Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p26183-cd

Quantitative Risk Assessment for Maritime Autonomous Navigation using AI Perception Metrics and Bayesian Networks

Simon Kupjetz
System Reliability, Fraunhofer Institute for Structural Durability and System Reliability LBF, Germany.
E-mail: simon.matthias.kupjetz@lbf.fraunhofer.de

Dr. Jürgen Nuffer
System Reliability, Fraunhofer LBF, Germany. E-mail: juergen.nuffer@lbf.fraunhofer.de

Jonathan Millitzer
System Dynamics, Fraunhofer LBF, Germany. E-mail: jonathan.millitzer@lbf.fraunhofer.de

Nico Jungbauer
Research & Technology, TKMS ATLAS ELEKTRONIK GmbH, Germany.
E-mail: nico.jungbauer@tkmsgroup.com

Dr. Jan Isermann
System Safety Management, TKMS ATLAS ELEKTRONIK GmbH, Germany.
E-mail: jan.isermann@tkmsgroup.com

Claas Rostock
Simulation Technologies, DNV SE, Germany. E-mail: claas.rostock@dnv.com

Daliya Issac
Artificial Intelligence for Smart Sensors and Actuators, Deggendorf Institute of Technology, Germany.
E-mail: daliyaissac@gmail.com

As maritime navigation increasingly shifts towards autonomy, ensuring the safety of AI-enabled perception systems under complex and uncertain conditions is a critical challenge. Traditional safety analysis methods often fail to capture the variability and uncertainty introduced by AI components in maritime environments. In this study, we investigate whether system-level risk for autonomous maritime navigation can be quantified by embedding object-classification performance metrics into a Failure Mode and Effects Analysis (FMEA)-linked Bayesian Network (BN). This network allows us to estimate the probability of failures, and to describe risk using these calculated probabilities together with a scenario-based model of severity. The novel framework integrates empirically derived AI performance data, specifically confusion-matrix probabilities from a computer-vision model, directly into the conditional probability tables of a Bayesian Network for Maritime Autonomous Surface Ships (MASS). In addition to internal AI components, the framework also models external factors, such as adverse weather conditions, and evaluates the impact of misclassifications on navigational safety. Initial results confirm the feasibility of the integrated framework and its capability to propagate AI misclassification probabilities through the Bayesian Network to derive a system-level risk consistent with MIL-STD-882E. By parameterizing the network using confusion-matrix information and reusing Bayesian structures across scenarios, the approach provides, for the first time in a maritime context, a transparent and scalable basis for risk-based assurance of AI-enabled maritime systems. Our main goal is to provide developers with confidence that architectural and design choices lead to systems that are free from unacceptable risks by quantifying uncertainties and allowing for uncertainty-aware safety claims, thereby yielding evidence that these claims hold. The proposed method bridges qualitative and quantitative safety perspectives by combining structured FMEA reasoning with Bayesian probabilistic inference. In future work, we will validate the approach in autonomous maritime trials to establish a robust foundation for assessing AI-enabled navigation safety.

Keywords: Probabilistic risk assessment, AI perception systems, Maritime autonomous navigation, Maritime Autonomous Surface Ships (MASS), Bayesian Networks

1. Introduction

The sector of Maritime Autonomous Surface Ships (MASS) continues to grow steadily. Mallam et al. (2024) provides a comprehensive review and highlights that these technologies are receiving increasing attention and are expected to transform future maritime transport operations as automation levels increase. To leverage the potential of rising automation in practice, though, operators must keep operations free from unacceptable risk, and developers must apply compliant Artificial Intelligence (AI) software development practices.

Bolbot et al. (2025) reviews safety acceptance criteria for MASS in the maritime sector, distinguishing three types: safety equivalence, the As Low As Reasonably Practicable (ALARP) principle, and equivalence to well-proven regulations, practices, and functions. Autonomous and remotely operated vessels are emerging in a regulatory landscape where rules from the International Maritime Organization (IMO) for conventional, crewed maritime operation do not yet address these new concepts. In the absence of mature, AI-specific safety regulations for MASS, researchers and regulators naturally look to established system safety standards from other safety-critical domains as a reference framework. The MIL-STD-882E (U.S. Department of Defense, 2012) does not provide guidance on conducting risk analyses for safety-critical systems under the influence of AI. A revised edition (U.S. Department of Defense, 2021) that explicitly considers higher levels of automation currently exists only as a draft (version F).

