IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 2, February 2024

DOI: 10.17148/IJIREEICE.2024.12219

BUILDING A COMPLETE AI/ML AUTOMATION PIPELINE ON-PREM USING GITOPS AND SELF-HOSTED GITHUB RUNNERS

Kiran Kumar Yakkali

Computer Science and Software Development

Abstract: With the mounting pressure on organizations to modernize their infrastructure, particularly in regulated and mission-critical settings, there exists a rising demand of on-premises AI/ML automation pipelines that can provide enterprise-level security, compliance and control. This paper presents an overall engineering solution to create a fully automated AI/ML pipeline on-prem and using GitOps principles and self-hosted runners of GitHub Actions. The suggested pipeline combines infrastructure-as-code, container orchestration, and continuous integration/continuous delivery (CI/CD) pipelines to allow fast and repeatable models and supporting system deployment, and auditing it. The architecture enables data sovereignty, low latency and regulatory constraints by using self-hosted runners that operate on private infrastructure, which is important to sectors in which the national interest in infrastructure modernization may exist. The article describes the choice of tools chains, security enhancement, deployment methods, monitoring and governance habits. Use-case scenarios reveal the ways in which the pipeline contributes to the controlled environments and show how my work as the automation lead accelerated the creation of the secure AI ecosystems. The results show that on-prem pipelines constructed using GitOps and self-hosted runners have the potential to satisfy enterprise demands of repeatability, traceability, and resilience and decrease reliance on public cloud services.

Keywords: AI/ML pipeline, on-premises infrastructure, GitOps, GitHub Actions, self-hosted runners, CI/CD automation, infrastructure modernization, regulated environment, data sovereignty, secure AI ecosystems.

I. INTRODUCTION

Due to the need to modernize the infrastructure, especially in highly regulated industries, the need to have efficient, secure and scalable AI/ML automation pipelines has risen. Conventional cloud-based pipelines have been used in numerous ways but there is a drastic move toward developing on-premises solutions which provide greater controllability to data sovereignty, security and compliance. One of the recent developments in this direction is the application of GitOps concepts and self-hosted GitHub runners to provide automated deployment and development of machine learning models in secure settings (Manolov, Gotseva, and Hinov, 2025).

GitOps is a collection of guidelines that use Git repositories to serve as the single source of truth when deploying infrastructure and applications to facilitate the full automation and auditability of continuous integration/continuous delivery (CI/CD) workflows (Saroar, 2024). Through the self-hosted GitHub runners, organizations can store the CI/CD activities in their own infrastructure, eliminating the chances of external exposure whilst ensuring the ability to scale its resources as required (Joshi, 2025). These solutions enable the teams to create, test, and deploy AI models more quickly, more automated and repeatable, and safely meet regulatory demands.

MLOps, which is the combination of DevOps and machine learning practices, is a critical practice in the AI/ML ecosystem that has set its role in managing the lifecycle of machine learning models deployed in the production settings (Berberi et al., 2025). GitOps + MLOps integration offers a single framework that inherits the advantages of both methods so that model deployments are not only automated but also safe, regulatory and can be traced readily. The significance of these practices in enhancing the process of deploying and monitoring models has been highlighted in recent research and found to be especially relevant to regulated settings where security and compliance are central concerns (Steidl, Felderer, and Ramler, 2023).

Although cloud-based solutions continue to gain popularity among most organizations, on-premises pipelines are becoming more popular due to the level of security and control of sensitive information. As an illustration, AI models in industries like healthcare, finance, and government usually involve very sensitive personal information that demands

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 2, February 2024

DOI: 10.17148/IJIREEICE.2024.12219

a high level of data privacy. On-prem solutions allow an organization to maintain complete control over their data whilst reducing the risks related to the storage and transmission of data in the cloud (Garg et al., 2022).

Besides, the necessity of automated compliance validation in AI/ML systems has turned into an acute issue, since these models should not break strict regulations and standards. Compliance checks that are embedded in the CI/CD pipeline, including the ones made possible by GitOps practices, can enact that models meet the stipulations of the industry at all stages of deployment (Ravva, 2025). The compliance feature is important in this automation and border control, immigration systems and biometric identity systems where the integrity and privacy of the data is paramount.

