International Journal of Advance and Innovative Research
Volume 13, Issue 1 (XIII): January – March 2026
ISSN 2394 - 7780

A UNIFIED AI-DRIVEN KNOWLEDGE GRAPH PROTOTYPE FOR ENHANCING MARINE ECOLOGICAL RESILIENCE IN THE INDIAN EEZ

Ojaswini Nair[1], Prachi Mahajan[2], Sandhya Petchimuthu[3] and Meghana Nandala[4]
[1]Bachelor's in Data Science, Vidyalankar School of Information Technology, ojaswini.nair@vsit.edu.in
[2]Assistant Professor, Vidyalankar School of Information Technology, India, prachi.mahajan@vsit.edu.in
[3]Bachelor's in data science, Vidyalankar School of Information Technology, India, sandhya.petchimuthu@vsit.edu.in
[4]Bachelor's in Data Science, Vidyalankar School of Information Technology, India, meghana.nandala@vsit.edu.in

ABSTRACT
Even small increases in sea surface temperature (SST) are reshaping the distribution of commercially important fish species such as the Indian oil sardine and Indian mackerel, highlighting the growing ecological vulnerability of India's marine ecosystems under climate variability. Although India possesses extensive oceanographic observations and biodiversity records, much of this information remains fragmented across isolated repositories, limiting its integrated use for ecosystem-level assessment and policy planning.

This study presents a prototype Marine Knowledge Graph framework designed to bridge these data silos through semantic integration of heterogeneous marine datasets. The system incorporates real biodiversity and environmental DNA (eDNA) records alongside controlled environmental scenarios developed to evaluate system functionality and scalability. A modular extract–transform–load (ETL) pipeline standardizes oceanographic variables, taxonomic metadata, and molecular sequences before transforming them into an interconnected graph structure linking species, habitats, monitoring nodes, and climate parameters.

Species identification is implemented using a k-mer–based genomic similarity approach combined with L2 normalization, UMAP dimensionality reduction, and FAISS nearest-neighbor indexing. Instead of probabilistic classification, detection is performed through embedding-distance similarity matching. Functional validation confirms effective semantic querying, relational linkage, and genomic similarity retrieval within a unified computational environment.

The framework provides a scalable foundation for integrative marine biodiversity monitoring and climate-responsive ecosystem management.

Keywords: *Marine Knowledge Graph, eDNA, k-mer Analysis, UMAP, FAISS, Genomic Similarity Search, Data Integration, Climate Variability, Marine Biodiversity Monitoring*

I. INTRODUCTION
Marine ecosystems across the Indian Exclusive Economic Zone (EEZ) are increasingly influenced by climate variability, particularly rising sea surface temperatures (SST), which have been associated with shifts in species distribution and changes in ecosystem stability. With more than 7,500 kilometers of coastline and millions of people dependent on marine resources for livelihood and food security, responding to these environmental changes has become both an ecological and socio-economic priority.

India possesses one of the largest collections of marine datasets in the region, including extensive oceanographic observations, Argo float profiles, taxonomic inventories, and environmental DNA (eDNA) records. However, these datasets remain distributed across heterogeneous repositories, institutional silos, and incompatible data formats. This fragmentation limits cross-domain integration and restricts the ability to analyze relationships between physical oceanographic variability, molecular biodiversity signals, and species-level responses within a unified analytical framework.

Traditional relational database systems are effective for structured storage but are not optimized for representing highly interconnected ecological relationships. Queries linking climate metrics, habitat characteristics, and species associations often require computationally intensive joins, limiting flexibility and scalability. Knowledge graph architectures offer an alternative paradigm by modeling data as entities and relationships, enabling semantic interoperability and efficient traversal of complex ecological networks.

This study presents a prototype of Marine Knowledge Graph framework that integrates heterogeneous marine datasets into a unified semantic structure. The architecture combines real biodiversity and molecular records with controlled environmental scenarios to evaluate system integration and scalability. Genomic similarity-based species identification, implemented through k-mer feature extraction, dimensionality reduction, and nearest-neighbor search, is embedded within the graph environment to enable unified ecological exploration.

ISSN 2394 - 7780

The proposed framework establishes a computational foundation for climate-responsive marine monitoring and interoperable Blue Economy data systems.