Rule-based approaches to autonomy pose major methodological, technological, and regulatory challenges across industries, *including* maritime. Consequently, classification societies and regulators rely on risk-based guidelines and trial frameworks to ensure safety levels at least equivalent to traditional operations. The Autonomous and Remotely Operated Ships (AROS) Additional Class Notations (DNV, 2024b) and the class guideline DNV-CG-0264 (DNV, 2024a) provide a goal- and risk-based framework for qualifying autonomous and remotely operated vessels. In the absence of specific MASS rules and standards, authorities use alternative design approaches, and differing flag-state interpretations can lead to varying safety levels. To enhance harmonization, European Maritime Safety Agency (2024) developed a Risk Based Assessment Tool (RBAT) and MASS-specific software to support the analysis and approval of preliminary MASS designs.

Under the EU AI Act (European Union, 2024), providers of high-risk AI systems must establish and continuously maintain a documented risk management system throughout the product lifecycle that systematically identifies known and reasonably foreseeable risks to safety according to *Article 9 – Risk management system*. *Article 10 – data and data governance* requires an assessment of the availability, quantity, and suitability of datasets. For AI systems such as a MASS, *Article 13 – Transparency and Provision of Information to Deployers* stipulates that providers must ensure that system operation remains sufficiently transparent to enable deployers to interpret the system's output and use it appropriately. The risk management system is not a one-time assessment but rather operates as a continuous iterative process that providers plan and run throughout the entire lifecycle of the high-risk AI system and that requires regular systematic review and updating.

The ISO/PAS 8800 standard (ISO, 2024), titled "Road Vehicles—Safety and Artificial Intelligence," addresses the safe integration of AI into automotive safety functions. It focuses on systematic errors, hardware faults, and performance insufficiencies of AI elements (e.g., bias, insufficient data quality) and introduces an AI safety lifecycle that complements ISO 26262 (Functional Safety) (ISO, 2018) and ISO 21448 (Safety of Intended Functionality) (ISO, 2022). The standard sets additional AI- and ML-specific requirements for safety-critical applications and provides industry guidance on using AI in safety-related functions for road vehicles. It also highlights safety analysis techniques suitable for AI systems, including established methods such as Fault Tree Analysis (FTA) and Failure Mode and Effects Analysis (FMEA), as well as newer methods such as System-Theoretic Process Analysis (STPA) and

Bayesian Networks. Bayesian Networks are probabilistic graphical models where nodes represent random variables and directed edges represent conditional dependencies (Ben-Gal, 2008). Adee et al. (2021) illustrates how they can model perception performance limiting conditions in automated driving.

Rauschenbach (2017), Kulkarni et al. (2021) and Rastayesh et al. (2019) demonstrate how to link FMEA and Bayesian Networks. Kabir and Papadopoulos (2019) give an overview of applications of Bayesian Networks in safety.

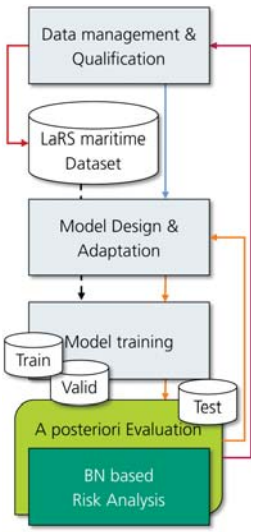

Fig. 1. Workflow of the BN-based maritime risk analysis based on (MLEAP Consortium, 2024).

To the best of our knowledge, no established method currently integrates such AI performance data into a system-level BN that is linked to FMEA and yields risk measures compatible with standards such as MIL-STD-882E.

2. Methodology

To develop a risk-based methodology for analyzing AI object classification performance, we first defined a functional system architecture for a Maritime Autonomous Surface Ship tasked with sailing independently to a predetermined target location. For collision and grounding avoidance and Convention on the International Regulations for Preventing Collisions at Sea (COLREGs)-compliant behavior, the ship uses nautical charts, Global Navigation Satellite System (GNSS), a stereoscopic camera for object detection with depth, and Radio Detection and Ranging (RADAR) for adverse weather. The system combines position, image, and radar data through sensor fusion (not detailed in this paper) to enable mutual plausibility checks and quantitatively *reduce* uncertainty in environment recognition and object detection, thereby improving safety.

To demonstrate how the methodology works, this paper focuses on one part of the signal processing chain, namely the computer-aided object classification of a camera module.

