SPHEREx image of Cygnus-X Star Forming Region

Classifying SPHEREx Sources Using Machine Learning

Harvard & the Legacy of Slavery Initiative

Presented by: Abdul-Majeed Hamidu

Mentor: Joseph Hora

Program: DBSP/CFA

SPHEREx Mission

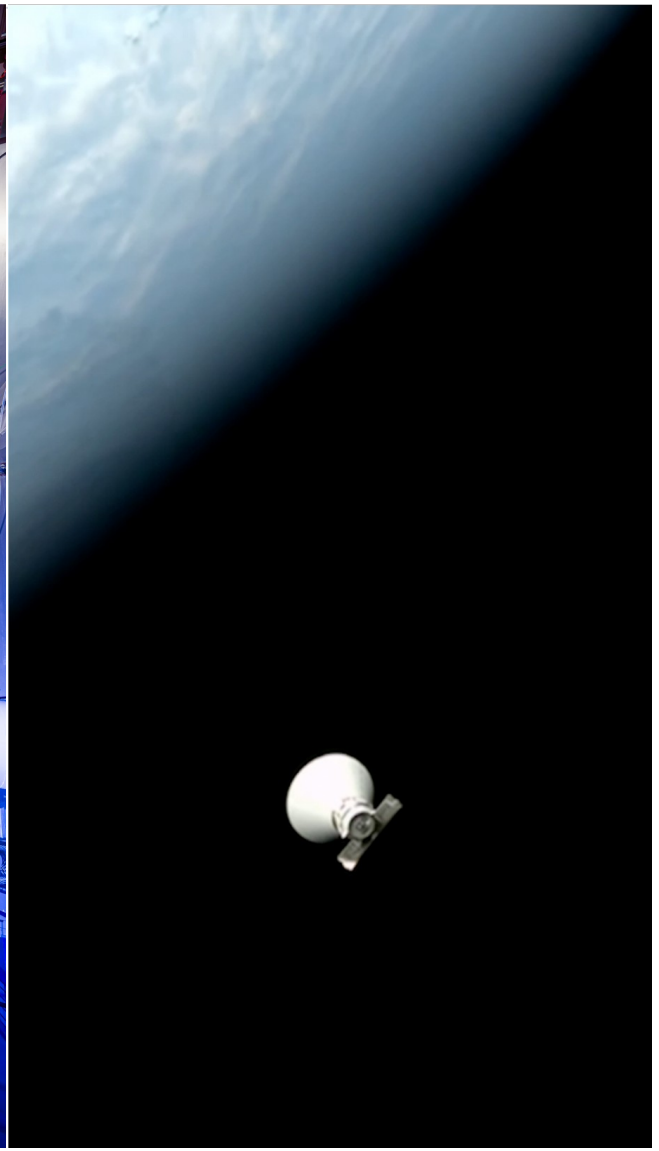
What SPHEREx Does

- Maps the whole sky
- near-infrared light
- Studies galaxies and stars
- Measures spectra

Why it matters here

- Finds young stars
- Helps tell stars apart

SPHEREx Before Launch



SPHEREx in Space
(Launched March 11, 2025)