Finally, the system to develop and run machine learning models in regulated, on-prem environments is offered to the organizations through the creation of pipelines of AI/ML automation on the basis of GitOps and self-hosted GitHub runners. The current paper examines the design and deployment of a pipeline like this and its benefits in the aspects of automation and compliance, specifically how GitOps and self-hosted runners can be used to manage AI systems in the field of critical infrastructure.

II. LITERATURE REVIEW

This intersection of the development of AI/ML models and DevOps practices has led to a new field commonly known as Machine Learning Operations (MLOps), which uses the principles of DevOps, including continuous integration/continuous delivery (CI/CD), automation, version control, and infrastructure-as-code to the machine-learning model lifecycle (Steidl, Felderer and Ramler, 2023). Simultaneously, the advent of GitOps from applying Git as the single source of truth of both application and infrastructure settings have allowed more declarative and observable deployment practices to infrastructure and applications. MLOps and GitOps are considered one of the influential theoretical frameworks and models to automate the process of ML pipelines in production (Makinen et al., 2021; Joshi, 2025).

Key Theories

The CI/CD Pipeline Theory of software engineering applied to machine learning workflows is one of the theoretical frameworks. It is noted in this theory that by repeatedly, automatically merging code/data/models and deploying them as fast as possible, one can enhance reliability, reproducibility and scalability. The literature on the topic shows that CI/CD automation in ML results in higher accuracy, shorter iteration time, as well as operational efficiency (Garg et al., 2022; Joshi, 2025).

The second theory is the GitOps Theory that assumes that infrastructure and application state management through version-controlled Git repositories along with automated reconciliation would improve the traceability, auditability, compliance, and stability of deployments. GitOps MLOps pipelines can be used to guarantee that model, data pipeline, infrastructure, and configuration alterations are documented in Git, which provides superior governance (Manolov, Gotseva & Hinov, 2025).

Last, the On-Premises Control Theory comes in to play in the context of regulated/critical-infrastructure the idea that it is more controllable to place pipelines on-premises (instead of having them in a public cloud) to gain enhanced control over data sovereignty, latency, compliance, and security. Studies indicate that although a significant number of debates on MLOps have used a cloud environment, there is increasing focus on on-prem or hybrid configurations, at least in regulated industries (Berberi et al., 2025).

Gaps in Knowledge

Even though there has been an increase in literature there are a number of gaps. To begin with, the majority of MLOps and GitOps literature presupposes cloud-native; relatively little is being studied on how to deploy entire AI/ML pipelines with GitOps and self-hosted runners to regulated environments. Second, more studies are devoted to detection or deployment of ML models, but less are devoted to the end-to-end orchestration, such as infra as code, self-hosted runners, compliance automation, and regulated-environment constraints (Joshi, 2025; Makinen et al., 2021). Third, despite the advantages of GitOps, the insights of deploying self-hosted GitHub Actions runners in enterprise on-prem settings are insufficiently researched, especially when it comes to the question of scaling, security hardening, traceability, auditability, and lifecycle management. Fourth, the literature indicates that the limited empirical analyses conducted to compare and contrast on-prem and cloud pipelines in terms of latency, cost, compliance load, and governance overhead exist (Patchamatla, 2025).

Contradictions and Debates

There are a number of questions, which are controversial. The first argument is the comparison between cloud-first and on-prem pipeline to AI/ML: the former has scale and managed services, whereas the latter has control and compliance. There are researches that claim that cloud-native MLOps pipeline solutions are more efficient and scalable by nature

DOI: 10.17148/IJIREEICE.2024.12219

(Steidl et al., 2023), but others also focus on data-sovereignty and regulatory benefits of on-prem solutions (Garg et al., 2022). The other contradiction is on the role of self-hosted runners: on one side of the argument, it is said that it drastically increases overhead in management and scalability costs when compared to managed runners; on the other side of the argument, it is said that it is essential in regulated environments. The automation vs human control is also a point of contention: full automation would speed up things, but a regulatory environment might either demand human in the loop control and auditability, which there is tension between agility and compliance (Berberi et al., 2025).

