Out of Distribution Detection in Self-adaptive Robots with AI-powered Digital Twins

Erblin Isaku*†, Hassan Sartaj*, Shaukat Ali*, Beatriz Sanguino‡, Tongtong Wang‡, Guoyuan Li‡,
Houxiang Zhang‡, and Thomas Peyrucain§
*Simula Research Laboratory, Oslo, Norway
{erblin, hassan, shaukat}@simula.no
†University of Oslo, Oslo, Norway
‡Norwegian University of Science and Technology, Ålesund, Norway
{beatriz.i.s.c.d.c.sanguino, tongtong.wang, guoyuan.li, hozh}@ntnu.no
§PAL Robotics, Barcelona, Spain
thomas.peyrucain@pal-robotics.com

***Abstract*—Self-adaptive robots (SARs) in complex, uncertain environments must proactively detect and address abnormal behaviors, including out-of-distribution (OOD) cases. To this end, digital twins offer a valuable solution for OOD detection. Thus, we present a digital twin-based approach for OOD detection (ODiSAR) in SARs. ODiSAR uses a Transformer-based digital twin to forecast SAR states and employs reconstruction error and Monte Carlo dropout for uncertainty quantification. By combining reconstruction error with predictive variance, the digital twin effectively detects OOD behaviors, even in previously unseen conditions. The digital twin also includes an explainability layer that links potential OOD to specific SAR states, offering insights for self-adaptation. We evaluated ODiSAR by creating digital twins of two industrial robots: one navigating an office environment, and another performing maritime ship navigation. In both cases, ODiSAR forecasts SAR behaviors (i.e., robot trajectories and vessel motion) and proactively detects OOD events. Our results showed that ODiSAR achieved high detection performance—up to 98% AUROC, 96% TNR@TPR95, and 95% F1-score—while providing interpretable insights to support self-adaptation.**

***Index Terms*—Self-Adaptive Robots, Digital Twins, Out-of-Distribution Detection, Uncertainty Quantification, Explainability**

I. Introduction

Autonomous robots are increasingly deployed in dynamic and unpredictable environments, where real-time adaptation to changes in environmental and internal states, as well as mission goals, is critical. This ability, called *self-adaptation*, is key to ensure their dependability and long-term autonomy [11]. Robots with this capability are known as Self-Adaptive Robots (SARs).

A key challenge in enabling self-adaptation is the timely detection of anomalous conditions. Among these, out-of-distribution (OOD), i.e., cases where data deviates from the training distribution [14, 22], are critical. While initially studied in image and text classification, OOD is also important in robotics, where data-driven models are often used for perception, planning, and control. OOD cases may arise from scenarios related to sensor faults, actuator drift, or environmental changes on which the robot wasn't trained [4, 9]. Thus, early OOD detection is essential for dependable robot behavior in dynamic environments.

Although well-studied in machine learning and vision tasks [14, 24, 30, 38], OOD still has gaps in the SAR context. (1) Most OOD approaches are reactive, detecting anomalies only after they occur. In SARs, proactive detection, i.e., forecasting future SAR states and identifying OOD before they occur, is critical for timely planning and safe adaptation [18]. (2) OOD approaches often give binary outputs without explaining why a sample is anomalous, limiting trust and hindering integration with self-adaptation frameworks, e.g., MAPLE-K [20], which may require reasoning about the cause before triggering adaptation. (3) Most approaches rely on forecasting error, reconstruction error, or uncertainty, whereas combining these can yield robust OOD detection, as emphasized in [14, 24]. (4) Though increasingly used in robotics, Digital Twins (DTs) often serve as passive simulators. Few works explore their role in adaptive decision-making [7, 25], while almost no work is on interpretable and proactive OOD detection using DTs.

To address these gaps, we present ODiSAR, a DT-based OOD detection approach integrated with the MAPLE-K self-adaptive architecture [20]. ODiSAR combines sequence-to-sequence forecasting, reconstruction-based error, and predictive uncertainty via Monte Carlo Dropout for proactive OOD detection of anomalous future states. It also supports state-level attribution, identifying which SAR states are most associated with anomalous behavior for interpretability.

We evaluate ODiSAR on two industrial SAR case studies from the European `RoboSAPIENS` project [2]. The first, from NTNU's industrial partner—Kongsberg Maritime AS[1]—a Norwegian company, involves autonomous vessels navigating at sea and using motion prediction to support onboard decision-making for safe and efficient operation. The second case from PAL Robotics[2] involves indoor robots autonomously navigating dynamic environments subject to blocked paths, changing layouts, and unexpected human presence. In both, ODiSAR

[1] https://www.kongsberg.com/maritime/
[2] https://pal-robotics.com/

2643-1572/25/$31.00 ©2025 IEEE
DOI 10.1109/ASE63991.2025.00281

forecasts system behaviors (i.e., robot trajectories and vessel motion states) and proactively detects OOD cases.

Our results show that ODiSAR consistently outperforms a baseline (i.e., RMSE) that relies on forecasting error as the only indicator for OOD detection, achieving up to 98% AUROC, 96% TNR@TPR95, and 95% OOD F1-score for maritime vessel, and 96%, 94%, and 89%, respectively, for mobile robot—while also providing confidence-aware predictions and interpretable attributions at the SAR state level. In most maritime vessel scenarios, the majority of predictions fall into either the IND Confident (59–62%) or OOD Uncertain (33–39%) categories. In contrast, the mobile robot case study shows an almost equal split between IND Confident (48.75%) and OOD Confident (51.25%), indicating potential overconfidence. These differences highlight an opportunity for future work on balancing model confidence through uncertainty regularization or calibration techniques.

II. Application Context and Industrial Case Studies

This work is part of a European project, RoboSAPIENS [2], whose objective is to build autonomous SARs that can operate in unknown environments and adapt their behavior in response to unforeseen situations. To this end, RoboSAPIENS proposes MAPLE-K architecture [20] for implementing self-adaptations in robots, also shown in Figure 1, which is an extension of the classical MAPE-K architecture [5]. In MAPLE-K, the **M**onitor component continuously monitors the state of the robot as well as its environment and analyzes it using methods implemented in the **A**nalyze component. If the analysis results reveal that adaptation is necessary (e.g., an anomaly is detected), the **P**lan component generates a set of adaptation plans to be executed and selects the most suitable one. The **L**egitimate component, newly introduced in MAPLE-K, verifies whether the adaptation is safe to execute. Once this verification is passed, the adaptation is executed by the **E**xecute component. The **K**nowledge component collects relevant information from all the MAPLE-K components.

The dynamic nature of SARs, driven by various changes such as structural (e.g., adding or upgrading sensors), functional (e.g., updating navigation algorithms), and environmental (e.g., diverse weather and human-robot interactions), introduces significant complexity in implementing self-adaptation and can cause the adaptation space to expand drastically [20]. This presents a critical challenge for industry practitioners to effectively monitor and analyze the behavior of SAR under continuously changing conditions. To address this, we propose a digital twin-based approach that enables real-time detection and adaptive response to OOD cases under varying unpredictable operational contexts.

