RESEARCH ARTICLE

Drift detection on feature attributions for monitoring visual reinforcement learning models in maritime port surveillance

[version 1; peer review: 3 approved]

Francisco Javier Iriarte[1,2], Beatrice Azoubel[1], Adrián Carrizo-Pérez[1], Andrés Chica Linares[3], Luis Unzueta[1], Ignacio Arganda-Carreras[2,4-6]

[1]Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Donostia/San Sebastián, Gipuzkoa, 20009, Spain
[2]Department of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU), Bilbao, Basque Country, 20018, Spain
[3]INDRA Sistemas S.A., Alcobendas, 28108, Spain
[4]Donostia International Physics Center, San Sebastián, Basque Country, 20018, Spain
[5]Biofisika Institute (CSIC-UPV/EHU), Leioa, Bizkaia, 48940, Spain
[6]Ikerbasque, Bilbao, Basque Country, 48009, Spain

v1 **First published:** 02 Jan 2026, **6**:2
https://doi.org/10.12688/openreseurope.22116.1
Latest published: 02 Jan 2026, **6**:2
https://doi.org/10.12688/openreseurope.22116.1

Abstract

Background

Maritime activity is expanding globally, increasing the demand for robust port security systems capable of detecting illegal trafficking. Due to the growing sophistication of smuggling methods, law enforcement agencies require advanced surveillance and prevention technologies such as those developed in the SMAUG project. In this context, initiatives such as the SMAUG project aim to deliver integrated surveillance capabilities coordinated by a high-level deep reinforcement learning (DRL) decision-making system that operates on image-based environmental representations. Despite their effectiveness, DRL models are closed-boxes, complicating continuous model monitoring (CMM). Conventional drift detection captures shifts in input or output distributions yet often fails to explain underlying problems. Explainable AI (XAI) techniques can provide a complementary approach with insights into the agent's inner workings, enabling monitoring of the concept rather than just the data.

Methods

We propose FADMON, an XAI-driven concept drift detection method for image-based models. FADMON performs statistical drift tests on

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 1 02 Jan 2026	✓ view	✓ view	✓ view

1. **Alexandru Pohontu**, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania
2. **Adeola Oluwatoyin OSUNDIRAN**, University of South Africa, Pretoria, South Africa
3. **Ayoola Babatunde Fadola**, Washington College of Law, American University, Washington, USA

Any reports and responses or comments on the article can be found at the end of the article.

feature attributions to detect deviations in learned policies. We demonstrate how FADMON can enhance CMM with a three-stage model monitoring architecture that enables semi-supervised explainable model monitoring. We validate our approach with SMAUG's decision-making DRL model on a simulated maritime port surveillance environment under multiple unforeseen scenarios.

Results

FADMON consistently flags drift on all drifted scenarios with mean p-values of 0.000 with no variance trough 30 repetitions, with lower mean p-values (0.553±0.215) on non-drifted scenarios with respect to other established drift detection methodologies such as prior probability shift detection (0.65 ± 0.000), though well above the standard 0.05 threshold.

Conclusions

FADMON can add an explainability layer to the monitoring system while also supporting detection of changes in the underlying interpretation of the input data by the model, monitoring the concept rather than the data, while matching established drift detection methods metrics-wise.

Plain Language Summary

This article shows how it is possible to monitor an AI model interpretation of the data it is being fed, rather than monitoring only the data itself, to improve our capability to understand how an AI model is behaving while it works. This allows us to detect changes in the way it is behaving, as well as understand these changes better, in order to detect more quickly when the model is not behaving correctly and fix it. In this article, we apply this approach to a maritime port surveillance environment, testing it under different simulated scenarios, such as overabundance of obstacles in the port or malicious attacks on the vessels, to see if the method meets its potential. We conclude that it does, with certain caveats such as the need of stronger hardware, resulting in a good balance between detection of irregular AI model behavior and interpretability of said behavior.

Keywords

Visual Reinforcement Learning, Drift Detection, Explainable AI, Continuous Model Monitoring, Maritime Port Surveillance

HE

This article is included in the Horizon Europe gateway.

Corresponding author: Francisco Javier Iriarte (iripatx@gmail.com)

Author roles: Iriarte FJ: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Writing – Original Draft Preparation; **Azoubel B**: Formal Analysis, Investigation, Methodology, Software; **Carrizo-Pérez A**: Investigation, Validation, Writing – Original Draft Preparation; **Chica Linares A**: Conceptualization, Writing – Review & Editing; **Unzueta L**: Conceptualization, Investigation, Writing – Review & Editing; **Arganda-Carreras I**: Conceptualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No [101121129] (Smart Maritime and Underwater Guardian - [SMAUG]).

How to cite this article: Iriarte FJ, Azoubel B, Carrizo-Pérez A *et al.* **Drift detection on feature attributions for monitoring visual reinforcement learning models in maritime port surveillance [version 1; peer review: 3 approved]** Open Research Europe 2026, **6**:2 https://doi.org/10.12688/openreseurope.22116.1

First published: 02 Jan 2026, **6**:2 https://doi.org/10.12688/openreseurope.22116.1

Introduction

Global maritime activity continues to expand in both scale and complexity, with extended trade routes and increased port throughput across major global hubs[1]. This sustained expansion, however, has been accompanied by a surge in maritime trafficking and smuggling operations. Recent studies by Europol and the European Monitoring Centre for Drugs and Drug Addiction[2] highlight a marked increase in the use of maritime routes, particularly containerized cargo and coastal vessels, for the trafficking of drugs and other illicit goods into Europe.

In response to these growing challenges, the European Union has supported several research and innovation initiatives aimed at enhancing maritime domain awareness and port security. The COMPASS2020 project demonstrated the operational integration of manned and unmanned platforms to extend surveillance coverage and reduce reaction times in coastal environments[3]. Similarly, the RAPID project advanced the concept of risk-aware, autonomous inspection by combining unmanned surface and aerial vehicles for real-time hull and infrastructure assessment[4]. Ongoing projects like UNDERSEC further contribute to this line of research by developing a modular underwater security architecture leveraging robotic assets for persistent port protection[5]. Building upon these efforts, the SMAUG project introduces a comprehensive framework that integrates acoustic sensing, rapid sonar hull scanning, and high-resolution underwater inspection to detect and characterize concealed threats such as submersible vessels or smuggling devices[6].

SMAUG's integrated architecture includes a high-level decision-making module that manages the system in real time. Deep reinforcement learning (DRL) has facilitated the design of these systems, in which RL agents learn to make optimal decisions through interaction with complex, partially observable environments. In particular, visual RL, where policies are trained end-to-end from pixel inputs, is being increasingly adopted in safety-critical and high-stakes domains, including robotics, autonomous navigation, and maritime port surveillance[7]. However, DRL models, like any other deep neural network (DNN) model, are often opaque and treated as closed boxes[8]. This opacity hinders the ability of developers and stakeholders to diagnose errors, increasing the risks and consequences of policy failures in high-risk applications.