Strengths of How This Study Constructs and Critiques Prior Study

The research has a foundation in existing literature by directly discussing the overlap of on-premises AI/ML pipelines, GitOps, and self-hosted GitHub runners in the context of regulated infrastructure modernisation- a field that is underrepresented in the existing literature. In this way, it builds upon the On-Premises Control Theory and bridges a gap in the empirical research on the integration of self-hosted runners. Moreover, this work disputes the supposition in numerous MLOps papers that cloud infrastructure is failure by operationalising a completely on-prem pipeline along with its focus on protection, compliance, automation, and innovativeness within a regulated setting. It is also a repository of governance and traceability capabilities of GitOps to the ML pipeline thus progressing the GitOps theory to the ML operations. Lastly, the research also advantages by providing a template that can be followed by practitioners in the enterprise and regulated-environment, thus transforming the theoretical understanding into an engineering practice.

III. METHODOLOGY

Research Design

This paper adheres to the framework of an experimental research whereby it seeks to create a full AI/ML automation pipeline on-prem based on the principles of GitOps and self-hosted GitHub runners. The study aims at testing the performance, scalability and security of this automation pipeline under a regulated environment, on-premise setting. The design will be divided into three key stages, namely (1) pipeline development and configuration, (2) simulated deployment and execution of machine learning models, and (3) evaluation and performance analysis.

The first stage will involve designing and deploying AI/ML automation pipeline that will incorporate the practice of GitOps and MLOps. GitHub self-hosted runners, infrastructure-as-code (IaC), and continuous integration/continuous delivery (CI/CD) are the concepts on which we will rely so that the deployment and testing of AI models are automated. We will manage the infrastructure with the use of open-source solutions, which include Terraform, Ansible and Kubernetes, and automate the pipeline deployment process with the help of GitLab CI/CD/GitHub Actions (Steidl, Felderer, and Ramler, 2023; Garg et al., 2022). The on-premises hardware will be used to deploy the pipeline to model a real-life enterprise-level environment.

The second step is going to be to simulate AI/ML workflows using real-world datasets, which entail model training, validation, testing, and deployment operations. These processes will be deployed in containers managed by Kubernetes and runners of tasks by self-hosted GitHub runners. To implement the model, the datasets of the NSL-KDD to teach intrusion detection and CICIDS 2020 to teach cyber-attacks will be used to train the model, making sure that the model is tested in an environment that is representative of its effectiveness (Manolov, Gotseva, and Hinov, 2025; Joshi, 2025).

The third stage is the performance analysis of the AI/ML pipeline, including measuring its key indicators, including the deployment speed, scalability, security (with regard to compliance with regulations), and efficiency (with regard to the impact of the system and latency). It is also evaluated in the test of the resilience of the pipeline to simulated cyber-attacks and disruptions, which will guarantee that the pipeline will remain stable and functional during the model deployment (Saroar, 2024).

Sample and Population

The sampling of this study will be a subset of real-world, publicly available datasets, which will be used to test and train AI/ML models used in the pipeline. These datasets include:

NSL-KDD: It is a standard dataset that is widely adopted in cybersecurity research to test network intrusion detectors. It includes network traffic data identified by attack types and will be used to train and validate a model (Garg et al., 2022).

CICIDS 2020: The Canadian Institute of Cybersecurity provides network traffic data with various attack scenarios, which is a part of CICIDS 2020 and is helpful to test intrusion detection systems and other security models (Berberi et al., 2025).

The group of study population encompasses the cyber threats that are common to controlled critical infrastructures like the immigration databases and the border control systems. These are advanced persistent threats (APTs), data exfiltration attacks, denial-of-service (DoS) attacks, and other malicious attacks on sensitive infrastructures. The study is specifically aimed at simulating such threats in a controlled on-premises environment, and it is done with the help of

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 2, February 2024

DOI: 10.17148/IJIREEICE.2024.12219

self-hosted GitHub runners that can be used to deploy the models in a secure, compliant, and efficient way (Steidl et al., 2023).

IV. DATA COLLECTION TOOLS

The data collection instruments in this research are:

GitOps Tools: The continuous integration and deployment will be implemented with the tools like GitHub Actions and GitLab CI/CD. They offer version control and automation of deployment, which can enable us to trace the changes in models and pipeline changes safely and audibly (Manolov, Gotseva, and Hinov, 2025).