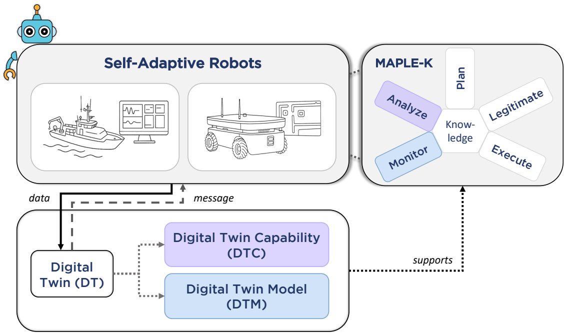


Figure 1: ODiSAR in the context of the MAPLE-K loop—implemented inside self-adaptive robots. The DT, composed of DTM and DTC components, supports the Monitor and Analyze phases (shown in light colors), respectively, enabling interpretable OOD detection and adaptation within the MAPLE-K framework.

A digital twin, in our case, consists of two components (see Figure 1): a digital twin model (DTM), which is a replica of a SAR, and a digital twin capability (DTC), which is an additional feature provided by the digital twin (e.g., OOD) that uses the DTM along with the most up-to-date state of the robot to perform the capability. Once the DTC detects unexpected behavior (e.g., a potential OOD likely to occur in the near future, in our context), it sends a signal to the SAR to initiate adaptation planning using the *Plan* component in MAPLE-K. Essentially, the digital twin in our setup supports two components of MAPLE-K implemented in SARs: continuous monitoring of the SARs' states with the *Monitor* component and analyzing them to signal any potential OOD with the *Analyze* component.

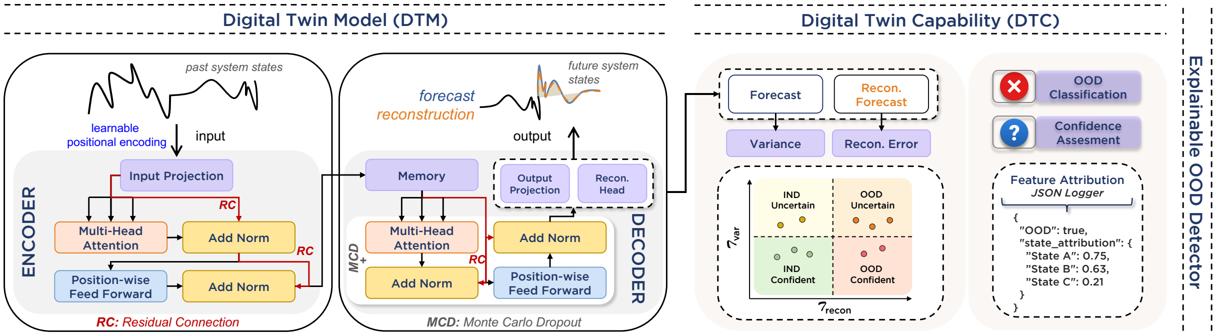
In the context of this paper, we evaluate ODiSAR using two industrial SAR case studies from RoboSAPIENS. The first case study focuses on autonomous vessels (AVs) provided by Kongsberg Maritime that is a global leader in building maritime applications. In this case, we explore how AVs' digital twins can enable them to detect OOD while navigating at sea. The second case study is provided by PAL Robotics (Spain), a world leader in building service robots. Here, we demonstrate the application of ODiSAR with a robot operating in an office environment. Its digital twin helps OOD detection.

III. Approach

Figure 2 presents a high-level overview of ODiSAR, our proactive and interpretable OOD detection approach based on the principles of the MAPLE-K loop for enabling self-adaptation. The approach is composed of two key components: (i) the digital twin model (DTM) responsible for forecasting the expected robot behavior, and (ii) the digital twin capability (DTC) responsible for detecting and interpreting deviations from the normal operational behavior. The full implementation of the proposed approach is publicly available on GitHub.[3]

ODiSAR takes a sequence of past system states as input (e.g., sensor readings and control commands) and creates the DTM consisting of an encoder and decoder to forecast future system states while simultaneously reconstructing them. The encoder processes the input using self-attention mechanisms

[3] https://github.com/Simula-COMPLEX/ODiSAR



```
{
 "OOD": true,
 "state_attribution": {
   "State A": 0.75,
   "State B": 0.63,
   "State C": 0.21
 }
}
```

Figure 2: Overview of the proposed Digital Twin-based approach (ODiSAR) for proactive OOD detection. The Transformer-based Digital Twin Model (DTM) takes historical input and produces both forecasted and reconstructed future system states. The Digital Twin Capability (DTC) analyzes these outputs using reconstruction error and uncertainty (via MC Dropout) to detect potential OOD instances. The Explainable OOD Detector combines this analysis to flag OOD states and attributes them to the most contributing system features.

and positional encoding to capture temporal dependencies. This encoded representation is passed to a decoder, which generates both forecasted outputs and their reconstructions through two parallel heads. To estimate predictive uncertainty, Monte Carlo (MC) Dropout is applied at inference, producing multiple stochastic forward passes. The resulting forecasts are post-processed by the DTC, which computes the reconstruction error and the forecast variance. These two scores are compared against learned thresholds to classify each prediction as either in-distribution or out-of-distribution, and as confident or uncertain. Finally, for predictions flagged as OOD, the approach provides a state-level attribution, identifying the most contributing states and exporting this information in a JSON format. This resulting information enables domain experts to continuously monitor and analyze abnormal behaviors of SARs and take proactive measures to ensure their dependability.

A. SAR Data and Model Settings

To determine the most suitable architecture for the DTM, we conducted an exploratory pilot study comparing several time series forecasting models, including recurrent neural networks (RNNs), variational autoencoders (VAEs), generative adversarial networks (GANs), and transformer neural networks. Based on preliminary training stability and forecasting results, the transformer model consistently outperformed the other models. Therefore, we adopted a vanilla transformer as the base model for the DTM and extended it with a reconstruction head to support proactive OOD detection.

The input and output features are selected based on domain-specific knowledge and task, specifically motion and trajectory prediction. For the autonomous maritime vessel case study, we forecast ship dynamics, including surge/sway velocity, yaw rate, and roll angle/rate. For the autonomous mobile robot case, we forecast the robot's future position states. To optimize both model architecture and training performance, we used the Optuna framework [3] for automated hyperparameter tuning. Table I presents the final configurations derived from the results of the pilot experiment for each case study.