The adapted framework excerpted from the final Machine Learning Application Approval (MLEAP) report (MLEAP Consortium, 2024), as shown in Figure 1, comprises data management and qualification, construction of the Lakes, Rivers and Seas (LaRS) maritime dataset, model design and adaptation, and model training using training, validation, and test subsets. We performed an a posteriori evaluation to derive Bayesian Network-based risk analysis. Arrows indicate data flow and feedback between stages.

To simulate object classification in the maritime domain, we used the LaRS dataset to train a computer vision model. LaRS is a large panoptic maritime obstacle detection dataset with over 4,000 manually labeled frames containing water, sky, static obstacles, and dynamic obstacle categories such as boat, row boat, paddle board, buoy, swimmer, animal, float, plus an open-world other class (Žust et al., 2023).

Based on this dataset, we trained a real-time object detection model from the You Only Look Once (YOLO) family, which frames detection as a single-pass regression problem to jointly optimize localization and classification with low latency (Redmon et al., 2016). Common metrics and statistical tests for machine learning are summarized in Rainio et al. (2024).

We evaluate multi-class classification performance using a confusion matrix with counts of True Positives (TP), True Negatives (TN), False Negatives (FN), and False Positives (FP).

Table 1 shows a row-normalized confusion matrix generated from the LaRS dataset using a YOLOv8 model trained and validated on this dataset. The bold diagonal values would be 1.00 in the case of ideal object classification.

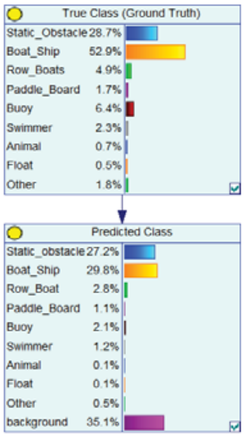

Fig. 2. Bayesian Network with occurrence rate and classification rate from confusion matrix.

Objects present in the ground truth but not detected by the model are represented as background in the predicted class of the confusion matrix. For static objects (e.g. shores, piers), the model correctly classified 95 % of cases, misclassified 1 % as boat/ship, and missed 4 % (background). For dynamic objects, performance was notably worse: buoys were correctly recognized in only 32 % of test cases, with 2 % misclassified as swimmers, 2 % as other, and 65 % not classified (background).

The normalized confusion-matrix values populate the conditional probability table of a Bayesian Network (see Figure 2) to visualize the computer vision performance metrics. The networks were created and visualized using GeNIe Modeler from BayesFusion (Druzdzel, 1999). The entries represent the relative frequencies of the eight true classes in the data and their corresponding detection rates (predicted classes). To reflect the susceptibility of optical image recognition to disturbances (see Figure 3), we model impairments due to restricted visibility (weather, sea state, ship

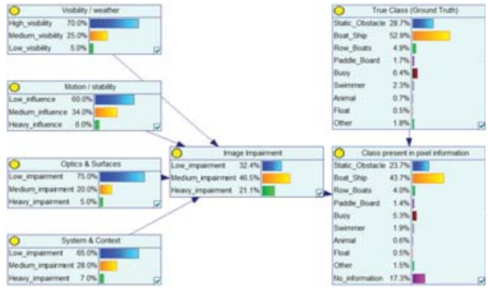

Fig. 3. Effect of image impairment on True Class (pixel information).

motion), optical effects (lens condensation, water reflections), and system/context factors such as insufficient light or domain shift in the training data.

The influence categories and percentage values are illustrative. Realistic values must be obtained via experiments or simulations and will depend on the operational domain (e.g. sea state). Severe impairments may even prevent the sensor from capturing usable image information. To assess object-recognition reliability, we must therefore not only quantify classification probabilities but also estimate the severity of resulting misclassifications.

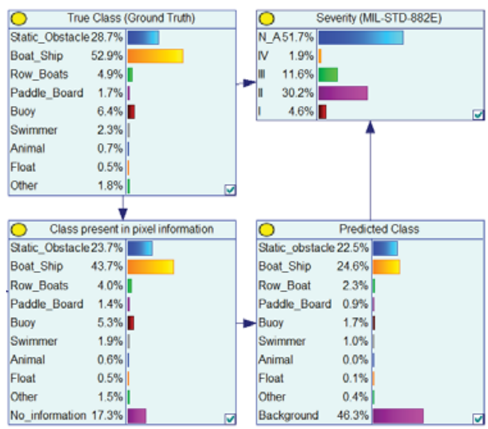

Fig. 4. Severity of misclassification modeled with a Bayesian Network.