These consequences are further amplified because DNNs are prone to performance degradation after deployment. This is commonly caused by shifts in the environment compared to the conditions seen during training, a phenomenon referred to as data drift. Moreover, the relationship between observations and the RL agent's chosen actions may also evolve over time, either because of data drift or due to the emergence of new behavioral patterns in the environment. This is typically described as concept drift[9]. Both forms of drift undermine the reliability of RL agents, particularly in dynamic real-world scenarios such as maritime ports.

The MLOps paradigm introduces Continuous Model Monitoring (CMM) as a mechanism to track deployed models and identify failure cases in order to update policies and sustain performance[10]. CMM is an important procedure that helps ensuring robust and trustworthy AI systems. For image-based models such as visual RL, however, this process is even more challenging than for regular supervised models, since reward signals are sparse, action distributions are high-dimensional, and access to ground-truth performance is limited[11].

Classical drift detection methods such as the Kolmogorov–Smirnov (KS) test[12] or Maximum Mean Discrepancy (MMD)[13] enable unsupervised monitoring by analyzing shifts in the input or output distributions. Yet, these techniques only consider environment states or policy outputs in isolation and do monitor nor provide visibility into the RL agent's internal decision-making process, making concept drift detection more challenging. Furthermore, fields like maritime port surveillance fall into the law enforcement and critical infrastructure categories under the European Commission's AI Act[14] and thus could be considered as high-risk AI systems if they meet certain conditions such as it profiling individuals or replacing human assessment. Therefore, any AI system used in these scenarios should comply with robustness and transparency standards that are not feasibly achieved with drift detection methods alone.

Explainable AI (XAI) methods such as SHAP[15] or Integrated Gradients[16] can provide feature attribution-based explanations that make policy decisions more transparent. By exposing the specific parts of an observation that drove an RL agent's action, XAI offers the necessary visibility into the decision boundary that is entirely missing from classical drift detection methods (like KS or MMD). As such, combining classical statistical drift detection with feature attributions holds significant potential to detect concept drift and improve compliance in high-risk applications. Recent articles have examined this combination for tabular models[17–19], but this approach is yet to be adapted to image-based models.

In this paper we propose FADMON (Feature Attribution Drift MONitoring), a semi-supervised, explainability-driven approach for concept drift detection of visual RL agents. By combining statistical drift detection with feature attribution-based explanations, FADMON supports unsupervised detection of concept drift while also enabling health reports and diagnostics of the model's decision-making process. We validate our approach directly under SMAUG's decision-making environment and RL agent, testing out methodology on a representative component of a maritime port surveillance system and verifying improvements on the robustness, transparency and maintainability of the system. Our contributions are as follows:

- FADMON, a semi-supervised concept drift detection method for visual RL that integrates statistical tests with explainability techniques.
- An XAI-based semi-supervised model monitoring architecture that integrates FADMON to raise automatic alerts of data, concept and probability drift while offering online visualization of policy explanations.
- An empirical validation of the method through experiments on a simulated visual RL-based maritime surveillance environment under diverse scenarios.

Related work

DRL combines reinforcement learning with DNNs to learn policies and value functions directly from high-dimensional inputs, enabling end-to-end decision-making systems from images or raw sensors. Early breakthroughs like Deep Q-Network (DQN) used convolutional networks with experience replay and target networks for discrete control, establishing practical stability tricks that informed later algorithms[20]. This was followed by policy-gradient and actor–critic families for continuous action spaces, such as Asynchronous Advantage Actor-Critic (A3C), which leverages parallel actors and advantage estimation for more stable updates[21] and Proximal Policy Optimization (PPO), which refines on-policy learning with a clipped surrogate objective to prevent destructive step sizes while retaining sample efficiency[22]. In visual domains, representation learning with Convolutional Neural networks (CNN) and, increasingly, Vision Transformers, together with auxiliary or self-supervised objectives, improves feature quality, stabilizes training, and reduces data requirements. Temporal encodings, frame stacking, and lightweight recurrence are often combined to exploit dynamics in video-based inputs[7].

Deep reinforcement learning is increasingly deployed in high-risk, safety-critical domains. Regarding maritime surveillance and law enforcement, DRL has been explored for port and open-sea coverage to detect abnormal vessel behaviors, support wide-area search, track targets, and better allocate resources under challenging sensing conditions[23]. Beyond the maritime domain, DRL coordinates multi-Unmanned Aerial Vehicle (UAV) surveillance by learning cooperative policies that improve reliability and coverage despite communication and dynamics uncertainties[24] as well as forensic investigation, where it has been proposed to streamline workflows by prioritizing evidence and adapting search strategies in complex scenes[25]. In autonomous driving, recent surveys document DRL's role in the control and decision-making systems at the core of self-driving vehicles[26].

Production-level use of DRL (as well as AI systems in general) has drawn attention to certain drawbacks that DNNs present after being deployed, which directly threaten their trustworthiness and viability: DNNs tend to lose accuracy over time[27], and are not understandable by humans due to their complexity and "black-box" behavior[28].

The phenomenon of models performing worse over time is often referred to as model drift, which is often caused by two key factors: data drift (or covariance shift), which refers to the distribution of the observed data changing over time, and concept drift, referring to changes in the relationship between input and output data of the model[29]. Drift detection is the process that aims to detect model drift when it occurs. It has many implementations, such as statistics-based tests like the Kolmogorov-Smirnov test[12], Cramér-von Mises[30], or Maximum Mean Discrepancy (MMD)[13], or model-based techniques such as ADWIN or drift detection method (DDM)[31]. Drift detection methods can be supervised, unsupervised or semi-supervised depending on whether labelled data (and thus manual annotation) is required, not required or partially required respectively[32].

Drift detection is a key component of CMM. Unsupervised drift detection is key in production environments; as it is impractical to label great amounts of operational data continuously and in short time. Still, while useful, drift detection methods do not provide insight into how a model responds to changes in data, making them unable to address the difficulty of comprehending why models make mistakes. Performance metrics such as accuracy, precision, recall, and F1-score offer some level of understanding, but as models increase in size and complexity, these metrics alone may be inadequate, particularly for DNNs[33].

To better understand the reasoning behind a DNN's decisions and diagnose the factors that lead to mistakes, XAI methods aim to provide additional information from the model via explanations. These explanations can be either intrinsic from inherently interpretable models[34,35] or be provided from external methods that generate them in a post-hoc manner[36]. They can be also classified as model-specific[34,35] or model-agnostic[31,37] if the method is applicable to a specific architecture or to any architecture, respectively. Finally, XAI methods can vary in the scope of their explanations: global methods aim to explain the model's general behavior[34,35], while local methods generate explanations for specific predictions, giving insight on why the model reached that answer[36,37]. XAI methods have been applied successfully to many Computer Vision architectures, e.g., CNNs[38] and Vision Transformers (ViT)[39].

