On the Importance of Domain Knowledge for Real-Time Event Detection

Janina Schneider*†, Daniel Lukats†‡, Elmar Berghöfer†, Iring Paulenz§, Lars Nolle§†,
Frederic Stahl† and Jochen Wollschläger*
*Institute for Chemistry and Biology of the Marine Environment,
Carl von Ossietzky Universität Oldenburg, Wilhelmshaven, Germany
Email: janina.schneider@uol.de
†Marine Perception, German Research Center for Artificial Intelligence, Oldenburg, Germany
‡Department of Computing Science, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany
§Department of Engineering Sciences, Jade University of Applied Sciences, Wilhelmshaven, Germany

***Abstract*—Due to the complexity and dynamics of marine ecosystems, real-time monitoring and possibilities for anticipatory intervention are required. This study emphasises the crucial role of domain (expert) knowledge in developing effective real-time event detection systems. Focusing on the Change Event based Sensor Sampling (ChESS) project, which utilises methods based on Artificial Intelligence (AI) for the analysis of marine data streams, the research integrates stakeholder input and domain knowledge into the development process. Using data from the Time-Series Station (TSS) Spiekeroog, domain knowledge influences the design of software architecture, algorithms' goals and constraints, data retrieval, and related actions as an automated triggering of a water sampler. Lessons learned stress the importance of early domain expert involvement and highlight the value of combining technological advancements with domain insights. Participatory approaches enhance the development of monitoring systems, supporting informed decision-making for sustainable marine ecosystem management.**

***Index Terms*—data stream mining, marine ecosystems, expert knowledge**

I. Introduction

Marine ecosystems are dynamic and complex, characterised by interconnections among various biological, chemical, and physical parameters. Real-time monitoring of marine ecosystems is essential to enhance ocean understanding and secure a sustainable blue economy [1]. Besides the provision of data streams and offline standard statistical analysis methods, monitoring efforts can benefit from automated real-time event detection systems, which can support natural scientists with information about specific events and with the possibility to act or automatically trigger an action, for example, sampling. Due to the interdisciplinary backgrounds of the different actors, i.e. developers and application domain experts, the development of such systems should ideally be implemented with co-design between these actors. The experts provide their expert knowledge, here referred to as domain knowledge.

Real-time event detection involves the timely identification and assessment of anomalies and drifts in the data through data stream mining techniques [2], which in marine ecosystems can occur due to, e.g., harmful algal blooms, temperature fluctuations, or pollution. Depending on the domain expert's interests or the intended use case, multiple events may or may not occur simultaneously. Likewise, the description of events can differ, as events can be described as anomalies occurring at a single time step or as abnormal time intervals [3]. If these events cause a lasting change in the observed data patterns, they are called concept drift in the data mining literature [4], [5]. While advances in sensor technology and machine learning algorithms have improved the capabilities of event detection systems, the integration of domain knowledge remains essential [6], [7]. This knowledge encompasses an understanding of the specific characteristics, behaviours, and interactions within marine ecosystems. It enables the development of context-aware algorithms and models, ensuring that detected events are not only accurate but also meaningful in the broader ecological context.

In this work, the retrieval and inclusion of domain knowledge and data is studied during the development of a real-time event detection system, based on data stream mining and Artificial Intelligence (AI) techniques, for marine data streams within the ChESS (Change Event based Sensor Sampling) project. The involvement of domain knowledge exists in the stages of architecture development, algorithm development and proof-of-concept study, e.g., requirement analysis, data retrieval, visualisation, data annotation, adaptation, and evaluation.

A time-series station within a coastal observatory, the Time-Series Station (TSS) Spiekeroog, serves as data stream provider for this use case. It continuously records oceanographic, meteorological, and biogeochemical data in a tidal channel close to the island of Spiekeroog in the German Bight (southern North Sea). Within the use case, the automated triggering of a water sampler by the event detection system is examined. This work highlights the importance of domain knowledge in the development of effective and accurate real-time event detection systems in these ecosystems and demonstrates the lessons learned on this topic within the project ChESS.

Funded by the Ministry of Science and Culture of Lower Saxony and the VolkswagenStiftung through "Niedersächsisches Vorab" (ZN3683).