Source Classes

Key terms

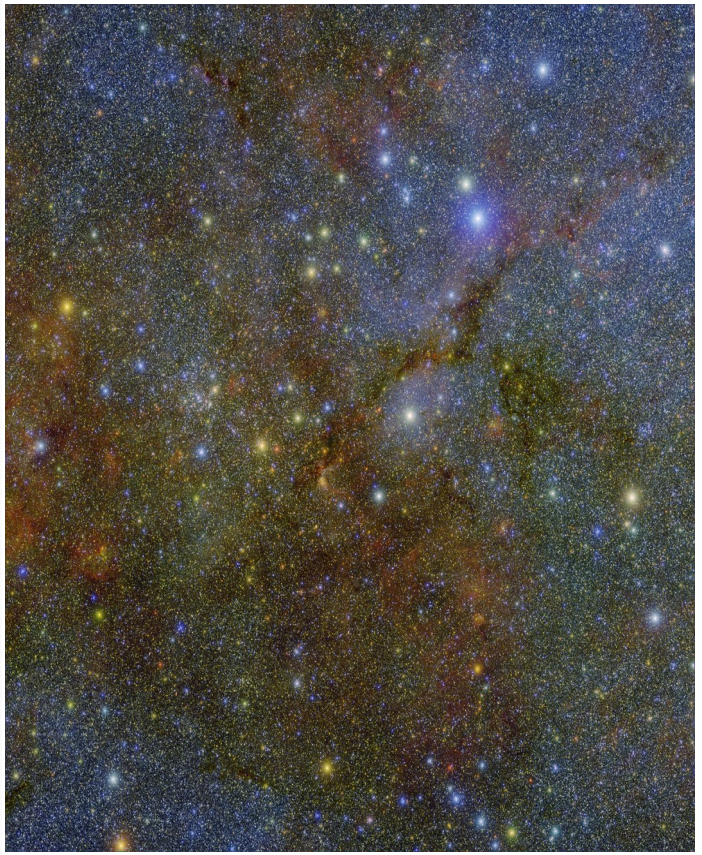
- **Young Stellar Object (YSO):**

  A young star forming inside gas and dust
- **Active Galactic Nucleus (AGN):**

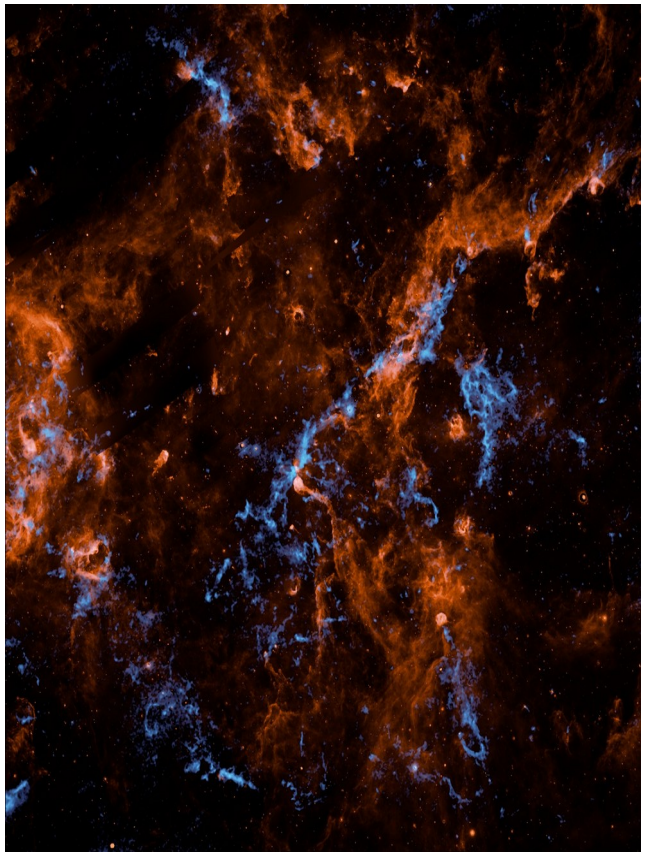
  A bright galaxy center powered by a black hole
- **Asymptotic Giant Branch Star (AGB):**

