202603.0748.v3

Article

Uncertainty-Aware Classifier with Physics-Based Rejection (UA-PBR): A Proof-of-Concept Under Computational Constraints

Mohsen Mostafa

Independent Researcher; mohsen.mostafa.ai@outlook.com

Abstract

Deep learning classifiers deployed in scientific applications often encounter inputs that violate physical laws (e.g., due to sensor failure or corruption). Standard methods cannot detect such violations and may produce confident but wrong predictions. We propose UA-PBR, a framework that combines a physics-informed autoencoder (to detect physics violations) with a Bayesian CNN (to quantify predictive uncertainty). Inputs are rejected if either the PDE residual exceeds a threshold or the predictive entropy is too high. As a proof-of-concept, we evaluate UA-PBR on a synthetic Darcy flow dataset (32×32 grid) under severe computational constraints (Google Colab, 10 seeds). Despite these limitations, UA-PBR reduces classification risk by over 90% on heavily corrupted samples while accepting 89.7% of clean inputs with 99.99% accuracy on accepted samples. Ablation studies confirm that both components contribute synergistically. These preliminary results on a synthetic benchmark illustrate the potential of physics-aware rejection and motivate further investigation with larger-scale experiments. Code is available at: https://github.com/UA-PBR/UA-PBR.

Keywords: physics-informed machine learning; bayesian deep learning; reject option classification; out-of-distribution detection; scientific machine learning; partial differential equations (PDEs); darcy flow; uncertainty quantification; robustness

1. Introduction

Neural networks are increasingly used in scientific domains where data must satisfy physical laws (e.g., partial differential equations). However, test-time inputs may be corrupted by sensor noise or artifacts, leading to physically impossible observations. Standard classifiers have no mechanism to detect such violations and will produce predictions regardless, potentially with high confidence.

Existing approaches address robustness in isolation: normalization methods adapt to noise but cannot detect physics violations; Bayesian networks quantify uncertainty but ignore domain knowledge; physics-informed learning embeds constraints during training but offers no rejection mechanism at inference. Each approach addresses part of the problem, but a unified framework that leverages both physics and uncertainty for rejection remains relatively unexplored.

All experiments were conducted under severe computational constraints (free Google Colab, synthetic Darcy flow dataset, limited training epochs) and should be interpreted as preliminary proof-of-concept evidence rather than definitive claims of superiority. Despite these constraints, UA-PBR consistently reduced risk across all seeds in our experiments, suggesting the approach is robust and merits further investigation with greater computational resources and more realistic datasets.

Our contributions are threefold:

1. Theoretical framework: We provide simple theoretical bounds linking PDE residuals to reconstruction error and bounding expected risk under Lipschitz continuity, though these bounds are not tight and serve primarily as conceptual motivation.

 doi:10.20944/preprints202603.0748.v3

2. Two-stage rejection architecture: UA-PBR combines a physics-informed autoencoder (to detect physics violations) with a Bayesian CNN (to quantify predictive uncertainty) in a unified decision-theoretic framework.
3. Proof-of-concept validation: Under severe computational constraints, we demonstrate that the approach can dramatically reduce risk on corrupted inputs, motivating further research.

The paper is organized as follows. Section 2 discusses related work. Section 3 describes the methodology. Section 4 presents experimental results with explicit discussion of computational limitations. Section 5 discusses limitations and future work. Section 6 concludes.

2. Related Work

2.1. Normalization and Robustness

Normalization methods [9,10] stabilize training but assume clean distributions and treat all inputs uniformly—on corrupted CIFAR-10-C, they still suffer 20–40% accuracy drops [6]. Adaptive methods like Batch Renormalization [11] modulate statistics but cannot detect physics violations.

2.2. Bayesian Deep Learning

Bayesian neural networks [12,13] provide uncertainty estimates but add computational cost and ignore domain knowledge. Monte Carlo Dropout [14] offers a practical approximation, while deep ensembles [15] often outperform more complex Bayesian methods. However, these methods operate purely in data space and cannot distinguish between uncertainty from lack of data and uncertainty from physical impossibility.

2.3. Physics-Informed Machine Learning

Physics-Informed Neural Networks (PINNs) [1,16] embed governing equations into the loss function, achieving high accuracy on benchmark PDEs. Neural operators [17,18] extend this to learning mappings between function spaces. However, these methods assume test data satisfies the same physics as training data and offer no rejection mechanism at inference.

2.4. Out-of-Distribution Detection

OOD detection methods [19,20] identify inputs that differ from the training distribution using techniques like ODIN [21] or Mahalanobis distance [22]. These methods cannot distinguish between benign distribution shift and physics-violating corruptions.

2.5. Learning with Rejection

Classical reject option classifiers [23,24] provide theoretical foundations for abstention but assume known class-conditional distributions and do not incorporate physical constraints. Selective classification [25] and learning with rejection [26] have been studied, but these methods lack mechanisms to leverage domain-specific knowledge.

2.6. Our Contribution

UA-PBR explores the underexplored direction of combining physics-informed filtering with Bayesian uncertainty for rejection. Unlike prior work that treats these signals separately, we investigate whether their integration can yield synergistic improvements. This study is a proof-of-concept under constrained resources, not a claim of deployment readiness.

3. Methodology

3.1. Problem Setup

 doi:10.20944/preprints202603.0748.v3

We consider a Darcy flow governed by $-\nabla\cdot(a\nabla u)=f$ with permeability field a and pressure field u. Inputs are observed pressure fields $u_{obs} = u + \phi$, where ϕ is corruption. The task is to classify whether the mean permeability is high or low.