II. Use case and approaches

A. Use case description

The ChESS project aims to develop an event detection system for marine sensor data streams to support natural scientists. The natural sciences' main goal is to establish a coherent model, which can describe the subject under study, its interrelations and correlations. It is a repetitive and indolent process in which scientists share a common knowledge based comprising facts, hypotheses and rules. This repetitive process is depicted in Fig. 1; scientists start with the available information from the body of knowledge, in order to identify gaps, anomalies, discrepancies or lack of explanation. Once this gap is identified, this motivates the definition of a new research hypothesis. Next, scientists develop ideas/theories how the hypothesis could be proven right or wrong leading to the development of a research methodology, i.e. design of experiments to generate research data. The collected data are then subsequently processed (organised, curated, analysed) to create information. From this information knowledge is generated, i.e. by extracting new patterns, correlations, interrelations, anomalies, etc. This newly generated knowledge is then documented into scientific papers, which are peer-reviewed and, if of sufficient quality, these papers are subsequently added to the public body of knowledge through publication.

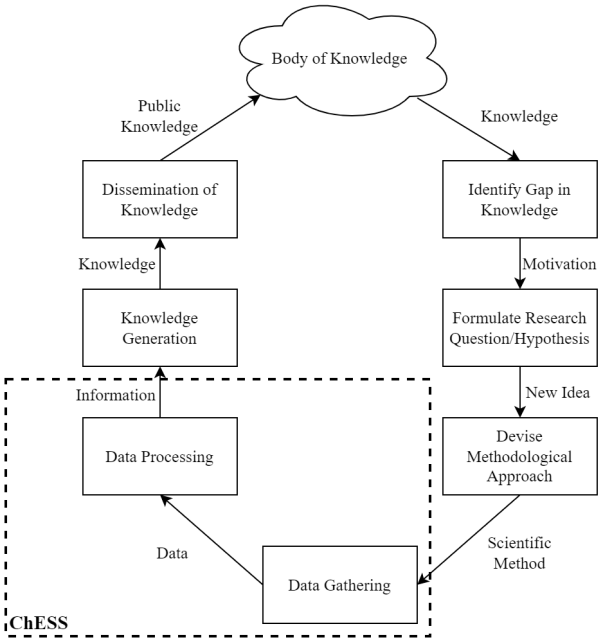

Fig. 1. Knowledge generation as motivation for the development of a real-time event detection system. The ChESS system automates the data gathering and data processing.

Automatisation of knowledge generation through targeted exploitation of digital technologies (e.g. AI) could greatly accelerate knowledge generation in natural sciences. Thus, the ChESS research project aspires to automate parts of the life cycle. The parts being automatised are the data gathering and data processing through the ChESS methodology, as indicated by the dashed bounding box in Fig. 1. Nevertheless, a human domain expert is still necessary for the system configuration, i.e. sensor maintenance, calibration and rescuing. The domain expert also derives meta-data from the information provided by the event detection.

