SMART INDIA HACKATHON 2025

TITLE PAGE

- **Problem Statement ID –** SIH25042
- **Problem Statement Title-** Identifying Taxonomy and Assessing Biodiversity from eDNA Datasets
- **Theme-** Miscellaneous
- **PS Category-** Software
- **Team ID-** 114829
- **Team Name-** The Deep Divers

IDEA TITLE

Problem Definition:

- ❑ The Deep sea eDNA Analysis is hindered by **poor reference database** coverage and slow, **alignment/mapping based tools** leading to **misclassification** and underestimation of biodiversity.
- ❑ Current pipelines like **QIIME2 , DADA2** and Mothur **fails** to detect **novel taxa** efficiently due to **poor** system **architecture.**

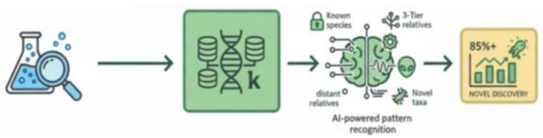
Unique Value Propositions (UVP) :

- ❖ **Faster processing -** MinHash + GPU(AI) acceleration
- ❖ **Deep-sea curated databases -** DeepSeaDB integration
- ❖ **Auto contaminant & bias correction** - Bayesian normalization
- ❖ **Dual deployment: Web + Python package**
- ❖ **interactive dashboards** – Plotly + Dash visualization
- ❖ **Pphylogenetic validation for novel taxa** - EPA-ng placement

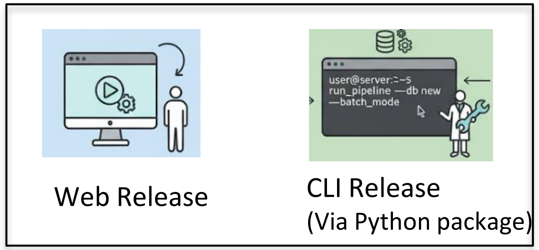
IDEA/Solution:

Implementation of an AI-driven pipeline for analyzing environmental DNA (eDNA) from deep-sea ecosystems

- ❑ Our solution uses **k-mer based database matching** with **AI-powered pattern recognition** to identify both known and novel deep-sea organisms from eDNA samples.
- ❑ The system uses a **3-tier approach**: **exact matching** for common species, **machine learning embeddings** for distant relatives, and **phylogenetic analysis** to characterize truly novel taxa— Can archive 85% accuracy with good novel detection rate than QIIME2.
- ❑ The system can offer dual deployment: a **Web app** for easy access and a **CLI support**(Via Python Package) for advanced, scalable analyses.



1.Flow of idea



```
user@server:~$
run_pipeline —db new
—batch_mode
```

Web Release

CLI Release
(Via Python package)