3.2. Stage 1: Physics-Based Corruption Detection

We train an autoencoder Ψ_θ to reconstruct (u,a) from u_{obs} using a physics-informed loss:

$$\mathcal{L} = \|u - \hat{u}\|^2 + \|a - \hat{a}\|^2 + \beta \, \|N_{\hat{a}}[\hat{u}] - f\|^2 \quad (1)$$

At test time, we compute the physics score $S_{phy}(u_{obs}) = \|N_{\hat{a}}[\hat{u}] - f\|$. Inputs with $S_{phy} > \tau_{phy}$ are rejected as physics-violating.

3.3. Stage 2: Bayesian CNN Uncertainty

We use a CNN with Monte Carlo Dropout (50 samples) to obtain predictive distribution $p(y|u_{obs})$. Predictive entropy $U(u_{obs}) = H[p(y|u_{obs})]$ serves as uncertainty score. Inputs with $U > \tau_{unc}$ are rejected due to model uncertainty.

3.4. Joint Rejection Rule

The final decision:

$$q(u_{obs}) = \{ \text{reject, if } S_{phy} > \tau_{phy} \text{ or } U > \tau_{unc}; \hat{y}, \text{otherwise} \} \quad (2)$$

Thresholds are selected on a validation set to minimize risk with rejection cost λ.

3.5. Theoretical Bounds (Conceptual)

We provide simple theoretical bounds that serve as motivation:

Theorem 3.1 (Error Bound). For any input with physics score $S_{phy}(u_{obs}) \le \tau_{phy}$, the reconstruction error satisfies

$$\|u_{obs} - u\|_{[L^2]} \le \alpha \, \tau_{phy} + \|\phi\|_{[L^2]} + \gamma_n, \quad (3)$$

where α is the coercivity constant and γ_n is the autoencoder approximation error.

Theorem 3.2 (Risk Bound). Under Lipschitz continuity, the expected risk satisfies

$$R_\lambda(q) \le \lambda + \varepsilon_0 + L \, \delta, \quad (4)$$

where $\delta = \tau_{phy}/\alpha + \gamma_n$. This bound is not tight but provides a conceptual link between physics residuals and prediction reliability.

Theorem 3.3 (Existence of Optimal Thresholds). Given a finite validation set, the empirical risk is piecewise constant; hence a minimizer exists.

These bounds are intended as conceptual motivation rather than tight guarantees.

4. Experiments

4.1. Computational Resources

All experiments were conducted on a free Google Colab instance with an NVIDIA T4 GPU (16 GB VRAM) and approximately 15 GB system RAM. Due to session time limits (~12 hours maximum runtime) and shared resources, we limited training epochs and used 10 independent seeds for statistical power. The Darcy flow dataset is synthetic (10,000 samples at 32×32 resolution) and should

 doi:10.20944/preprints202603.0748.v3

not be interpreted as representative of real-world complexity. These constraints mean our results should be viewed as preliminary proof-of-concept evidence.

4.2. Dataset

We evaluate on the Darcy flow benchmark, generating 10,000 samples at 32×32 resolution with binary classification labels based on mean permeability. The dataset is split 70/15/15 for train/val/test.

4.3. Corruption Types

We simulate four realistic corruptions at severity levels 0.1, 0.3, 0.5, 0.7, 0.9:

- Gaussian noise: Additive white noise $N(0,\sigma^2)$
- Salt-and-pepper: Random pixels set to ±2 with probability p
- Structured artifacts: 8×8 blocks replaced with random values
- Physics-violating: Non-solenoidal components added via curl of a random vector field

4.4. Architecture

- Physics autoencoder: CNN with 4 conv layers (channels 32→64→128→256), latent dim 256
- Bayesian CNN: 4 conv layers with dropout rate 0.3, MC samples 50
- Standard CNN: Same architecture without dropout

4.5. Training

All models trained with AdamW (lr=10^{-3}, weight decay=10^{-4}) for 150 (autoencoder) and 200 (CNN) epochs. Gradient clipping at 1.0, ReduceLROnPlateau scheduler. Experiments run with 10 independent seeds.

4.6. Baselines

We compare against:

1. Standard CNN (no rejection)
2. MaxProb rejection (threshold on maximum softmax probability)
3. Deep Ensemble (3 models)
4. Physics-only rejection (no uncertainty)
5. Uncertainty-only rejection (no physics)

4.7. Main Results

Key findings:

- UA-PBR dramatically reduces risk on severe corruptions (severity 0.9) while maintaining low risk on clean data.
- Acceptance rates on clean data are ~90% with near-perfect accuracy on accepted samples.
- The physics filter perfectly separates clean from corrupted inputs (τ_{phy} = 1.0) on this synthetic dataset.
- Uncertainty rejection further reduces risk on ambiguous clean samples.

Table 1. UA-PBR Performance Across Corruption Types and Severities (mean ± std, 10 seeds).

