Field Validation of Hydrotwin for Automated Vessel and Dolphin Detection

Joachim Vanneste
jvanneste@blueoasis.pt

Tiago Gomes
tgomes@blueoasis.pt

Pedro Raposo
praposo@blueoasis.pt

Astrid van Toor
avtoor@blueoasis.pt

Diogo Valdrez
dvaldrez@blueoasis.pt

Guilherme Beleza Vaz
gvaz@blueoasis.pt

***Abstract*—This paper presents a field validation of Hydrotwin, a passive acoustic monitoring system combining edge-deployed AI with real-time IoT connectivity, for automated vessel and dolphin detection. Performance was evaluated against ground truth derived from AIS and STANAG 4817 vessel tracks, and expert-annotated dolphin whistle labels, collected during the NATO REPMUS 2025 exercise off the Portuguese coast. Results show that Hydrotwin achieves high detection accuracy across diverse environmental conditions, demonstrating robust performance in both offshore (HT-S) and coastal (HT-C) configurations. These findings support its suitability for applications in environmental monitoring and security operations.**

I. Introduction

Reliable detection of acoustically active entities is fundamental to understanding and managing the maritime domain. Both surface vessels and marine mammals generate distinctive underwater acoustic signatures that provide information for applications ranging from maritime security to biodiversity protection [1]. The Hydrotwin[1] system provides passive, real-time AI-based vessel and mammal detection in three operational configurations [2], differing in deployment context, power supply and acoustic front-end hardware. This study evaluates the performance of two configurations — Hydrotwin-Spotter (HT-S) and Hydrotwin-Cable (HT-C) — using manual labels and vessel positioning data from AIS and STANAG 4817 messages collected during the NATO REPMUS 2025 exercise. By comparing Hydrotwin detections with these assembled ground truths, we assess detection accuracy, reliability, and suitability for operational deployment in environmental monitoring and security applications.

II. Hydrotwin System

All Hydrotwin configurations operate as IoT devices, enabling near-real-time monitoring through cellular or satellite connectivity. The three configurations share a common AI processing framework and overall system architecture but differ in deployment context, power supply, and acoustic front-end hardware:

- **Hydrotwin-Spotter (HT-S, Fig. 3)**: designed for offshore deployments, relying on the Sofar Ocean[2] Spotter buoy for power and telemetry,
- **Hydrotwin-Cable (HT-C)**: optimised for coastal environments, operating via a direct cabled power supply,
- **Hydrotwin-Vessel (HT-V)**: installed directly on vessels and powered by the vessel's electrical system.

[1]https://hydrotwin.ai/
[2]https://www.sofarocean.com/

Fig. 1: Hydrotwin-Spotter for offshore deployments.

Fig. 2: Hydrotwin-Cable for cabled environments.

Fig. 3: Hydrotwin-Vessel for autonomous patrolling.

A. Hardware

HT-V and HT-C share the same in-house developed acoustic front-end, including the hydrophone and analog-to-digital conversion (ADC) stage, ensuring consistent signal acquisition across both systems. HT-C is installed as a fixed seabed unit with the sensing assembly positioned on the ocean floor, whereas HT-V operates with a 15 m towed line deployed from the host vessel. HT-S is built around the commercially available BOREALIS[3] system, which provides the complete acoustic hardware stack. Despite this difference in sensing hardware, HT-S shares the same computing unit and AI processing framework as the other configurations.

B. AI

Raw acoustic data is processed locally using onboard CNN-based models that detect and classify underwater acoustic sources directly on the device. Operating on spectrograms — 2D time-frequency representations of the acoustic signal — rather than raw waveforms reduces computational load while preserving the features relevant for classification, making real-time inference feasible in a low-power edge-computing setting.

The architecture has been optimised for resource-constrained deployment: it is up to 90% less resource-intensive than comparable edge-optimised models such as YOLO Nano [6] and EfficientNet [4], while retaining the ability to identify a wide variety of underwater sound sources including dolphin whistles, whale calls, and multiple vessel classes. Models were trained on expert-labelled internal and external acoustic datasets [5], and all configurations run identical firmware on a shared computing unit, ensuring consistency across hardware variants. An example of the Hydrotwin AI system is action can be seen here: https://youtu.be/jc5txdZ1y_I.