XAI methods can convey explanations in multiple formats, such as feature attributions, which assign each input feature an importance score for how much it influenced a given output, or counterfactual explanations, which generate slightly modified inputs that radically change the original output[40]. Extending feature attributions, attribution (or saliency) maps visualize how input features contribute to an image-based model's prediction, typically as pixel-level heatmaps. Gradient-based methods include Integrated Gradients[16], which estimate contributions by accumulating integrals through the network. G-DeepSHAP[41] propagates Shapley-value attributions through a composition of differentiable layers/components, providing efficient attributions for neural networks and model stacks. Some methods like Grad-CAM[36] and its follow-ups (e.g., Grad-CAM++[42], Eigen-CAM[43]) specialize in CNN architectures to produce attribution maps for specific output classes. For model-agnostic analysis, KernelSHAP[15] can compute attribution maps by computing the model's output over a set of modified inputs that omit specific features, aggregating output changes with Shapley's game theory. B-cos makes attributions interpretable by design by replacing standard linear layers with a B-cos transform that enforces input–weight alignment, so a model's forward pass yields faithful, human-readable saliency that coherently sums contributions across layers[34].

However, feature attribution methods have limitations independently on the method used to approximate them. SHAP values - same as its inspiration, the Shapley values - can be misinterpreted, and it's even possible to create intentionally misleading interpretations, reducing trustworthiness of explanations that are implemented to improve it in the first place.

These methods are also known to make feature independence assumptions, as well as only highlight correlations that do not imply causality. Finally, deterministic implementations like KernelSHAP are slow and unsuitable for real-time monitoring. Global explanations are particularly affected, since generating them would require computing several local values[44].

Feature attributions could be used to monitor the model's interpretation of the input data, complementing monitoring of the data itself. Recent studies have successfully implemented drift detection on SHAP values to monitor ML models for tabular data[17–19] but, to the best of our knowledge, its applications for image-based models such as visual RL have not yet been explored.

Methodology

Figure 1 shows the flow diagram of FADMON, our proposed attribution-based concept drift detector. The method takes pixel-level feature attributions of a reference dataset and a batch of test data, previously computed using methods such as SHAP or Integrated Gradients. Then, min–max normalization is applied on all attributions using only the reference set's extremes. This sets all values to a fixed scale based on the reference attributions, providing a stable baseline and amplifying deviations in the test attributions. Next, to reduce the influence of the pixels with near-zero activations which dilute distributional differences, we perform top-bottom percentile selection. We retain only the upper and lower τ-th percentiles of this feature attributions, where $\tau \in (0,1)$ is the percentile threshold. These values are sorted to form fixed-length feature vectors, discarding spatial information since it is not relevant for feature attributions. Finally, we compute the Maximum Mean Discrepancy (MMD) between the reference and test vectors, yielding a statistically meaningful metric of the similarity of the distributions from which reference and test data came from. This value is compared against a predefined threshold to create concept drift alerts: If the similarity of the original distributions is lower than the threshold, FADMON flags concept drift on the current test data batch.

MMD is an integral probability metric that measures the distance between the distributions of two samples by mapping them with a kernel into a reproducing kernel Hilbert space (RKHS) and computing the norm of the difference between their mean embeddings[13]. In practice, it evaluates pairwise kernel similarities within and across the two samples; with a characteristic kernel (e.g., Gaussian RBF), MMD equals zero if and only if the underlying distributions are identical. MMD is multivariate and non-parametric, making it well suited to high-dimensional signals such as images or attribution vectors, and is widely used both as a loss/cost for training machine learning models (e.g., GANs)[45] and for drift detection. Because the method is kernel-based, one can tailor sensitivity to different forms of shift by choosing the kernel and its hyperparameters (e.g., bandwidth in the Gaussian RBF kernel). These properties make MMD a good fit for FADMON, though any statistical metric that measures distributional differences and supports high-dimensional feature vectors can replace or complement MMD.

To better specify how FADMON could enhance CMM, we will explain next how to integrate it in a model monitoring architecture. Figure 2 illustrates this three-stage architecture that enables XAI-based monitoring of a deployed DRL agent. Stage 1 builds reference monitoring data, which contains the values corresponding to the expected behavior of the model. Stage 2 takes incoming state batches and performs online model drift detection including concept drift detection using FADMON. The feature vectors computed for FADMON are also leveraged to create user-friendly model health reports. Stage 3 closes the loop with semi-supervised feedback from an expert user, potentially scheduling a model retraining.

The workflow is kickstarted by a new model version. The process begins with the reference monitoring data generation stage, where a working copy of the model as well as a feature attribution extractor process the training data on which the model was originally trained. This generates a pack of data consisting of three types of monitoring signals: input states (images), model outputs (probability vectors), and pixel-level feature attributions. Though any feature extraction method can potentially work, a low latency method like Integrated Gradients or an interpretable-by-design model such as B-cos are recommended, since the same method is also deployed in near-real-time at the next stage. Stage 1 is done offline, either in parallel or before the model is deployed, and is only activated once per model version since the reference monitoring data needs to be generated only the first time.

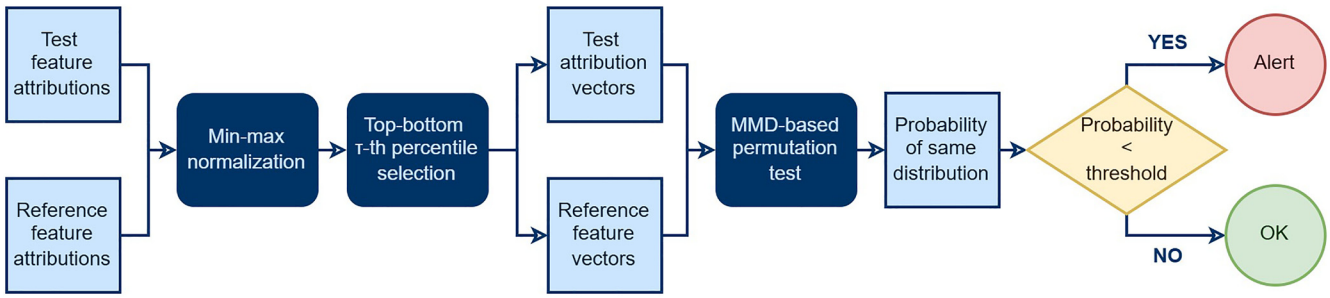

Figure 1. Flow diagram of FADMON, illustrating how it detects concept drift by performing drift detection on feature attributions.