II. LITERATURE REVIEW

A. Climate-Driven Volatility in the Indian Ocean

The increasing temperatures in the Indian Ocean are now considered a primary factor contributing to ecological instability. The Indian Council of Agricultural Research and the Central Marine Fisheries Research Institute conducted a study that found that the Sea Surface temperature along the Indian coastline has increased by .2 to .3 degrees Celsius over the past 45 years. Although these slight changes may seem insignificant, they can have serious impacts on marine organisms due to the fact that their internal body temperature is regulated by the temperature of the surrounding water. The basic metabolism, reproductive cycles, and geographical habitats of marine organisms can all be significantly disrupted if there are even minor variations in temperature.

Studies that have been specifically conducted on the Indian oil sardine have shown that there is a significant northward shift (from the traditional Malabar coast) to the waters off the coasts of Maharashtra and Gujarat.[16] In addition, there have been reports of the Indian mackerel moving into deeper waters in order to avoid the warmer surface temperatures. However, the limitation of existing studies on these shifts is that they are mostly based on retrospective data; therefore, there is a lack of real time predictive modelling for these types of movements combined with physical oceanographic data (i.e. coastal upwelling, mixed layer depth, etc.) that is resulting in increased ecological instability.

B. The Technical Crisis of Data Siloing

One prominent subject found in the literature since 2020 concerning Data Science within the ocean sector continues to be addressed; this is also known as ‘Silo Effect’. For example, while National Oceanographic Data Centre (NODC) and MoES each manage very large archives of information, often this data remains ‘trapped’ away in the various departmental databases [14]. The theoretical model of General Data Management describes the negative consequences of silos as ‘Data Decay’, where the same information (e.g. weather report) is maintained in some departments while remaining out of date within other departments, and ‘Redundant Effort’, whereby multiple groups within an organization are analysing and manipulating identical sets of data yet remain unaware of each other’s efforts [13]. Specifically with regards to the Indian maritime setting, there is further detrimental fragmentation as a result. A specific example could be, in the case of a researcher studying a sudden fish kill, they may have access to monitoring equipment that produces water quality readings, however the researcher may not know where to find the eDNA markers that can be used to identify toxic algae blooms [8]. The problem is compounded by traditional RDBMS or relational database management systems, as they require that all relationships between one million records must be established through many different join operations which can take up to several minutes, even hours, to complete; this time delay creates a significant barrier for achieving the real-time monitoring capabilities outlined in the objectives of The Blue Economy initiative [13].

C. Environmental DNA (eDNA) and Molecular Monitoring

The advent of eDNA metabarcoding has dramatically changed the way we evaluate biodiversity; this is accomplished by detecting species without capturing them physically. By collecting and analysing the genetic material from water samples, it is possible to detect rare, invasive and commercially important species with an exceptionally high degree of sensitivity [6]. Nonetheless, the 'Knowledge Gap' described by researchers at CMFRI is due to the absence of fully referenced eDNA databases specifically for the Indian Exclusive Economic Zone (EEZ) [16]. Furthermore, eDNA is typically treated as an independent molecular record, rather than being considered as a dynamic factor that changes according to variations in physical factors such as salinity and temperature. Our prototype attempts to use eDNA "fingerprints" as common points of access to a greater ecological network.

III. METHODOLOGY

A. System Architecture Overview

The proposed framework as indicated in **Figure 1** follows a multi-layered architecture designed to integrate fragmented marine datasets into a unified semantic structure. The system consists of four primary components: (i) Data Ingestion Layer, (ii) Knowledge Graph Engine, (iii) Analytics Layer, and (iv) Unified Interface.

