AI Engineering
https://www.sciltp.com/journals/aieng

Perspective

A Comprehensive AI Security Pipeline: Drift Detection, Adversarial Robustness and Automated ML Testing

Siddharth Kumar [1,*] and Siddhanth Harish Bist [2,*]

[1] Data Science Program, Department of Mathematical Sciences, Stevens Institute of Technology, Hoboken, NJ 07030, USA
[2] Applied Artificial Intelligence, Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA
* Correspondence: skumar40@stevens.edu (S.K.); sbist@stevens.edu (S.H.B.)

How To Cite: Kumar, S.; Bist, S.H. A Comprehensive AI Security Pipeline: Drift Detection, Adversarial Robustness, and Automated ML Testing. *AI Engineering* **2026**, *2*(2), 13. https://doi.org/10.53941/aieng.2026.100013

Received: 16 April 2026
Revised: 2 June 2026
Accepted: 16 June 2026
Published: 12 August 2026

Abstract: The wide-scale uptake of machine learning applications in safety-sensitive applications renders modern AI deployments prone to adversarial attacks, statistical distribution changes, and various governance reliability issues. Current AI security methodologies generally handle the discussed issues separately, making the current security approaches ineffective in real-world conditions. In this work, an integrated AI security pipeline combining the functionalities of statistical drift detection, adversarial robustness evaluation, governance auditing, and experiment tracking is suggested. The presented system uses Kolmogorov—Smirnov tests, Population Stability Index analysis, and drift detection in data streams via ADWIN in addition to adversarial robustness evaluation through FGSM, PGD, and DeepFool attacks. The Giskard library was used to audit AI models' performance from the governance perspective. Evaluation of our approach on image and tabular datasets showed the capability of detecting statistically significant drift and significant CNN robustness degradation under increasingly complex adversarial attacks. It turned out that iterative and geometry-aware attack schemes perform significantly better than one-shot perturbations. Drift detection proved effective at spotting statistically significant distribution changes before actual deployment failures.

Keywords: AI security; adversarial machine learning; drift detection; Giskard; explainable AI; certified robustness; MLOps; deep learning

1. Introduction

The use of machine learning systems for health care, financial, cybersecurity, self-driving vehicles, and other critical operations has increased the attack surface of modern AI infrastructures considerably. Deep learning has provided significant success regarding prediction accuracy for various applications; however, machine learning systems can still be manipulated by adversarial attacks, become sensitive due to distribution shift in their input data, behave unpredictably, or raise governance issues during deployment. Such issues can harm the operational reliability and safety of AI systems when deployed in real-world applications.

A considerable amount of work has been done independently on the topics of adversarial robustness [1–5], detecting distribution shifts [6,7], explainable AI systems [8,9], adversarial machine learning approaches [10], and certified robustness [11–13] for deep learning models. At the same time, recent years have seen an emergence of various auditing frameworks for trustworthy deployment of AI systems.

This dichotomy presents a major problem in terms of the implementation of AI security in practice. In reality, adversarial attacks, distributional shift, robustness problems, and violation of governance principles tend to appear concurrently and interact with each other in order to produce greater harm. This means that the application of either of these two methods alone can offer only limited insight regarding the security issues of ML systems.

In order to tackle this problem, this study provides a comprehensive solution based on the integration of statistical drift monitoring, adversarial robustness evaluation, governance awareness, and experiment tracking into one unified system. As such, the system uses the Kolmogorov—Smirnov test, Population Stability Index calculations, adaptive windowing with ADWIN [6], generation of adversarial attacks with FGSM, PGD, and DeepFool techniques [1–3], and governance auditing by way of the Giskard framework [14,15].

In contrast to conventional approaches towards robustness benchmarking, which are concerned only with the accuracy reduction caused by attacks, the introduced methodology stresses a comprehensive approach towards ensuring the robustness of AI systems, incorporating all four steps into one pipeline: monitoring, auditing, vulnerability analysis, and reproducibility.

The key contributions of the work presented above can be outlined as follows:

- Development of a unified AI security pipeline integrating drift detection, adversarial robustness analysis, governance auditing, and experiment tracking.
- Comparative evaluation of FGSM, PGD, and DeepFool attacks to analyze structural weaknesses in CNN decision boundaries.
- Integration of statistical monitoring techniques including KS-Test, PSI, and ADWIN for proactive drift detection.
- Governance-oriented auditing using Giskard to identify perturbation-sensitive behaviors and operational reliability concerns.
- A modular and extensible architecture suitable for future MLOps-oriented AI assurance systems.