Figure 4 illustrates how a Bayesian Network models the probability and severity of misclas-

Table 1. Row-normalized confusion matrix (condensed). Values indicate the fraction of predictions per class.

True ↓ Predicted →	Static	Boat/Ship	Buoy	Swimmer	Other	Background
Static	**0.95**	–	–	–	–	0.13
Boat/ship	0.01	**0.54**	–	0.05	0.06	0.70
Buoy	–	–	**0.32**	0.03	0.01	0.04
Swimmer	–	–	0.02	**0.52**	–	0.01
Other	–	0.01	0.02	0.05	**0.38**	0.12
Background	0.04	0.44	0.65	0.35	0.54	–

Note: Relatively rare classes (row boats, paddle board, animal, float, other) are aggregated into the “Other” category.

sifications in a vision-based object classifier on a MASS. The bottom-left node, *class present in pixel information*, indicates which object class (boat, buoy, swimmer, etc.) is actually visible in the image or whether effectively *no information* is available, allowing the BN to distinguish visibility issues (occlusion, limited field of view, adverse weather) from pure classifier errors. An exemplary misclassification severity matrix assigns MIL-STD-882E severity categories (I–IV) to perception errors, from Category I (catastrophic) and Category II (critical) to Category III (marginal) and Category IV (negligible), reflecting the worst credible consequence under the current design and mitigations. Correct classifications are denoted as not applicable (N/A). For example, misclassifying a buoy as a boat mainly causes inefficient avoidance, whereas misclassifying a swimmer as background can lead to an undetected collision risk and is therefore assigned Severity Category I (catastrophic).

3. Results

The Bayesian Network supports several types of analysis. It can quantify and visualize the expected reliability of image recognition based on training and validation data.

Fig. 6 shows how a more comprehensive overview of risk modeling can be generated. In addition, given suitable data, we can use it to simulate how deteriorating weather conditions may influence object classification. The MIL-STD-882E severity category shown in the fictitious example indicates that 51.7 % of classifications operate as intended (N/A). Analogous to the classification

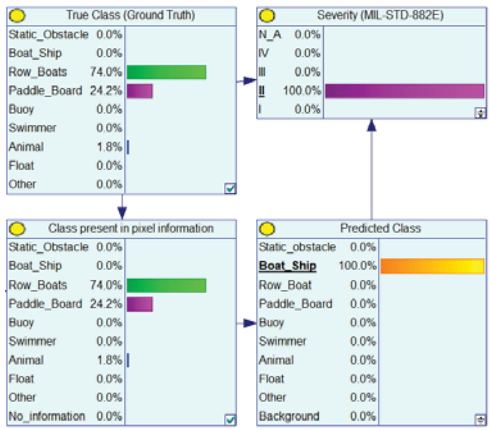

Fig. 5. Severity of a misclassification of a ship.

of severity according to MIL-STD-882E from I–IV, we can also use the classification of severity according to an FMEA from 1–10.

To analyze hypothetical observational scenarios, we condition the Bayesian Network on specified observed variables. Specifically, we instantiate the corresponding nodes with their hypothesized values, and perform probabilistic inference to obtain the posterior distribution over all remaining variables (see Fig. 5). Here, we manually set that a severity level II (critical) was present and that the system had predicted a boat/ship. In cases where there was no impact on visibility (image impairment = low), the labeled data actually showed a row boat (74.0 %), a paddle boat (24.2 %), or an animal (1.8 %).

Table 2 shows how, for example, we can implement an improvement in the true positive rate.

