Hackathon Submission

1) Idea Title

Avalanche eDNA: A Hybrid AI + Bioinformatics Platform for Fast, Trustworthy Biodiversity Assessment

2) Idea Description

Environmental DNA (eDNA) enables the detection of organisms from trace genetic material in water, soil, and sediment samples. Yet today's eDNA workflows are either slow (manual BLAST-only pipelines) or opaque (ML-only black boxes). Avalanche eDNA bridges this gap with a hybrid platform that is fast, trustworthy, and easy to use.

What we built

- End-to-end pipeline that ingests FASTA/FASTQ, generates sequence embeddings using a state-of-the-art DNA language model (DNABERT■2), clusters similar sequences, assigns taxonomy via a KNN+LCA approach, and falls back to BLAST with lineage enrichment through NCBI taxdump.

- A modern Streamlit web UI with a Home page, Results Viewer, Run Browser, and Taxonomy Viewer, so biologists and stakeholders can explore results without command■line tools.

- Storage conventions and run management (F:\AvalancheData\datasets and F:\AvalancheData\runs) for reproducible, shareable analyses.

Why it matters

- Speed: ML embeddings accelerate similarity search and taxonomy suggestions; BLAST is invoked selectively, reducing overall runtime.

- Trust: When BLAST provides a taxid, the system prioritizes that lineage and records tie■breaks between KNN and BLAST, exposing conflicts rather than hiding them.

- Usability: A clear UI, recent■runs shortcuts, and a run browser turn complex analyses into a guided experience.

Differentiators

- Hybrid accuracy: Combines ML nearest■neighbor evidence (KNN+LCA) with authoritative BLAST taxids, producing richer and more reliable lineage.

- Explainability: Tie■breaker reporting and lineage provenance show why a call was made and which data source won.

- Scalability path: CPU■only works today; GPU or ONNX Runtime upgrades bring 10–100× embedding throughput for large datasets.

Impact & use cases

- Marine, freshwater, and sediment biodiversity surveys

- Invasive species monitoring and conservation prioritization

- Rapid triage of large field datasets with sharable, auditable outputs

Current status

- Fully working prototype on Windows; web UI launched locally; subset run identified 61 unique taxa with enriched lineage. GPU acceleration and run■to■run comparison are the next milestones.

3) Abstract/Summary

Avalanche eDNA is a hybrid AI + bioinformatics system for environmental DNA analysis that balances speed, transparency, and scientific rigor. The pipeline converts DNA sequences into numeric embeddings with a pretrained transformer (DNABERT■2■117M), enabling efficient similarity search and clustering. For taxonomy, the platform uses a KNN + lowest common ancestor approach over reference embeddings and integrates BLAST as a targeted fallback. When BLAST returns a taxonomic identifier (taxid), Avalanche resolves the full scientific lineage via NCBI taxdump and prioritizes this lineage over name■only ML matches; otherwise, ML assignments are used with conservative consensus. The system records tie■break decisions and flags potential conflicts, improving explainability.

A Streamlit web application makes results accessible: a Home page with navigation tiles and recent runs, a Results Viewer for pipeline summaries and visuals, a Taxonomy Viewer with filtering and downloads, and a Run Browser for fast discovery of analyses. Data and outputs are organized in standardized directories for reproducibility.

In tests on a 2,000■sequence subset from SRR35551197, the taxonomy step completed in ~31 seconds and identified 61 unique taxa. Embedding throughput on CPU measured ~3.8 sequences/second on a 512■sequence benchmark; GPU or ONNX Runtime can significantly accelerate this. Avalanche eDNA reduces time■to■insight for biodiversity assessments while maintaining lineage traceability and user■friendly review, supporting applications in monitoring, conservation, and rapid environmental triage.

4) Technology Bucket

- AI/ML: Transformer■based sequence embeddings (Hugging Face, PyTorch), KNN + LCA taxonomy
- Bioinformatics: NCBI BLAST fallback, taxid extraction, NCBI taxdump lineage resolver
- Data/Systems: FAISS (CPU) similarity index, reproducible run storage (datasets/runs)
- Web/UI: Streamlit front■end with Results Viewer, Taxonomy Viewer, and Run Browser
- Performance/Scale (roadmap): GPU mixed precision, ONNX Runtime (CPU), representative■only embedding, caching/deduplication