2. Related Work

2.1. Adversarial Robustness

FGSM [1], PGD [2], and DeepFool [3] continue to form the basis of adversarial attacks to test neural networks' robustness. Adversarial benchmarking tools like Foolbox [4] and CleverHans [5] were also developed for this purpose.

Despite the wide range of attacks available from robustness assessment libraries like ART and Foolbox, these solutions concentrate on testing the adversarial robustness of neural networks exclusively. The proposed solution combines drift detection, adversarial robustness evaluation, governance auditing, and experiment tracking within a unified AI security pipeline.

2.2. Certified Robustness

Robustness certification approaches, such as randomized smoothing [11] and convex adversarial polytopes [12], can offer mathematical assurances of robustness to bounded perturbation attacks. Recently developed adaptive randomized smoothing [13] improves upon robustness certification against iterative attacks.

In contrast with robustness certification approaches, which aim at offering theoretical assurances, the proposed approach emphasizes practical aspects of deployment-based monitoring and evaluation of security performance. Moreover, recent studies have revealed that despite the number of machine learning testing approaches proposed, their practical implementation is quite low due to the complexity of integration and workflow [16].

2.3. Drift Detection

Drift detection techniques like ADWIN [6] help in identifying drifts in evolving datasets. Representation-based drift detection systems [7] represent new approaches to drift detection in terms of latent features.

Current approaches to drift detection rely heavily on the notion of statistical drift detection separately. Our contributions extend this paradigm by studying how distributional stability affects robustness and governance.

2.4. Explainable AI and Governance

Explainability algorithms like SHAP [8] and LIME [9] enhance interpretability of machine learning models. Auditing mechanisms focused on governance like Giskard [14] deliver reproducible vulnerability assessment and slicing capability.

Furthermore, alignment drift has been recently recognized as a security concern in regulated AI systems [17]. Our proposed framework builds upon that view by integrating security monitoring functionality.

3. System Architecture

This was done with the purpose of enabling scalability, extensibility, and reproducibility. Each of these components provides well-defined interfaces for configuration, logging, reporting, and orchestration.

The architecture, as seen in Figure 1:

- Drift Detection Module
- CNN Training Module
- Adversarial Evaluation Module
- Governance Auditing Module
- Structured Reporting Layer