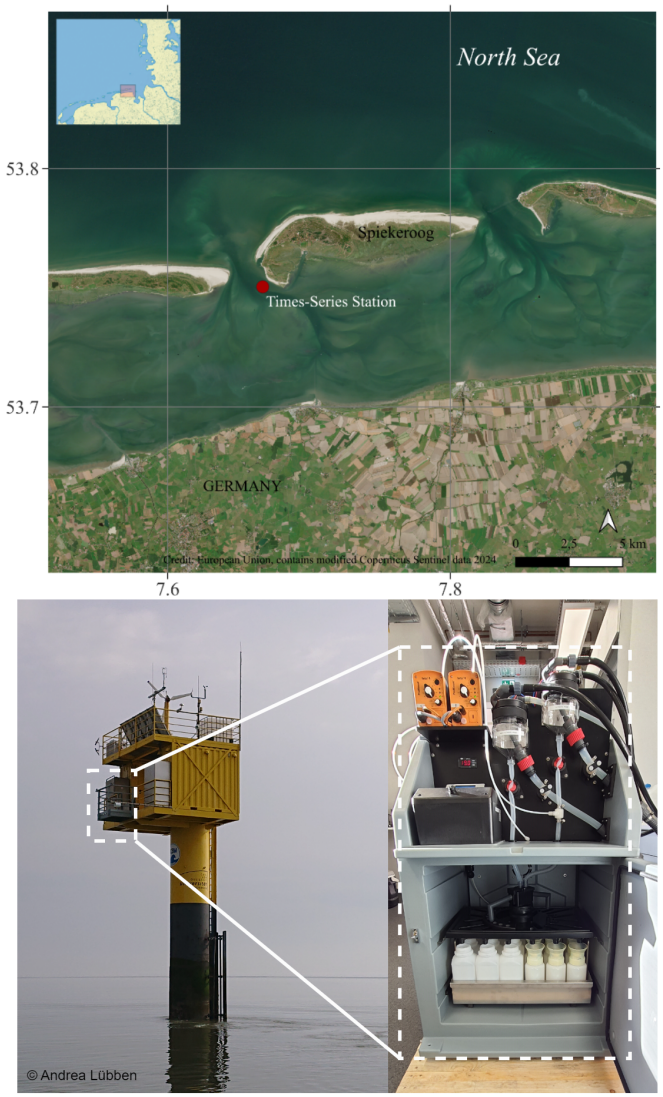

Fig. 2. Above: Map section of the German Bight (southern North Sea) indicating the geographical location of the TSS close to the island of Spiekeroog. Below: TSS Spiekeroog (left) with automated water sampler (right).

The proof-of-concept study is placed within the Spiekeroog Coastal Observatory [8], which incorporates different measurement locations on and around the island of Spiekeroog in the German Bight, North Sea. The TSS Spiekeroog [9] serves as data stream provider for this use case, as it continuously records oceanographic, meteorological, and biogeochemical data in a tidal channel close to the island of Spiekeroog. During maintenance visits, water samples are drawn. Within the ChESS project, a water sampler was acquired and installed at the TSS (Fig. 2), by which an automated sampling can be

TABLE I
TASKS WITH SPECIFIC APPROACHES FOR DOMAIN KNOWLEDGE INVOLVEMENT

Goal	Approach	Outcome
Requirement analysis	Stakeholder interviews	Interdisciplinary requirements for software architecture
Data understanding	Expert interviews	Data access, additional information and list of parameters
Use case definition	Expert interviews	Event definitions, specific algorithm/software requirements, sampling strategy design
Algorithm development	Data annotation	Labeled data set for training and testing

triggered. The goal is to use the developed ChESS system as shown in Fig. 3 (green bold arrows):

- The TSS provides the marine data stream for the ChESS system.
- Within the ChESS system, the data are analysed and finally events are detected using data stream mining algorithms. This triggers (potentially with a predefined threshold [10]) the water sampler at the TSS, and passes information or possible warnings to the domain experts (natural scientists).
- With this information, the expert can schedule actions, e.g. additional maintenance. With the next visit, the water samples are collected and can be analysed later on in a lab.

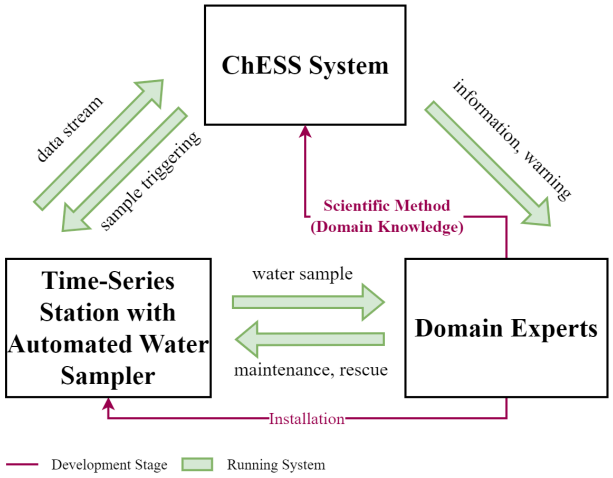

Fig. 3. Relations between the ChESS system, the TSS as sensor system and the domain experts during the development stage (see purple thin arrows) and in the envisioned live system (see green bold arrows).

B. Domain knowledge inclusion

As shown in Fig. 3, the domain experts are already required in the development stage (purple thin arrows). Firstly, the installation and integration of the acquired automated water sampler into the TSS is the responsibility of the domain experts operating the TSS. Secondly, the input of domain knowledge for the development of the ChESS system is required. Fig. 4 shows potential areas within the system development, where this knowledge may be necessary. This includes the building of a knowledge base, the software and algorithm development, and the application and adaptation.

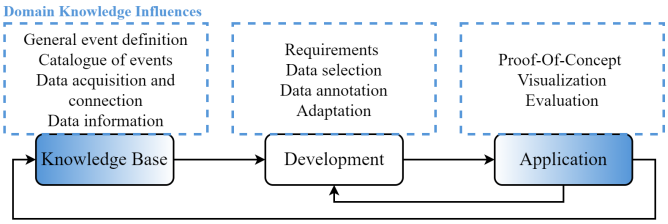

Fig. 4. Areas of influence for domain knowledge input within the process from the knowledge base, over the development, application and adaptions, to the generation of new knowledge.