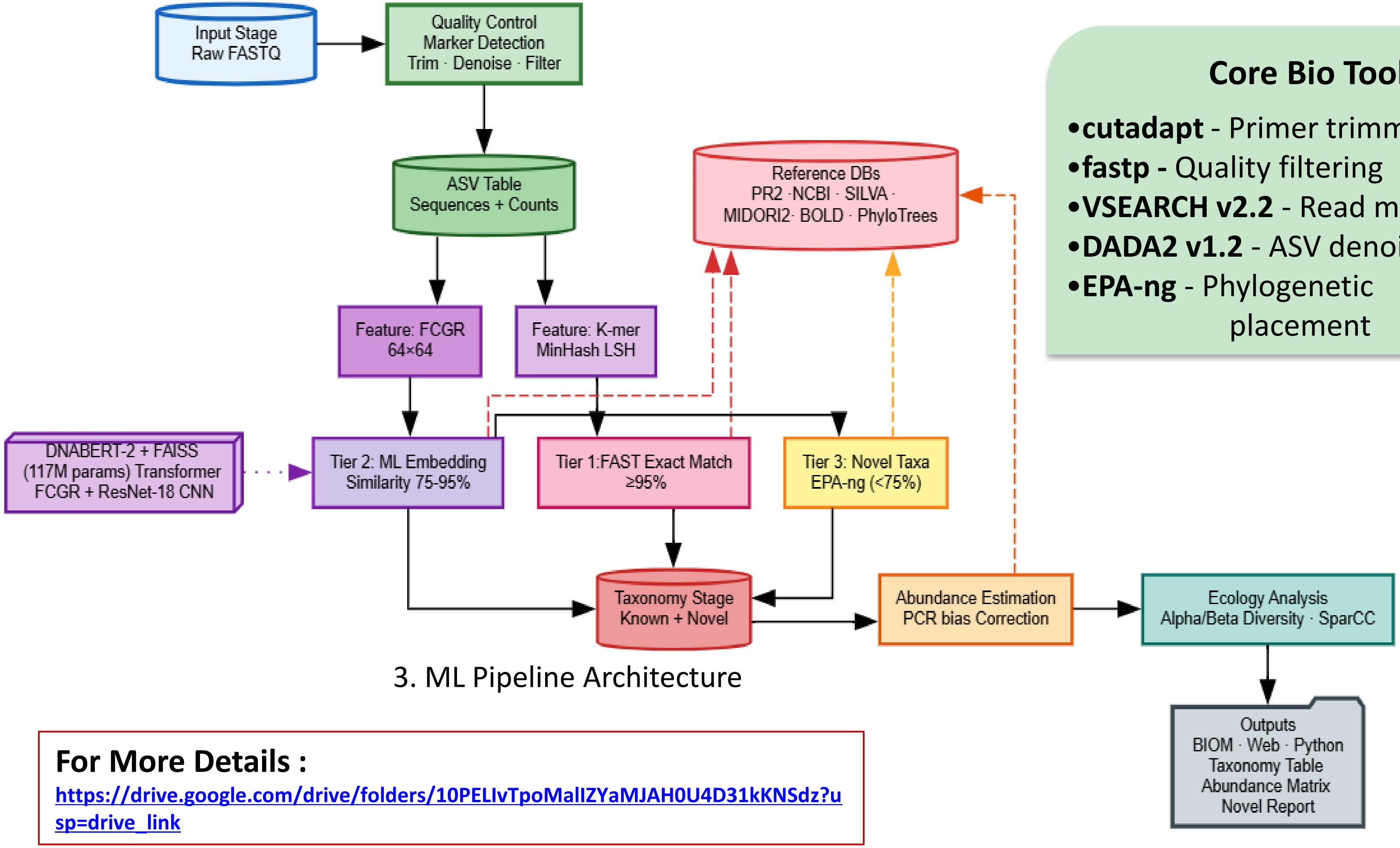
2. Release Methods

DEEP DIVERS

TECHNICAL APPROACH

SMART INDIA HACKATHON 2025

Process flow Architecture:



Input Stage
Raw FASTQ

Quality Control
Marker Detection
Trim · Denoise · Filter

ASV Table
Sequences + Counts

Reference DBs
PR2 ·NCBI · SILVA ·
MIDORI2· BOLD · PhyloTrees

Feature: FCGR
64×64

Feature: K-mer
MinHash LSH

DNABERT-2 + FAISS
(117M params) Transformer
FCGR + ResNet-18 CNN

Tier 2: ML Embedding
Similarity 75-95%

Tier 1:FAST Exact Match
≥95%

Tier 3: Novel Taxa
EPA-ng (<75%)

Taxonomy Stage
Known + Novel

Abundance Estimation
PCR bias Correction

Ecology Analysis
Alpha/Beta Diversity · SparCC

Outputs
BIOM · Web · Python
Taxonomy Table
Abundance Matrix
Novel Report

3. ML Pipeline Architecture

For More Details :
https://drive.google.com/drive/folders/10PELIvTpoMalIZYaMJAH0U4D31kKNSdz?usp=drive_link

Tech Stack:

Core Bio Tools

- **cutadapt** - Primer trimming
- **fastp -** Quality filtering
- **VSEARCH v2.2** - Read merging
- **DADA2 v1.2** - ASV denoising
- **EPA-ng** - Phylogenetic placement

AI/ML Framework

- **PyTorch 2.0+** -Deep Learning
- **Transformers 4.3** - DNABERT-2
- **FAISS 1.7.4** – Vector Search
- **scikit-learn 1.3.2** - ML utilities
- **R 4.3+** - Statistical analysis