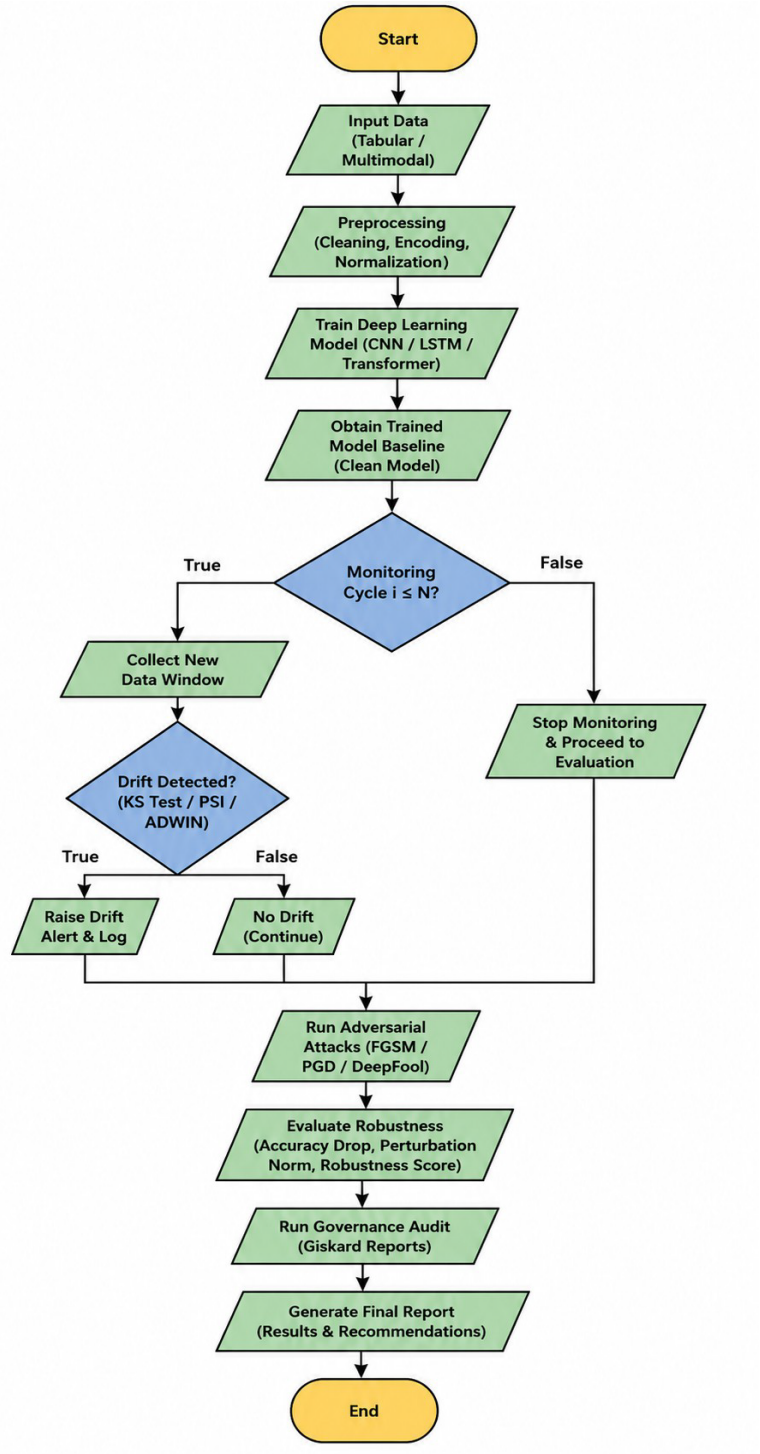

Figure 1. Unified AI security pipeline architecture.

4. Methodology

MNIST was used to evaluate the adversarial attacks' robustness, and the Iris dataset [18] was utilized for the controlled drift experiments.

4.1. Drift Injection

In order to simulate the drifting process in the latter half of the stream, the sepal length feature value was increased by an additional offset of 0.8.

For stream-oriented drift monitoring, the ADWIN [6] algorithm was used as the primary adaptive windowing mechanism. Concepts from streaming learning literature inspired the overall monitoring design philosophy.

4.2. Adversarial Attacks

FGSM attack was considered with $\epsilon = 0.2$.

For the PGD attack, the following settings were chosen:

- $\epsilon = 0.3$
- step size $\alpha = 0.01$
- number of iterations = 20

DeepFool parameters were selected as follows:

- max iterations = 15
- overshoot = 0.02

The assessments for PGD and DeepFool attacks were conducted using random samples of 1000 MNIST test images.

5. Results

5.1. Drift Detection Results

The distributional drift detection algorithm successfully detected statistically significant changes in the distribution of the Iris dataset following an injected drift into the dataset. A constant drift value of +0.8 was introduced into the sepal length attribute for about 50% of the input samples. The specific drift value was deliberately chosen to mimic a realistic scenario where feature distribution deviates because of sensor drift, changes in demographics, seasonality, and other factors. As observed in the results, the KS test returned a very low *p*-value of 0.000331, clearly less than the conventional threshold of 0.05, which implies that the altered distribution is statistically different from the original distribution. On the other hand, the Population Stability Index (PSI), a commonly used metric for monitoring distribution changes [6], was well above the standard threshold of 0.25. Regular monitoring of already-deployed ML models is an essential element of ML testing as the behavior of the models and the distribution of data could change after deployment, requiring regular verification [19].

Figure 2 provides visual proof of the gap created due to drift in terms of the original and drifted distribution of the features. It is clearly seen that the drifted distribution is noticeably moved towards greater feature values, thereby proving that the synthetic drift creation was effective.

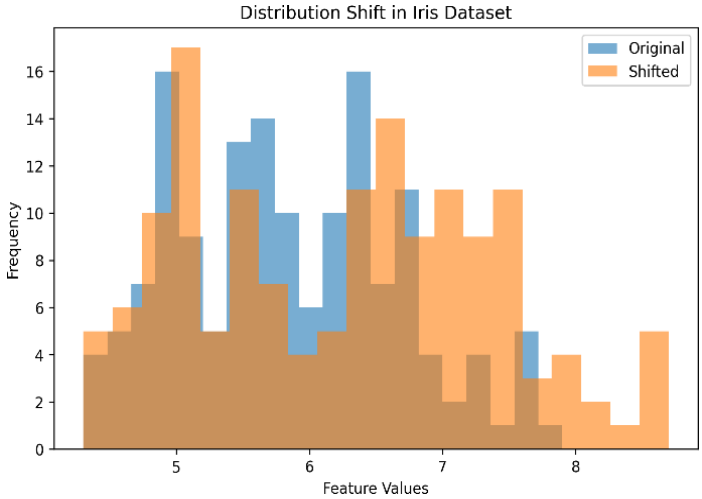

Figure 2. Controlled drift injection in the iris dataset.