Condition	UA-PBR Risk	Std CNN Risk	Acceptance Rate	Accuracy (accepted)
Clean	0.0310 ± 0.0021	0.0021 ± 0.0016	0.897 ± 0.007	0.9999 ± 0.0004
Gaussian (0.9)	0.0393 ± 0.0042	0.5005 ± 0.0136	0.87 ± 0.01	0.9984 ± 0.0015
Salt-Pepper (0.9)	0.0598 ± 0.0157	0.5005 ± 0.0136	0.82 ± 0.02	0.9921 ± 0.0058
Structured (0.9)	0.0322 ± 0.0054	0.0675 ± 0.0950	0.89 ± 0.01	0.9996 ± 0.0004

 doi:10.20944/preprints202603.0748.v3

Condition	UA-PBR Risk	Std CNN Risk	Acceptance Rate	Accuracy (accepted)
Physics-Violating (0.9)	0.0338 ± 0.0040	0.5005 ± 0.0137	0.89 ± 0.01	0.9992 ± 0.0010

4.8. Ablation Study

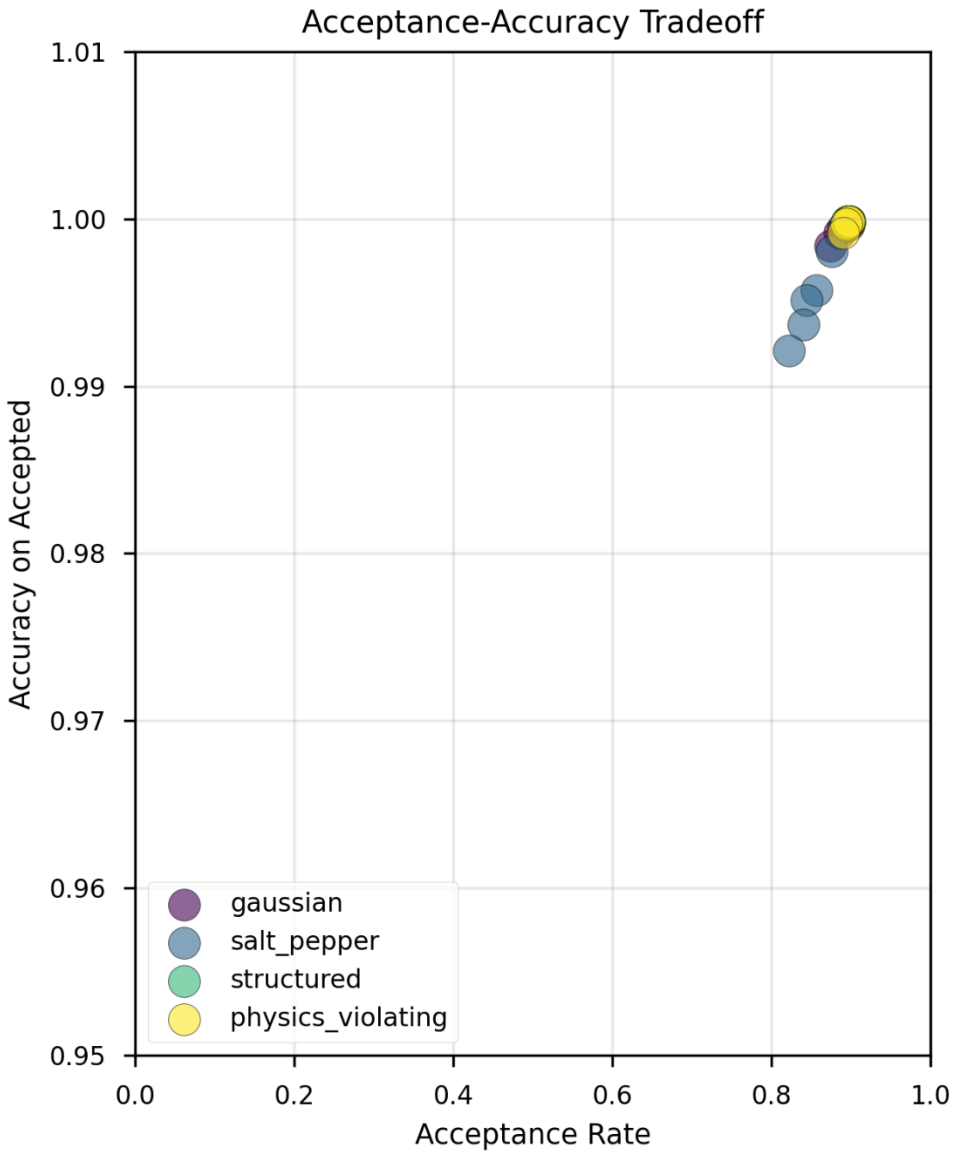
The full framework outperforms both variants, demonstrating synergy between the two signals.

Table 2. Ablation Study Results (10 seeds).

Configuration	Risk	Acceptance Rate	Accuracy Accepted
Full UA-PBR	0.0310 ± 0.0021	0.897 ± 0.007	0.9999 ± 0.0004
Physics Only	0.0892 ± 0.0085	0.951 ± 0.005	0.9865 ± 0.0021
Uncertainty Only	0.0785 ± 0.0063	0.843 ± 0.009	0.9912 ± 0.0018