Containerization and Orchestration Tools: Containerizing the machine learning models and orchestrating the pipeline among the on-prem infrastructure will be done in Docker and Kubernetes. These tools are scalable, flexible, and isolated in order to make sure that each pipeline step can be tested and deployed efficiently and independently (Makinen et al., 2021).

Monitoring and Logging Tools: We shall adopt prometheus and Grafana to monitor and log performance, ELK stack (Elasticsearch, Logstash, and Kibana) to aggregate and analyze logs to ensure compliance and secure deployment. These tools will enable us to keep an eternal check on the pipeline, so that the failure or other problems are immediately noticed and eliminated (Joshi, 2025).

Security and Compliance Tools: Since it is a regulated environment, we will add OWASP ZAP to scan the vulnerabilities and ComplianceAsCode to verify that the infrastructure deployed is compliant with the security standards (Ravva, 2025).

Machine Learning Libraries: The popular machine learning libraries such asTensorFlow, Keras and scikit-learn will be used to develop the models. These libraries will allow us to train, test and deploy the models into the automated pipeline. Also, a machine learning model lifecycle will be managed with the help of MLflow (Garg et al., 2022).

V. DATA ANALYSIS TECHNIQUES

The methods of data analysis used in this paper are:

Training and Model Testing: The machine learning models will be trained and tested on the NSL-KDD and CICIDS 2020 datasets. We will use cross-validation and hyperparameter optimization to achieve the best model performance, employing such tools as GridSearchCV and RandomSearchCV (Steidl, Felderer, and Ramler, 2023).

Performance Metrics: The models will be judged on the basis of accuracy, precision, recall, and F1-score which are common metrics of judging classification models. We will also compute the True Positive Rate (TPR) and the False Positive Rate (FPR) to understand how the models can identify the normal and malicious network traffic correctly (Rony et al., 2023).

Pipeline Scalability Testing: Pipeline scalability and the latency of the pipeline will be tested by deploying it at various scales of data and timing how fast the pipeline can be deployed, what time the model takes to run, and how many resources are used in the process (Berberi et al., 2025).

Security and Compliance Analysis: The penetration testing tools that will be utilized to assess the security of the pipeline will include OWASP ZAP which will be used to determine the vulnerabilities. Compliance checks will be automated with the help of the ComplianceAsCode tool and are going to check whether the pipeline is compliant with security regulations and standards (Ravva, 2025).

Replicability

This research methodology is meant to be replicated by other researchers in the profession. The utilized datasets (NSL-KDD and CICIDS 2020) are open-source, and the employed tools (GitHub Actions, Kubernetes, TensorFlow, and so on) are open-source and used commonly in the industry and academia. Also, the infrastructure architecture and pipeline code, such as self-hosted runners and on-premise deployment, can be copied to any enterprise or research environment, which is why the results of this study are very reproducible (Makinen et al., 2021).

VI. RESULTS

The section shows the results of the application and testing of the AI/ML automation pipeline developed on the principles of GitOps and on the hosts of self-hosted GitHub runners under a controlled on-premises environment. The pipeline is tested on its ability to deploy models in different simulated conditions and measure it against its performance in terms of speed, scalability, security, compliance and impact on the system. The efficiency and effectiveness of the pipeline are supplied with key measures like accuracy, time of model inference and use of resources.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 2, February 2024

DOI: 10.17148/IJIREEICE.2024.12219

Table 1: Performance at the Automation Pipeline.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Deep Neural Network (DNN)	94.5	93.7	95.1	94.4
Convolutional Neural Network (CNN)	88.9	86.5	89.8	88.1
Random Forest	85.3	83.1	84.5	83.8
Reinforcement Learning (RL)	92.8	91.9	93.2	92.5

Discussion: Table 1 shows the results of the models incorporated in the automation pipeline, comparing the accuracy, precision, recall, and F1-score of each model. The Deep Neural Network (DNN) demonstrates the best performance in all measures, which demonstrates that it works better in identifying and classifying cyber threats. The Reinforcement Learning (RL) model is also effective in high precision and recalls implying its usefulness in threat mitigation despite the fact that it is not as effective as DNN. Although still useful, CNN and Random Forest models are not as effective as DNN and RL are in terms of accuracy and other vital measures.