Analysis Library

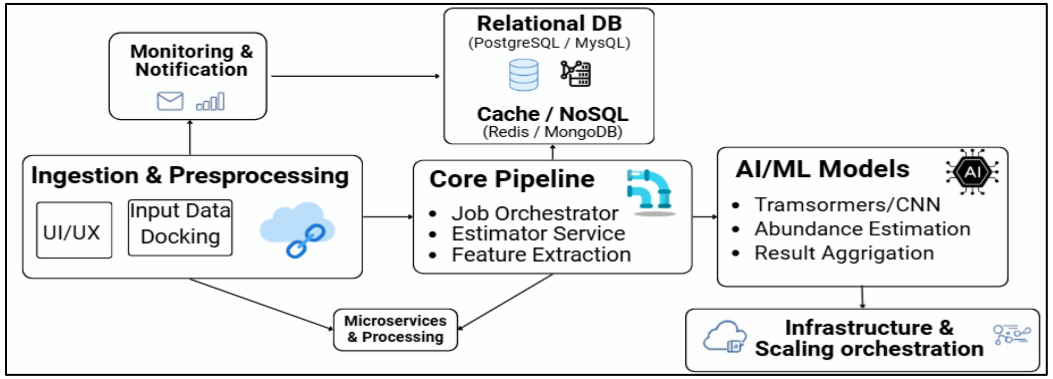
- **scipy** - Scientific computing
- **Statsmodels-** Statistical modeling
- **skbio** – SparCC
- **matplotlib** - Static figures
- **seaborn** - Statistical graphics

Model Support:. DNABERT-2 (GPU use, Max accuracy) or ResNet-18/FCGR CNN (CPU, lightweight)—ensuring accessibility from low-resource to High end devices.

FEASIBILITY AND VIABILITY

Analysis of the feasibility of the idea:

- ❑ **Technical**: Built on established bioinformatics tools (DADA2, VSEARCH) + modern AI (DNABERT-2/ResNet-18).
- ❑ **Accessibility :** Web app for easy access + CLI for advanced users who wants Custom Solutions; GPU or CPU fallback.
- ❑ **Scientific**: Validated eDNA + deep learning unlock hidden biodiversity patterns.
- ❑ **Market**: High-demand tool for research, conservation, and biotech sectors. Enables rapid deep-sea biodiversity assessment for CMLRE operations.
- ❑ **Impact**: Drives deep-sea discovery, conservation, innovation, and climate insights.



Monitoring & Notification

Relational DB (PostgreSQL / MySQL)

Cache / NoSQL (Redis / MongoDB)

Ingestion & Presprocessing
UI/UX
Input Data Docking

Core Pipeline
- Job Orchestrator
- Estimator Service
- Feature Extraction

AI/ML Models
- Tramsormers/CNN
- Abundance Estimation
- Result Aggregation

Microservices & Processing

Infrastructure & Scaling orchestration

4. Web Diagram

Challenges and risks:

Incomplete reference databases: Many deep-sea sequences remain unclassified.

High error & noise in eDNA samples: Risk of misidentification.

Insufficient labeled deep-sea **data** for robust model training.

Computational complexity: Processing massive datasets in real-time is challenging.

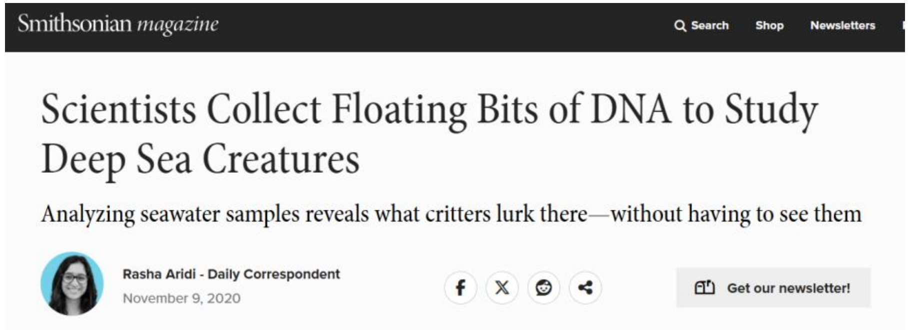
Strategies for overcoming these challenges:

- ❑ **3 tier AI-driven unsupervised learning** to reduce reliance on incomplete reference databases .
- ❑ **Noise filtering & error correction algorithms** to improve accuracy of eDNA reads.
- ❑ Multi-database curation + balanced sampling + data augmentation.
- ❑ **Optimized pipelines & cloud computing** for faster, scalable data processing. Python Release promotes customization on pieline(like QIIME 2)
- ❑ **Dual Support :**ResNet-18/FCGR (16-core CPU, 32GB RAM) or DNABERT-2 (32-core CPU + RTX 4090)—workstation to cloud scalable. - scalable from workstations to cloud.

IMPACT AND BENEFITS

Potential impact on the target audience:

1. **Researchers & Scientists :** Achieve faster discoveries with **minute-level eDNA analysis** and scalable workflows.
2. **Conservationists & Bio prospectors :** Unlock **novel biodiversity** to identify new species, genes, and bioresources.
3. **Policymakers & Regulators :** Use **bias-corrected, reliable data** for informed marine conservation policies.
4. **Deep-Sea Industries & Navigators :** Gain **validated biodiversity insights** for safer, data-driven operations.
operational reliability, environmental compliance, and resource planning.
5. **Research Innovation & Field Contribution :** Advance **deep-sea genomics and AI-biodiversity analytics**, setting new standards for marine ecosystem research.