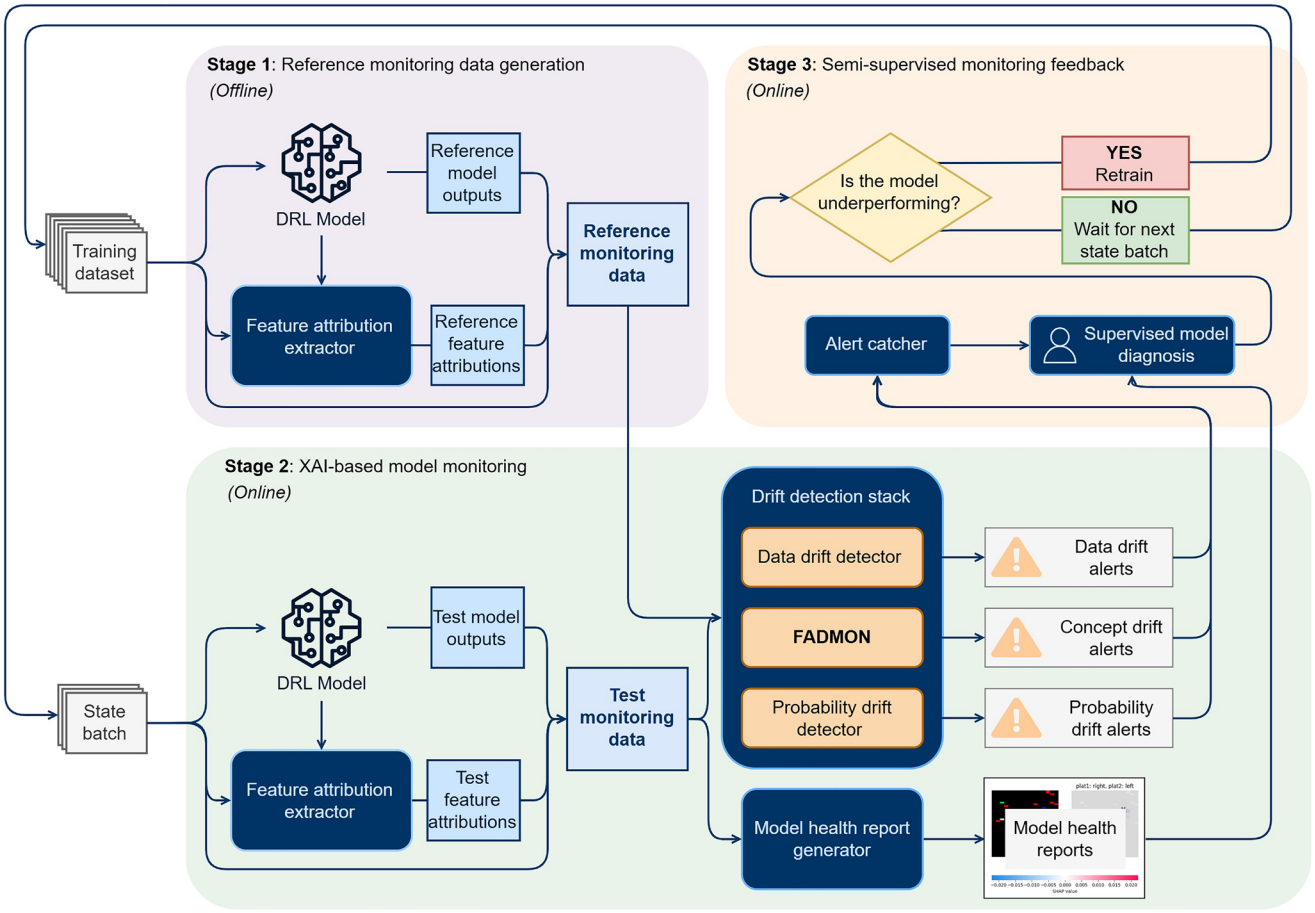

Figure 2. XAI-based semi-supervised model monitoring architecture. Stage 1 is done offline and once per model version, generating reference monitoring data. Stage 2, the main stage, performs online monitoring that creates drift alerts and user-friendly model health reports. Stage 3 involves a user that analyzes monitoring data in a semi-supervised way and decides if model retraining is necessary.

The second stage, XAI-based model monitoring, is the main stage of the architecture and the one integrating FADMON. This stage periodically takes batches of current states the DRL model is exposed to, processing them with a working copy of the model and the feature extraction method, thus obtaining test monitoring data with the same structure as the reference. This data is passed through the drift detection stack, containing three drift detectors that detect distributional changes: The data drift detector, which analyzes input images, the probability drift detector, which analyzes probability vectors, and FADMON, which analyzes feature attributions. This stack generates soft-real-time alerts, enabling swift detection of data, probability, and concept drift. In parallel, the test monitoring data is reused to create model health reports: User-friendly visualizations of every piece of relevant information for model monitoring. These reports include the original input, output, and attribution maps, as well as any additional information relevant to the real-time monitoring of the model, such as date, time and/or operation status. Since local feature attributions are used, a report can be generated for each observed state, as well as aggregated reports for global explanations.

Finally, the semi-supervised monitoring feedback stage introduces a semi-supervised feedback loop in which human experts can react to alerts and diagnose model performance by examining statistical metrics and the model health reports, being able to confidently determine whether readjustment actions are required. All alerts produced in Stage 2 are routed to an alert catcher module that filters, aggregates, and prioritizes them. When a relevant incident is detected, the alert catcher triggers a supervised model diagnosis step. If the diagnosis concludes that the DRL policy is underperforming or that the detected drift poses a considerable risk, model retraining is kickstarted, involving collection of the data that triggered the drift alert in the first place and returning to Step 1 after model and training data are updated. If the analysis concludes that the model is still able to function correctly, the architecture returns to the start of Step 2, waiting for a new state batch to arrive to start the monitoring process again.

Experimental results

Visual RL model

For our experiments, we make use of the Proximal Policy Optimization (PPO) model architecture. PPO is an on-policy,

actor–critic reinforcement learning algorithm that updates the policy by maximizing a clipped surrogate objective, which constrains each gradient step to stay close to the previous policy. This stabilizes training while still allowing relatively large policy updates and has become a standard baseline for continuous and discrete control tasks[22].

In SMAUG, PPO is deployed as a high-level decision-making module that coordinates a fleet of surveillance vessels within a port. The policy recommends routes for each vessel that aim to minimize traversal time and fuel consumption while avoiding static obstacles and other moving vessels. The system operates in a target analysis mode, in which vessels are required to move towards specific locations to perform detailed inspection of regions of interest in the port.

SMAUG's decision-making system operates on pixel level representations of the port's state. Following this specification, we train and evaluate an instance of an image-based PPO architecture that leverages a convolutional backbone to extract features from pixel-level representations, outputting a probability vector of discrete motion commands (up, down, left, right), indicating the direction in which the vessel should move at the next decision step.

Dataset

We generate a series of synthetic datasets of the model's expected pixel-level states in multiple scenarios, both under normal conditions and under unexpected situations. We create these datasets by simulating vessel positions, obstacles, and regions of interest and render them into image-like states that match the input format of the PPO agents. In these states, blue pixels represent obstacles that are not traversable and should be avoided, red pixels represent targets the vessels need to reach, green pixels represent vessels, and black pixels represent traversable areas (in this case, obstacle-free water). To introduce realistic model drift in maritime port environments, we create two unexpected or drifted scenarios. These drifted configurations are used to generate states that depart from the reference distribution, emulating operational changes in the environment that the model has not seen during training. All scenarios are simulated during a total of 300 timesteps, resulting in 300 pixel-based representations, for a total of 2100 samples. Figure 3 illustrates examples of both reference and drifted states for the target analysis model.