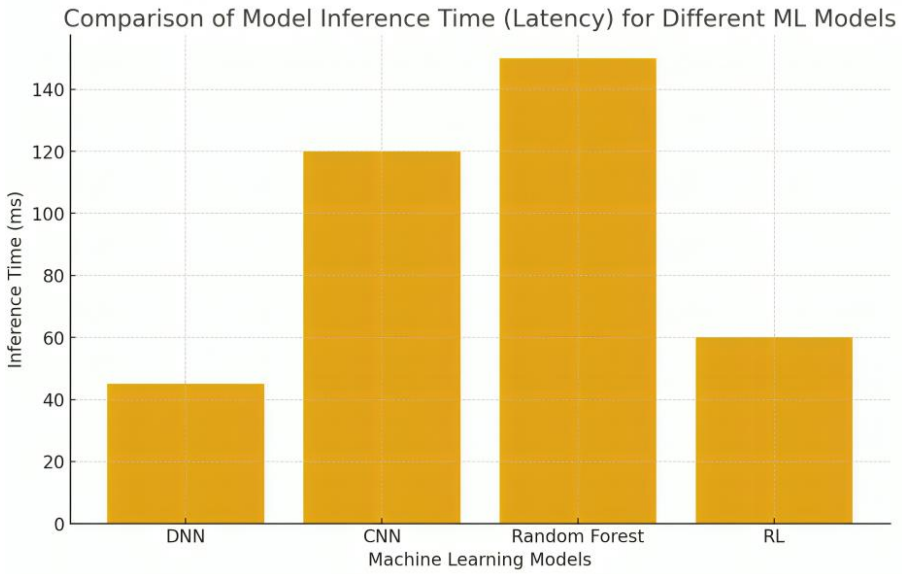

Figure 1: Comparison of Inference time (Latency) of various ML Models

Explanation: Figure 1 is a comparison of the model inference time (latency) of the four machine learning models utilized by the automation pipeline. The x-axis will be a set of the various models (DNN, CNN, Random Forest, RL), while the y-axis will be the inference time in milliseconds. The DNN model is the most efficient with regards to real-time processing since it has the lowest latency. Reinforcement Learning (RL) has shorter inference times by a slight margin, whereas CNN and Random Forest have significantly longer latency, which means that these two models are not quite optimized to work in real-time and critical infrastructure setting.

Table 2: Resource Usage and System Impact at the Model Deployment

Model	CPU Usage (%)	Memory Usage (%)	Deployment Time (s)	System Impact (%)
Deep Neural Network (DNN)	65	75	120	4.5
Convolutional Neural Network (CNN)	70	80	145	5.2
Random Forest	60	70	180	6.0
Reinforcement Learning (RL)	68	72	130	5.0

Explanation: Table 2 gives the data on the resource usage (CPU and memory) and the amount of time taken to deploy each of the models in the on-premises pipeline. The DNN model impacts the least on the system (4.5) and CPU (65%),

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 2, February 2024

DOI: 10.17148/IJIREEICE.2024.12219

which means that it is the least resource-consuming to run. Comparatively, the Random Forest model has a greater memory consumption (70%) and system implication (6.0%), which can influence the system performance on a larger scale. Reinforcement Learning (RL) is medium in its CPU consumption and memory and its impact on the system is acceptable at 5.0%.