ISSN 2394 - 7780

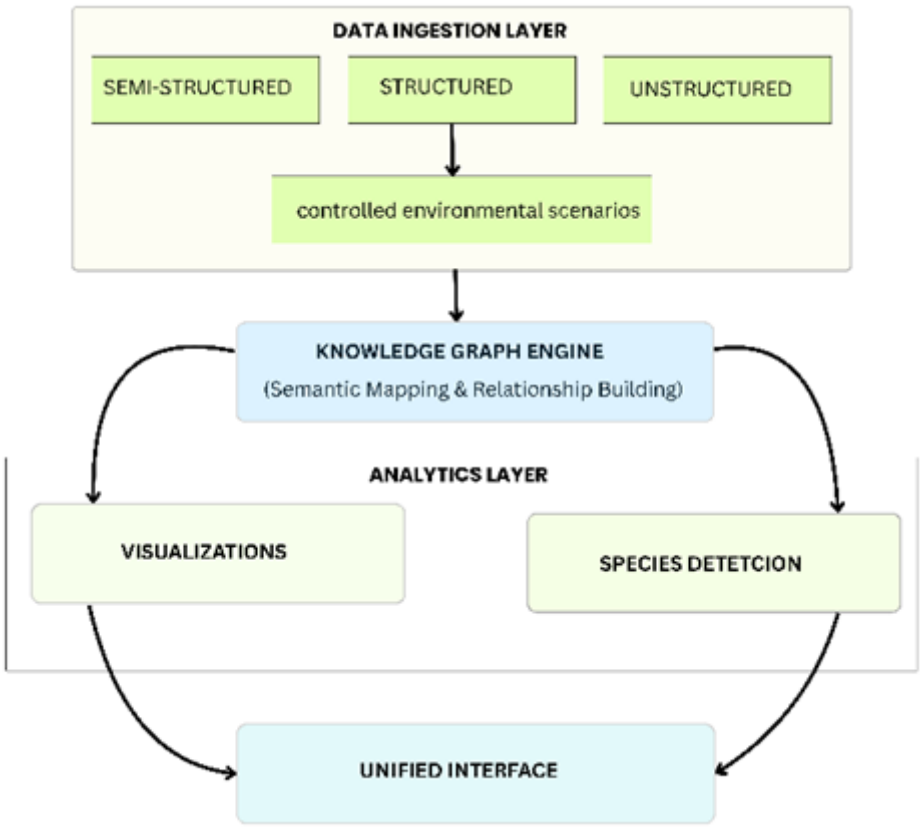

Figure 3: Model Framework

B. Data Ingestion Layer

The Data Ingestion Layer consolidates heterogeneous marine data sources, including biodiversity, molecular, and environmental datasets, into a unified framework. As these datasets originate from different repositories and formats, they exhibit structural inconsistencies that hinder integrated analysis. This layer standardizes, validates, and harmonizes data prior to integration into the Knowledge Graph.

Dataset Composition and Integration

The framework integrates real marine biodiversity datasets along with controlled environmental scenarios designed for system validation. This hybrid approach enables evaluation of integration capacity, relational updates, and stress-detection workflows while maintaining ecological plausibility.

Biological and Molecular Data

The following real-world datasets were incorporated:

- Environmental DNA (eDNA) sequence records[22]
- Taxonomic and species occurrence metadata[21][22]

Species identifiers were standardized to prevent duplication and resolve nomenclature inconsistencies. Data cleaning included format normalization and metadata alignment to ensure semantic consistency before graph construction.

Environmental and Spatial Data

Due to limited access to synchronized real-time monitoring systems, selected environmental parameters were generated to simulate realistic marine conditions for architectural validation. Controlled variables included:

- Sea Surface Temperature (SST) within realistic Indian Ocean ranges (26–30°C)
- Spatial redistribution patterns of species
- GPS-tagged sensor node coordinates
- Variations in species abundance for scalability testing

Incremental temperature anomalies were introduced to evaluate stress-response behavior. These controlled datasets were used solely to assess system robustness, scalability, and integration performance, not for empirical ecological conclusions.

ISSN 2394 - 7780

C. Knowledge Graph Engine

The Knowledge Graph Engine forms the core integration layer of the framework, transforming processed datasets into an interconnected semantic network. Instead of storing information in isolated tables, the system represents entities as nodes and their associations as relationships, enabling explicit modelling of ecological linkages.