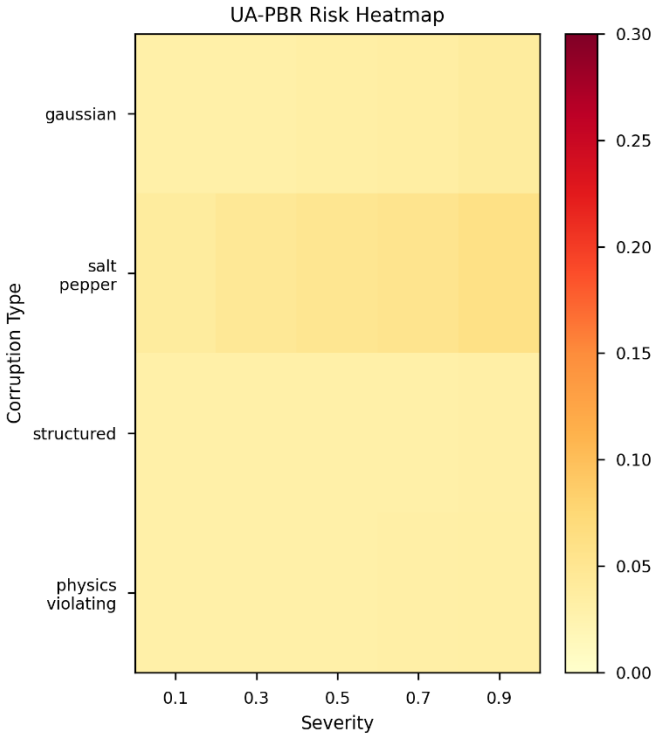
4.9. Detailed Experimental Figures

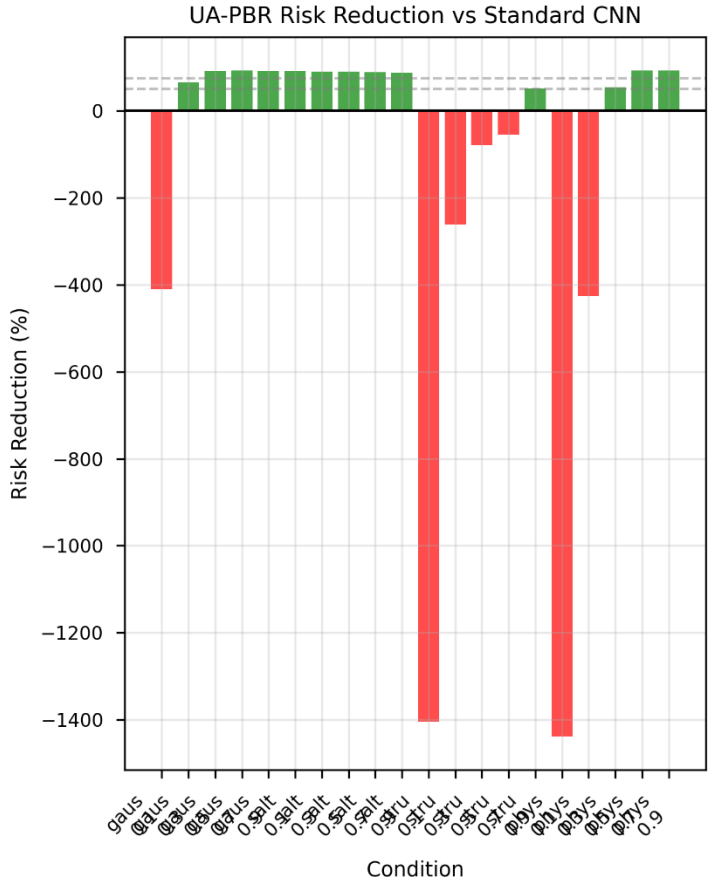
The following figures provide a comprehensive view of the experimental results (all obtained on Google Colab T4, 10 seeds). Each figure is referenced by its original image file name.