  An older, cool star
- **Main-Sequence Star Temperature Template (ms_t):**

  Normal star templates labeled by temperature

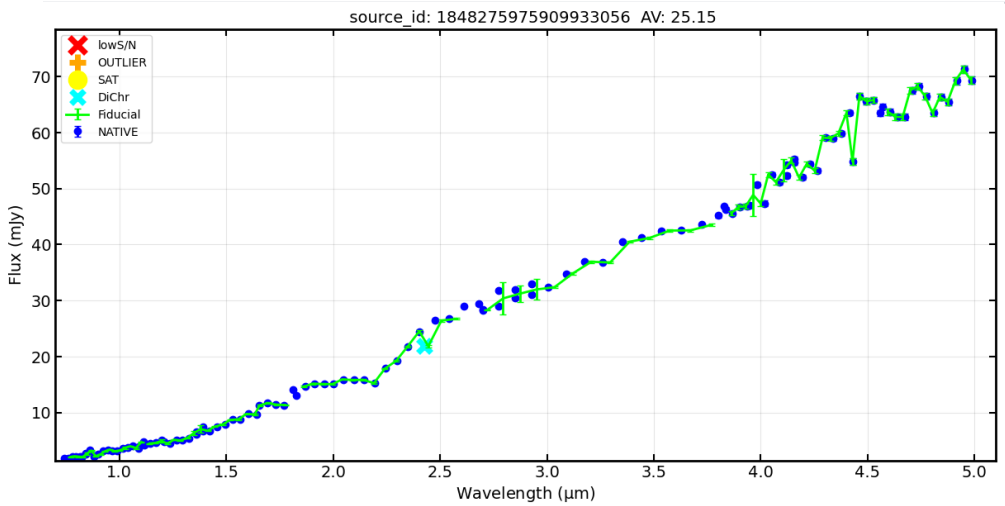


SPHEREx image of cygnus-x star forming region

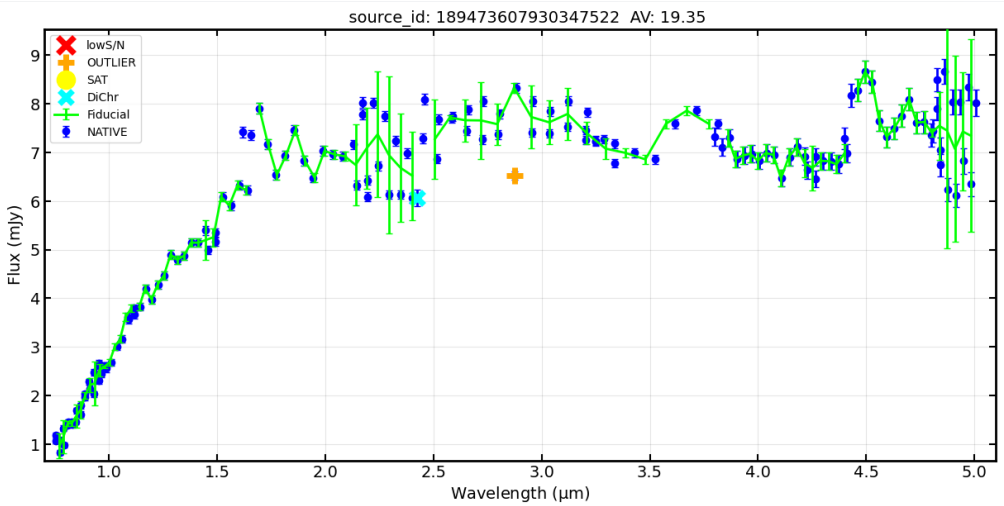


SPHEREx image of interstellar Ice in Milky Way

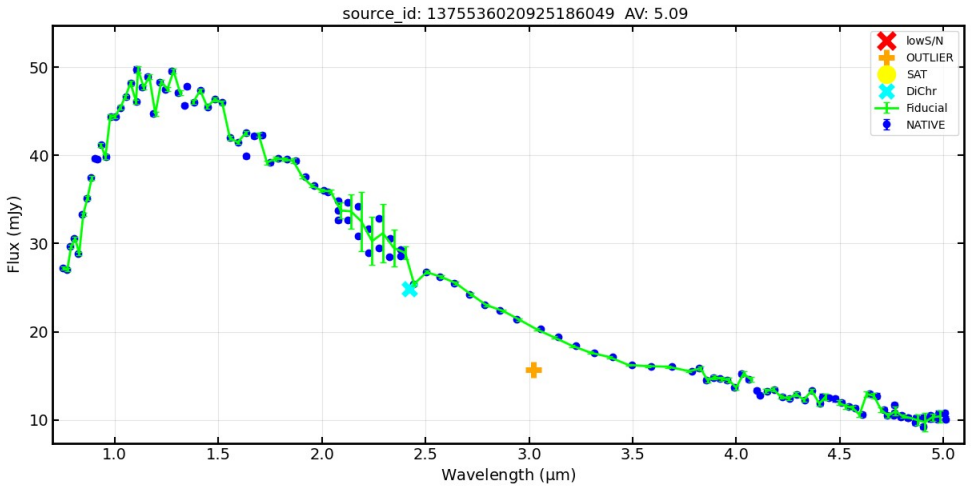
Representative Spectra Used for Classification I