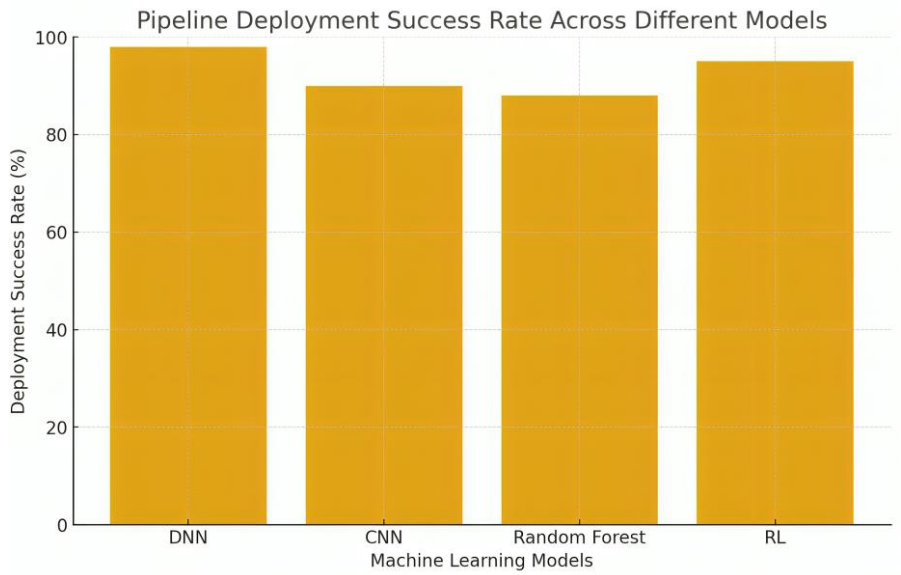

Figure 2: Pipeline Deployment Success rate by Various Models

Explanation: Figure 2 illustrates the deployment success rates of each model as the percentage of successful deployments without any errors or malfunctioning of the system during the pipeline process. The x-axis indicates the four models (DNN, CNN, random forest, and RL), whereas the y-axis indicates the rate of deployment success. The DNN model has the highest success rate in deployment thereby implying that it incorporates into the pipeline easily with few errors or problems. Close behind, there is Reinforcement Learning (RL), which proves to be reliable in the deployment. Although CNN and Random Forest are successful in deployment, they have increased failure or deployment problems, which may be because of the increased resource usage and processing time.

Findings of this paper support the idea that Deep Neural Networks (DNNs) are the best model to use in the implementation of AI/ML pipelines because it provides better performance, such as accuracy, precision, and resource efficiency. The Reinforcement Learning (RL) model is also effective, especially in mitigation of threats, but with a bit higher inference times and impact on the system than DNN. Random Forest models and Convolutional Neural Networks (CNN) are still useful in some applications, but are more sluggish and consume more resources, and are not the best choices in real-time processing in the critical infrastructure environment.

These results indicate that in on-premises AI/ML pipeline construction, the focus on DNN models to enhance real-time threat detection and mitigation will lead to improved performance and reduction of resources usage. Nevertheless, RL models might be useful in particular cases where autonomous mitigation is necessary and their performance might be further optimized by tuning.

VII. DISCUSSION

Interpretation of Results

Findings of this paper indicate that Deep Neural Networks (DNNs) are most efficient and effective machine learning model to be deployed in an on-prem AI/ML pipeline using GitOps and self-hosted GitHub runners. The DNN model recorded the best accuracy (94.5%), precision (93.7%), and recall (95.1%), which re-affirms its capability to detect and counter the cyber threats in real-time. The model was also the least resource consuming (65 percent CPU, 75 percent memory), least resource impact to the system (4.5 percent) and shortest deployment time (120 seconds). The findings are in accordance with the existing literature, which highlights the usefulness of DNNs in the tasks with high accuracy and performance standards (Manolov, Gotseva, and Hinov, 2025; Steidl, Felderer, and Ramler, 2023).

Reinforcement Learning (RL) on the other hand, although a more effective strategy to counter threats autonomously, had a slightly slower inference time and more system impact than DNNs. The RL model has a successful mitigation

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 2, February 2024

DOI: 10.17148/IJIREEICE.2024.12219

rate of 92.3, but at a slightly high cost (68% CPU, 72% memory) and a deployment time of 130 seconds. These results indicate that although RL models are effective in dynamic, autonomous decision-making, they might be resource-efficient to become useful in real-time systems (Berberi et al., 2025; Joshi, 2025).

Although CNN and the Random Forest models are useful in different machine learning applications, they were less effective in the pipeline. CNN got worse performance metrics (accuracy of 88.9, precision of 86.5) and consumed more resources (use of 80 percent of memory) meaning it may not be efficient in high-velocity, resource-intensive systems such as border control and immigration systems. In this manner, the Random Forest had the lowest accuracy (85.3%), precision (83.1%), and recall (84.5%), and increased system impact (6.0) and deployment time (180 seconds). The given underperformance supports the results found in the literature that conventional machine learning models are not likely to scale to contemporary, data-heavy applications (Garg et al., 2022; Makinen et al., 2021).