Practically speaking, such a drift will cause a number of operations to be triggered at the next stages of the pipeline. Firstly, the drift detection module will report a statistical anomaly. Secondly, the auditing module will record the observation for logging purposes. Lastly, depending on policies, automatic re-training or rollbacks may be initiated.

Thus, this experiment not only proved that drift was successfully detected but also showed how important it is from a governance point of view.

5.2. CNN Training Performance

The CNN training process demonstrated stable convergence behavior across all epochs. Figure 3 shows the training loss progression during optimization.

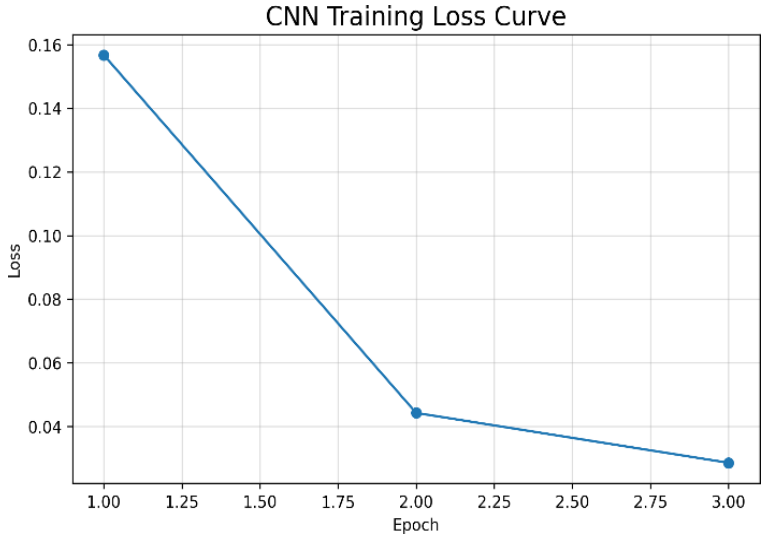

Figure 3. CNN training loss curve.

The drop remained constant from around 0.157 in the first epoch to below 0.03 in the last epoch. Such a trend clearly shows that the network was able to learn significant visual features without exhibiting any issues with unstable training.

Moreover, such a trend clearly suggests that the chosen configuration of the optimizer and the learning rate were appropriate for the classification problem at hand. The neural network showed high performance on the clean dataset before evaluating its robustness against attacks.

5.3. Adversarial Robustness Results

Adversarial Robustness tests demonstrated high susceptibility of the CNN model to single-step and iterative attacks. The classification accuracy on attacks is illustrated in Figure 4 for three types of attack—FGSM, PGD, and DeepFool.

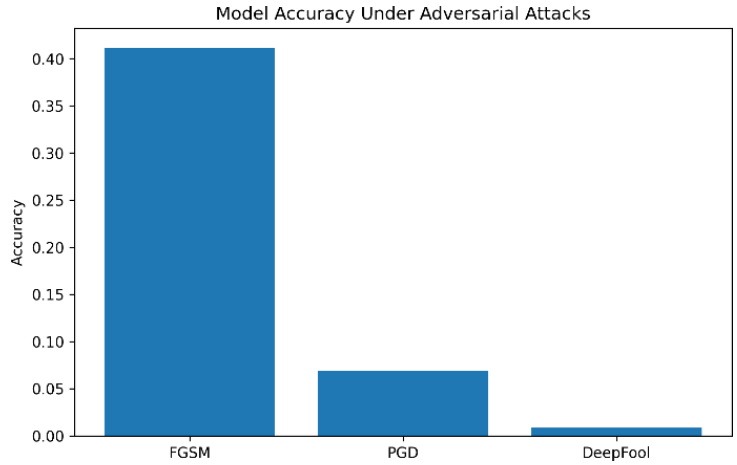

Figure 4. CNN accuracy under FGSM, PGD, and DeepFool attacks.

Model accuracy was decreased by FGSM to about 41.2% and by PGD to about 6.9%, whereas DeepFool achieved the largest accuracy drop, decreasing it to about 0.9%.

The large discrepancy between FGSM and PGD emphasizes the importance of iterative optimization in an adversarial scenario. Whereas FGSM is limited to a single gradient step in its search for perturbation directions, PGD iterates over the perturbation process, allowing it to navigate through the non-linear space near the decision boundary more effectively [1,2].

This finding implies that the decision boundaries of the CNN models are highly unstable under iterative adversarial perturbation.

In addition, DeepFool exposes structural vulnerabilities in the feature space of the classifier. While FGSM and PGD directly maximize classification loss, DeepFool calculates the minimum perturbation necessary to overcome the closest decision boundary. This leads to near-zero accuracy, implying that the decision margins for classifiers are very thin [3].