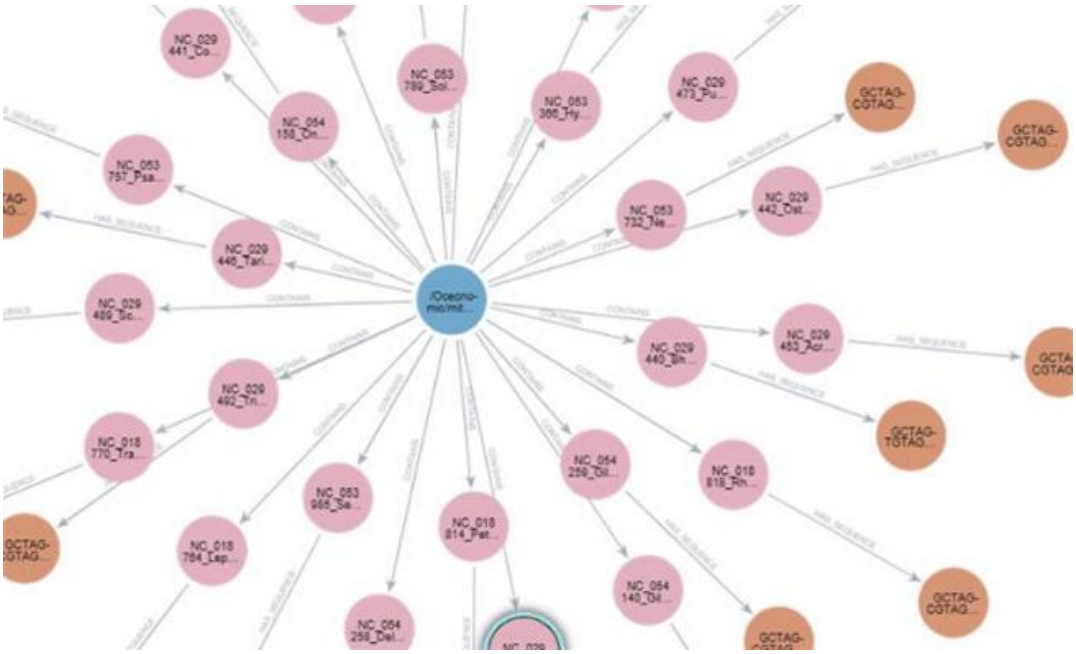

Figure 4: Conceptual Framework of Knowledge Graph

In the implemented graph (**Figure 2**), a central marine ecosystem entity ("Ocean Environment") serves as the core contextual node. Surrounding nodes represent species-level genomic records (NC_ identifiers), which are connected through relationships such as *CONTAINS* and *HAS_SEQUENCE*. Molecular sequence fragments are modelled as distinct nodes linked to their corresponding species records, enabling traceability from environmental context to genomic detail.

This structure demonstrates how biodiversity entities and molecular components coexist within a unified graph framework. Relationships encode ecological containment, sequence association, and hierarchical linkages, allowing intuitive traversal from ecosystem-level context to species and molecular attributes.

Although the current implementation primarily integrates structured biodiversity and molecular datasets, the architecture remains extensible to incorporate environmental variables and unstructured textual sources through entity extraction and semantic mapping.

D. Analytics Layer

The Analytics Layer combines two key components: a visualization module for exploratory analysis and a genomic similarity module for automated species identification. Together, these components allow the system to support both human interpretation and computational intelligence within the Knowledge Graph environment.

Visualization Module

The visualization module translates graph-derived data into an interactive network representation, making complex relationships easier to interpret. Data extracted directly from the Knowledge Graph are displayed as interconnected nodes and edges, representing ecosystem entities, species-level genomic records, and associated molecular sequences.