In addition to regular exchange with the domain experts about general topics, certain tasks need explicit attention. The overview of approaches for domain knowledge inclusion used in this study for specific tasks is shown in Table I. First, general stakeholders (future users) for the proposed system are identified without focusing on specific domains. With stakeholder interviews, the requirements for the software architecture are collected. As a case study within the Spiekeroog Coastal Observatory is the application case of ChESS, further development steps are tailored for this. Therefore, the available data and their management is of utmost interest. To specify the use case, definitions of the term event and explicit descriptions need to be gathered. These steps are conducted through extensive exchange with the domain experts. Without domain expert feedback, it is unlikely that concept drift of anomaly detection algorithm will detect relevant events, as the algorithms require proper configuration and training. Hence, a ground truth data set must be annotated by the domain experts with information on real events to enable both development and optimisation of the event detection algorithm. Finally, a sampling strategy needs to be designed with respect to the domain expert's interests.

III. INSIGHTS GATHERED THROUGH THE DESIGN PROCESS OF CHESS

First, stakeholder interviews were conducted to analyse requirements of such event detection systems for architecture development [11]. The identified stakeholders covered not only the marine domain, e.g., scientists of the TSS, but stakeholders with broader backgrounds, e.g. industry. This is crucial to achieve a universal system, as event detection is not restricted to the marine domain and can benefit a wide range of applications. Stakeholder requirements differ in data frequency, response time (time needed to detect events), data pre-processing and other application specific requirements.

Therefore, a modular approach of the software is the desired outcome, to enable customisation of the software to a specific application at hand. The analysis also revealed that there is a need for (near-)real-time systems, as potential follow-up actions must take place as quickly as possible to match the time frame of the event. For the TSS use case, the handling of missing data and differences in temporal resolution need to be considered. A particular requirement of the scientists of the TSS for the system is a graphical interface for the exchange and update of information of the live system, e.g. to visualise the live data and the timestamps of detected events, but also to annotate previously undetected events. The desired action to follow is triggering the automated water sampler. In general, with proper stakeholder identification beforehand, not only are the requirements more defined for general usage, but can additional use cases be identified. The following insights were gained during the requirement analysis:

- Involvement of stakeholders with broad and diverse backgrounds enable generalisation requirements.
- Modularity of the system allows diverse applications.
- Thorough stakeholder identification creates possibilities for additional use cases.

A plausability study on meteorological data from a similar station within the Spiekeroog Coastal Observatory identified both anomalies and concept drift [12]. Here, domain knowledge was crucial for the identification of relevant target variables for the event detection, relevant correlations and for the evaluation of the model predictions. The same applies to ongoing analysis of data from the TSS. The domain experts involved with the TSS are scientists and engineers with backgrounds in oceanography, chemistry and marine biology, operating the TSS and working with its data. The parameters preliminary identified for this study can be found in Table II. Reasons for the selection are availability, research interest and parameter correlations.

TABLE II
PRELIMINARY PARAMETERS OF TSS SPIEKEROOG IDENTIFIED BY THE DOMAIN EXPERTS FOR THE USE CASE

Parameter	Temporal Resolution
Nitrite	1 h
Nitrite and Nitrate	1 h
Phosphate	1 h
Silicate	1 h
Solar irradiance	5 min
Air temperature	1 min[a]
Atmospheric pressure	1 min[a]
Relative humidity	1 min[a]
Wind direction	1 min[a]
Wind speed	1 min[a]
Water pressure	1 min[a]
Water temperature	1 min[a]
Salinity	1 min[a]
Oxygen	1 min[a]

[a] Average of 5 sec measurement intervals.