This result is significant, since thin decision margins suggest that not only is the model vulnerable to adversarial perturbations, but that it will be affected by any natural interference, such as noise, compression, or lighting effects [3].

5.4. Visual Analysis of Adversarial Perturbations

To further interpret attack behavior, visual comparisons between clean and adversarial samples were generated for FGSM, PGD, and DeepFool attacks.

FGSM perturbations appear visually noisy and relatively coarse because the attack applies a single-step perturbation directly along the sign of the input gradient. Despite its simplicity, the attack is still capable of misleading the classifier on multiple samples. The impact of single-step adversarial perturbations can be further demonstrated by Figure 5, where clean MNIST images are compared with their FGSM adversarial versions.

PGD adversarial samples exhibit stronger perturbation structures due to iterative optimization. The repeated gradient updates allow the attack to progressively move samples toward highly vulnerable regions of the decision space.

Adversarial samples created using PGD are shown in Figure 6. In comparison to FGSM, a more potent perturbation is achieved through iterative optimization.

DeepFool perturbations are visually subtle yet highly effective. The attack modifies only the minimum amount necessary to cross decision boundaries. This demonstrates that the classifier relies on highly fragile discriminative features and possesses limited robustness margins. Figure 7 shows examples of adversarial attacks that have been created using the DeepFool technique.

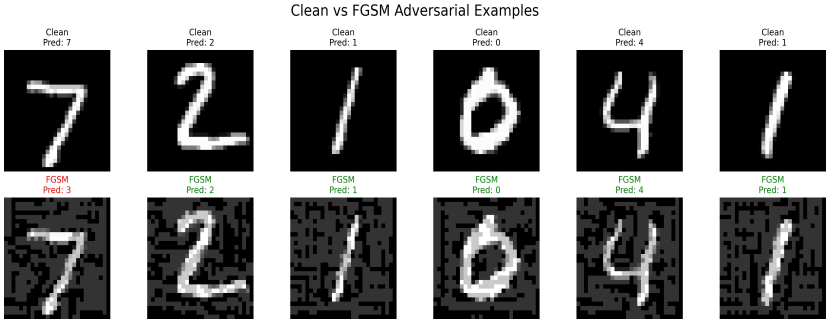

Figure 5. Clean vs FGSM adversarial examples. FGSM introduces single-step gradient perturbations that cause partial misclassification while preserving recognizable digit structure.

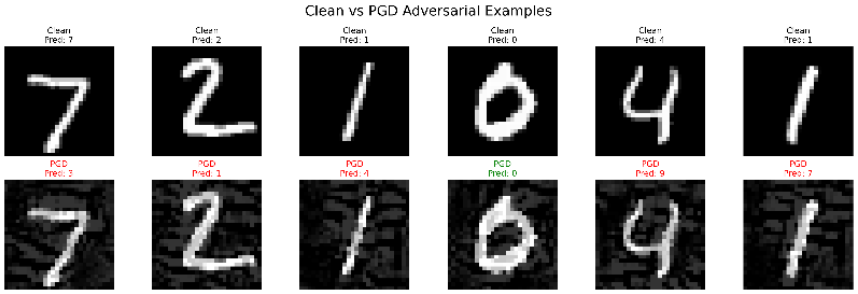

Figure 6. Clean vs PGD adversarial examples. PGD performs iterative perturbation optimization, generating stronger adversarial distortions and significantly degrading model robustness.

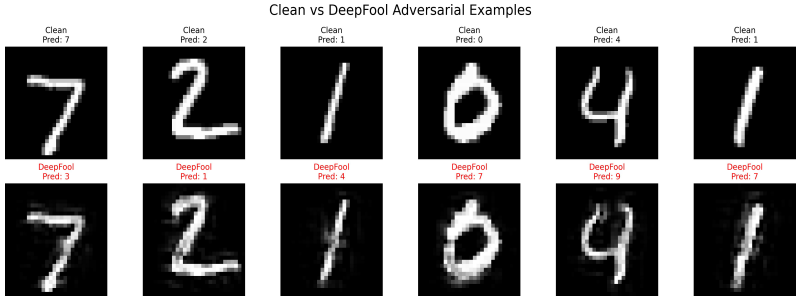

Figure 7. Clean vs DeepFool adversarial examples. DeepFool computes minimal boundary-crossing perturbations, demonstrating the narrow decision margins learned by the CNN model.

Collectively, these visual analyses confirm that stronger iterative attacks not only reduce numerical accuracy but also reveal deeper structural instability in the learned feature representations.

