Extension of Hyperlyse for Scalable, ML-Supported Spectral Similarity Search

Table of Contents

1. Purpose of This Document

This specification defines the **scope, requirements, implementation order, and completion criteria** of the project extension for Hyperlyse.

Objectives

This document serves as a shared reference for all parties involved and establishes a common understanding of:

- **What** is to be implemented
- **In which order** it is implemented
- **How completion** is defined

2. Current State

Current Capabilities

The open-source software **Hyperlyse** currently supports:

- Loading individual hyperspectral image datasets (HSI cubes)
- Selecting individual pixels or image regions within a cube
- Extracting and visualizing spectral signatures from selected locations
- Creating and maintaining a **manually curated spectral reference database** by:
    - exporting selected spectra,
    - storing them in open spectral formats (e.g., `.jdx` files),
    - organizing them in a user-defined directory structure
- Comparing a selected spectrum with:
    - all pixels of the currently loaded HSI cube, or
    - spectra stored in the manually curated reference database

Identified Limitations

While scientifically valid, the current implementation has the following constraints:

- **Scalability**:
  Spectral comparison is not optimized for large numbers of spectra or large collections of image data.
- **Efficiency**:
  Reference spectra stored in the database are repeatedly loaded and reprocessed for each comparison.
- **Scope**:
  Similarity evaluation is limited to:
  - the currently loaded HSI cube, and
  - a manually curated reference database,
    without the ability to efficiently compare across multiple image datasets.

3. Target State (Objectives)

Project Goals

The objective of this project is to extend Hyperlyse to achieve the following:

- **Efficiency**: Spectral similarity computations become efficient and reusable
- **Scalability**: Spectral comparison across multiple HSI cubes becomes possible
- **Intelligence**: Machine learning is applied in a supporting and interpretable role to improve spectral similarity search
- **Continuity**: The existing user workflow remains unchanged

4. Guiding Principles

The project is governed by the following core principles:

Principle	Description
Reproducibility	Raw spectral data remains unchanged and authoritative
Transparency	All processing steps are explicit and documented
Machine Learning Usage	ML methods support and improve similarity search while remaining interpretable and methodologically transparent (e.g., dimensionality reduction or feature-space transformation)
Workflow Stability	No changes to the user interface or interaction model
Interpretability	Results must remain scientifically understandable

5. Phased Implementation Plan

Implementation is carried out **sequentially** in clearly defined phases to ensure systematic progress and quality control.

Phase 0: Alignment & Baseline

Objectives:

- Align scope and responsibilities with all parties
- Freeze the current Hyperlyse version as reference
- Document baseline computation time and behavior

Phase 1: Architectural Preparation

Objectives:

- Analyze existing comparison and data-flow logic
- Introduce a centralized spectral preprocessing function
- **No functional behavior changes**

Phase 2: Reuse & Caching

Objectives:

- Preprocess reference spectra once
- Reuse processed representations across queries
- Demonstrate measurable performance improvements

Expected Outcome: Improved efficiency through computation reuse

Phase 3: ML-Supported Feature Space Transformation

Objectives:

- Introduce an interpretable machine-learning-based transformation of the spectral feature space
- Apply similarity search within the transformed feature representation
- Document the effects on robustness, comparability, and runtime behavior

Expected Outcome: Enhanced similarity search accuracy and robustness

Phase 4: Cross-Cube Comparison

Objectives:

- Prepare a shared search structure across multiple HSI cubes
- Enable comparison of one selected spectrum against many scans

Expected Outcome: Multi-dataset analysis capability

Phase 5: Validation & Documentation

Objectives:

- Perform before/after performance benchmarks
- Produce technical documentation
- Prepare clean project handover

Expected Outcome: Complete project documentation and validated improvements

6. Functional Requirements (FR)

ID	Requirement	Description
FR1	Canonical Spectrum Representation	The system **shall** provide a unified preprocessing pipeline that converts raw spectra into fixed-length numeric vectors
FR2	Reuse of Reference Spectra	Reference spectra **shall not** be reprocessed for every query; preprocessing results must be reused
FR3	Formal Similarity Search	Spectral similarity search **shall** be formulated as a distance-based, KNN-style process
FR4	ML-Based Evaluation	An interpretable ML-based transformation of the spectral feature representation shall be applied to support and improve spectral similarity search
FR5	Cross-Cube Comparison	A selected spectrum **shall** be comparable against spectra originating from multiple HSI cubes
FR6	Measurable Time Savings	The system **shall** demonstrate a measurable reduction in computation time for repeated similarity queries

7. Non-Functional Requirements (NFR)

ID	Requirement	Description
NFR1	No UI Changes	The existing user interface and interaction model shall remain unchanged
NFR2	Scientific Traceability	All processing steps must remain documented and scientifically explainable
NFR3	Data Integrity	Original `.jdx` files remain the authoritative data source; derived representations are secondary artifacts

8. Acceptance Criteria

The project is considered complete when **all** of the following criteria are met:

- Reference spectra are no longer recomputed per query
- A centralized spectral preprocessing pipeline exists
- An interpretable ML-based feature transformation is applied to support similarity search
- Cross-cube spectral comparison is technically possible
- Computation time savings are measured and documented