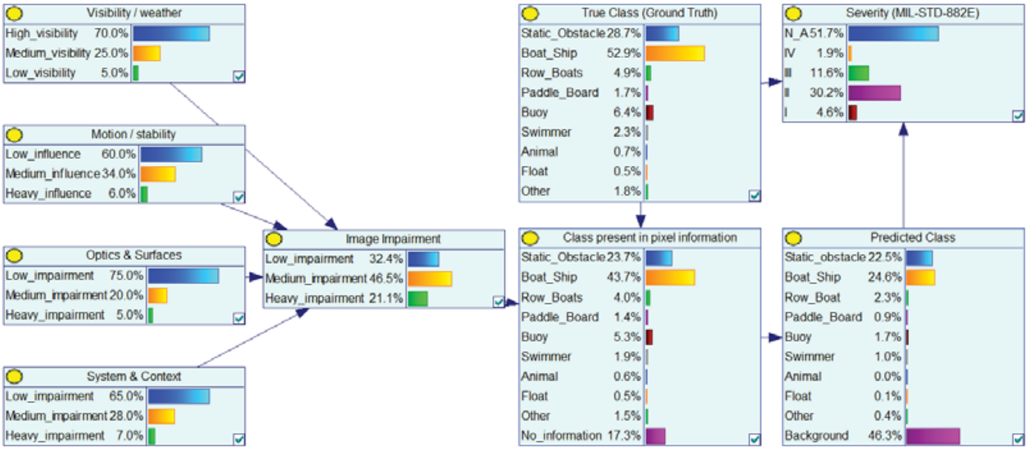

Fig. 6. Bayesian Network for safety modeling of the object classification.

Table 2. Modification of the confusion matrix to simulate an improvement in the rate of a correct classification of a boat/ship.

True ↓ Predicted →	Baseline	Improved
Boat/Ship	54 %	64 %
Other	1 %	1 %
Background / not class.	45 %	35 %

In this example, the rate of correct detection for a boat/ship has increased from 54 % to 64 % while the rate of false detection as background has decreased from 45 % to 35 %.

Table 3. Effect on the severity distribution for a misclassification.

Severity	Baseline	Improved
N/A	**62.5 %**	**67.8 %**
IV	1.9 %	1.9 %
III	6.6 %	6.6 %
II	**25.3 %**	**20.0 %**
I	3.7 %	3.7 %

Table 3 shows the effect of the change in classification rates on the distribution of severity. The rate for a critical severity decreases from 25.3 % to 20.0 %.

By interactively manipulating probability values and conditioning on observed variables, analysts can quickly test and visualize various what-if scenarios (e.g. best-case and worst-case). This step does not require in-depth domain knowledge and allows analysts to quickly generate and communicate insights to various stakeholders. With reasonable effort, the analysis can, for instance, show which parts of the AI system are involved in a potential safety goal violation, and to what extent.

Developers can use the methodology throughout the development process to identify weaknesses in the system at various stages of the software development cycle. For example, it can help teams ensure that they select a sufficiently large and representative set of data to achieve reliable object classification. Once the team determines the number of objects to classify, they can update the Bayesian Network as soon as they generate a new confusion matrix, enabling risk assessment throughout the entire lifecycle.

4. Conclusion and future work

In this work, we present a framework for quantifying system-level risk in autonomous maritime navigation by embedding AI perception performance into an FMEA-linked Bayesian Network. Parameterizing the network with confusion-matrix data from an object-classification model and a scenario-based severity model allows us to capture how misclassifications and environmental factors propagate through the architecture to affect navigational safety. The explicit mapping from FMEA entries into the Bayesian Network makes decision pathways visible, as it traces how specific perception outcomes and failure modes contribute to hazardous system states and overall risk. Our approach thus combines qualitative safety reasoning from FMEA with quantitative probabilistic inference, enabling a transparent and structured assessment of AI-enabled components.

Compared with traditional safety analyses and state-of-the-art AI performance evaluations that stop at component-level accuracy metrics, the methodology explicitly links perception performance to system-level consequences. This delta to existing practice enables a more realistic quantification of how AI misclassifications contribute to overall risk, supports a more defensible determination of the achieved safety level, and ultimately helps increase the safety level of AI-enabled maritime systems. The results provide structured evidence for uncertainty-aware safety cases and design trade-offs, enhancing the credibility of safety arguments for both developers and assessors.

The methodology contributes to ongoing research on risk-based strategies in guidelines, standards, and policy development. It supports the development of safety assurance frameworks that incorporate uncertainty, as it can yield evidence for uncertainty-aware safety claims and help quantify residual risk. Developers can also use the results in early project phases to argue risk acceptance for both the design and operation of a product, and thereby support product acceptance.

The current case study is limited to object classification and does not model risk mitigation measures, which constrains its informative value. The next step is to model sensor fusion, linking image information with other sensors such as RADAR, as well as implementing further mitigation measures to increase robustness. We plan simulations and experimental tests to assess whether the methodology is suitable for modeling a real maritime (sub)system and to validate its practical benefit for design decisions and operational safety management.