The first drift scenario is obstacle density increase, where the port environment is filled with more obstacles than the model has seen during training. This scenario immediately introduces

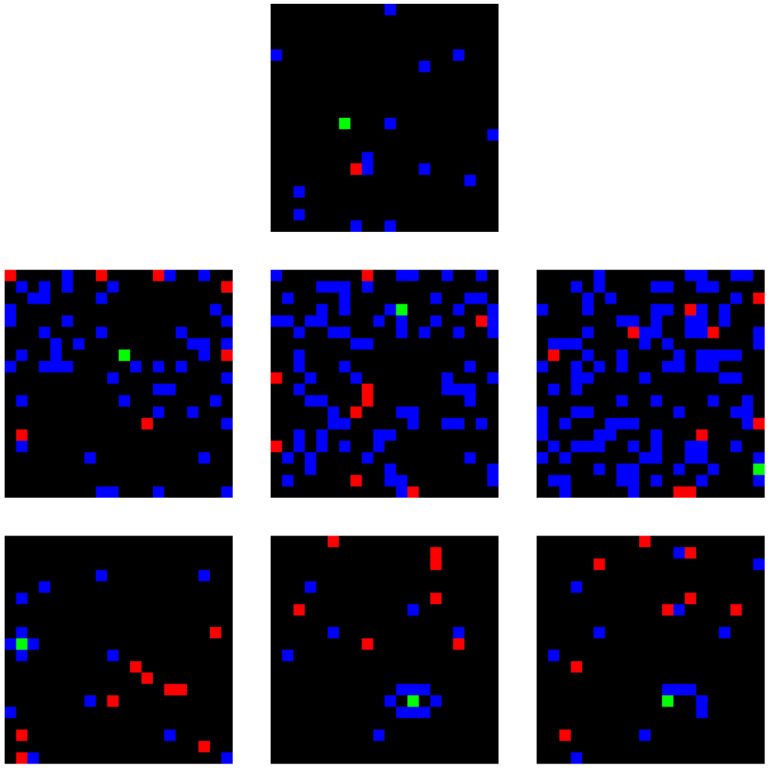

Figure 3. Examples of the pixel-level representations that serve as model input. Reference (top), obstacle density increase (middle), and vessel confinement (bottom). Blue pixels represent obstacles, red pixels represent targets, green pixels represent vessels, and black pixels represent traversable areas.

realistic data drift, as the drifted input representations differ from the ones used for model training. This data drift causes, in turn, changes in the optimal policy, since older optimal routes may be riskier, more costly, or even impossible if objective are unreachable, inducing concept drift. We create three datasets of this drifted scenario with variable increases in obstacle density. Three variants of this scenario have been creating, each with increasing obstacle density.

The second drift scenario, vessel confinement, involves partially or completely surrounding vessels with obstacles and effectively locking them in place or greatly restricting its movement options, simulating an accident with other elements of the port or an intentional attack. While this scenario also introduces data drift in the increase of targets over time, this represents a sharper alteration in the expected reward of the decisions taken by the model: actions that were previously safe become impossible to make or even dangerous, leading to a stronger concept drift in the optimal behavior. Three variants have been created as well: One where the vessel is fully surrounded with no possible movement; a second where it is fully surrounded but retains a narrow area to move; and a third where it is not fully surrounded, but obstacles are arranged so that the policy cannot reliably guide the vessel out of the blocked region.

Feature attribution extraction method

We make use of Expected Gradients, commonly known as GradientSHAP, to approximate SHAP values for DRL models. SHAP values quantify how much each feature contributes (positively or negatively) to a model's prediction, thus serving as feature attributions. The method makes use of a background set of input data, which constitutes the method's baseline samples. These samples are then used to generate interpolated inputs and average the output's gradients with respect to theirs, resulting in approximate SHAP values. GradientSHAP is notably fast, making it especially suited for image-based deep differentiable models, such as the ones use in Visual RL that comparatively have a greater number of input features than tabular models. This method is also known to compute smooth attributions that are consistent between executions, if a correct background is provided.

The size of the background used by GradientSHAP is unique for each case and needs to be calibrated. This background must be sufficiently large to maintain the original distribution of the training dataset, while not being excessively large, as this can significantly increase the computational cost of the explanation process. We estimate a suitable background size as follows: For a given model and candidate background size, we repeatedly (100 times) build a background by randomly sampling training instances. We then compute feature attributions using GradientSHAP, yielding a set of attributions that we normalize and use to compute a variance matrix over features and pixel positions. We then average this matrix to obtain a single mean variance score for that background size. Comparing these scores across sizes allows us to assess how background size affects the stability of the feature attributions and find the value that yields the highest precision over computational cost. Based on this analysis, a background size of 200 was selected for the experiments.

Experimental workflow

To test FADMON under a relevant scenario and compare it to other established drift detection procedures, we compare the obtained p-values and MMD distances against a data drift and a probability drift method. All drift detectors are tested on the three variants of the drifted scenarios described in the Dataset section, as well as on an undrifted dataset with similar distribution to the training dataset to evaluate the method's robustness.

We configure FADMON with a percentile threshold τ of 0.025 to receive feature attributions from a GradientSHAP instance with a background size of 200. To create our data drift detection workflow, we follow Rabanser *et. al.*'s proposed methodology, consisting of an untrained autoencoder to reduce input dimensionality followed by a statistical test on the resulting feature vectors[46]. Probability drift is detected by directly performing the statistical test on the probability vectors that the model outputs. To compare the three approaches fairly, we employ MMD with an RBF kernel as the statistical test for all three drift detectors. Equity is also ensured by running the experiments for all detectors under multiple kernel bandwidth values, ranging from 0.01 to 1000, and selecting the value that maximizes the p-value difference of undrifted vs drifted datasets. Due to the randomness introduced by UAEs for data drift detection and by feature attribution approximations of GradientSHAP, each drift detection experiment is further repeated 30 times to obtain the mean and variance of each detector's p-values.

Results

The results of the experiment are illustrated in Table 1.

All methods correctly identify the undrifted dataset as originating from the same distribution as the reference, yielding p-values well above the standard 0.05 significance threshold. Among them, FADMON exhibits the lowest mean p-value on undrifted data which, although still safely non-significant, suggesting a certain oversensitivity to minor distributional changes. Still, FADMON consistently detects all drift scenarios with zero variance across runs, with all reported p-values numerically equal to 0 with no variance. This behavior represents a clear improvement over the data drift detector, which fails to flag drift in scenarios where changes in the input distribution are more subtle, such as the vessel confinement scenarios. These observations indicate that FADMON is more focused on the underlying concept drift, relying less on direct input distribution comparison and more on changes in the model's attribution patterns.