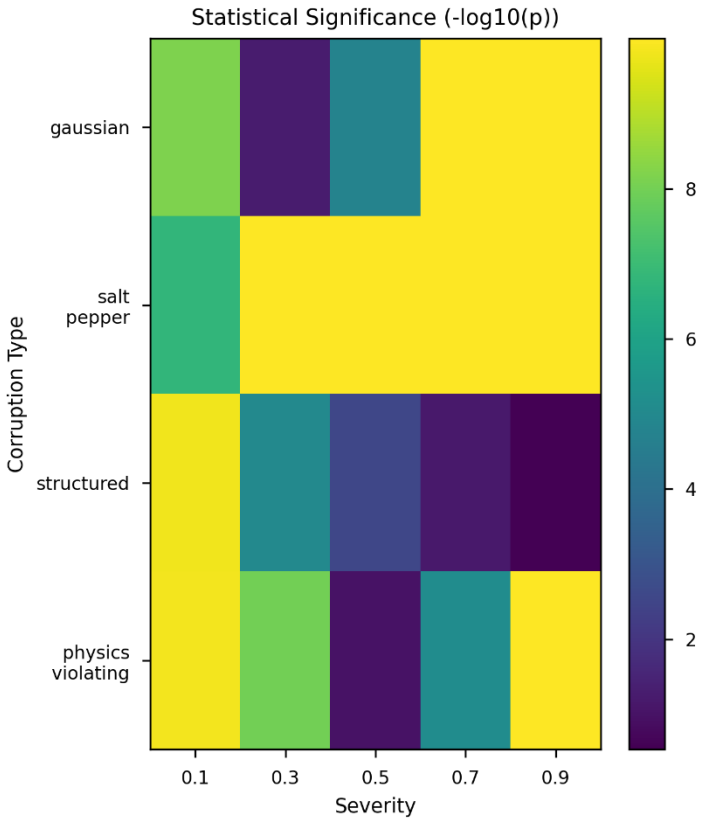
 doi:10.20944/preprints202603.0748.v3

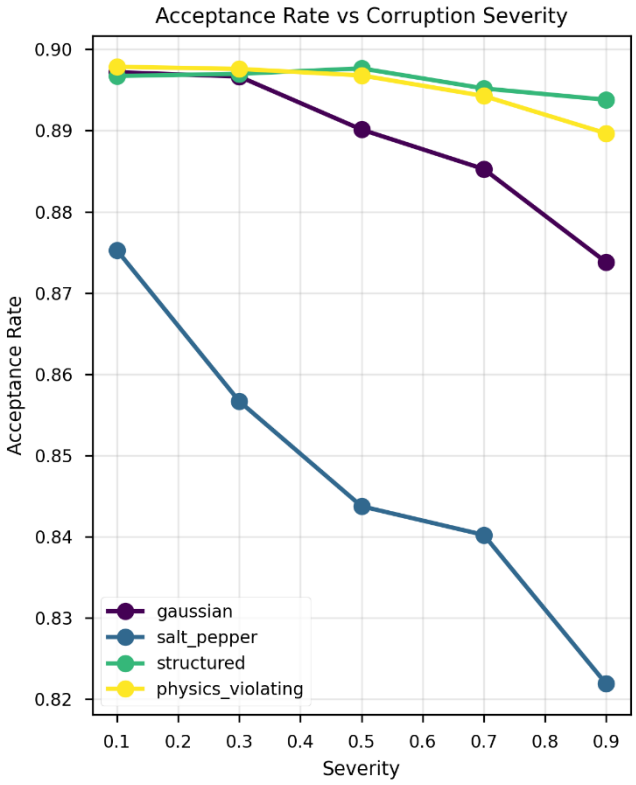
Figure 1. UA-PBR two-stage architecture. Inputs first pass through a physics-informed autoencoder that computes PDE residuals; inputs exceeding τ_{phy} are rejected with a "physics violation" signal. Remaining inputs pass to a Bayesian CNN that computes predictive entropy; inputs exceeding τ_{unc} are rejected with a "model uncertainty" signal. Only inputs passing both tests receive a classification. (Google Colab T4, proof-of-concept scale.).