Table I: ODiSAR: Model and Training Parameters

Parameter	Maritime Vessel	Mobile Robot
Model Dimension (d_{model})	64	64
Number of Attention Heads	4	4
Feedforward Dimension	128	128
Dropout Rate	0.1	0.2
Batch Size	16	16
Learning Rate	0.0001	0.0001
Epochs	200	200

B. Creating DTM

The purpose of DTM is to learn the normal behavior of the target system through time series forecasting, thereby supporting the *Monitor* phase in the MAPLE-K control model by generating future predictions of key system variables (e.g., yaw rate, velocity, or position) and enabling continuous observation of expected behavior. Therefore, we formulate this forecasting task as a sequence-to-sequence prediction problem. Given a sequence of past observations $\mathbf{X}_{t-w+1:t} = [\mathbf{x}_{t-w+1}, \ldots, \mathbf{x}_t]$ where t denotes the current time step, the goal is to predict the corresponding future sequence of system states $\hat{\mathbf{Y}}_{t+1:t+h} = [\hat{\mathbf{y}}_{t+1}, \ldots, \hat{\mathbf{y}}_{t+h}]$, where w is the input window size and h is the forecast horizon.

We design the DTM using a Transformer-based architecture composed of an encoder and decoder with multi-head attention and feedforward layers. The model learns how the system behaves/maneuvers over time by analyzing patterns in past data and uses this understanding to predict future states. In addition to forecasting, the model is trained to reconstruct its own forecasted outputs. This is achieved via a lightweight multi-layer perceptron (MLP) reconstruction head attached to the decoder output. By reconstructing what the model has just predicted, it learns to form a robust internal representation of what "normal" predictions look like. This representation is later used by the DTC to proactively detect abnormal patterns during deployment, based on how well the model reconstructs its own state predictions.

As part of training, the model minimizes two objectives. The first is a *forecasting loss* that compares the predicted future sequence $\hat{\mathbf{Y}}_{t+1:t+h}$ to the ground truth future values $\mathbf{Y}_{t+1:t+h}$. This loss is typically computed using Mean Squared Error (MSE):

$$\mathcal{L}_{\text{Forecast}} = \frac{1}{h}\sum_{i=1}^{h} \|\hat{\mathbf{y}}_{t+i} - \mathbf{y}_{t+i}\|^2 \quad (1)$$

The second is a *reconstruction loss* between the forecasted outputs and their reconstructed versions $\tilde{\mathbf{Y}}_{t+1:t+h} = [\tilde{\mathbf{y}}_{t+1}, \ldots, \tilde{\mathbf{y}}_{t+h}]$. The reconstruction loss, similar to forecasting, is computed as:

$$\mathcal{L}_{\text{Recon}} = \frac{1}{h}\sum_{i=1}^{h} \|\tilde{\mathbf{y}}_{t+i} - \hat{\mathbf{y}}_{t+i}\|^2 \quad (2)$$

The overall training objective combines both the forecasting and reconstruction losses. Specifically, the total loss is calculated using:

$$\mathcal{L}_{\text{Total}} = \underbrace{\mathcal{L}_{\text{Forecast}}}_{\text{Forecasting Loss}} + \underbrace{\mathcal{L}_{\text{Recon}}}_{\text{Reconstruction Loss}} \quad (3)$$

This training setup enables the DTM to learn to forecast future system states and reconstruct those jointly. The dual objective ensures that the model develops a robust internal representation of normal system behavior, which is needed for supporting proactive OOD detection at deployment time.

C. Creating DTC

The DTC enables proactive OOD detection by analyzing the forecasted outputs of the DTM. It includes two key metrics—reconstruction error (Eq. (2)) and predictive uncertainty (Eq. (4)), i.e., variance—to identify deviations from expected behavior. At inference time, the DTC evaluates how well the model can reconstruct its own predictions. A high reconstruction error indicates that the forecasted sequence lies outside the distribution the model has learned, suggesting potential OOD behavior. In addition, we estimate predictive uncertainty using MC Dropout [12]. During inference, we perform multiple stochastic forward passes through the DTM with dropout enabled. Each pass produces a slightly different forecast due to the randomness introduced by dropout. By computing the variance across these sampled forecasts using Equation (4), we capture the model's epistemic uncertainty. In Equation (4), $\hat{\mathbf{y}}_{t+i}^{(n)}$ denotes the forecasted output at time step $t+i$ from the n^{th} stochastic forward pass, and $\bar{\mathbf{y}}_{t+i}$ is the mean of these forecasts over N passes.

$$\text{Var}(\hat{\mathbf{y}}_{t+i}) = \frac{1}{N}\sum_{n=1}^{N} \left(\hat{\mathbf{y}}_{t+i}^{(n)} - \bar{\mathbf{y}}_{t+i}\right)^2 \quad (4)$$

To support proactive detection, the DTC defines two thresholds in Equation (5) based on statistics from in-distribution validation data. Here, μ and σ are the mean and standard deviation of the reconstruction errors and variances, and k is a tunable sensitivity parameter (e.g., $k = 3$ for a 3-sigma threshold [29]).

$$\tau_{\text{recon}} = \mu_{\text{recon}} + k \cdot \sigma_{\text{recon}}, \quad \tau_{\text{var}} = \mu_{\text{var}} + k \cdot \sigma_{\text{var}} \quad (5)$$

Each forecast window—i.e., the entire sequence of predicted future values over the forecast horizon—is then assigned to one of four categories (illustrated inside DTC block in Figure 2) based on these classification categories: (i) *IND and Confident:* Low reconstruction error, low uncertainty (ii) *IND and Uncertain:* Low reconstruction error, high uncertainty (iii) *OOD and Uncertain:* High reconstruction error, high uncertainty (iv) *OOD and Confident:* High reconstruction error, low uncertainty.

This categorization supports interpretable, confidence-aware, and actionable detection of anomalous behavior by not only flagging potential OOD events but also indicating the ODiSAR's level of certainty. For instance, uncertain cases can undergo further monitoring or human review, while confident OOD detections can trigger immediate adaptive responses. Similarly, uncertain IND cases might cause warrant caution or continued observation.

D. Explainable OOD Detection

To complement detection with interpretability, we enhance the DTC with an explainable mechanism that identifies which system features contributed most to the OOD classification. This provides actionable insight during deployment and supports human operators in understanding anomalous behavior.

For each forecast window flagged as OOD (i.e., when the reconstruction error exceeds the threshold), we compute the feature-wise root mean squared error (RMSE) between the model's forecast $\hat{\mathbf{y}}$ and its reconstruction $\tilde{\mathbf{y}}$ using Equation (6). Here, h is the forecast horizon length, and $\mathbf{e}_{\text{feat}} \in \mathbb{R}^d$ is a vector of reconstruction errors for each of the d output features.

$$e_{\text{feat}}^{(i)} = \sqrt{\frac{1}{h}\sum_{j=1}^{h} \left(\tilde{y}_j^{(i)} - \hat{y}_j^{(i)}\right)^2}, \quad \text{for } i = 1, \ldots, d \quad (6)$$

We then rank the features (i.e., states) by their RMSE values and report the top three as an interpretable attribution of which system states showed the strongest anomalous responses. These attributions are saved in a structured JSON file alongside metadata such as the sequence index, start/end time steps, reconstruction error, uncertainty variance, and assigned OOD category.