C. Dashboard

Processed detections are transmitted to a cloud platform and presented through an interactive web-based dashboard. The dashboard supports both real-time monitoring and retrospective analysis of historical datasets. It provides operators with several complementary views: a detection and sound statistic timeline, a map with geolocated detections, a spectrogram viewer for manual inspection of flagged events, and aggregated statistics on detected source classes. This combination of spatial, temporal, and spectral context is intended to support both environmental monitoring workflows and maritime security applications requiring rapid situational awareness.

[3]https://www.appliedoceansciences.com/borealis

III. REPMUS Exercise

A. Overview

The NATO REPMUS (Robotic Experimentation and Prototyping with Maritime Unmanned Systems) 2025 exercise provided a unique opportunity to validate Hydrotwin's detection models under mixed traffic and environmental conditions. REPMUS 2025 took place between Sesimbra and the Tróia peninsula between September 8th and September 25th 2025 involving vessel activity across multiple operational areas.
A key feature of this exercise was the availability of a Automatic Identification System (AIS) dataset; an internationally mandated system requiring certain vessels to broadcast their identity and positions at regular intervals via radio signals.

Additionally, the coastal waters south of Sesimbra host resident bottlenose dolphin populations. Dolphin acoustic activity observed during the exercise enabled concurrent validation of Hydrotwin's cetacean detection models using expert-annotated whistle labels as ground truth.

B. Deployments

During the exercise, six Hydrotwin units were deployed: five HT-S units and one HT-C across different locations and depths. Table I summarises the deployed systems.

Hydrotwin ID	Lat. (°)	Long. (°)	Depth (m)	Water depth (m)
HT-S-007	38.4000	-9.0300	10	99
HT-S-008	38.3700	-9.0300	10	110
HT-S-009	38.3625	-9.0750	65	151
HT-S-010	38.3560	-9.0750	65	197
HT-S-011	38.3560	-9.0670	65	185
HT-C-003	38.4779	-8.8719	19	20

TABLE I: Deployment details of the six Hydrotwin systems during the REPMUS 2025 exercise. Five HT-S units were positioned offshore at varying depths and one HT-C unit was installed in the coastal waters of Tróia Bay. Coordinates and depths refer to sensor positions measured at deployment.

Two of the HT-S units (*HT-S-007* and *HT-S-008*) were deployed at a depth of 10 m, where acoustic propagation conditions are more favorable. The remaining three HT-S units were deployed at deeper locations (65 m) to assess the impact of depth on detection performance. Figure 4 shows the deployment locations of all Hydrotwin units.

IV. Methodology

A. Ground truth construction

Dolphin detection performance was evaluated using a 10-hour subset of HT-C-003 recordings corresponding to the period

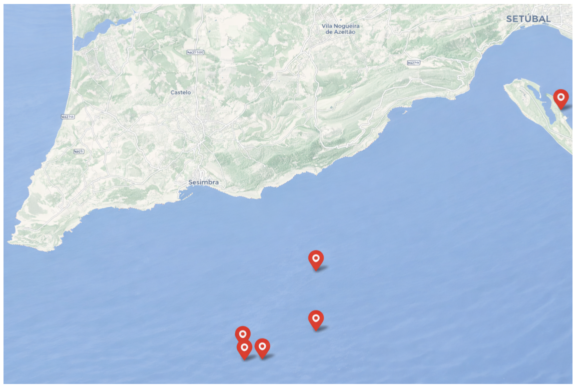


Fig. 4: Map of Hydrotwin deployments during REPMUS 2025, showing the five offshore HT-S units and one coastal HT-C system.

with the highest concentration of AI detections. This data was manually annotated by in-house acoustic experts and used as ground truth for validating the Hydrotwin dolphin detection model. AIS data was used as ground truth for vessel detection. Prior to validation, AIS coverage was assessed using expert-labelled acoustic data from September 23–25, 2025 for three units (*HT-C-003*, *HT-S-007*, and *HT-S-010*). Of 236 labelled vessel events, 217 (91.9%) had corresponding AIS targets within 5 km, while the remaining 19 were associated with AIS vessels at 5–10 km range. This confirmed AIS reliability within the 5 km operational detection range.

B. Validation procedure

This procedure applies to **vessel detections only**; dolphin detections were validated separately against expert-labelled annotations.