The probability drift detector matches FADMON's detection performance on the drifted scenarios and yields slightly higher mean p-values, with zero variance, on the undrifted dataset. This suggests that probability drift detection remains a strong baseline in terms of pure drift detection performance when only output distributions are considered. However, probability vectors are not inherently interpretable and therefore provide limited support for comprehensive supervised model monitoring. In contrast, the feature attributions fed to FADMON can expose the spatial patterns driving the detected drift

through attribution visualizations, at the cost of higher computational overhead. Overall, these results show that FADMON offers a favorable trade-off between detection capability and interpretability, complementing data and probability-based detectors for a comprehensive, more explainable model monitoring procedure. Figure 4 illustrates how the feature attributions FADMON uses can be visualized for better model interpretation. In drifted scenarios, the attribution maps show overall weaker activations, indicating that the model struggles more to identify strong evidence on which to base its decisions. The maps also expose the model's limitations in multi-target scenarios: the model receives multiple strong, yet conflicting, positive and negative attributions across all objectives, suggesting that it attempts to satisfy all targets simultaneously and therefore considers multiple, potentially inconsistent, directions.

Table 1. Mean and variance of the p-values for each drift detection method, obtained by performing each experiment 30 times.

	Undrifted dataset	Obstacle density drifted scenario 1	Obstacle density drifted scenario 2	Obstacle density drifted scenario 3	Vessel confinement scenario 1	Vessel confinement scenario 2	Vessel confinement scenario 3
Data drift detection	0.757±0.217	0.059±0.086	0.007±0.02	0.003±0.006	0.132±0.93	0.066±0.093	0.209±0.135
Probability drift detection	**0.65±0.000**	0.008±0.000	**0.000±0.000**	**0.000±0.000**	**0.000±0.000**	**0.000±0.000**	**0.000±0.000**
FADMON (ours)	0.553±0.215	**0.000±0.000**	**0.000±0.000**	**0.000±0.000**	**0.000±0.000**	**0.000±0.000**	**0.000±0.000**

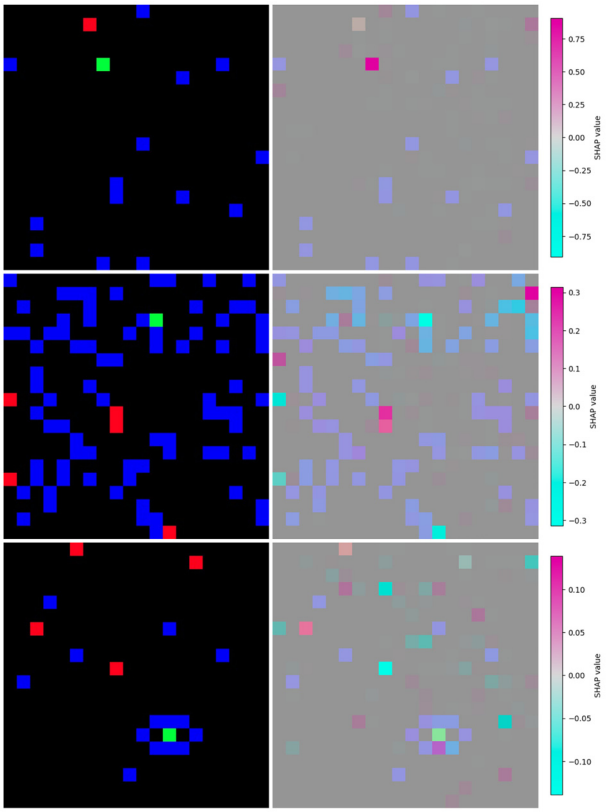

Figure 4. Examples of image-based model inputs (left) and visualizations of feature attributions computed by GradientSHAP (right) for the undrifted scenario (up), obstacle density increase scenario (middle) and vessel confinement scenario (bottom). Feature attributions show how the models interprets the data differently under the drifted scenarios, showing less overall feature activation strength as well as mixed positive-negative activations on targets.

Conclusions

In this paper we have presented FADMON, a semi-supervised concept drift detection method for visual RL. By integrating statistical tests with feature attribution techniques, FADMON enables unsupervised concept drift detection as well as supervised comprehensive model diagnosis using visual explanations. We have shown how FADMON can be integrated into the CMM cycle of DRL models such as the high-level decision-making system of SMAUG's maritime port autonomous surveillance system. Our experiments suggest that, while incurring additional computational costs that don´t directly improve drift detection precision, FADMON can add an explainability layer to the monitoring system while also supporting detection of changes in the underlying interpretation of the input data by the DRL model, monitoring the concept rather than the data. Future work could explore the proposed methodology on other feature attribution methods or interpretable-by-design models that generate feature attributions without additional operations. Other statistical metrics could also be tested. FADMON could also be tested in other image-based DNN models such as image classifiers, detectors or generators. Finally, our method could be tested on other high-risk scenarios where continuous explainable monitoring is beneficial or even mandatory. Overall, this work shows how the combination of drift detection and XAI can enhance model monitoring in scenarios where understanding why a model fails is as important as when.

Ethics and consent

Ethical approval and consent were not required.

Data availability

Underlying data

No source data were used for this article.

Software availability

The synthetic data used in the experiments of this article can be accessed at the following repository:

Zenodo: Generated synthetic dataset of pixel-level representations of maritime port surveillance vessels under multiple drift scenarios

https://doi.org/10.5281/zenodo.17793696[47]

This project contains the following underlying data:

- background_ag1.pkl (Background of training data provided to GradientSHAP)
- reference_data_ag1.pkl (Reference input data)
- test_data_ag1.pkl (Undrifted test input data)
- data_drift_opt1_ag1.pkl (Drifted input data from the obstacle density increase scenario, variant 1)
- data_drift_opt2_ag1.pkl (Drifted input data from the obstacle density increase scenario, variant 2)
- data_drift_opt3_ag1.pkl (Drifted input data from the obstacle density increase scenario, variant 3)
- data_prob_drift_opt1_ag1.pkl (Drifted input data from the vessel confinement scenario, variant 1)
- data_prob_drift_opt2_ag1.pkl (Drifted input data from the vessel confinement scenario, variant 2)
- data_prob_drift_opt3_ag1.pkl (Drifted input data from the vessel confinement scenario, variant 3)
- ref_list_probs_ag1.pkl (Reference probability vectors)
- test_list_probs_ag1.pkl (Undrifted test probability vectors)
- list_probs_datadrift_opt1_ag1.pkl (Drifted probability vectors from the obstacle density increase scenario, variant 1)
- list_probs_datadrift_opt2_ag1.pkl (Drifted probability vectors from the obstacle density increase scenario, variant 2)
- list_probs_datadrift_opt3_ag1.pkl (Drifted probability vectors from the obstacle density increase scenario, variant 3)
- list_probs_probdrift_opt1_ag1.pkl (Drifted probability vectors from the vessel confinement scenario, variant 1)
- list_probs_probdrift_opt2_ag1.pkl (Drifted probability vectors from the vessel confinement scenario, variant 2)
- list_probs_probdrift_opt3_ag1.pkl (Drifted probability vectors from the vessel confinement scenario, variant 3)

Acknowledgements

No further acknowledgements are stated.