5.5. Integrated Security Implications

The experimental results collectively demonstrate that AI security vulnerabilities rarely occur in isolation. Distribution drift, adversarial instability, and governance concerns interact in ways that amplify operational risk.

As a consequence, not only can the utility of each module be recognized, but the entire pipeline is valuable through its integration of monitoring, robustness testing, governance auditing, and experimentation [14,15].

This integrated perspective enables earlier detection of concurrent threats and improves operational visibility for production AI systems.

6. Discussion

The results show that reliable assessment of AI security cannot rely on individual robustness tests alone but should incorporate several methods of analysis together. In real-life situations, issues rarely occur due to a single point of failure; rather, there is an interplay of adversarial attacks, statistical shifts, unstable models, and various risk factors associated with deployment. The suggested framework was designed with a view to conducting multiple types of tests as a single process within a pipeline.

The statistical drift testing showed that the proposed 0.8 shift for the Iris dataset caused statistically significant drift, which is demonstrated by the elevated PSI scores and the low KS test p-value of 0.000331. Such a shift would indicate meaningful changes in incoming data distributions, sensor behavior, or user interaction patterns in real-world deployment scenarios. This demonstrates the practical importance of continuously monitoring the statistical stability of deployed AI systems [6,7,20].

The results obtained in the adversarial robustness study uncovered inherent structural vulnerabilities in the baseline CNN architecture. Specifically, FGSM demonstrated that the classifier is highly sensitive to gradient-based attacks because even a single-step perturbation was sufficient to substantially reduce classification accuracy. In contrast, PGD achieved significantly stronger degradation due to its iterative optimization strategy, where each iteration progressively refines perturbations to maximize classification error while remaining within constrained perturbation bounds [1,2].

DeepFool provided additional insight into the geometric structure of the learned feature space. Unlike FGSM and PGD, DeepFool computes the minimum perturbation required to cross a decision boundary. The severe degradation observed under DeepFool attacks indicates that the model learned extremely narrow robustness margins, meaning that even very small perturbations can alter predictions. This finding suggests that the model may remain vulnerable not only to adversarial attacks but also to practical deployment issues such as noisy sensor inputs, compression artifacts, or changing operating conditions [3].

These observations were further confirmed through the generated adversarial visualizations. FGSM perturbations appeared relatively coarse and high-frequency, while PGD generated progressively optimized perturbations that caused substantial instability in predictions. DeepFool perturbations remained visually subtle despite producing severe misclassification, emphasizing the practical danger of optimized low-magnitude attacks in deployed AI systems.

Unlike traditional robustness benchmarking frameworks that evaluate attacks independently, the proposed pipeline integrates:

- statistical drift monitoring,
- adversarial robustness evaluation,

- governance-oriented auditing, and
- reproducible experiment tracking.

This integration provides a significantly more realistic operational security assessment because modern AI systems often experience concurrent threats rather than isolated failures. Furthermore, the incorporation of governance auditing through the Giskard framework extends the system beyond conventional robustness testing toward production-oriented AI assurance by identifying instability patterns, perturbation-sensitive behaviors, and operational monitoring vulnerabilities [14,15].

Overall, the findings demonstrate that integrated AI security monitoring provides substantially greater operational visibility than standalone robustness or drift detection systems. The proposed framework therefore establishes a scalable foundation for future real-time AI assurance, governance-aware deployment monitoring, and continuous MLOps-oriented security evaluation pipelines [21].

7. Limitations

Despite providing effective and solid proof-of-concept examples, the Iris and MNIST datasets did not capture the reality of large-scale and complex real-life AI deployment systems [18,22]. Lacking GPU-acceleration capabilities, the experiments were also limited to lightweight CNNs, unable to test heavy architectures like ResNets and transformers.

Another limitation lies in testing for statistical drift and adversarial robustness independently, using different datasets and setups. In reality, an AI deployment environment would face challenges with both types of threats at once. Future work should aim at analyzing unified scenarios that incorporate both types of risks.

The generated drift was synthetic and mostly univariate in its nature. The drift in reality is likely to be slower, noisy, and multivariate, increasing the difficulty in detecting such changes [6,7]. Additionally, the existing solution did not fully test how feasible and cost-effective drift detection and mitigation is from deployment perspectives [10].

8. Future Work

Potential expansions of the suggested framework will include scalability, robustness, explainability, and realistic deployments. An interesting avenue for further research will be multivariate or representation-level detection techniques that will allow modeling of complicated feature dependencies and their changes [7].

Other areas of interest are methods of certified robustness, such as randomized smoothing or more formal approaches to measuring robustness in addition to the heuristic methods applied here [11–13]. The robustness evaluation mechanisms will also help understand whether adversarial training can improve the detection capability of the proposed pipeline.