Each forecast window is categorized into one of four semantic quadrants (e.g., IND Confident or OOD Uncertain) based on its reconstruction error and uncertainty level. For instance, in the example shown in Figure 3, the window is labeled as `"red"`, corresponding to the *OOD & Confident* region in the quadrant plot. The `state_attribution` field highlights the top three states—*Surge Speed*, *Sway Speed*, and

Yaw Rate—that show the highest reconstruction errors and were most impacted by the underlying anomaly.

```
{
  "sequence_index": 3,
  "start_time_step": 420,
  "end_time_step": 479,
  "is_OOD": true,
  "reconstruction_error": 0.17066404223442078,
  "uncertainty_variance": 0.018417222425341606,
  "recon_exceeds_threshold": true,
  "uncertainty_exceeds_threshold": false,
  "category": "red",
  "state_attribution": {
    "Surge Speed": 0.26233699917793274,
    "Sway Speed": 0.21531985700130463,
    "Yaw Rate": 0.13875150680541992
  }
}
```

Figure 3: Example of structured JSON output for a forecast window (autonomous maritime vessel case study) flagged as OOD. The top-3 contributing states are provided under `state_attribution`.

While our approach does not follow formal explainability frameworks (e.g., SHAP [23] or LIME [31]), it aims to improve practical interpretability by highlighting the most affected system states showing potential anomalous behavior. This lightweight diagnostic capability is designed to assist human operators and practitioners in tracing the root of behavioral deviations in SARs.

IV. Experiment Design

A. Simulation Environment & Robotic System

Autonomous Maritime Vessel: The data used in this study was collected using a professional navigation bridge simulator, called K-Sim Navigation, manufactured by Kongsberg Maritime AS [1]. This simulator is highly accurate and realistic, allows for the simulation of a wide range of environmental conditions, and is widely used for training nautical students and professional captains. For our experiments, we used a Ro-Ro ferry (Ferry Basto Fosen) model equipped with two azimuth thrusters – one at the bow and one at the stern. During simulations, thrusters operated at a constant speed of 206 RPM with a fixed propeller pitch of 80%, and only the thruster angles were adjusted to perform maneuvers.

Autonomous Mobile Robot: We used the open-source PAL Robotics OMNI Base Simulation environment [28], which integrates with ROS 2 and Gazebo to provide a realistic 3D simulation of the TIAGo OMNI Base robot [35]. This robot features omnidirectional 3-DOF planar motion (x, y, θ) and is equipped with two LIDAR sensors that provide a full 360º field of view for real-time obstacle detection and autonomous indoor navigation tasks. It is widely used in research and industry applications such as office automation, healthcare, hospitality, and logistics, due to its precise navigation and versatile sensing capabilities. In our study, the simulation was conducted within PAL's office map environment, where the robot performed waypoint-based navigation.

B. Maneuvers and Disturbance Scenarios

Autonomous Maritime Vessel: A variety of maneuvers were conducted and recorded over a 20-minute period for each maneuver, with data sampled at a frequency of 1 Hz. As shown in Table II, these maneuvers included Zigzag, Turning, and Random patterns, each with several variants. To simulate OOD conditions, environmental disturbances such as wind, waves, and currents were introduced. The recorded data captures key navigational variables, including surge and sway velocities, yaw rate, roll dynamics, and environmental factors like wind, waves, and currents.

To evaluate the ODiSAR's ability to detect different types of OOD events, three environmental disturbance cases were designed. In each case, disturbances were introduced gradually between minutes 7 and 14 to simulate realistic transitions, designated as the OOD event. In Case 1 and Case 3, the initial 7 minutes had no disturbances. Starting from minute 7, wind (e.g., 18 knots from 45° with gusts and directional variation), waves, and current (e.g., 6 knots from 205°) were applied, with smooth transitions in direction (90°/min) and speed (35–40 knots/min). After minute 14, only wind gusts (Case 1) or no disturbances (Case 3) remained. In Case 2, the simulation began with light wind and waves, intensified mid-way to stronger wind and current, and returned to initial conditions after minute 14.

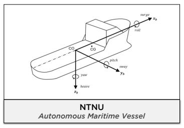
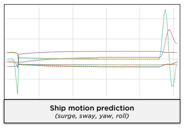
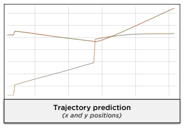
Autonomous Mobile Robot: The simulated robot followed a series of waypoint-based navigation tasks within the PAL Robotics office environment using the ROS 2 Nav2 stack. For each task, five spatially separated waypoints were sampled from the map, ensuring minimum clearance from obstacles and coverage across the environment. The robot navigated sequentially to each waypoint while maintaining 3-DOF planar motion.

To simulate sensor degradation and evaluate how well ODiSAR performs under such situations, noise was added to the robot's odometry after reaching the first waypoint. Specifically, Gaussian noise with zero mean was applied to the position, orientation, and velocity values in the $/mobile\_base\_controller/odom$ topic. This noisy period lasted for 80 seconds and is treated as the OOD event. During the entire navigation task, odometry data were recorded at 10 Hz.

C. Evaluation Metrics

To evaluate the performance of ODiSAR, we use three commonly adopted metrics in OOD detection [10, 15]: the Area Under the Receiver Operating Characteristic (AUROC), the True Negative Rate at 95% True Positive Rate (TNR@TPR95), and the F1-score. These metrics allow us to assess both the general separability of in-distribution (IND) vs. OOD samples and the model's behavior at specific decision thresholds.

Table II: Overview of the experimental setup across both cases studies. Maneuver types marked with † indicate cases where disturbances were added to generate OOD data for inference and model evaluation.

Use Case	Task	Disturbance/OOD Source	Maneuver Type	Variants
 surge x_b roll CO CG pitch sway y_b yaw heave z_b NTNU Autonomous Maritime Vessel		Environmental disturbance (wind, waves, currents)	Zigzag†	10°, 15°, 20°, 30°
			Random†	1, low, high
			Turning†	10°, 15°, 20°, 30°
		Sensor noise ($/odom$ topic)	Waypoint†	–

AUROC: AUROC evaluates the model's ability to distinguish IND from OOD samples across all thresholds, measuring the trade-off between true positive rate (TPR) and false positive rate (FPR):

$$\text{AUROC} = \int_0^1 \text{TPR}(\alpha)\, d(\text{FPR}(\alpha)), \tag{7}$$

where α is the threshold. A value closer to 1.0 indicates better separability.

TNR@TPR95: This metric evaluates how well the model reduces false positives while maintaining a TPR of 95%. The threshold α^* is computed as:

$$\alpha^* = \arg\min_{\alpha} \left|\text{TPR}(\alpha) - 0.95\right|, \tag{8}$$

$$\text{TNR@TPR95} = 1 - \text{FPR}(\alpha^*). \tag{9}$$

F1-score: The F1-score balances precision and recall, and is defined as:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

where:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

A high F1-score indicates that the model performs well in detecting OOD instances while minimizing false positives and false negatives.

D. Research Questions

Our experiment investigates the following research questions (RQs).