Use of generative AI

Generative AI assistance was provided by ChatGPT (OpenAI, GPT-5.1) to help review and refine the manuscript text and to assist in generating and debugging code for the experiments. All outputs were checked and validated by the authors.

References

1. United Nations Conference on Trade and Development: **Review of maritime transport 2023**. 2023; [cited 2025 Nov 5].
Reference Source
2. European Monitoring Centre for Drugs and Drug Addiction: **European drug report 2022: trends and developments**. LU: Publications Office, 2022; [cited 2025 Nov 5].
Publisher Full Text
3. CORDIS | European Commission: **Coordination Of maritime assets for Persistent And Systematic Surveillance | COMPASS2020 | project | fact sheet | H2020**. [cited 2025 Nov 5].
Publisher Full Text
4. CORDIS | European Commission: **Risk-aware Automated Port Inspection Drone(s) | RAPID | project | fact sheet | H2020**. [cited 2025 Nov 5].
Publisher Full Text
5. CORDIS | European Commission: **Underwater security | UnderSec | project | fact sheet | HORIZON.** [cited 2025 Nov 5].
Publisher Full Text
6. CORDIS | European Commission: **Smart maritime and Underwater Guardian | SMAUG | project | fact sheet | HORIZON**. [cited 2025 Nov 5].
Publisher Full Text
7. Wu W, Gao C, Chen J, *et al.*: **Reinforcement learning in vision: a survey**. arXiv. 2025; [cited 2025 Nov 5].
Publisher Full Text
8. Gikay AA, Lau PL, Sengul C, *et al.*: **High-risk Artificial Intelligence systems under the European Union's AI Act: systemic flaws and practical challenges.** SSRN, 2023.
Publisher Full Text
9. Patchipala SG: **Tackling data and model drift in AI: strategies for maintaining accuracy during ML model inference.** *Int J Sci Res Arch.* 2023; **10**(2): 1198–1209.
Publisher Full Text
10. Testi M, Ballabio M, Frontoni E, *et al.*: **MLOps: a taxonomy and a methodology.** *IEEE Access.* 2022; **10**(3): 63606–63618.
Publisher Full Text
11. Irpan A, Rao K, Bousmalis K, *et al.*: **Off-Policy evaluation via Off-Policy classification**. In: *Advances in Neural Information Processing Systems.* Curran Associates, Inc, 2019; [cited 2025 Dec 2].
Publisher Full Text
12. Massey FJ: **The kolmogorov-smirnov test for goodness of fit.** *J Am Stat Assoc.* 1951; **46**(253): 68–78.
Publisher Full Text
13. Gretton A, Borgwardt KM, Rasch MJ, *et al.*: **A kernel two-sample test.** *J Mach Learn Res.* 2012; **13**(25): 723–773.
Reference Source
14. **Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)**. June 13, 2024.
Reference Source
15. Lundberg SM, Lee SI: **A unified approach to interpreting model predictions.** In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, *et al.*, editors, *Adv Neural Inf Process Syst*. 2017.
Publisher Full Text
16. Sundararajan M, Taly A, Yan Q: **Axiomatic attribution for deep networks**. In: *International Conference on Machine Learning.* PMLR, 2017; 3319–3328.
Publisher Full Text
17. Lee Y, Lee Y, Lee E, *et al.*: **Explainable Artificial Intelligence-based model drift detection applicable to unsupervised environments.** *Comput Mater Contin.* 2023; **76**(2): 1701–1719.
Publisher Full Text
18. Lee YE, Lee TJ: **A study on efficient ai model drift detection methods for MLOps.** *J Internet Comput Serv.* 2023; **24**(5): 17–27.
Publisher Full Text
19. Zimmermann B, Boussard M: **Improving drift detection by monitoring shapley loss values**. In: El Yacoubi M, Granger E, Yuen PC, Pal U, Vincent N, editors. *Pattern Recognition and Artificial Intelligence*. Cham: Springer International Publishing, 2022; 455–466.
Publisher Full Text
20. Mnih V, Kavukcuoglu K, Silver D, *et al.*: **Playing atari with deep reinforcement learning.** arXiv. 2013; [cited 2025 Dec 2].
Publisher Full Text
21. Mnih V, Badia AP, Mirza M, *et al.*: **Asynchronous methods for deep reinforcement learning.** arXiv. 2016; [cited 2025 Nov 10].
Publisher Full Text
22. Schulman J, Wolski F, Dhariwal P, *et al.*: **Proximal policy optimization algorithms.** arXiv. 2017; [cited 2025 Nov 10].
Publisher Full Text
23. Gamage C, Dinalankara R, Samarabandu J, *et al.*: **A comprehensive survey on the applications of machine learning techniques on maritime surveillance to detect abnormal maritime vessel behaviors.** *WMU J Marit Aff.* 2023; **22**(4): 447–477.
Publisher Full Text
24. Yun WJ, Park S, Kim J, *et al.*: **Cooperative multiagent deep reinforcement learning for reliable surveillance via autonomous multi-UAV control.** *IEEE Trans Ind Inform.* 2022; **18**(10): 7086–7096.
Publisher Full Text
25. Nandhini TJ, Thinakaran K: **Optimizing forensic investigation and security surveillance with deep reinforcement learning techniques.** In: *2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI).* Chennai, India: IEEE, 2023; 1–5.
Publisher Full Text
26. Zhao R, Li Y, Fan Y, *et al.*: **A survey on recent advancements in autonomous driving using deep reinforcement learning: applications, challenges, and solutions.** *IEEE Trans Intell Transp Syst.* 2024; **25**(12): 19365–19398.
Publisher Full Text
27. Vela D, Sharp A, Zhang R, *et al.*: **Temporal quality degradation in AI models.** *Sci Rep.* 2022; **12**(1): 11654.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
28. Hassija V, Chamola V, Mahapatra A, *et al.*: **Interpreting black-box models: a review on explainable Artificial Intelligence.** *Cogn Comput.* 2024; **16**(1): 45–74.
Publisher Full Text
29. Sharief F, Ijaz H, Shojafar M, *et al.*: **Multi-class imbalanced data handling with concept drift in fog computing: a taxonomy, review, and future directions.** *ACM Comput Surv.* 2025; **57**(1): 1–48.
Publisher Full Text
30. Anderson TW: **On the distribution of the two-sample Cramer-von mises criterion.** *Ann Math Stat.* 1962; **33**(3): 1148–1159.
Publisher Full Text
31. Agrahari S, Singh AK: **Concept drift detection in data stream mining : a literature review.** *J King Saud Univ Comput Inf Sci.* 2022; **34**(10): 9523–9540.
Publisher Full Text
32. Hovakimyan G, Bravo JM: **Evolving strategies in machine learning: a systematic review of concept drift detection.** *Information.* 2024; **15**(12): 786.
Publisher Full Text
33. Bodor A, Hnida M, Daoudi N: **Machine learning models monitoring in MLOps context: metrics and tools.** *Int J Interact Mob Technol.* 2023; **17**(23): 125–139.
Publisher Full Text
34. Böhle M, Singh N, Fritz M, *et al.*: **B-Cos alignment for inherently interpretable CNNs and vision transformers.** *IEEE Trans Pattern Anal Mach Intell.* 2024; **46**(6): 4504–4518.
PubMed Abstract | **Publisher Full Text**
35. Rodriguez AM, Unzueta L, Geradts Z, *et al.*: **Multi-task explainable quality networks for large-scale forensic facial recognition.** *IEEE J Sel Top Signal Process.* 2023; **17**(3): 612–623.
Publisher Full Text
36. Selvaraju RR, Cogswell M, Das A, *et al.*: **Grad-CAM: visual explanations from deep networks via gradient-based localization**. In: *2017 IEEE International Conference on Computer Vision (ICCV).* Venice: IEEE, 2017; 618–626.
Publisher Full Text
37. Bora RP, Terhorst P, Veldhuis R, *et al.*: **SLICE: Stabilized LIME for Consistent Explanations for Image Classification**. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* Seattle, WA, USA: IEEE, 2024; 10988–10996.
Publisher Full Text
38. Ibrahim R, Shafiq MO: **Explainable convolutional neural networks: a taxonomy, review, and future directions.** *ACM Comput Surv.* 2023; **55**(10): 1–37.
Publisher Full Text
39. Fantozzi P, Naldi M: **The explainability of transformers: current status and directions.** *Computers.* 2024; **13**(4): 92.
Publisher Full Text
40. Arrieta AB, Díaz-Rodríguez N, Ser J, *et al.*: **Explainable Artificial Intelligence (Xai): concepts, taxonomies, opportunities and challenges toward responsible Ai.** *Inf Fusion.* 2020; **58**: 82–115.
Publisher Full Text
41. Chen H, Lundberg SM, Lee SI: **Explaining a series of models by propagating shapley values.** *Nat Commun.* 2022; **13**(1): 4512.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
42. Chattopadhay A, Sarkar A, Howlader P, *et al.*: **Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks**. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV).* Lake Tahoe, NV: IEEE, 2018; 839–847.
Publisher Full Text
43. Muhammad MB, Yeasin M: **Eigen-CAM: class activation map using principal components.** In: *2020 International Joint Conference on Neural Networks (IJCNN).*