ISSN 2394 - 7780

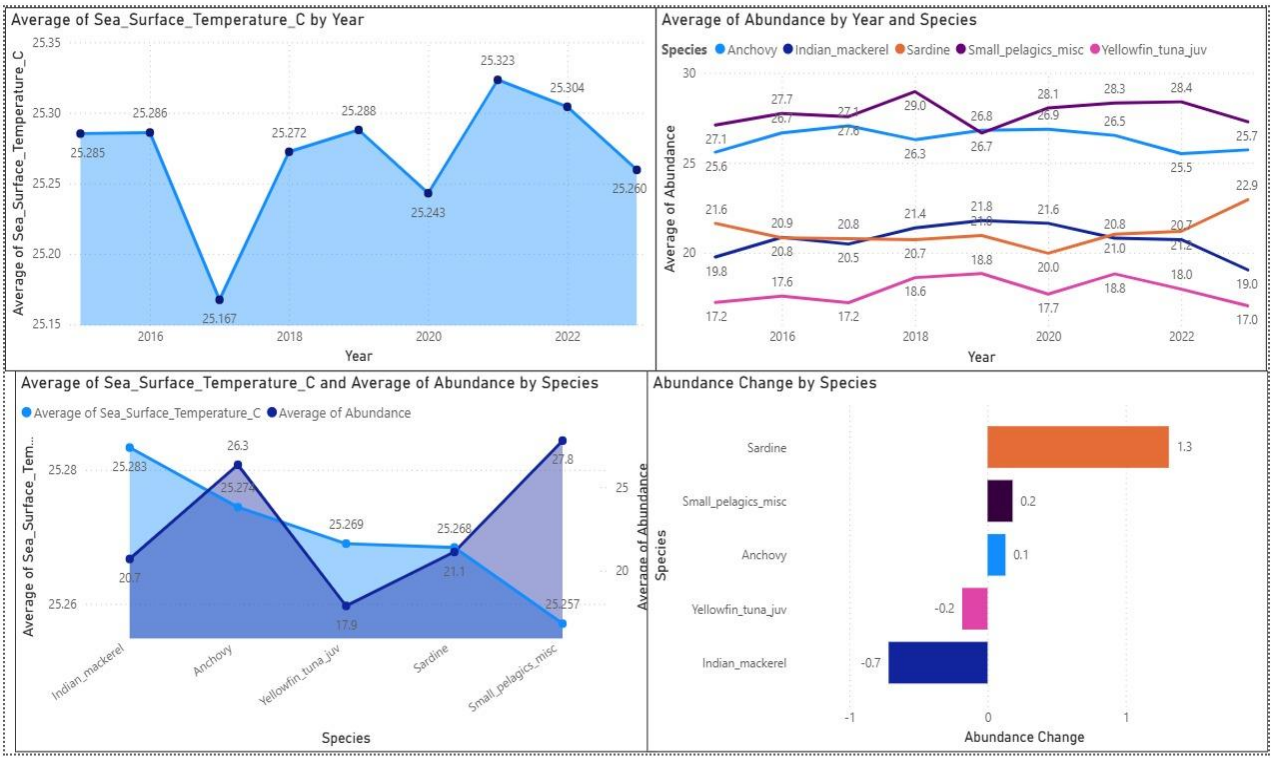

Figure 5: Visualization Generated

This visual representation helps in understanding how species and sequences are linked, examining connectivity patterns, and validating the overall graph structure. **Figure 3** illustrates the generated Knowledge Graph visualization

Genomic Similarity–Based Species Identification

Species identification was implemented using a similarity-based genomic framework integrated within the knowledge graph environment. Instead of supervised classification, detection was performed through sequence similarity in a learned embedding space. Each genomic sequence was decomposed into overlapping k-mers (k = 6) to generate frequency-based feature vectors using a global k-mer vocabulary. The vectors were L2-normalized to account for sequence length variability. Dimensionality reduction was applied using Uniform Manifold Approximation and Projection (UMAP) with cosine distance to preserve similarity relationships while reducing computational complexity. The resulting embeddings were indexed using FAISS with an L2 distance metric to enable efficient nearest-neighbour retrieval. For a query sequence, k-mer features were extracted, normalized, projected into embedding space, and compared against the indexed dataset. Species identity was assigned based on minimum embedding distance. Functional validation was conducted through query-based testing, confirming successful embedding transformation and nearest-neighbour retrieval within the indexed repository. These tests verified the operational integration of genomic similarity search within the broader knowledge graph framework.

E. Unified Interface Layer

The Unified Interface Layer provides a single platform, to view and interpret the system's outputs. It brings together the Knowledge Graph visualization (Oceanography Dashboard as shown in Figure 4) and the genomic similarity results (Species Prediction Model as shown in Figure 4) in one place.

The Oceanography Dashboard provides visual insights into marine environmental data derived from the Knowledge Graph, while the Species Prediction Model enables similarity-based genomic identification. By presenting both environmental visualization and molecular analysis in one place, the interface makes the system easier to navigate and interpret.

Figure 4 shows the interface, highlighting the integration of ecological monitoring and species prediction within a unified platform.