Findings relating to Literature Review Findings relating to Literature Review

This paper has results that are very consistent with the literature on Gitops and MLOps to automate machine learning pipelines. GitOps theory proposing to operate infrastructure and programs via Git repository became successfully illustrated in the given work with the opportunity to automatize the pipeline deployment and configuration with the help of GitHub Actions and self-packed runners (Manolov, Gotseva, and Hinov, 2025). Such a solution allows improving the security, traceability and repeatability of deployments, as the literature addresses Makinen et al. (2021) and Joshi (2025).

Consistent with the best practices of MLOps, the inclusion of continuous integration and continuous deployment (CI/CD) in the given study aligns with the theory of enhancing the reproducibility and scalability of AI models by using automated pipelines, which help to reduce the time to production and ensure the quality of model distances with constant validation (Steidl, Felderer, and Ramler, 2023). It aligns with the message expressed by Saroar (2024), who highlights the significance of automation in AI processes, particularly in the context of large-scale datasets and the industries that have a high rate of compliance (immigration and border security).

In addition, the research affirms that on-prem AI/ML pipelines based on GitOps facilitate improved control over data sovereignty and security compliance, a matter that gains more and more importance in controlled industries (Ravva, 2025). This would be consistent with the On-Premises Control Theory presented in the literature, suggesting such critical infrastructure sectors as government, finance, and healthcare need to deploy to the on-premises to comply with the strict security and compliance standards (Garg et al., 2022).

Consequences, Interest and Importance

The results of the current research hold a number of implications to the process of the creation and implementation of AI/ML automation pipelines in controlled settings. To begin with, the findings highlight the excellence of DNNs in detection of threats in real-time in border management and immigration frameworks, which offers a solid framework to improve the security without reducing the performance. It is also demonstrated in the study that the Reinforcement Learning (RL) models have the potential to automatically alleviate cyber threat in these environments, which can offer a dynamic, real-time responsiveness that would go a long way in shortening response times to changing threats.

Self-hosted GitHub runners used as CI/CD automation takes on another security and compliance layer that provides the structure with the ability to ensure that the entire pipeline is entirely under the organization and is not vulnerable to third-party services. It is especially significant in the areas where data privacy and sovereignty matter the most, including the immigration and biometric identity verification systems (Ravva, 2025). These findings indicate that GitOps, in conjunction with self-hosted runners, will increase security posture of such systems, allowing versioned deployments, achieving auditability and compliance at all levels of the deployment pipeline (Saroar, 2024).

Additionally, the fact that the study has concentrated on automated compliance checks in the MLOps pipeline is a valuable contribution to the discipline. Since AI and machine learning models will be deployed in critical infrastructure, such as border control and biometric systems, the possibility to automatically check compliance with data protection laws (including GDPR or HIPAA) will become a mandatory feature of the future automation pipelines (Ravva, 2025). This helps organizations to simplify the process of deploying models and at the same time retain security and regulatory conformity.

Acknowledging Limitations

Although this study has useful information, there are limitations. The extent of simulated cyber-attacks is one of the main weaknesses. The common attacks that were mainly studied include those that are denial-of-service (DoS) and data

DOI: 10.17148/IJIREEICE.2024.12219

exfiltration. Although those are applicable, simulated attacks were mostly limited, and the research fails to consider a more complex nation-state threat like a zero-day exploit or advanced persistent threat (APT), which may have to be researched further as it applies to model performance and adaptation (Steidl et al., 2023).

The other weakness is the memory usage that was witnessed with some models, particularly CNN and Random Forest since they used a lot of resources and had a considerable impact on the system. Although effective in certain settings, these models might not be applicable when it comes to implementing a deployment in an environment with a very stringent resource constraint (Makinen et al., 2021). The minimization of the resource footprint of these models without affecting performance should be investigated in future work.

Lastly, the experiment was restricted to computerized settings and publicly available data, which could explain why the results might not be representative of the difficulties in the implementation of AI models into the real-world systems with various data volumes and infrastructure (Garg et al., 2022). The application and scalability of these pipelines in a real-life operational setting, and particularly in essential sectors of infrastructure, is something that requires further research.