RQ1 *How effective is ODiSAR in detecting out-of-distribution behavior under varying disturbance sources?*

RQ2 *How confident is ODiSAR in its predictions under different operational conditions?*

RQ3 *How does ODiSAR compare with a forecasting-error-only OOD detection method?*

RQ1 assesses whether ODiSAR can reliably detect out-of-distribution (OOD) behavior under varying disturbance sources, both environmental (e.g., wind, waves, and currents) and sensor-based (e.g., odometry topic). While the autonomous mobile robot case lacks data diversity, we conduct a small-scale robustness study in the autonomous maritime vessel case by training on one representative scenario (i.e., random maneuvering) and evaluating across others.

RQ2 explores how confidently ODiSAR operates under different conditions. Although the model doesn't produce explicit confidence scores, we estimate predictive uncertainty using Monte Carlo Dropout by measuring the variance of forecasted outputs. This enables us to assess the reliability of predictions and support more risk-aware OOD detection.

RQ3 compares ODiSAR with a baseline that uses only forecasting error (RMSE) for OOD detection. This helps determine whether integrating forecasting, reconstruction, and uncertainty estimation leads to measurable gains over simpler, threshold-based alternative.

V. Results and Analyses

a) RQ1 - Effectiveness: Table III presents the OOD detection performance of ODiSAR across two case studies. For the Autonomous Maritime Vessel, the evaluation includes various maneuver types such as Zigzag, Turning, and Random, each tested under multiple disturbance scenarios. For the Autonomous Mobile Robot, performance is assessed in a Waypoint-based indoor navigation setting under sensor noise perturbations.

When trained on all maneuvers, ODiSAR achieves strong and consistent results across both OOD cases and maneuver types. AUROC scores remain above 97%, OOD F1-scores range from 93% to 95%, and TNR@TPR95 stays above 94%. These results highlight the model's effectiveness across diverse and unseen disturbance patterns. The overall combined performance (AUROC 97.53%, OOD F1-score 94%, TNR@TPR95 95.1%) further confirms its effectiveness under varied test conditions.

To assess how well ODiSAR generalizes from a limited but diverse training set, we also consider a setup where the model is trained solely on random maneuvers, which include a mix of dynamic behaviors without following a fixed pattern. Even under this constraint, the model maintains high detection performance, with AUROC values consistently above 97% and

Table III: OOD detection performance across training setups and evaluation scenarios. The results illustrate the effectiveness of ODiSAR (RQ1) and its comparison with an RMSE-based baseline (RQ3), using AUROC, TNR@TPR95, and F1-scores for IND and OOD samples.

Trained On	**Evaluated On**		**ODiSAR**				**RMSE Baseline**			
	Category	Scenario	AUROC	TNR@TPR95	F1-Score		AUROC	TNR@TPR95	F1-Score	
					IND	OOD			IND	OOD
All Maneuvers	*OOD Case*	Case 1	97.88%	95.55%	96%	93%	94.42%	64.19%	43%	59%
		Case 2	98.62%	96.73%	96%	95%	97%	92.2%	50%	66%
		Case 3	97.42%	94.95%	97%	94%	94.51%	93.34%	46%	62%
	Maneuver	Zigzag	98.45%	96.38%	97%	95%	96.55%	94.91%	88%	85%
		Random	97.83%	96.15%	97%	94%	95.62%	91.4%	57%	65%
		Turning	97.55%	95.11%	97%	95%	96.6%	93.81%	10%	55%
	Overall	All data combined	97.53%	95.1%	96%	94%	95.43%	92.69%	54%	64%
Random	*OOD Case*	Case 1	98.57%	95.55%	94%	90%	95.37%	85.99%	30%	56%
		Case 2	98.33%	93.68%	96%	94%	97.1%	92.2%	11%	58%
		Case 3	97.45%	91.9%	95%	92%	93.82%	93.34%	16%	55%
	Maneuver	Zigzag	98.25%	94.91%	97%	95%	94.15%	93.58%	29%	60%
		Random	98.73%	92.71%	92%	89%	95.91%	91.4%	0%	53%
		Turning	98.02%	93.94%	95%	93%	96.82%	93.81%	12%	55%
	Overall	All data combined	97.96%	92.86%	94%	90%	94.23%	92.54%	22%	56%
Waypoint	*Maneuver*	Waypoint	96.52%	94.4%	91%	89%	96.56%	88.04%	84%	84%

OOD F1-scores between 89% and 94%, and TNR@TPR95 ranging from 91.9% to 94.9%. This demonstrates that ODiSAR can generalize well to previously unseen maneuver types, despite a narrower training distribution.

For the waypoint navigation scenario (PAL Robotics), training and evaluation are both limited to a single maneuver type. Here, ODiSAR achieves an AUROC of 96.52%, TNR@TPR95 of 94.4%, and an OOD F1-score of 89%. While generalization cannot be assessed in this setting, the results confirm reliable performance within the scenario.

Overall, the results indicate that ODiSAR is highly effective in detecting OOD behavior under various disturbances. Generalization is strongest when training data covers a diverse set of maneuvers, but the approach still shows strong performance even under more constrained training setups.

> ODiSAR achieves high OOD detection performance across varied disturbance sources, with AUROC scores above 97%, TNR@TPR95 up to 96%, and OOD F1-scores up to 95% when trained on diverse maneuvers. It maintains robust performance even under limited training diversity. In the PAL Robotics use case, results remain solid, though generalizability is less conclusive.

b) RQ2 - Confidence: To address RQ2, we examine how confidently the model behaves under varying operational conditions. Using predictive variance estimated via MC Dropout, we categorize model outputs into four quadrants based on whether reconstruction error and forecast variance exceed their respective thresholds: IND Confident, IND Uncertain, OOD Uncertain, and OOD Confident. This categorization provides deeper insights into the reliability and consistency of the model's predictions across scenarios.

Figure 4 presents the distribution of these confidence categories across several scenarios. The autonomous maritime vessel case study includes Cases 1–3 (as OOD sources), Zigzag, Random, and Turning (as maneuver types), while the Waypoint scenario originates from the autonomous mobile robot case.

Across the vessel scenarios, most predictions fall into the IND Confident category ($\geq 57\%$), aligning with expectations, as the model is trained on these in-distribution data. In contrast, predictions labeled as OOD are predominantly uncertain, with OOD Uncertain rates ranging between 33-39%, while OOD Confident predictions remain consistently low ($< 5\%$). This pattern highlights a key insight: uncertainty is effective at flagging a lack of confidence but insufficient on its own for reliable OOD detection. In most OOD cases, the model correctly flags anomalies but expresses low certainty – showing that reconstruction error remains the more decisive indicator. Therefore, uncertainty is best used as a supportive metric, not a standalone criterion.