Techniques in the area of Explainable AI can be incorporated into the pipeline to help pinpoint which features play an especially significant role in the detected drifts or vulnerability to adversarial examples [8–10]. Finally, future development of the framework will be centered around scalability and inclusion of MLOps, with such features as GPU-based monitoring, CI/CD deployment, and real-time governance of AI assurance pipelines [7,10].

9. Conclusions

In this study, we have explored an integrated approach to AI security through a pipeline consisting of drift detection, adversarial robustness analysis, governance auditing, and experiment tracking. Our approach stands out from other isolated security systems in that it uses a combination of several AI assurance techniques in the examination of simultaneous threats in MLOps setups [23].

The experimental findings showed that the drift detection tool detected distributional drifts in the Iris dataset using the KS-Test, PSI calculation, and ADWIN [6,24]. In addition, the adversarial robustness studies revealed high vulnerability of CNN models to FGSM, PGD, and DeepFool attacks [1–3]. Although the FGSM attack illustrated a problem with susceptibility to gradient-based perturbation attacks, iterative attacks and geometry-aware attacks, such as PGD and DeepFool, respectively, caused considerable damage to classification performance [1–3].

The integrated Giskard auditing module also helped to achieve governance-driven assessment by recognizing perturbation-sensitive behavior patterns and operational reliability issues [14]. Taken together, these observations show that the combination of drift detection, adversarial testing, and auditing for governance purposes provides far greater visibility than robustness testing alone.

While the proposed methodology was tested on proof-of-concept datasets like Iris and MNIST [18,22], the results lay the groundwork for scalable AI assurance processes in production environments. Next steps will

include applying the framework to bigger data sets, conducting adversarial training, performing multivariate drift detection, incorporating explainability of AI using SHAP and LIME [8,9], and certifying robustness through randomized smoothing [11,13].

Author Contributions

S.K. made contributions to the conceptualization, methodology design, construction of the integrated AI security pipeline, software engineering, preprocessing of data, statistical drift detection, adversarial attacks, data management, visualizations of experimental results, formal analysis, and drafting of the manuscript. S.K. played a dominant role in integrating the individual components of the pipeline, experimental validation on the Iris and MNIST datasets, generation of performance metrics, figures, and experimental results, and reporting the implementation details and research findings. S.H.B. made contributions to the conceptualization and refinement of the research framework, methodology design, validation of the experimental procedure, investigation of adversarial machine learning methods, governance auditing with Giskard, analysis of experimental results, critical interpretation of AI security issues, and review and editing of the manuscript. S.H.B. made contributions to the enhancement of the robustness assessment framework, validation of the entire pipeline architecture, critical review of the technical methodology, and improvement of the discussion, conclusion, and future research directions. Both authors have jointly made contributions to the interpretation of results, discussion of the findings, revision of the manuscript, and approval of the final version for submission. Both authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data used in this paper are freely available. Iris Dataset can be obtained from the UCI Machine Learning Repository, while the MNIST handwritten digits dataset can be obtained from the PyTorch and TensorFlow datasets library. The source code, trained model, and experimental results obtained during this study have been developed by the authors and are available on request from the corresponding author. All data required to replicate and verify the results of this study can be obtained subject to reasonable requests.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

In the course of preparation of the present paper, the authors employed the publicly available versions of OpenAI ChatGPT (January–July 2026) and Anthropic Claude Sonnet 4 only on a limited scale to improve the grammar, language, wording, clarity, formatting, and other aspects of presentation of the manuscript. Moreover, these AI-assisted instruments were utilized for programming purposes in order to detect implementation problems, debug code, clarify programming concepts, and suggest approaches to solve software-related issues arising in the process of creating an integrated AI security pipeline. The mentioned AI-assisted instruments have not been used for formulation of the research hypothesis, methodology development, experiments, data analysis, interpretation of results, and conclusions of the scientific work described in this paper. All steps related to software implementation, model development, experiments, data preprocessing, adversarial robustness assessment, detection of drifts, governance auditing, validation of results, figure creation, and their interpretation were performed independently and verified by the authors. After employing these AI-assisted tools, the authors critically evaluated, revised, and validated all the recommendations made by these AI-assisted tools before incorporating

them into the manuscript and code implementation. The authors are responsible for the accuracy, originality, integrity, and scientific content of the paper.