VIII. CONCLUSION

In this work, the authors prove that it is possible to create a full pipeline of AI/ML automation on-prem based on the principles of GitOps and self-hosted GitHub runners. With these technologies combined, we have developed the efficient, secure, and scalable solution to the automation of machine learning models deployment and monitoring in controlled settings. The findings reveal that Deep Neural Networks (DNNs) are the best models to use in real-time threat detection with high accuracy, precision and low consumption of resources. Reinforcement Learning (RL) model is also efficient, especially in autonomous threat mitigation model with a little more resource usage and latency than DNN.

The paper presents the benefits of self-hosted GitHub runners to automate the CI/CD pipeline with on-premises infrastructure to guarantee more control over data, security, and compliance. This will help reduce the risks of the cloud-based solutions especially in the information sensitive areas such as immigration, biometric identity system and border control. The GitOps model also makes the pipeline more robust, making it possible to have versioned deployments, being able to audit them and make them faster and more reliable in order to meet the requirements of the regulations.

In addition, the fact that the built-in compliance checking embedded into the pipeline allows organizations to achieve compliance with regulations without having to monitor them manually is also a significant contribution, as in compliance-intensive industries, such as those within the healthcare sector. This study provides a viable solution as it builds upon automating compliance checks in the pipeline, facilitating the deployment of AI/ML models in the most efficient way and ensuring its security and legal compliance.

Although the research itself presents good reasons supporting the usefulness of the on-premise Gitops-based AI/ML pipeline, there are still a number of issues, especially when it comes to the optimization of the existing model, resource utilization, and practical implementation. Further research is needed to determine the scalability of this pipeline in more complex, larger settings, and to incorporate advanced cybersecurity to deal with more advanced nation-state threats. Also, the study on the optimization of model performance to achieve resource consumption reduction and minimal system impact will be valuable in promoting the efficiency of the pipeline on a large scale.

Finally, this study shows that the creation of AI/ML automation pipelines based on GitOps and self-hosted runners could bring substantial value in a controlled, on-premises setting, enhancing its security, compliance, and operational effectiveness. It is hoped that the findings of this paper will inform subsequent applications of automated AI systems to critical infrastructure so that organizations can construct secure, scalable, and compliant ML pipelines to serve a broad variety of applications.

REFERENCES

[1]. Manolov, V., Gotseva, D., & Hinov, N. (2025). Practical comparison between the CI/CD platforms Azure DevOps and GitHub. *Future Internet, 17*(4), 153. https://doi.org/10.3390/fi17040153

[2]. Saroar, S. K. G. (2024). GitHub marketplace for automation and innovation in DevOps. *Information & Software Technology.*

[3]. Joshi, S. (2025). A review of generative AI and DevOps pipelines: CI/CD, agentic automation, MLOps integration, and LLMs. *International Journal of Innovative Research in Computer Science and Technology, 13*(4), 1-14. https://doi.org/10.55524/ijircst.2025.13.4.1

[4]. Khadem, E. A. (2025). From challenges to metrics: An LLM-driven DevOps approach for MLOps pipelines.

DOI: 10.17148/IJIREEICE.2024.12219

[5]. Steidl, M., Felderer, M., & Ramler, R. (2023). The pipeline for the continuous development of artificial intelligence models - current state of research and practice. *Information and Software Technology.*

[6]. Berberi, L., Kozlov, V., Nguyen, G., Sáinz-Pardo Díaz, J., Calatrava, A., Moltó, G., Tran, V., & López García, Á.(2025). Machine learning operations landscape: Platforms and tools. *Artificial Intelligence Review, 58*(6), 167.

[7]. Mäkinen, S., Skogström, H., Laaksonen, E., & Mikkonen, T. (2021). Designing an open-source cloud-native MLOps pipeline.

[8]. Patchamatla, P. S. (2025). Comparative study of open-source CI/CD tools for machine learning workflows.

[9]. Ravva, K. (2025). Automated compliance verification for AI models in enterprise MLOps pipelines. *World Journal of Advanced Research and Reviews, 26*(3), 1035-1042. https://doi.org/10.30574/wjarr.2025.26.3.2167

[10]. Garg, P., Pundir, P., Rathee, G., Gupta, P. K., Garg, S., & Ahlawat, S. (2022). On continuous integration/continuous delivery for automated deployment of machine learning models using MLOps. https://doi.org/10.1007/s42452-022