As the main goal is to gain the most value from this event detection system for the users, a thorough determination of concepts was conducted a priori. The general definition of an event for this use case varies ranging from specific concepts, e.g. a deviation of 2σ (with σ being the standard deviation), to simple experience of a domain expert, e.g. if data reach a specific threshold or if rapid changes occur, which cannot be transferred to explicit rules. The preliminary event catalogue contains events such as storms and heavy rainfalls or algal bloom and tipping points. A sampling strategy must ensure comparable samples by sampling during matching tidal phases, but it must also respect further requirements. Most importantly, this applies to aspects such as the relationship between an event and a water sample—should the sample be drawn during the event or after it concluded—as well as resource constraints. Since the automated water sampler contains a finite amount of bottles and sample retrieval is tied to maintenance operations, the number of samples that can be obtained is limited. Further limitations are placed on the sampling strategy, as domain experts asked for the option to obtain water samples at scheduled times or manually. As the run-time of the water sampler exceeds one hour, the event detection and the following triggering signal is not time-critical in terms of minutes. This also defines the term real-time for the action intended in the application. In addition, the knowledge of the domain experts contributed to the handling of measurement artefacts and general data issues for data management. Data related insights and best-practices regarding domain knowledge are:

- Plan ahead and involve the domain experts from the start.
- Clear and frequent communication with domain experts ensures management of expectations on both sides, prevents potential labour-intensive issues, e.g. due to different standards in data management such as used date time format or file encoding, and fosters accurate implementation.
- Preparation of definitions is necessary: Accurate definitions of events in general, the specific event catalogue and determination of time frames for corresponding actions after a detection are important.
- Finally, both domain experts and developers of AI models must be certain that the available data is sufficient for the task at hand both in terms of quality and quantity, as the data needs to be accurate, reliable and have a good (spatio-)temporal resolution.

For the following adaptation and optimisation of the event detection algorithm, a data set of the TSS is currently being annotated by the domain experts. Events or areas of interest are assigned to certain features present in the data. This annotation process is performed using a Grafana dashboard [13]. Grafana is an open-source software for monitoring and analytics, suitable for the visualisation of streaming data. Both, developers and domain experts, have requirements for this task, as on the one hand it should fit the standards and expectations of the domain experts, and on the other hand it must be useful for the algorithms' training processes. Especially, the diverse definition of an event (e.g., algal bloom) as the interpretation of relevance of an event varies between different domain experts

which needs to be considered. As shown exemplary in Fig. 5, multiple annotation can exist for the same event, caused by different annotated points in time or different parameters used as decision basis. Therefore, annotations of events should consist of a time window for flexibility in the evaluation, whereas desired sampling should be a point in time.

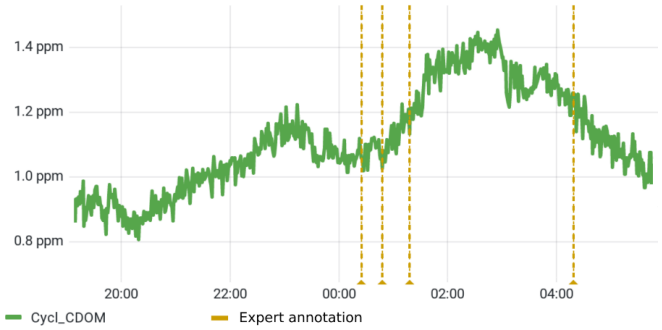

Fig. 5. Example of annotation results: Multiple expert annotations (dashed lines) while parameter value is increasing, which can be assigned to a single event.

For data annotation with domain experts, lessons learned are:

- Multiple domain experts for annotations are useful, as every person will annotate the same event differently, even if the experts originate from the same domain.
- Finding a suitable performance metric is essential: The trade-off between algorithm performance and domain expert requirements needs to be considered, e.g. if it is required to match an exact point in time (e.g., the beginning of an event) or if it is better to detect an event more reliable but less accurate in time within a certain margin.
- General annotations of events by multiple different experts are expected to have a quite large variance. This is due to the fact that each expert has their own bias towards their field of expertise.

IV. Conclusion

In an iterative manner, the involvement of domain knowledge for the ChESS system was engaged through a constant exchange of ideas, issues, and solutions in all stages of the development, paving the way for a joint evaluation of the proof-of-concept study.

Lessons learned within the development process include the need for extensive data knowledge and early domain expert inclusion. The benefits of participation should be clear on both sides from the beginning. Real-time event detection in marine data streams works with agile perception, which means measuring not only where it matters, but especially when it matters. This has to be extracted from domain knowledge and differs depending on each individual expert, even if they originate from the same domain.