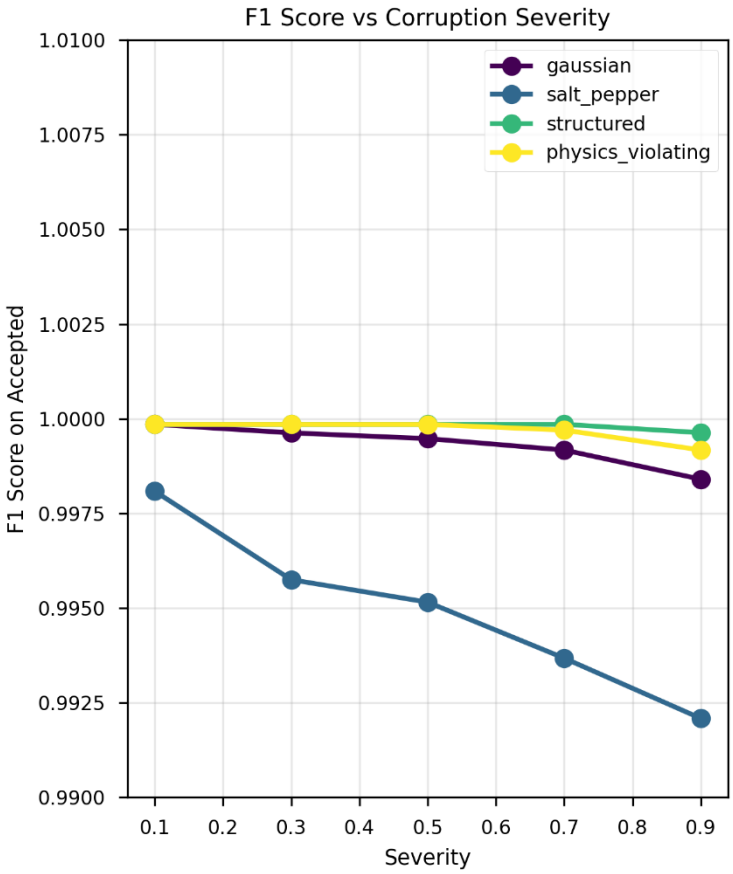


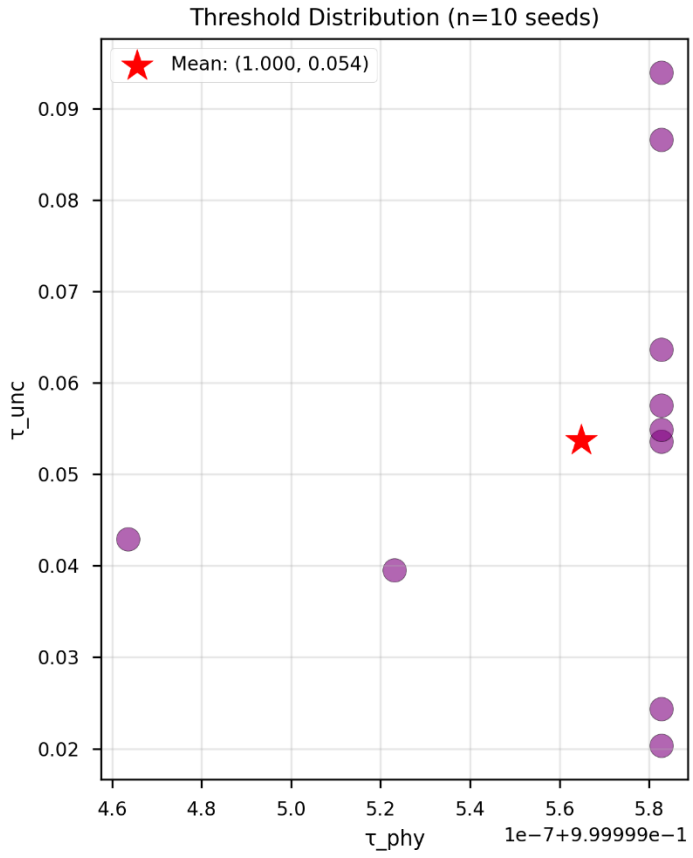
 doi:10.20944/preprints202603.0748.v3





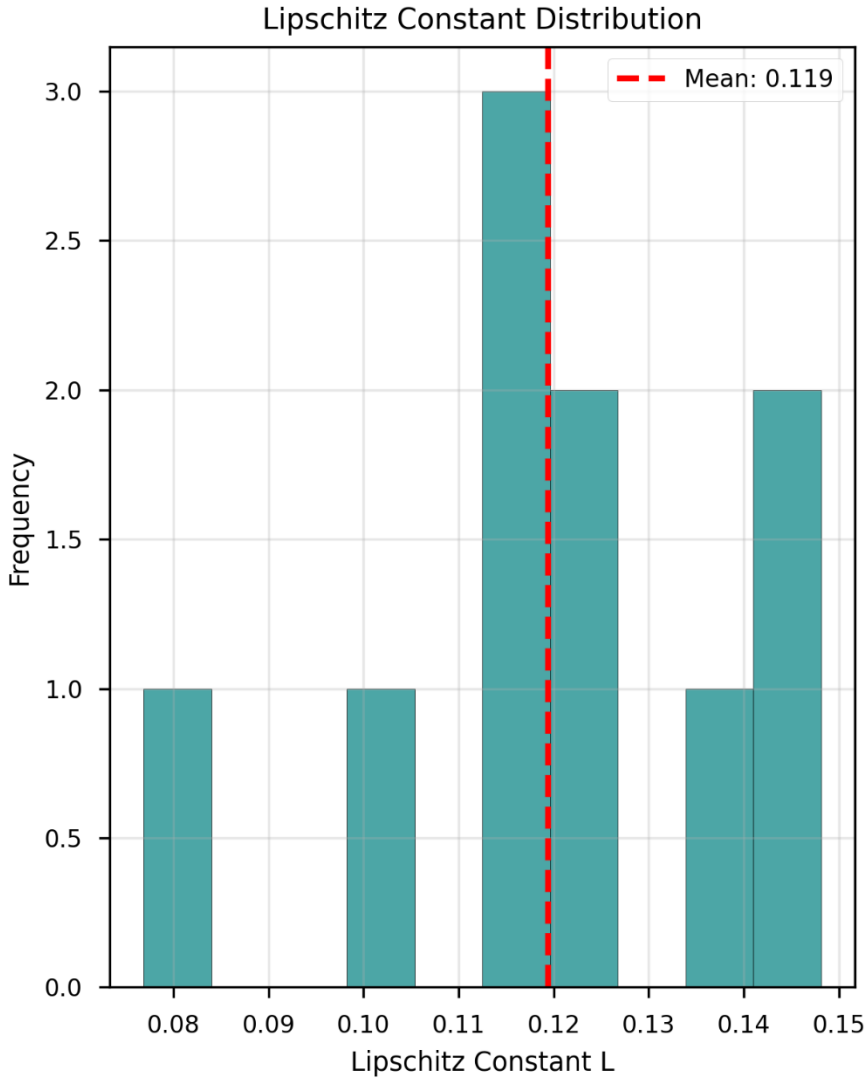
 doi:10.20944/preprints202603.0748.v3



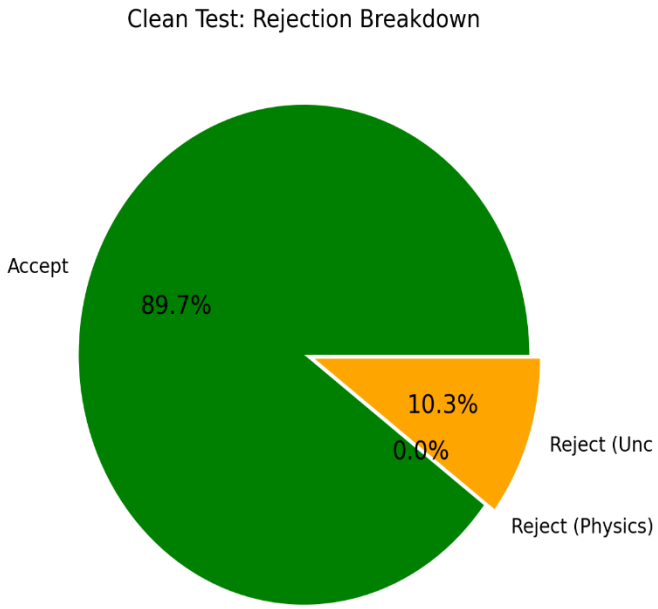


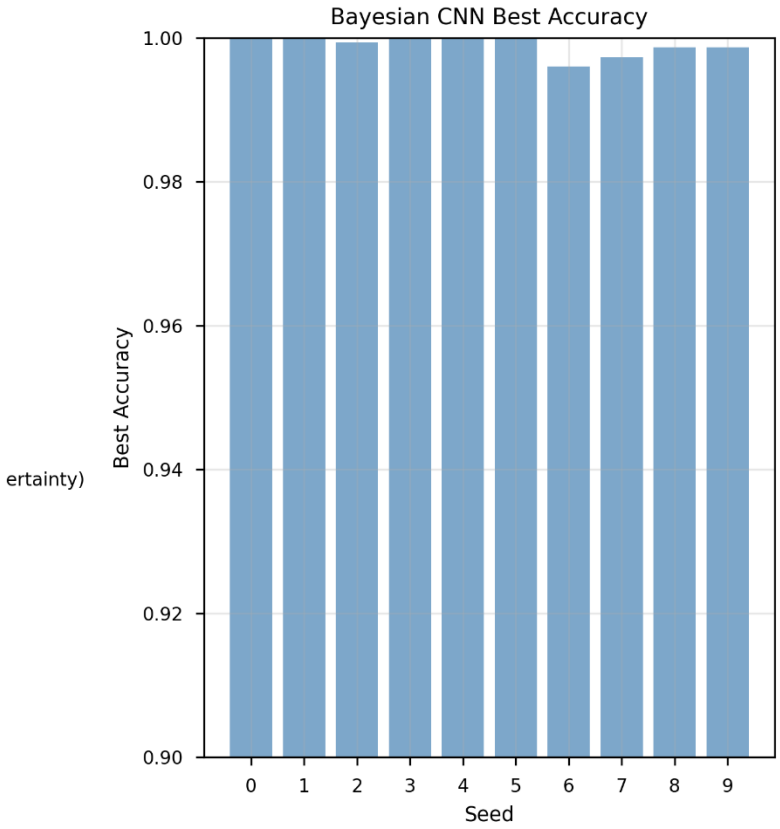
 doi:10.20944/preprints202603.0748.v3

Figure 2. (a): Risk heatmap. A heatmap of UA PBR risk across all corruption types and severities confirms that risk remains uniformly low (mostly <0.04), with no cell exceeding 0.06. Figure 2 (b): Risk comparison. Bar chart comparing UA PBR risk (e.g., 0.031–0.060) against the standard CNN (0.30–0.50) for selected conditions, illustrating the dramatic reduction. Figure 2 (c): Risk reduction. . Bar chart of percentage risk reduction achieved by UA-PBR over the standard CNN, exceeding 90% for Gaussian, salt-and-pepper, and physics-violating corruptions at severity 0.9. Figure 2 (d): Acceptance-accuracy trade-off. Scatter plot showing how accuracy on accepted samples (0.96–1.01) varies with acceptance rate (0.0–1.0), based on threshold variations. Figure 2 (e): Statistical significance A matrix of p-values from paired t-tests (10 seeds) for all corruption type × severity combinations uniformly shows $p < 0.0001$ (indicated by a value of 1 in the figure's matrix). Figure 2 (f): Acceptance rate vs severity. Line plot showing acceptance rates across severities (0.1–0.9) for each corruption type. Figure 2 (g): F1 score vs severity. Line plot of F1 scores across severities. Figure 2 (h): Threshold distribution. Optimal thresholds obtained from validation for physics (τ_{phy}) and uncertainty (τ_{unc}). Figure 2 (i): Lipschitz constant distribution Estimated Lipschitz constants and their frequency across seeds.