To account for duty-cycled recordings, AIS position reports were grouped into time windows matching each Hydrotwin recording interval. A binary vessel presence ground truth was defined for each window based on AIS activity. Hydrotwin-Vessel detections were similarly reduced to a binary indicator per window, and detection performance was evaluated through window-level comparison.

A vessel was considered present if an AIS position occurred within $\pm 10\,\mathrm{s}$ of the recording window, within 5 km of the sensor, and at a reported speed greater than 1 knot.

Detection performance was quantified using precision, recall, and F1-score. The same metrics were used for dolphin detection evaluation; however, in that case performance was assessed by comparing Hydrotwin detections against expert-labelled acoustic ground truth rather than AIS data. It is worth noting that independent visual observations (e.g., from shore-based cameras, drones, or satellite imagery) could potentially supplement validation efforts; however, since dolphin vocalisations originate underwater and individuals are not necessarily visible at the surface during sound production, imagery was not used to support this validation.

V. Results

The five HT-S units recorded 12 minutes out of every 30 to allow for continuous monitoring over extended periods while managing power constraints. In total, 8974 files were recorded totaling almost 1800 hours, 4387 of these files had vessel detections and 329 had dolphin detections. The HT-C unit operated in a continuous recording mode, capturing one-minute files thanks to its connection to power. In total, 36,937 files were recorded, corresponding to approximately 615 hours of acoustic data. Out of these, 15,416 files contained vessel detections and 95 files contained dolphin detections.

A. Dolphin detection

Across the selected 10-hour dataset, 282 dolphin whistles were identified across 74 consecutive one-minute recordings. The Hydrotwin model detected dolphin presence in 35 files, of which 29 were true positives. This corresponds to a precision of 81.8% and a recall of 47.3% at the file level. Figure 5 shows the temporal alignment between expert-labelled dolphin vocalisations and Hydrotwin detections. The results indicate a conservative detection behaviour, favouring high precision over recall, which is appropriate for long-term environmental monitoring applications where false positives are undesirable.

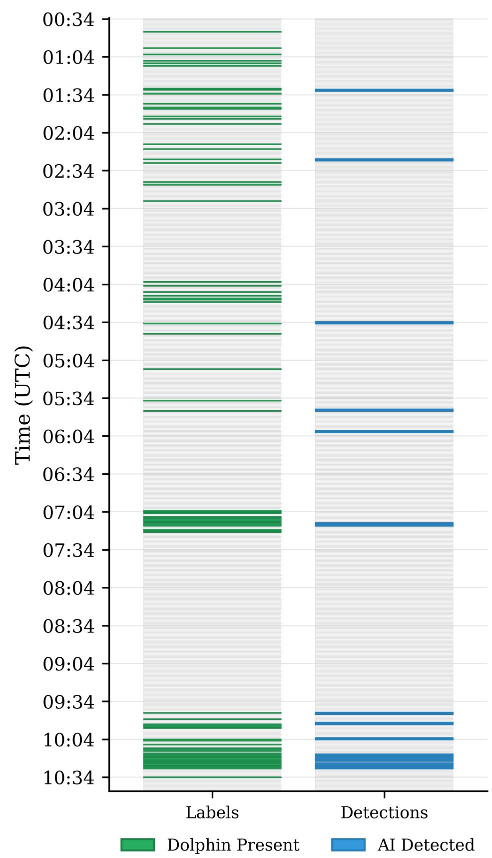
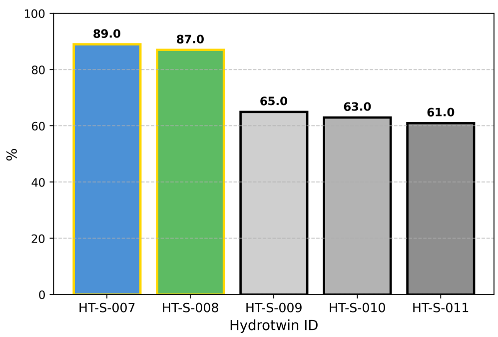


Fig. 5: Timeline comparison of dolphin labels and AI detections over a 10-hour recording period from HT-C-003. Green bars indicate dolphin vocalisations identified by expert labellers and blue bars show AI detections.