The Waypoint scenario (autonomous mobile robot) shows a contrasting trend – 51% of predictions are OOD Confident, and only 48.75% are IND Confident, with no OOD Uncertain or IND Uncertain cases. While this suggests high certainty across predictions, results from RQ1 reveal lower OOD detection performance for mobile robot data, implying that the DT model may be overconfident. This calls for a deeper investigation into the model's calibration and generalization under limited training diversity and unseen operational settings, highlighting

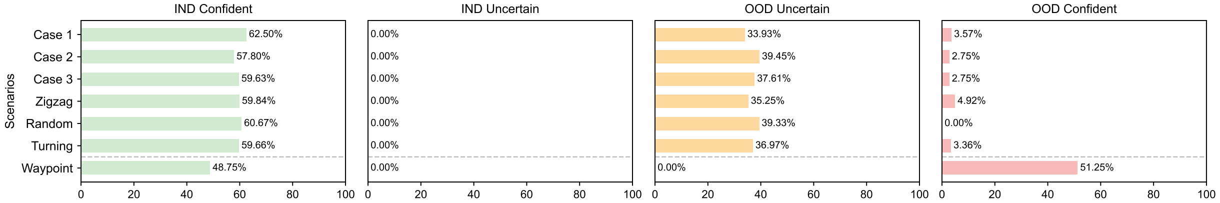


Figure 4: Distribution of model confidence across different operational scenarios. Each bar reflects the proportion of predictions falling into one of four categories—IND Confident, IND Uncertain, OOD Uncertain, and OOD Confident—based on reconstruction error and predictive variance. The $Waypoint$ scenario is from the autonomous mobile robot case study, while the remaining scenarios originate from the autonomous maritime vessel case. This highlights how the model's confidence varies with scenario type and detection outcome.

a potential direction for future work, such as uncertainty regularization or calibration methods.

> ODiSAR shows high confidence in in-distribution data, as expected given its training exposure. Uncertainty helps flag unreliable predictions but is insufficient alone for OOD detection, reinforcing the need for reconstruction error as the main indicator.

c) RQ3 - Baseline Comparison: We compare ODiSAR against a baseline that relies solely on forecasting error (RMSE) for OOD detection, across multiple scenarios and training configurations (see Table III).

At a high level, ODiSAR consistently outperforms the baseline across all evaluation dimensions. While the RMSE-based method shows relatively high AUROC scores (often >90%), its threshold-based metrics (TNR@TPR95 and F1-Score) are significantly lower and highly inconsistent – particularly for OOD detection. For example, under the "All Maneuvers" training setup, the RMSE baseline achieves an AUROC of 95.43% but an OOD F1-score of only 64%. This contrast highlights a key limitation of relying on forecasting error alone for OOD detection.

The discrepancy arises from how each metric operates. While AUROC reflects the ODiSAR's ability to discriminate between IND and OOD instances across all possible thresholds, metrics like F1-score and TNR@TPR95 evaluate performance at a fixed threshold. As such, AUROC represents the best case scenario assuming an optimal cutoff, whereas F1-score and TNR@TPR95 indicate how well the model performs with the actual decision boundary in use, being more representative of practical performance.

We notice that the RMSE baseline is highly sensitive to the choice of threshold, and its performance drops significantly in scenarios where RMSE values of IND and OOD samples are not well separated, or when there is class imbalance. For instance, in the "Random" training setup, OOD F1-score drops to 56% overall and even lower in specific cases.

In contrast, ODiSAR achieves consistently high performance across all metrics, with OOD F1-scores above 89% and TNR@TPR95 above 91% in all setups. This highlights its robustness, particularly in handling varied disturbances and training conditions.

Overall, while the RMSE-based baseline may appear promising in terms of AUROC, its poor calibration and reliance on error magnitudes make it unreliable without extensive threshold tuning. ODiSAR, by combining reconstruction and uncertainty thresholds, delivers both strong separability and practical detection reliability across all test conditions.

> ODiSAR consistently offers high performance across metrics and scenarios, in contrast to the forecasting-only baseline, which shows unstable results due to its sensitivity to threshold selection and class imbalance. This demonstrates the value of combining reconstruction error with uncertainty for robust and reliable OOD detection.

VI. Threats to Validity

This section discusses potential threats that may affect the validity of our results and findings.

Internal Validity. A potential internal validity threat can be due to the selection of hyperparameters. Small variations in these parameters—such as learning rate or dropout rate—can significantly affect forecasting accuracy and OOD detection outcomes. To mitigate this, we conducted a pilot study using automated hyperparameter optimization (i.e., Optuna), allowing us to systematically explore parameter space. Final configurations were selected based on consisten performance across validation sets.

External Validity. The generalizability of our findings is subject to constraints associated with our evaluation datasets. The datasets used were obtained from specific simulation scenarios involving maritime vessels and mobile robot systems. Although these datasets realistically represent self-adaptive robot systems, the observed results may vary when the method is applied to other domains or operational environments. Future research should investigate the applicability of our approach across broader contexts to strengthen external validity.

Construct Validity. Construct validity concerns the extent to which our metrics accurately measure proactive out-of-distribution detection capability and interpretability. We addressed this by selecting widely recognized and established metrics such as AUROC, F1-score, and TNRTPR95. However, alternative metrics or additional qualitative assessments might offer complementary insights into system performance and interpretability.

Conclusion Validity. Our conclusions are supported by consistent results across two diverse robotic systems and a range of OOD scenarios. We applied the same methodology and architecture with minimal tailoring to each case, and we report multiple performance metrics to capture different aspects of model effectiveness. Nevertheless, repeated evaluations, additional datasets, and anomaly scenarios could further validate and enhance the reliability of our results.

VII. Discussion and Lessons Learned

Our evaluation across two diverse robotic systems—an autonomous maritime vessel (NTNU) and an autonomous mobile robot (PAL Robotics)— demonstrates that ODiSAR offers an effective and robust approach for proactive OOD detection. Based on our analysis, we present the following key lessons learned.

a) Relevance to self-adaptive control loops: ODiSAR supports key components of self-adaptive robotic architectures, such as MAPE-K [5], AWARE [32], and MAPLE-K [20] in our context. Specifically, it enhances the Monitor phase through continuous forecasting and state reconstruction, and Analyze phase via proactive and interpretable OOD detection. Furthermore, by providing confidence-aware predictions, the proposed approach contributes to building actionable Knowledge, enabling informed adaptation decisions. For example, in the autonomous maritime vessel case study, we observed that roll-related states (i.e., roll angle and roll rate) were among the most frequently identified contributors to OOD classifications. This aligns well with the nature of the environmental disturbances, which include wind, waves, and currents—factors known to significantly affect a vessel's roll dynamics. In this regard, ODiSAR provides system-specific insights that can inform adaptation logic, such as adjusting controls or switching to disturbance-aware navigation strategies.