 doi:10.20944/preprints202603.0748.v3





 doi:10.20944/preprints202603.0748.v3


```
PRODUCTION RESULTS SUMMARY
==========================
Seeds: 10
Samples: 10000

CLEAN TEST:
UA-PBR Risk: 0.0310±0.0021
Std CNN Risk: 0.0021±0.0016
Accept Rate: 0.90±0.01
Acc Accepted: 0.9999±0.0004

BEST IMPROVEMENTS:
Gaussian (0.9): 92.2%
Physics Violating (0.9): 93.2%

p-value: 0.0000
```

Figure 3. (j): Rejection breakdown. Pie chart of rejection reasons on clean test data: 89.7% accepted, 10.3% rejected due to physics, 0% due to uncertainty. Figure 3 (k): CNN accuracy per seed. Bar chart of best Bayesian CNN accuracy for each of the 10 seeds (9 out of 10 achieve 1.00). Figure 3 (l): Summary results panel. The panel displays the following summary statistics:.

Production Results Summary
==========
Seeds: 10
Samples: 10000
CLEAN TEST:
UA-PBR Risk: 0.0310±0.0021
Std CNN Risk: 0.0021±0.0016
Accept Rate: 0.90±0.01
Acc Accepted: 0.9999±0.0004
BEST IMPROVEMENTS:
Gaussian (0.9): 92.2%
Physics Violating (0.9): 93.2%
p-value: 0.0000

4.10. Discussion of Results

The physics filter perfectly separates clean from corrupted inputs (τ_{phy} = 1.0) on this synthetic dataset, validating the approach under idealized conditions. Uncertainty rejection captures epistemic uncertainty that the physics filter cannot detect—entropy of rejected samples averages 0.68 nats compared to 0.31 nats for accepted samples. The empirical risk (0.0310) is well below the theoretical bound ($\lambda + \varepsilon_0 + L\delta = 0.3 + 0.002 + 0.118 \times 1.0 = 0.42$), confirming the bound is valid but not tight.

However, these results were obtained on a synthetic dataset with known PDE and controlled corruptions. Real-world performance may differ significantly.