Combining technological advancements with the insights of domain experts through participatory approaches is a key point, not only for the development of real-time event detection systems but also for the development of other components of monitoring systems. It fosters a sustainable management of marine ecosystems, as it provides contextual understanding, and can improve scalability and reliability [14]. Such participative approaches, and therefore their lessons learned, are also relevant in the context of Digital Twins of the Oceans, supporting the informed decision-making process [15]. Future research in this context could also examine the ensemble with the counterpart of domain knowledge inclusion, which is explainability provided by the system itself, enabling a give-and-take basis.

Acknowledgment

We would like to thank all stakeholders and domain experts involved in this project, especially the researchers and workshop team within the Spiekeroog Coastal Observatory. Special thanks to Andrea Lübben, Carola Lehners, Axel Braun and Corinna Mori.

References

[1] R. Venkatesan, A. Tandon, E. D'Asaro, and M. A. Atmanand, Eds., *Observing the Oceans in Real Time*, ser. Springer Oceanography. Cham: Springer International Publishing, 2018.

[2] M. M. Idrees, F. Stahl, and A. Badii, "Adaptive learning with extreme verification latency in non-stationary environments," *IEEE Access*, vol. 10, pp. 127 345–127 364, 2022.

[3] G. Li and J. J. Jung, "Deep learning for anomaly detection in multivariate time series: Approaches, applications, and challenges," *Information Fusion*, vol. 91, pp. 93–102, Mar. 2023.

[4] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–37, Apr. 2014.

[5] M. Hammoodi, F. Stahl, and A. Badii, "Real-time feature selection technique with concept drift detection using adaptive micro-clusters for data stream mining," *Knowledge-Based Systems*, vol. 161, Aug. 2018.

[6] J. Demšar and Z. Bosnić, "Detecting concept drift in data streams using model explanation," *Expert Systems with Applications*, vol. 92, pp. 546–559, Feb. 2018.

[7] P. Kumar and M. Sharma, "Data, machine learning, and human domain experts: None is better than their collaboration," *International Journal of Human–Computer Interaction*, vol. 38, no. 14, pp. 1307–1320, Aug. 2022.

[8] O. Zielinski, D. Pieck, J. Schulz, C. Thoelen, J. Wollschläger, M. Albinus, T. Badewien, A. Braun, B. Engelen, C. Feenders, S. Fock, C. Lehners, K. Lõhmus, A. Lübben, G. Massmann, J. Meyerjürgens, H. Nicolai, T. Pollmann, K. Schwalfenberg, and H. Winkler, "The Spiekeroog Coastal Observatory: A scientific infrastructur at the land-sea transition zone (southern North Sea)," *Frontiers in Marine Science*, vol. 8, Feb. 2022.

[9] R. Reuter, T. H. Badewien, A. Bartholomä, A. Braun, A. Lübben, and J. Rullkötter, "A hydrographic time series station in the Wadden Sea (southern North Sea)," *Ocean Dynamics*, vol. 59, no. 2, pp. 195–211, Apr. 2009.

[10] J. Zenisek, F. Holzinger, and M. Affenzeller, "Machine learning based concept drift detection for predictive maintenance," *Computers & Industrial Engineering*, vol. 137, p. 106031, Nov. 2019.

[11] I. Paulenz, D. Lukats, J. Schneider, E. Berghöfer, F. T. Stahl, L. Nolle, and O. Zielinski, "Requirement analysis for an automated event detection system in marine environments, (in German)," in *Umweltinformationssysteme – Vielfalt, Offenheit, Komplexität*, F. Fuchs-Kittowski, A. Abecker, F. Hosenfeld, H. Ortleb, and M. Klafft, Eds. Wiesbaden: Springer Fachmedien, 2022, pp. 149–165.

[12] D. Lukats, E. Berghöfer, F. Stahl, J. Schneider, D. Pieck, M. M. Idrees, L. Nolle, and O. Zielinski, "Towards concept change detection in marine ecosystems," in *OCEANS 2021: San Diego – Porto*, San Diego, CA, Sep. 2021, pp. 1–10.

[13] "Grafana Technical Documentation," https://grafana.com/docs/, (accessed: Jan. 29, 2024).
[14] D. Kerrigan, J. Hullman, and E. Bertini, "A survey of domain knowledge elicitation in applied machine learning," *Multimodal Technologies and Interaction*, vol. 5, no. 12, p. 73, Dec. 2021.
[15] J. Schneider, A. Klüner, and O. Zielinski, "Towards digital twins of the oceans: The potential of machine learning for monitoring the impacts of offshore wind farms on marine environments," *Sensors*, vol. 23, no. 10, p. 4581, May 2023.