ISSN 2394 - 7780

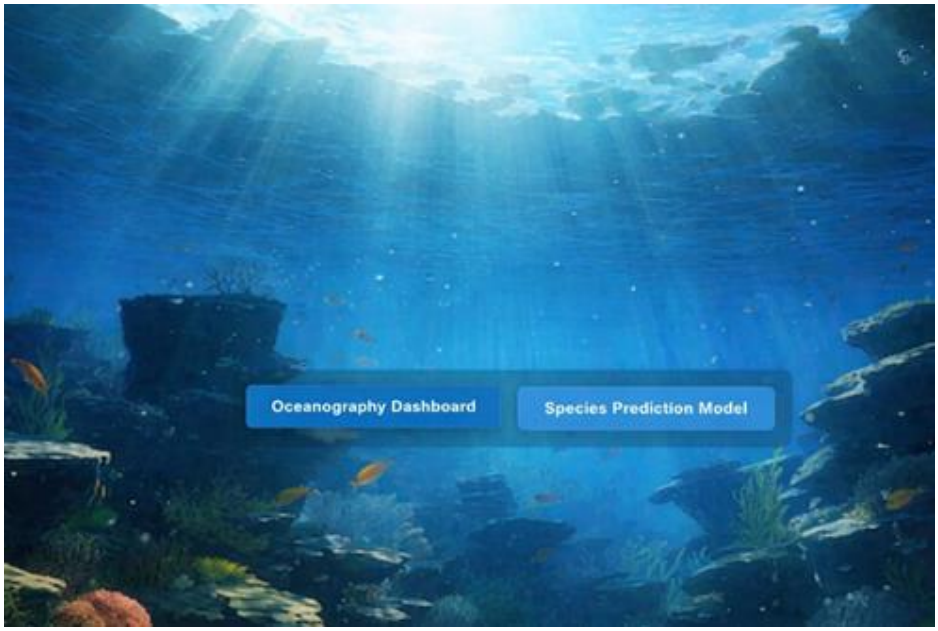

Figure 6: Unified Interface

V. RESULTS AND DISCUSSION

The proposed framework successfully integrates diverse marine datasets including biodiversity records, environmental variables, and genomic sequences into a unified Knowledge Graph structure. Through the ETL pipeline, entities such as species, habitats, monitoring stations, and climate parameters were standardized to ensure consistent representation and interoperability across domains. By modelling data as interconnected nodes and relationships, the graph enables multi-hop traversal across biological and environmental dimensions. For example, species occurrences can be examined alongside their associated Sea Surface Temperature (SST) exposure and linked genomic records within a single semantic environment. This interconnected structure demonstrates the practical advantage of graph-based modelling for ecological systems, where relationships are as important as the data itself. The framework also remains flexible, allowing additional datasets to be incorporated without extensive restructuring.

The genomic similarity module complements this structure by enabling molecular-level analysis within the same environment. Using k-mer–based feature extraction, genomic sequences were transformed into normalized embeddings, with UMAP preserving similarity relationships while reducing dimensionality. FAISS indexing supported efficient nearest-neighbor retrieval, allowing rapid identification of closest species matches based on embedding distance. Functional validation confirmed successful transformation, embedding, and retrieval of query sequences, demonstrating smooth integration of similarity-based detection within the Knowledge Graph. Although formal benchmarking was not conducted, operational testing verified consistent system performance.

Controlled SST scenarios were introduced to evaluate responsiveness to environmental variability. These simulations showed that the graph can dynamically represent changing climate conditions alongside species and genomic relationships. By modelling environmental variables as interconnected entities, the framework enables exploration of potential climate-linked redistribution patterns within a unified analytical space. Overall, the results demonstrate the feasibility of combining Knowledge Graph architecture with embedding-based genomic similarity search for integrated marine ecosystem analysis. While the current implementation serves as a prototype, it provides a scalable computational foundation for climate-aware marine monitoring and future Blue Economy data infrastructure.

VI. CONCLUSION

The “A Unified Marine Data Prototype” provides an integrated, intelligent ecosystem framework that supports our transition from compartmentalized data management to an interconnected, seamless marine data framework. The prototype utilizes a Knowledge Graph Architecture to integrate molecular biology, taxonomy, and physical oceanography into one massive database that captures the vast complexity and interactivity of the Indian Ocean. While still being developed, the integration of real-time environmental data with species identification offers an example of how to create an integrated national marine data infrastructure that can be utilized in a scalable manner by the Government of India to help manage marine biodiversity, support sustainable coastal economic growth, and face the increasing threats of climate change on coastal areas.