Smithsonian *magazine*

Scientists Collect Floating Bits of DNA to Study Deep Sea Creatures

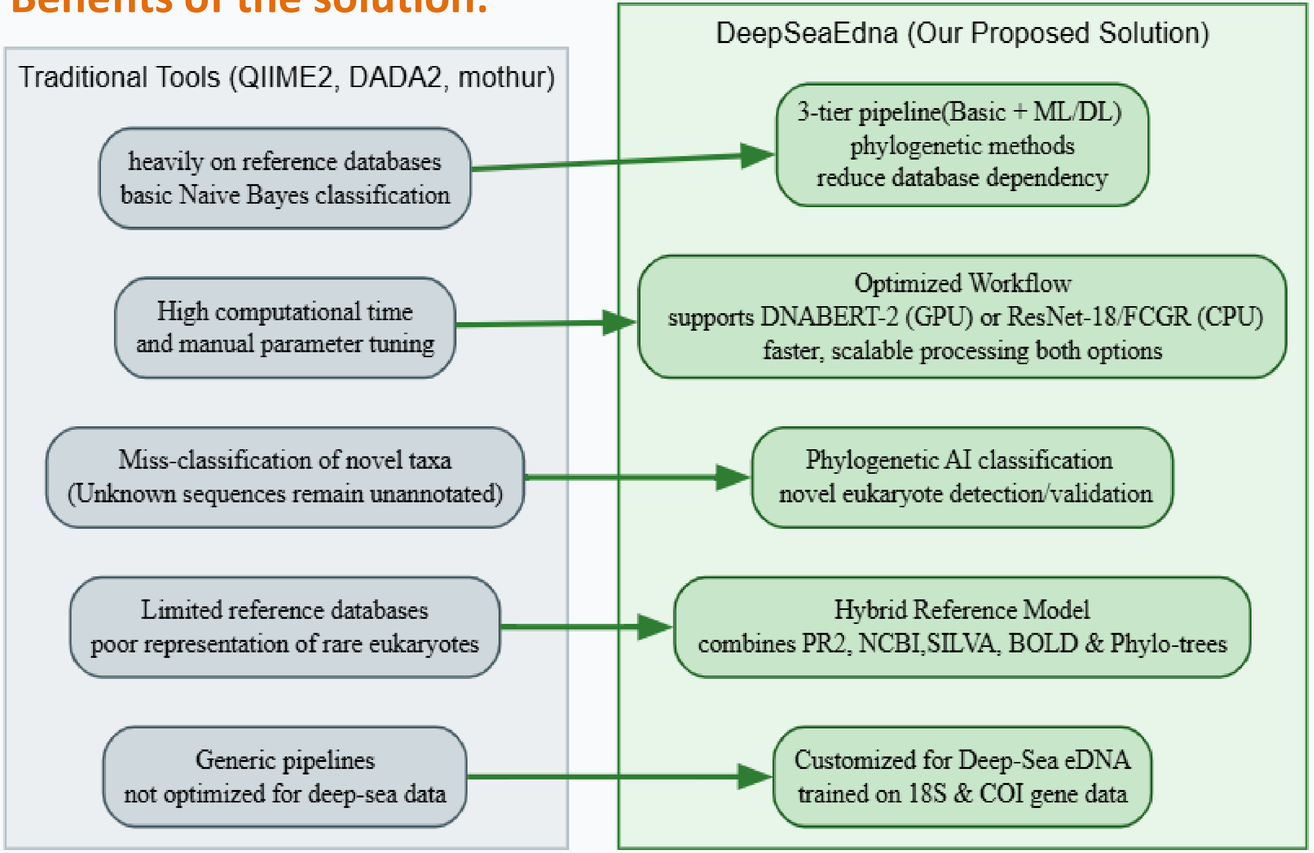
Analyzing seawater samples reveals what critters lurk there—without having to see them

Rasha Aridi - Daily Correspondent

November 9, 2020

5.Article

Benefits of the solution:



Traditional Tools (QIIME2, DADA2, mothur)
heavily on reference databases
basic Naive Bayes classification
High computational time
and manual parameter tuning
Miss-classification of novel taxa
(Unknown sequences remain unannotated)
Limited reference databases
poor representation of rare eukaryotes
Generic pipelines
not optimized for deep-sea data
DeepSeaEdna (Our Proposed Solution)
3-tier pipeline(Basic + ML/DL)
phylogenetic methods
reduce database dependency
Optimized Workflow
supports DNABERT-2 (GPU) or ResNet-18/FCGR (CPU)
faster, scalable processing both options
Phylogenetic AI classification
novel eukaryote detection/validation
Hybrid Reference Model
combines PR2, NCBI,SILVA, BOLD & Phylo-trees
Customized for Deep-Sea eDNA
trained on 18S & COI gene data

DEEP DIVERS

RESEARCH AND REFERENCES

SMART INDIA HACKATHON 2025

Our Work :

Github RepoLink : https://github.com/Jay9115/Deep-Divers-SIH-114829
Google Drive (Videos/Images of Implementation):
https://drive.google.com/drive/folders/1uEGvf2bzJVJp5MXNXttwbp4plhbcLvR9?usp=sharing

DataSet Links:

(1) Silva v138.2
(2) PR2 database v. 2.0.0
(3) EKOI
(4) MIDORI
(5) NCBI

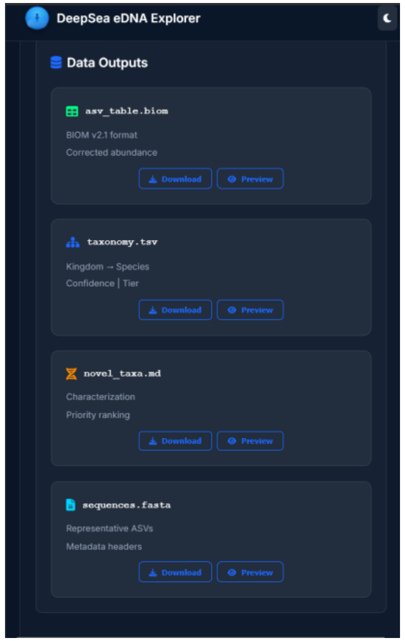
Research Papers:

(1) Aquatic environmental DNA: A review of the macro-organismal biomonitoring revolution
(2) North Atlantic deep-sea benthic biodiversity unveiled through sponge natural sampler DNA
(3) Unlocking natural history collections to improve eDNA reference databases
(4) Creating interpretable deep learning models to identify environmental DNA sequences
(5) Ji Y. et al. (2024) DNABERT-2: Transformer Models for Genomics.

Websites releted :

(1) https://www.unesco.org/en/edna-expeditions
(2) https://www.envirodna.com/solutions/biodiversity-assessments
(3) https://docs.qiime2.org/2024.10/tutorials/overview/- Background Study

Screenshots:


DeepSea eDNA Explorer

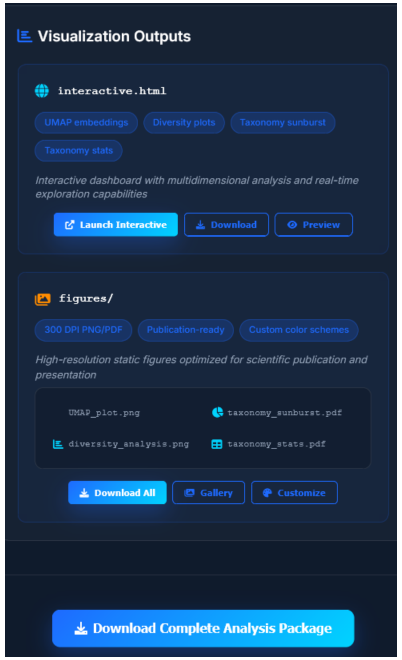
Data Outputs

asv_table.biom
BIOM v2.1 format
Corrected abundance

taxonomy.tsv
Kingdom → Species
Confidence | Tier

novel_taxa.md
Characterization
Priority ranking

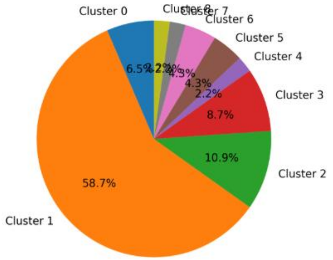
sequences.fasta
Representative ASVs
Metadata headers


Visualization Outputs

interactive.html
UMAP embeddings
Diversity plots
Taxonomy sunburst
Taxonomy stats
Interactive dashboard with multidimensional analysis and real-time exploration capabilities

figures/
300 DPI PNG/PDF
Publication-ready
Custom color schemes
High-resolution static figures optimized for scientific publication and presentation

UMAP_plot.png
taxonomy_sunburst.pdf
diversity_analysis.png
taxonomy_stats.pdf



- Prephase pipeline clustering Output for understanding