These capabilities demonstrate the potential of the ODiSAR to be integrated into feedback loops, improving resilience and environmental awareness in SARs. Future work will explore how ODiSAR can contribute to the other phases (e.g., Legitimate) of such control loops.

b) SAR applicability and generalizability: The results show that ODiSAR maintains strong performance across a range of operational conditions, including different maneuver types (zigzag, turning, random) and varying environmental disturbances (wind, waves, current) in the autonomous vessel setting, as well as indoor navigation with sensor degradation in the mobile robot setting. Notably, this was achieved without modifying the model architecture or tailoring it to each scenario. While hyperparameter tuning was performed, the resulting configurations were nearly identical across both case studies, indicating consistent behavior across applications. Despite the differences in dynamics, sensing modalities, and operating conditions, ODiSAR maintained high OOD detection performance. These findings highlight its broad applicability and strong potential for generalization to other robotic systems with minimal adaptation effort.

c) Confidence-aware interpretation: Our analysis revealed that the ODiSAR's confidence estimates—based on predictive variance from MC Dropout—provide meaningful insights into its behavior that can potentially inform decision-making in SARs. For example, in the autonomous mobile robot case, a final confidence-aware output might indicate and be interpreted as: *"In the following minute (from timestep 1200 to 1260), the ODiSAR is confident that the robot will exhibit abnormal behavior affecting mostly the x and y position states."* Similarly, in the maritime vessel case, the model may highlight roll rate, roll angle, and sway speed as the most affected states under uncertain conditions. These interpretable outputs offer domain experts a clearer understanding of how abnormal behavior correlates with potential OOD sources, such as environmental disturbances, enabling faster and more informed responses.

Based on the results, in the autonomous maritime vessel case, most OOD predictions were flagged as uncertain, which helps identify situations where the model is less reliable. However, in the autonomous mobile robot, predictions indicated potential overconfidence under limited training diversity. These findings suggest that while uncertainty can support interpretability, it shall be complemented by a more robust detection indicator (i.e., reconstruction error) and possibly improved with better calibration techniques (e.g., Bayesian inference). In the future, we plan to explore how different dropout rates affect the confidence estimates and their alignment with actual prediction reliability.

d) Challenges and insights from analyzing real robot data: In our preliminary experiments with the PAL Robotics case, we explored the use of real robot data. We collaborated with practitioners to collect data that represent both normal (IND) and OOD behaviors. Although collecting IND data was similar to the simulator-based setup, capturing OOD scenarios in real-world environments was challenging. Unlike simulation, it was difficult—and potentially unsafe—to introduce sensor or actuator noise on a physical robot, unless using faulty components, which requires strict safety precautions. An alternative is to create OOD scenarios through environmental disturbances for autonomous mobile robots, such as operating a robot on sandy or oily surfaces that affect robot motion, or introducing challenging objects like transparent glass obstructions. We noticed that these scenarios often involve multimodal data, such as combining camera images and lidar data. Given that our approach relies on a transformer-based architecture, extending it to support multimodal OOD detection is a promising direction for future research.

VIII. Related Works

We compare our approach to related work across four key aspects: the use of DTs for anomaly detection, proactive and confidence-aware OOD detection, embedded explainability, and conceptual advances over our prior work.

Digital Twins in Self-Adaptive Robots: Recent studies have applied DTs to support runtime monitoring, fault diagnosis, and control adaptation in self-adaptive robotic systems, including autonomous ships (Hasan et al. [13]), unmanned aerial vehicles (Song et al. [33]), self-driving cars (Xiong et al. [36]), and mobile robots (Betzer et al. [8]). These works demonstrate the potential of DTs for enhancing situational awareness and operational safety; however, these studies target specific DT applications in isolation—e.g., monitoring or fault diagnosis/detection—whereas our approach leverages a transformer-based model for both runtime behavior monitoring and OOD detection in one proactive and interpretable DT framework. This unified approach supports not only monitoring but also analysis of SAR behavior.

Digital Twins and Anomaly Detection: Recent advances in DT applications span various domains, including industrial wireless systems (Moharam et al. [26]), power-grid infrastructures (Idrisov et al. [16]), offshore wind turbines (Stadtmann and Rasheed [34]), and cyber-physical system frameworks enhanced by curriculum learning (Xu et al. [37]). These approaches have significantly improved real-time anomaly detection by enabling continuous system monitoring and fault diagnosis. However, the use of DTs for proactive out-of-distribution detection—particularly in the context of self-adaptive robots—remains largely unexplored. Our work addresses this gap by introducing a DT-based approach that anticipates anomalous behavior through forecasting and reconstruction, thus extending the role of DTs beyond monitoring toward early detection and analysis of unseen or unexpected system states.

Proactive and Confidence-Aware OOD Detection: Traditional OOD detection approaches typically operate post-hoc, flagging anomalies only after they have occurred (Hendrycks and Gimpel [14], Lee et al. [21]). In contrast, proactive methodologies leverage predictive uncertainty (Gal and Ghahramani [12]) to anticipate potential anomalies before they manifest. Our method advances this research by combining uncertainty quantification with reconstruction-based metrics, enabling proactive and confidence-aware OOD detection. In addition, we employ a Transformer-based DT, which improves the forecasting and detection capability of the system over traditional sequence models [24]. This integration allows the DT framework to detect deviations early while assessing its confidence, ultimately supporting more reliable monitoring and analysis in SAR.

Explainability in Machine Learning: Interpretability is increasingly critical, especially for safety-critical applications, facilitating human trust and intervention (Montavon et al. [27]; Arrieta et al. [6]). Widely used post-hoc explanation methods such as SHAP (Lundberg and Lee [23]) and LIME (Ribeiro et al. [31]) provide model-agnostic insights but are often computationally expensive and disconnected from the core detection pipeline. While our work does not propose a novel explainability technique in this regard, it incorporates a lightweight attribution mechanism directly within the proposed digital twin framework. This embedded mechanism identifies the most influential system states contributing to each OOD detection decision, thereby providing timely, context-specific interpretations and supporting more informed decision-making in SARs.

Relation to Prior Work: Our earlier work [17] also proposed a DT-based framework for OOD detection in maritime systems. While conceptually similar, the current study significantly advances that line of research in several ways: (i) it replaces the previous dual-model setup (RNN forecaster + autoencoder) with a single unified transformer model trained on a dual objective, simplifying integration and runtime execution; (ii) it extends applicability to two industrial case studies (autonomous maritime vessel and autonomous mobile robot), highlighting generalization across domains with diverse dynamics; (iii) it enhances interpretability by providing both confidence-aware and state attributions of OOD events, improving transparency for domain experts; and (iv) it aligns closely with the MAPLE-K loop by extending support from the *Monitor* phase (in prior work) to also include the *Analyze* phase.

IX. Conclusions and Future Work

This paper presented ODiSAR, a proactive out-of-distribution detection approach that integrates forecasting, reconstruction, and uncertainty estimation for self-adaptive robots. ODiSAR demonstrated high effectiveness across two real-world case studies, specifically autonomous maritime vessels (NTNU) and autonomous mobile robots (PAL Robotics). ODiSAR achieved up to 98% AUROC, 96% TNR@TPR95, and 95% OOD F1-score in the autonomous vessel case, and 96% AUROC, 94% TNR@TPR95, and 89% OOD F1-score in the mobile robot. The comparison results showed that ODiSAR outperforms the RMSE baseline using forecast error alone, highlighting the added value of combining reconstruction and uncertainty in detecting OOD events.