ISSN 2394 - 7780

VII. REFERENCES

1. Agarwala, N., & Saengsupavanich, C. (2023). Oceanic Environmental Impact in Seaports. *Oceans, 4*(4), 360-380. https://doi.org/10.3390/oceans4040025
2. Asif, M. (2023). Blue Economy and Power Politics in the Indian Ocean: Challenges and Opportunities. *Journal of Nautical Eye and Strategic Studies*, *2*, 2-37. https://doi.org/10.58932/mulg0003
3. Barbier, E. B. (2007). Valuing ecosystem services as productive inputs. *Economic Policy*, *22*(49), 177-229.
4. Bell, J. D. et al. (2013). Mixed responses of tropical Pacific fisheries and aquaculture to climate change. *Nature Climate Change*, *3*(6), 591-599.
5. Bindoff, N. L. et al. (2019). Changing Ocean, Marine Ecosystems, and Dependent Communities. *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*.
6. Chouhan, N., Dekari, D., Choudhary, B., Singh, A., & Gon Choudhury, T. (2023). Environmental DNA (eDNA) technology: Fisheries and aquaculture perspectives. *Indian Journal of Animal Health, Online*. https://doi.org/10.36062/ijah.2023.spl.02623
7. Convention on Biological Diversity. (2020). *Assessing Progress Towards Aichi Biodiversity Target 6 on Sustainable Marine Fisheries*. https://www.cbd.int/doc/publications/cbd-ts-87-en.pdf
8. Doyle, T. (2018). Blue Economy and the Indian Ocean Rim. *Journal of the Indian Ocean Region*, *14*(1), 1-6. https://doi.org/10.1080/19480881.2018.1421450
9. Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685-695.
10. Mahé, K., MacKenzie, K., Ider, D. et al. (2021). Directional Bilateral Asymmetry in Fish Otolith: A Potential Tool to Evaluate Stock Boundaries? *Symmetry*, *13*(6), 987. https://doi.org/10.3390/sym13060987
11. Mahé, K. et al. (2024). Otolith morphogenesis during the early life stages of fish is temperature-dependent. *Journal of Fish Biology*.
12. Mandal, A., & Ghosh, K. (2024). Artificial Intelligence in Aquatic Biodiversity Research. *Frontiers in Marine Science*.
13. Neo4j. (2024). *Graph Databases for Environmental Science*.
14. Rajurkar, T. (2025). *An Integrated Framework for Marine Biodiversity Assessment: Leveraging eDNA and AI*.
15. Shaikh, A., Rahman, W., Roksana, K., Islam, T., Rahman, M. M., Alshahrani, H., Sulaiman, A., & Reshan, M. S. A. (2025). An improved deep CNN-based freshwater fish classification with cascaded bio-inspired networks. *Automatika*, *66*(249-280). https://doi.org/10.1080/00051144.2025.2457803
16. Vivekanandan, E. (2011). *Impact of Climate Change on Indian Marine Fisheries and Options for Adaptation*. Eprints@CMFRI. https://eprints.cmfri.org.in/8432/1/VIVEKANANDAN_RALBAM_65-71.pdf
17. Wang, Y., Zhang, F., Geng, Z., Zhang, Y., Zhu, J., & Dai, X. (2023). Effects of Climate Variability on Two Commercial Tuna Species Abundance in the Indian Ocean. *Fishes*, *8*(2), 99. https://doi.org/10.3390/fishes8020099
18. Wee, A. K. S., Salmo III, S. G., Sivakumar, K. et al. (2023). Prospects and challenges of environmental DNA (eDNA) metabarcoding in mangrove restoration in Southeast Asia. *Frontiers in Marine Science*, *10*. https://doi.org/10.3389/fmars.2023.1033258
19. World Bank. (2011). *India marine fisheries: issues, opportunities and transitions for sustainable development*.
20. Xi, Y. et al. (2022). Knowledge graphs for scene understanding. *Taylor & Francis Online*. https://www.tandfonline.com/doi/full/10.1080/17538947.2025.2607168
21. Dataset Reference: https://ftp.ncbi.nlm.nih.gov/blast/db/
22. Dataset Reference: https://obis.org/data/access/