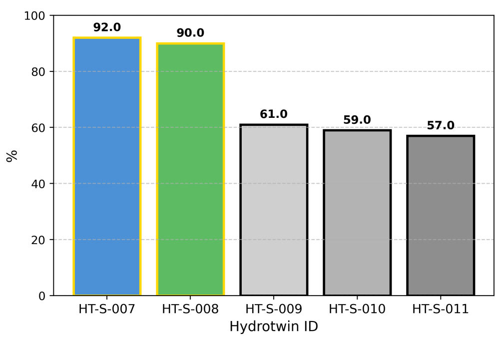
The detection performance could likely be improved by optimising the model's decision threshold, allowing a less conservative operating point and a more balanced trade-off between precision and recall.

B. Vessel detection

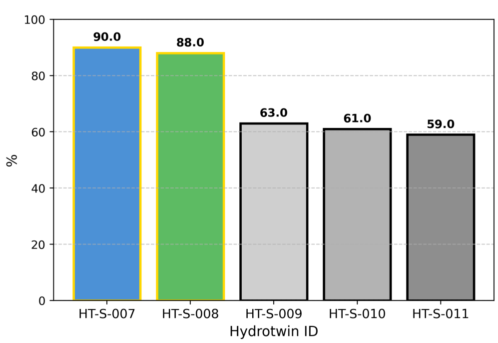
The detection performance of the HT-S units is averaged over the full duration of the REPMUS 2025 exercise. Figure 6 shows the precision, recall and F1 for the units. With an average F1-score of **89%**, the two shallow deployments (*HT-S-007* and *HT-S-008*, both at 10 m depth) consistently achieved higher performance across all metrics compared to the three deeper units (*HT-S-009*, *HT-S-010*, and *HT-S-011*, deployed at 65 m).



(a) Precision for HT-S deployments.



(b) Recall for HT-S deployments.



(c) F1-score for HT-S deployments.

Fig. 6: Precision, recall, and F1-score for the five HT-S systems during the REPMUS 2025 validation campaign, computed using AIS-derived vessel presence as ground truth. The two units deployed at 10 m depth achieved the highest overall performance, while deeper deployments (65 m) showed reduced detection accuracy.

This confirms the expected advantage of shallower placements, where acoustic propagation paths are less attenuated. In contrast, the deeper deployments showed reduced sensitivity, likely due to greater signal loss with depth and more complex propagation effects. As part of the validation analysis, the spatial matching radius was varied to test detection range limitations. Increasing the radius from 5 km to 15 km reduced F1-score by 45%, primarily due to increased false negatives — distant vessels present within the expanded radius but not detected acoustically. This confirms that system sensitivity is well-matched to the 5 km operational threshold, beyond which acoustic propagation losses reduce detection probability. This is supported by acoustic propagation modeling results in the region using RAINDROP [3]. The HT-C matched against AIS data achieved a precision of **77.0%**, a recall of **89.9%**, and an overall F1-score of **83.0%**. The high recall demonstrates the system's strong sensitivity to vessel activity, successfully identifying the majority of vessel events present during the exercise. While the HT-C achieved high recall, the precision metric was partially reduced by *apparent* (see next paragraph) false positive detections. Figure 7 shows the confusion matrix against AIS for this deployment.

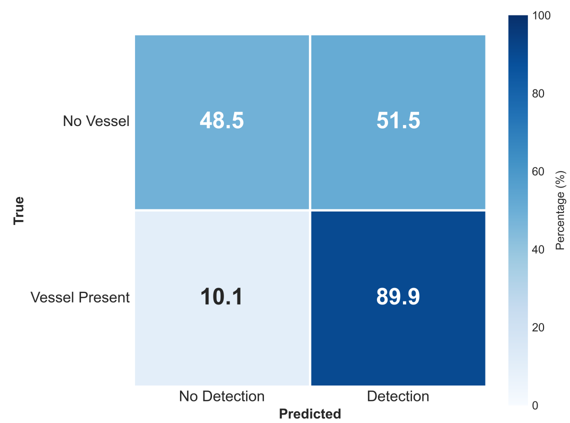


Fig. 7: Confusion matrix of HT-C vessel detections versus AIS ground truth. Percentages represent the proportion of analysed 1-minute files in each classification category.