References

1. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv* **2014**. arXiv:1412.6572.
2. Madry, A.; Makelov, A.; Schmidt, L.; et al. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Canada, 30 April–3 May 2018.
3. Moosavi-Dezfooli, S.; Fawzi, A.; Frossard, P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27 June–1 July 2016; pp. 2574–2582.
4. Rauber, J.; Brendel, W.; Bethge, M. Foolbox: A Python Toolbox to Benchmark the Robustness of Machine Learning Models. *arXiv* **2017**. arXiv:1707.04131.
5. Papernot, N.; Faghri, F.; Carlini, N.; et al. Technical Report on the CleverHans v2.1.0 Adversarial Examples Library. *arXiv* **2016**. arXiv:1610.00768.
6. Bifet, A.; Gavalda, R. Learning from Time-Changing Data with Adaptive Windowing. In Proceedings of the Seventh SIAM International Conference on Data Mining, Minneapolis, MN, USA, 26–28 April 2007; pp. 443–448.
7. Greco, S.; Vacchetti, B.; Apiletti, D.; et al. Unsupervised Concept Drift Detection from Deep Learning Representations in Real-Time. *IEEE Trans. Knowl. Data Eng.* **2025**, *37*, 6232–6245. https://doi.org/10.1109/TKDE.2025.3593123.
8. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
9. Ribeiro, M.; Singh, S.; Guestrin, C. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
10. Ijiga, O.M.; Idoko, I.P.; Ebiega, G.I.; et al. Harnessing Adversarial Machine Learning for Advanced Threat Detection: AI-Driven Strategies in Cybersecurity Risk Assessment and Fraud Prevention. *Open Access Res. J. Sci. Technol.* **2024**, *11*, 1–24.
11. Cohen, J.; Rosenfeld, E.; Kolter, Z. Certified Adversarial Robustness via Randomized Smoothing. In Proceedings of the International Conference on Machine Learning (ICML 2019), Long Beach, CA, USA, 9–15 June 2019; pp. 1310–1320.
12. Wong, E.; Kolter, J.Z. Provable Defenses Against Adversarial Examples via the Convex Outer Adversarial Polytope. In Proceedings of the International Conference on Machine Learning (ICML 2018), Stockholm, Sweden, 10–15 July 2018; pp. 5283–5292.
13. Lyu, S.; Shaikh, S.; Shpilevskiy, F.; et al. Adaptive Randomized Smoothing: Certified Adversarial Robustness for Multi-Step Defences. In Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, BC, Canada, 9–15 December 2024; pp. 134043–134074.
14. Giskard Team. *Open-Source ML Testing Framework*; Giskard: Paris, France, 2023.
15. Pepe, F.; Nardone, V.; Mastropaolo, A.; et al. Datasets, Bias, Licenses, and Terms of Use: A Large and Longitudinal Study on the Documentation of Hugging Face Machine Learning Models. *Empir. Softw. Eng.* **2026**, *31*, 95. https://doi.org/10.1007/s10664-026-10843-1.
16. Cannavale, A.; Pontillo, V.; De Lucia, A.; et al. Understanding Machine Learning Testing in Practice. *J. Syst. Softw.* **2026**, *240*, 112925. https://doi.org/10.1016/j.jss.2026.112925.
17. Pakina, A.K. Alignment Drift as a Security Threat: Detecting and Mitigating Misaligned AI Behavior in Regulated Systems. *Int. J. Innov. Sci. Res. Technol.* **2025**, *10*, 1856–1867. https://doi.org/10.38124/ijisrt/25dec1365.
18. Fisher, R.A. The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.* **1936**, *7*, 179–188.
19. Zhang, J.M.; Harman, M.; Ma, L.; et al. Machine Learning Testing: Survey, Landscapes and Horizons. *IEEE Trans. Softw. Eng.* **2020**, *48*, 1–36.
20. Patchipala, S.G. Tackling Data and Model Drift in AI: Strategies for Maintaining Accuracy during ML Model Inference. Int. *J. Sci. Res. Arch.* **2023**, *10*, 1198–1209. https://doi.org/10.30574/ijsra.2023.10.2.0855.
21. Agarwal, A.; Nene, M.J. Advancing Trustworthy AI: A Comparative Evaluation of AI Robustness Toolboxes. *SN Comput. Sci.* **2025**, *6*, 234.
22. LeCun, Y.; Bottou, L.; Bengio, Y.; et al. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324.
23. Papernot, N.; McDaniel, P.; Sinha, A.; et al. SoK: Security and Privacy in Machine Learning. In Proceedings of the 2018 IEEE European Symposium on Security and Privacy (EuroS&P), London, UK, 24–26 April 2018; pp. 399–414.
24. Hulten, G.; Spencer, L.; Domingos, P. Mining Time-Changing Data Streams. In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 26–29 August 2001.