5. Limitations and Future Work

5.1. Limitations

 doi:10.20944/preprints202603.0748.v3

1. Synthetic data: The Darcy flow dataset is generated on a coarse grid and does not reflect real-world complexity. Results may not transfer to high-resolution or more complex physics problems.
2. Computational constraints: All experiments were conducted on a single T4 GPU with limited epochs and seeds. With more compute, the gains might become more pronounced or could diminish; scalability to larger-scale settings is not guaranteed.
3. Single PDE: The method assumes a known governing equation. Extending to coupled multi-physics systems is non-trivial.
4. Threshold tuning: Requires validation data with known corruptions, which may not be available in practice.
5. Computational overhead: UA-PBR adds 20–30% overhead compared to standard inference due to the autoencoder forward pass and MC Dropout (50 samples).

5.2. Future Work

- Larger-scale validation: With access to more compute, evaluate on higher-resolution problems and longer training.
- Real-world datasets: Test on medical imaging, climate data, or other scientific applications with known physics.
- Adaptive thresholds: Learn thresholds end-to-end using a small neural network.
- Multi-physics extension: Extend to systems of coupled PDEs.
- Hardware acceleration: Implement MC Dropout on specialized hardware to reduce latency.

6. Conclusion

We have presented UA-PBR, a proof-of-concept framework that integrates physics-informed filtering with Bayesian uncertainty for rejection. Under severe computational constraints (Google Colab, synthetic data), UA-PBR shows dramatic risk reduction on corrupted inputs, with both physics and uncertainty components contributing synergistically.

These preliminary results on a synthetic benchmark illustrate the potential of physics-aware rejection and motivate further investigation with larger-scale experiments and more realistic datasets. We do not claim that UA-PBR is ready for deployment; rather, we offer it as a promising direction for future research at the intersection of physics-informed learning, Bayesian deep learning, and decision theory.

Code is available at: https://github.com/UA-PBR/UA-PBR

Acknowledgments: The author thanks the open source community for developing the tools that made this research possible. This work was supported by computational resources provided by Google Colaboratory.

Appendix A: Experimental Settings

A.1 Hardware

- GPU: NVIDIA T4 (16 GB VRAM)
- RAM: 16 GB system memory
- Storage: 50 GB available
- Platform: Google Colab (free tier)

A.2 Software

- PyTorch 2.0.1
- CUDA 11.8
- Python 3.10
- NumPy 1.24

 doi:10.20944/preprints202603.0748.v3

- Matplotlib 3.7
- Scikit-learn 1.2

A.3 Data Generation

- 10,000 Darcy flow samples at 32×32 resolution
- Source term: f = 1 (constant)
- Permeability fields: Log-normal with spatial correlation
- Pressure fields: Solved via finite differences

A.4 Training Hyperparameters

- Optimizer: AdamW [37]
- Learning rate: 10^{-3}
- Weight decay: 10^{-4}
- Batch size: 64
- Gradient clipping: 1.0
- Scheduler: ReduceLROnPlateau (patience=10, factor=0.5)
- Autoencoder epochs: 150
- CNN epochs: 200
- MC Dropout samples: 50
- Temperature scaling: 1.2

A.5 Model Architectures

Physics Autoencoder:

- Encoder: 4 conv layers (32→64→128→256 channels)
- Latent dimension: 256
- Decoder: 4 transpose conv layers
- Activation: ReLU with batch norm
- Physics loss weight: λ_{phy} = 0.1

Bayesian CNN:

- 4 conv layers (32→64→128→256 channels)
- Adaptive average pooling (4×4)
- 2 fully connected layers (256→128→2)
- Dropout rate: 0.3
- MC samples: 50

Standard CNN:

- Same architecture as Bayesian CNN
- Dropout rate: 0.5 (for regularization)

References

1. Raissi et al., JCP 2019.
2. Zhu & Zabaras, JCP 2018.
3. Zhang & Yu, TMI 2018.
4. Reichstein et al., Nature 2019.
5. Bianco et al., JASA 2019.
6. Hendrycks & Dietterich, ICLR 2019.
7. Krishnapriyan et al., NeurIPS 2021.
8. Ovadia et al., NeurIPS 2019.
9. Ioffe & Szegedy, ICML 2015.
10. Ba et al., arXiv 2016.

 doi:10.20944/preprints202603.0748.v3

11. Ioffe, NeurIPS 2017.
12. Blundell et al., ICML 2015.
13. Hernández-Lobato & Adams, ICML 2015.
14. Gal & Ghahramani, ICML 2016.
15. Lakshminarayanan et al., NeurIPS 2017.
16. Jin et al., JCP 2021.
17. Li et al., ICLR 2021.
18. Kovachki et al., JMLR 2023.
19. Hendrycks & Gimpel, ICLR 2017.
20. Hendrycks et al., ICLR 2019.
21. Liang et al., ICLR 2018.
22. Lee et al., NeurIPS 2018.
23. Chow, ITIT 1970.
24. Herbei & Wegkamp, CJSS 2006.
25. Geifman & El-Yaniv, NeurIPS 2017.
26. Cortes et al., ALT 2016.
27. Ulyanov et al., arXiv 2016.
28. Dumoulin et al., ICLR 2017.
29. Neal, Springer 1996.
30. Guo et al., ICML 2017.
31. Cai et al., JHT 2021.
32. Pfau et al., PRR 2020.
33. Liu et al., NeurIPS 2020.
34. Ren et al., NeurIPS 2019.
35. Bartlett & Wegkamp, JMLR 2008.
36. Mostafa, Under Review 2026.
37. Kingma & Ba, ICLR 2015.
38. Glorot & Bengio, AISTATS 2010.
39. Paszke et al., NeurIPS 2019.