Upon manual inspection of a subset of the recorded files, several contained clear acoustic evidence of vessel activity despite the absence of corresponding AIS transmissions. This indicates that a portion of the apparent false positives are in fact true detections of non-AIS or AIS-inactive vessels. For instance, between 00:38 and 00:42 UTC on September 19, the HT-C recorded a sequence of files containing strong broadband tonal signatures consistent with vessel passages, yet no AIS targets were present within the detection range. Figure 8 shows a screenshot from the Hydrotwin web application displaying the spectrograms of these five consecutive files, where the characteristic propeller and engine harmonics are clearly visible. Fig. 9 (Appendix) shows a close up view of a raw audio sample with vessel detections alongside dolphin detections in the Hydrotwin webapp dashboard. These findings highlight

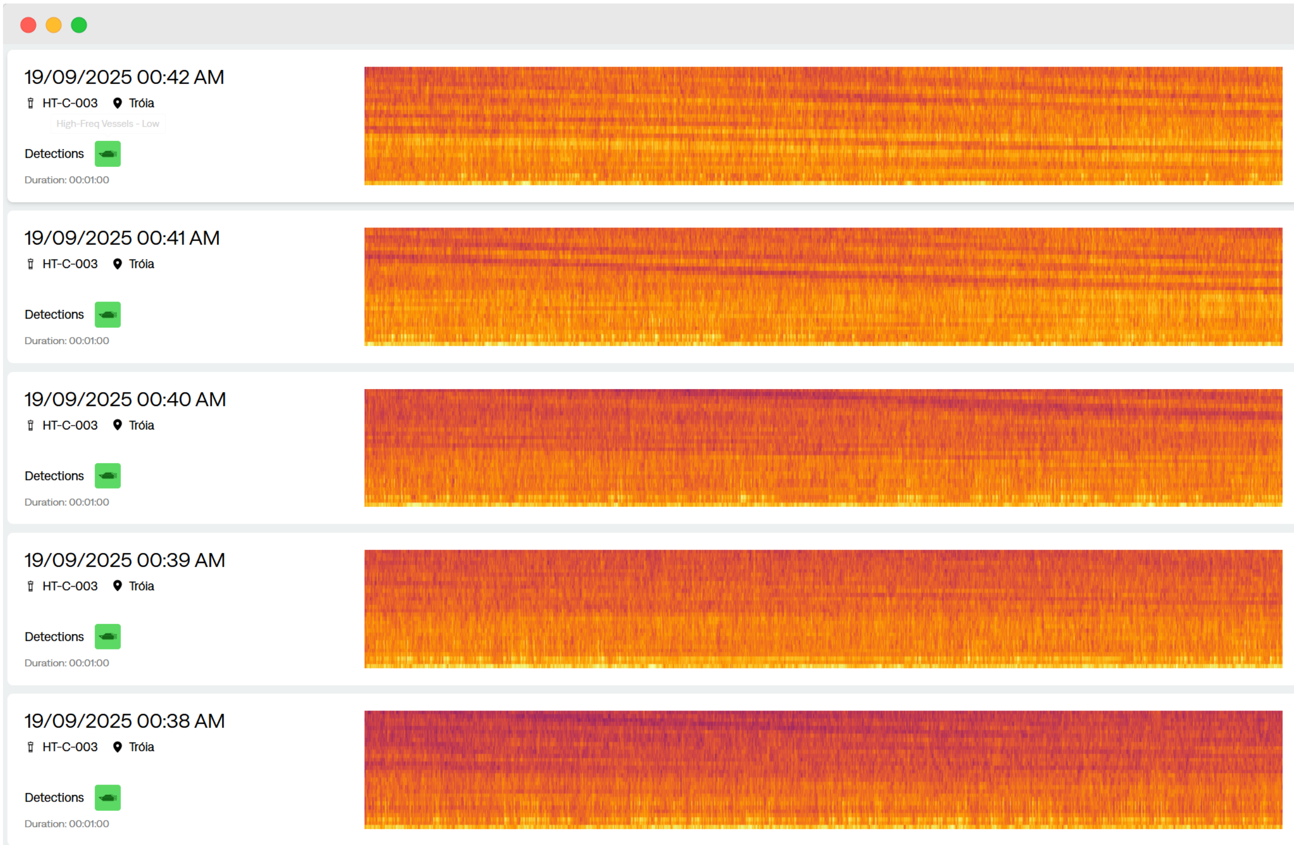


Fig. 8: Spectrograms from five consecutive HT-C recordings between 00:38 and 00:42 UTC on September 19, showing clear acoustic vessel signatures despite no corresponding AIS targets.

the system's capability to detect non-AIS or AIS-inactive vessels, which may include suspicious or unreported maritime activity; such detections are surfaced in the Hydrotwin web application, where users are automatically notified and can review the associated spectrograms and audio evidence for further assessment. Overall, the HT-C results confirm that the system performs reliably under continuous coastal operation. Its ability to maintain high recall and balanced precision over extended recording periods demonstrates both the robustness of the detection model and its suitability for long-term monitoring applications.