Given the results and lessons learned from this study, we identify the following directions for future work. (i) Extend the approach to support multimodal data fusion by including additional sensor inputs (e.g., LIDAR, camera), enabling richer representation of robotic behavior and context-aware OOD detection. (ii) Investigate uncertainty calibration techniques (e.g., Bayesian calibration [19]) and conduct empirical evaluations on how factors such as dropout rate affect the reliability and interpretability of confidence estimates. (iii) Apply the proposed method to additional self-adaptive robots to further reinforce its generalizability and scalability across domains. (iv) Conduct a direct empirical comparison with our prior DT-based approach under identical settings—e.g., same datasets, feature sets, and evaluation metrics—to better understand their relative performance and trade-offs.

Acknowledgments

This work is supported by the RoboSAPIENS project funded by the European Commission's Horizon Europe programme under grant agreement number 101133807.

References

[1] "K-Sim® Navigation," https://www.kongsberg.com/maritime/products/simulation/k-sim-navigation/, [Online; accessed 15-September-2025].

[2] "RoboSAPIENS," https://robosapiens-eu.tech/, [Online; accessed 25-March-2025].

[3] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[4] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[5] P. Arcaini, E. Riccobene, and P. Scandurra, "Modeling and analyzing MAPE-K feedback loops for self-adaptation," in *Proceedings of the 10th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, ser. SEAMS '15. IEEE Press, 2015, pp. 13–23.

[6] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[7] M. S. Azari, S. Santini, F. Edrisi, and F. Flammini, "Self-adaptive fault diagnosis for unseen working conditions based on digital twins and domain generalization," *Reliability Engineering & System Safety*, vol. 254, p. 110560, 2025.

[8] J. S. Betzer, J. Boudjadar, M. Frasheri, and P. Talasila, "Digital twin enabled runtime verification for autonomous mobile robots under uncertainty," in *2024 28th International Symposium on Distributed Simulation and Real Time Applications (DS-RT)*. IEEE, 2024, pp. 10–17.

[9] F. Cai and X. Koutsoukos, "Real-time out-of-distribution detection in learning-enabled cyber-physical systems," in *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 2020, pp. 174–183.

[10] T. Che, X. Liu, Y. Ge, R. Zhang, C. Xiong, and Y. Bengio, "Deep verifier networks: Verification of deep discriminative models with deep generative models," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 8, 2021, pp. 7002–7010.

[11] R. De Lemos, H. Giese, H. A. Müller, M. Shaw, J. Andersson, M. Litoiu, B. Schmerl, G. Tamura, N. M. Villegas, T. Vogel *et al.*, "Software engineering for self-adaptive systems: A second research roadmap," in *Software Engineering for Self-Adaptive Systems II: International Seminar, Dagstuhl Castle, Germany, October 24-29, 2010 Revised Selected and Invited Papers*. Springer, 2013, pp. 1–32.

[12] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[13] A. Hasan, T. Asfihani, O. Osen, and R. T. Bye, "Leveraging digital twins for fault diagnosis in autonomous ships," *Ocean Engineering*, vol. 292, p. 116546, 2024.

[14] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[15] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 951–10 960.

[16] I. N. Idrisov, D. Okeke, A. Albaseer, M. Abdallah, and F. M. Ibanez, "Leveraging digital twin and machine learning techniques for anomaly detection in power electronics dominated grid," *arXiv preprint arXiv:2501.13474*, 2025.

[17] E. Isaku, H. Sartaj, and S. Ali, "Digital twin-based out-of-distribution detection in autonomous vessels," *arXiv preprint arXiv:2504.19816*, 2025.

[18] T. Ji, A. N. Sivakumar, G. Chowdhary, and K. Driggs-Campbell, "Proactive anomaly detection for robot navigation with multi-sensor fusion," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4975–4982, 2022.

[19] V. Kuleshov, N. Fenner, and S. Ermon, "Accurate uncertainties for deep learning using calibrated regression," in *International conference on machine learning*. PMLR, 2018, pp. 2796–2804.

[20] P. G. Larsen, S. Ali, R. Behrens, A. Cavalcanti, C. Gomes, G. Li, P. De Meulenaere, M. L. Olsen, N. Passalis, T. Peyrucain *et al.*, "Robotic safe adaptation in unprecedented situations: the robosapiens project," *Research Directions: Cyber-Physical Systems*, vol. 2, p. e4, 2024.

[21] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, 2018.

[22] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint arXiv:1706.02690*, 2017.

[23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[24] M. Ma, L. Han, and C. Zhou, "Research and application of transformer based anomaly detection model: A literature review," *arXiv preprint arXiv:2402.08975*, 2024.

[25] F. Mo, H. U. Rehman, J. C. Chaplin, D. Sanderson, and S. Ratchev, "Digital twin-based self-learning decision-making framework for industrial robots in manufacturing," *The International Journal of Advanced Manufacturing Technology*, pp. 1–20, 2025.

[26] M. H. Moharam, O. Hany, A. Hany, A. Mahmoud, M. Mohamed, and S. Saeed, "Anomaly detection using machine learning and adopted digital twin concepts in radio environments," *Scientific Reports*, vol. 15, no. 1, p. 18352, 2025.

[27] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K. Müller, "Explainable ai: interpreting, explaining and visualizing deep learning," *Spring er LNCS*, vol. 11700, no. 1, 2019.

[28] PAL Robotics, "Tiago omni base ros 2 simulation," https://github.com/pal-robotics/omni_base_simulation, [Online; accessed 02-April-2025].

[29] F. Pukelsheim, "The three sigma rule," *The American Statistician*, vol. 48, no. 2, pp. 88–91, 1994.

[30] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," *Advances in neural information processing systems*, vol. 32, 2019.

[31] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[32] B. P. Sanwouo, P. Temple, and C. Quinton, "Breaking the loop: Aware is the new mape-k," in *FSE'25-International Conference on the Foundations of Software Engineering*, 2025.

[33] J. Song, D. Wang, Z. Chen, and K. Zhao, "Digital twin system for vtol uav fault diagnosis based on px4," in *International Conference on Autonomous Unmanned Systems*. Springer, 2022, pp. 2389–2401.

[34] F. Stadtmann and A. Rasheed, "Diagnostic digital twin for anomaly detection in floating offshore wind energy," in *International Conference on Offshore Mechanics and Arctic Engineering*, vol. 87851. American Society of Mechanical Engineers, 2024, p. V007T09A033.

[35] TIAGo OMNI Base, "Tiago omni base robot," https://wiki.ros.org/Robots/TIAGo-OMNI-base, [Online; accessed 28-April-2025].

[36] H. Xiong, Z. Wang, G. Wu, and Y. Pan, "Design and implementation of digital twin-assisted simulation method for autonomous vehicle in car-following scenario," *Journal of Sensors*, vol. 2022, no. 1, p. 4879490, 2022.

[37] Q. Xu, S. Ali, and T. Yue, "Digital twin-based anomaly detection with curriculum learning in cyber-physical systems," *ACM Transactions on Software Engineering and Methodology*, vol. 32, no. 5, pp. 1–32, 2023.

[38] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5635–5662, 2024.