Researchers can also apply the methodology to the quantitative risk analysis of other AI components, such as navigation and route planning. As the confusion matrix cannot directly express different types of uncertainty because it only summarizes classification performance without capturing model confidence or data reliability, further research is required to model and distinguish these effects. In addition to misclassification effects, we can incorporate other failure sources, such as camera hardware failures, and represent their interactions with AI components in the Bayesian Network. This will extend the framework into a more complete, decision-transparent basis for assessing, comparing, and improving the safety level of AI-enabled maritime systems beyond the current state of the art.

Acknowledgement

The authors gratefully acknowledge the financial support of the German Federal Ministry for Economic Affairs and Energy (BMWE) within the framework of the funded project SIMAS (Analysis and assessment of the safety of maritime autonomous systems under the influence of artificial intelligence). This work was supported under grant number 03SX618D. The authors also thank all project partners and contributors for their valuable collaboration and discussions.

Supported by:

Federal Ministry
for Economic Affairs
and Energy

on the basis of a decision
by the German Bundestag

References

Adee, A., R. Gansch, P. Liggesmeyer, C. Glaeser, and F. Drews (2021). Discovery of perception performance limiting triggering conditions in automated driving. In *2021 5th International Conference on System Reliability and Safety (ICSRS)*, pp. 248–257.

Ben-Gal, I. (2008). *Bayesian Networks*. John Wiley Sons, Ltd.

Bolbot, V., M. Bergström, M. Rahikainen, and O. A. Valdez Banda (2025). Investigation into safety acceptance principles for autonomous ships. *Reliability Engineering & System Safety 257*, 110810.

DNV (2024a). Class Guideline: Autonomous and remotely operated vessels. Class Guideline DNV-CG-0264, DNV. Edition December 2024.

DNV (2024b). DNV Rules for Classification: Ships—Part 6 Additional class notations, Chapter 12 Autonomy and Remote Operation. Rules for Classification DNV-RU-SHIP Pt.6 Ch.12, DNV.

Druzdzel, M. J. (1999). GeNIe: A development environment for graphical decision-analytic models. In *Proceedings of the Annual Symposium of the American Medical Informatics Association (AMIA)*, Washington, D.C., pp. 1206.

European Maritime Safety Agency (2024). RBAT – Method description. Technical report, European Maritime Safety Agency.

European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689, 12 July 2024.

ISO (2018). ISO 26262: Road vehicles—Functional safety. Edition 2 published 2018; standard remains current as of 2025.

ISO (2022). ISO 21448: Road vehicles—Safety of the intended functionality (SOTIF). Edition 1 published 2022; latest official version.

ISO (2024). Road vehicles—safety and artificial intelligence. Technical Report ISO/PAS 8800:2024, International Organization for Standardization, Geneva, Switzerland.

Kabir, S. and Y. Papadopoulos (2019). Applications of Bayesian networks and Petri nets in safety, reliability, and risk assessments: A review. *Safety Science 115*, 154–175.

Kulkarni, C. S., M. Corbetta, and E. I. Robinson (2021). Systems health monitoring: Integrating FMEA into Bayesian networks. In *42nd International IEEE Aerospace Conference*, Number 20205006373.

Mallam, S. L., T. Porathe, S. Nazir, et al. (2024). A human-centred review on maritime autonomous surface ships: impacts, responses, and future directions. *Transport Reviews 44*(3).

MLEAP Consortium (2024). EASA Research – Machine Learning Application Approval (MLEAP) Final Report. Final Report, Horizon Europe Research Project, European Union Aviation Safety Agency (EASA).

Rainio, O., J. Teuho, and R. Klén (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports 14*(1).

Rastayesh, S., S. Bahrebar, F. Blaabjerg, D. Zhou, H. Wang, and J. Dalsgaard Sørensen (2019). A System Engineering Approach Using FMEA and Bayesian Network for Risk Analysis—A Case Study. *Sustainability 12*(1), 77.

Rauschenbach, M. (2017). *Probabilistische Grundlage zur Darstellung integraler Mehrzustands-Fehlermodelle komplexer technischer Systeme*. Phd thesis, Technische Universität Darmstadt.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi (2016). You only look once: Unified, real-time object detection.

U.S. Department of Defense (2012). MIL-STD-882E: Standard Practice System Safety. U.S. Department of Defense Standard.

U.S. Department of Defense (2021). Draft MIL-STD-882F: Standard Practice System Safety. U.S. Department of Defense Standard.

Žust, L., J. Perš, and M. Kristan (2023). LaRS: A diverse panoptic maritime obstacle detection dataset and benchmark. In *International Conference on Computer Vision (ICCV)*.