VI. Discussion

The validation results demonstrate that Hydrotwin provides consistent and accurate vessel and dolphin detections across a range of operating conditions.

The dolphin detection results demonstrate that the Hydrotwin model operates conservatively, prioritising detection reliability over completeness. The high precision indicates that when dolphin presence is reported, it is typically supported by clear acoustic evidence, minimising false positive detections. This behaviour is particularly important for long-term passive acoustic monitoring, where spurious detections can significantly increase post-processing effort and reduce user trust.

Regarding the vessel detections, the systems deployed at 10 m depth achieved superior performance, with higher precision and F1 scores compared to the deeper (65 m) units. This depth dependence likely reflects the stronger signal-to-noise ratio for surface vessel noise in shallower deployments, where acoustic propagation loss is reduced. The presence of apparent false positives, particularly in the HT-C results, was further investigated through manual inspection. Several of these cases corresponded to clear acoustic signatures of vessels with no associated AIS transmissions, indicating that Hydrotwin successfully detected non-AIS or AIS-inactive traffic.

Overall, the results confirm Hydrotwin's capability to deliver reliable vessel and dolphon detections under diverse real-world conditions. Continued validation across additional sites and seasons will further refine model robustness and support broader operational deployment in environmental monitoring and maritime security contexts.

VII. Conclusion

The REPMUS 2025 field validation demonstrated that Hydrotwin delivers reliable vessel and dolphin detection under real operational conditions across two hardware configurations. For vessel detection, shallow HT-S deployments (10 m) achieved a mean F1-score of 89%, with performance declining at deeper placements (65 m), consistent with increased