Glasgow, United Kingdom: IEEE, 2020; 1–7.
Publisher Full Text

44. Molnar C: **Interpretable machine learning: a guide for making black box models explainable.** 3rd edn. 2025.
Reference Source

45. Niu L, Li Z, Li S: **MMD fence GAN unsupervised anomaly detection model based on maximum mean discrepancy.** *Int J Cogn Inform Nat Intell.* 2024; **18**(1): 1–13.
Publisher Full Text

46. Rabanser S, Günnemann S, Lipton ZC: **Failing loudly: an empirical study of methods for detecting dataset shift.** In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems.* Red Hook, NY, USA: Curran Associates Inc., 2019; 1396–1408.
Publisher Full Text

47. Iriarte FJ: **Generated synthetic dataset of pixel-level representations of maritime port surveillance vessels under multiple drift scenarios.** [Data set]. Zenodo. 2025.
http://www.doi.org/10.5281/zenodo.17793696

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 1

Reviewer Report 30 January 2026

https://doi.org/10.21956/openreseurope.23936.r67977

✔ **Ayoola Babatunde Fadola**
Washington College of Law, American University, Washington, D.C, USA

What works well

1. Clear motivation: Monitoring "concepts" via how a policy interprets inputs is compelling in safety/security contexts and aligns with transparency expectations under the EU AI Act.
2. Method simplicity: A clean pipeline that's easy to integrate into CMM.
3. Deployment-oriented architecture: The three-stage framework (reference data, online monitoring with health reports, and human-in-the-loop) is practical.
4. Qualitative insights: Attribution maps provide actionable diagnostics beyond a binary "drift/no-signal.

Key issues to address

1. Resolve data-availability inconsistency: The section states "No source data were used," yet links to Zenodo datasets used in the experiments. Please amend to accurately reflect that all synthetic source data are available and point to the precise DOI and file mapping used for each experiment in Table 1.

Closing note

This is a promising and timely contribution. With the statistical reporting tightened, FADMON could become a practical building block for explainable monitoring of visual RL in high-risk, safety-critical deployments.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My research focuses on maritime law, operations, and autonomous shipping. I am able to adequately review the conversation as to the layout and operability.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 30 January 2026

https://doi.org/10.21956/openreseurope.23936.r67508

Adeola Oluwatoyin OSUNDIRAN
University of South Africa, Pretoria, Gauteng, South Africa

The research article is relevant. The problem statement was well articulated, alluding to the incessant smuggling of drugs to Europe. I would like to suggest that, inasmuch as the FADMON was highly recommended, it would have been great for the authors to highlight the challenges or applicability for other nations.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Area of Expertise is on Port Efficiency and Operations. Maritime Supply Chain; Gender; Sustainability and Green Ports and Corridors

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 08 January 2026

https://doi.org/10.21956/openreseurope.23936.r67510

Alexandru Pohontu
National University of Science and Technology Politehnica Bucharest, Bucharest, Romania

This paper introduces FADMON, an explainability-driven drift detector for image-based deep reinforcement learning (DRL) models, motivated by the need for continuous model monitoring (CMM) in safety/security contexts such as maritime port surveillance (SMAUG). Instead of monitoring only input images or output action probabilities, FADMON monitors feature attribution maps (e.g., GradientSHAP) to detect changes in how the model "interprets" its inputs, framed as monitoring the concept rather than only the data.

Are sufficient details of methods and analysis provided to allow replication by others? ***Partly***
It would be helpful to specify how MMD p-values were computed. They came from a permutation test or an asymptotic approximation? What were the key settings (e.g., number of permutations, random seeds)?
Also,the method description says the MMD value is compared to a predefined threshold to raise alerts, but the paper should specify how that threshold is chosen, whether it's fixed across environments, and how false positive rates are controlled.

Finally, the paper is clear, academically meaningful, provides supporting evidence for its conclusions, and shares source artifacts for the experiments.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I am studying the potential of Artificial Intelligence an Machine Learning in maritime surveillance. My focus is on implementing AI in various Maritime Domain Awareness applications, including anomaly detection, risk assessment, route predictions, and collision avoidance. I review state-of-the-art methods and research novel approaches to enhance conventional maritime surveillance systems.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.