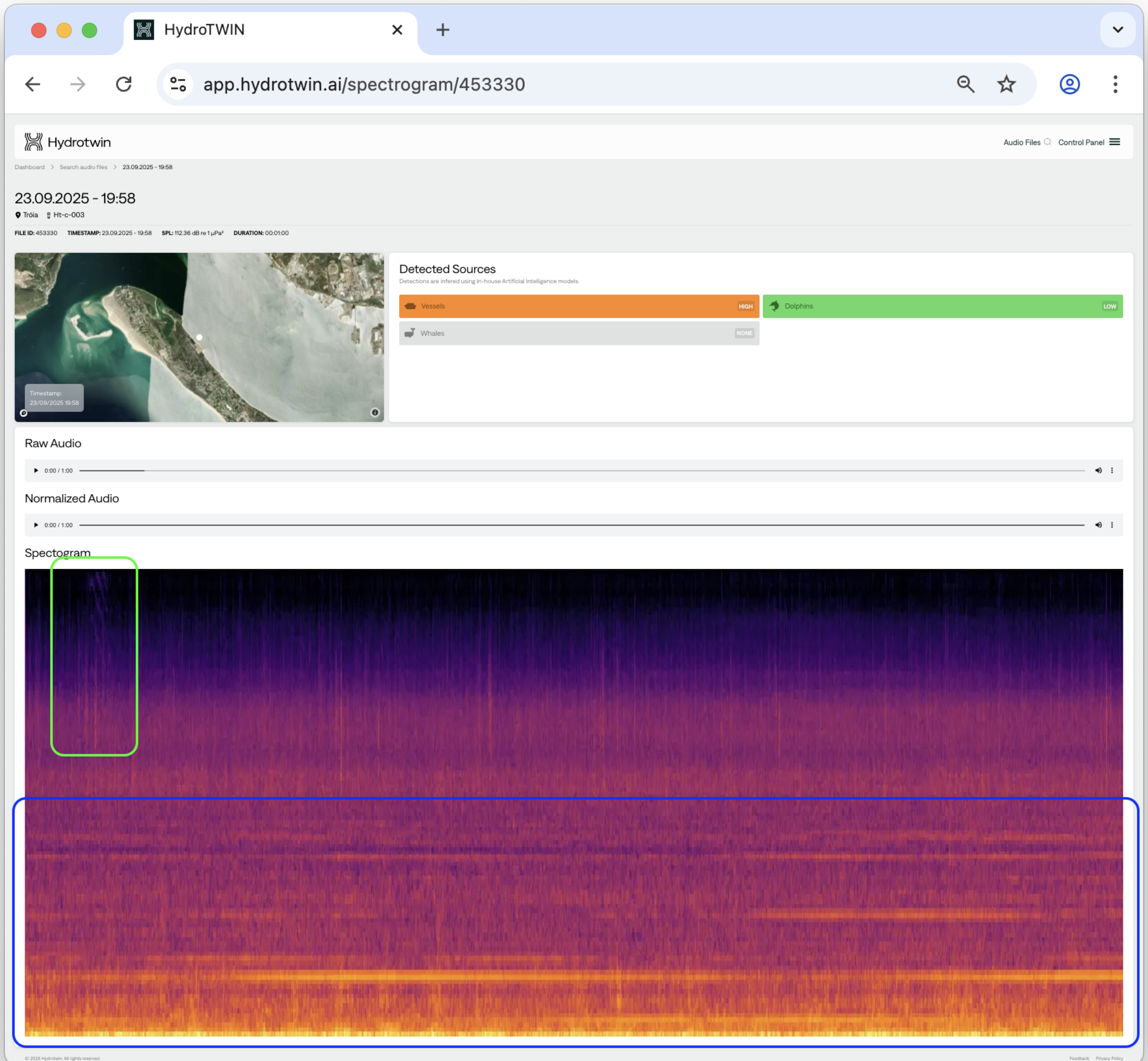


Fig. 9: Vessel and dolphin detection by HT-C during REPMUS in the Hydrotwin dashboard. The blue box indicates the boundaries of the vessel signature throughout, and the green box indicates the dolphin signature.

acoustic propagation loss. The HT-C achieved an F1-score of 83.0% under continuous coastal operation, with manual inspection confirming that a subset of apparent false positives corresponded to acoustically detectable vessels without active AIS transponders demonstrating situational awareness capability beyond what AIS alone can provide. Dolphin detection achieved 81.8% precision at a conservative operating point, appropriate for long-term passive acoustic monitoring where false positive minimisation is prioritised over recall.

Taken together, these results validate Hydrotwin's dual-use applicability for both environmental monitoring and maritime security, and establish concrete performance baselines for future development. Ongoing work focuses on expanding training datasets, refining detection thresholds, and advancing species- and vessel-class-level classification, while maintaining the energy-efficient edge-computing architecture required for offshore deployment.

The AI models are continuously improving as more data is labeled for training. Additionally, there is ongoing focus on advanced feature extraction to enhance the granularity of detections, such as vessel characteristics and marine mammal species. All these improvements will be implemented while keeping the same prioritisation of energy-efficient algorithms to support offshore edge-computing applications.

REFERENCES

[1] Muhammad Azeem Aslam, Lefang Zhang, Xin Liu, Muhammad Irfan, Yimei Xu, Na Li, Ping Zhang, Zheng Jiangbin, and Li Yaan. Underwater sound classification using learning based methods: A review. *Expert Systems with Applications*, 255:124498, 2024.

[2] Tiago Gomes, Alberto Santos, Joachim Vanneste, Astrid Toor, Carolina Ramos, Daniele Aromatario, Rita Carriço, Erica Cruz, and Guilherme Vaz. Hydrotwin: Capturing the heartbeat of the ocean. pages 235–237, 09 2025.

[3] Tiago Gomes and Guilherme Beleza Vaz. RAINDROP: A framework for real-time holistic underwater acoustic maps. In *Proceedings of the Underwater Acoustics Conference and Exhibit 2025*.

[4] Van-Thanh Hoang and Kang-Hyun Jo. Practical analysis on architecture of efficientnet. In *2021 14th International Conference on Human System Interaction (HSI)*, pages 1–4, 2021.

[5] A. van Toor. Cnn architecture diagram of "supervised detection of baleen whale calls on edge-compute" for dcase 2025. Zenodo, 2025. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) 2025, Barcelona, Spain.

[6] Alexander Wong, Mahmoud Famuori, Mohammad Javad Shafiee, Francis Li, Brendan Chwyl, and Jonathan Chung. Yolo nano: a highly compact you only look once convolutional neural network for object detection. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS Edition (EMC2-NIPS)*, pages 22–25, 2019.