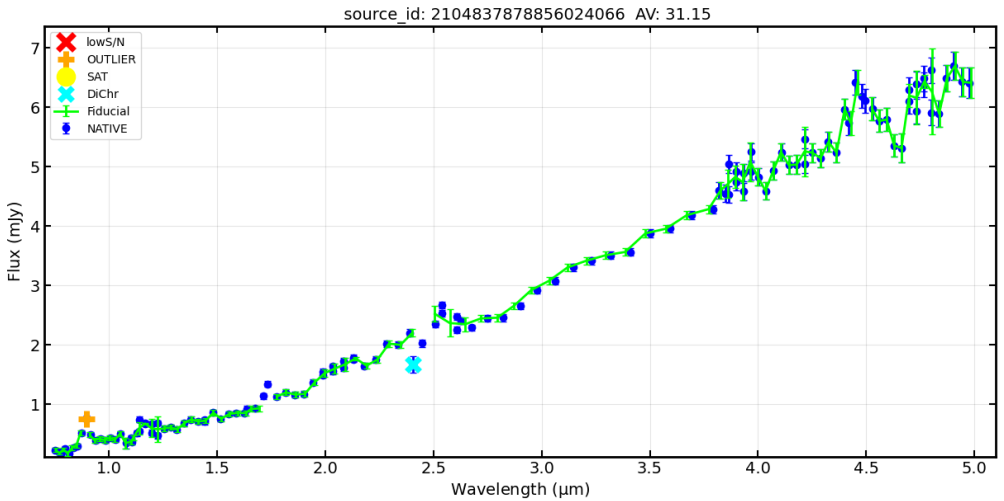
CIYSO Spectrum



C2YSO Spectrum

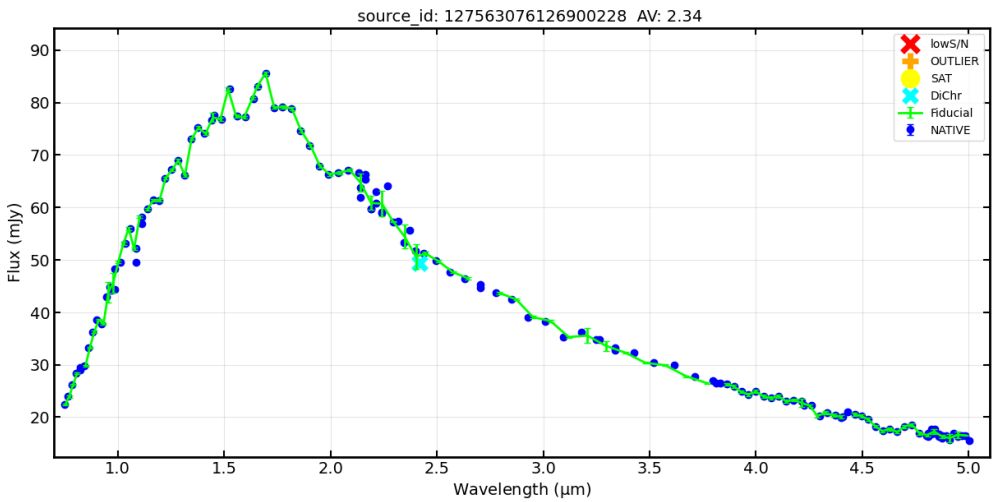


C3YSO Spectrum

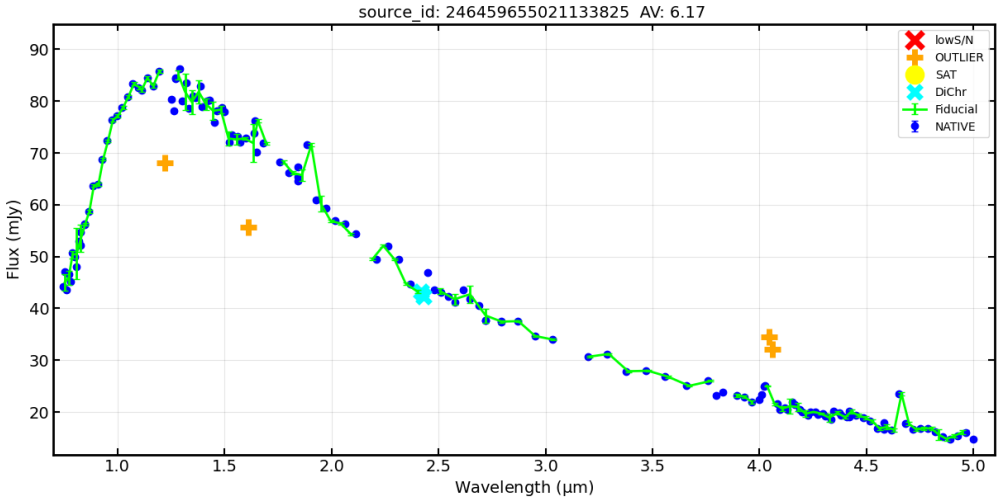


AGN Spectrum

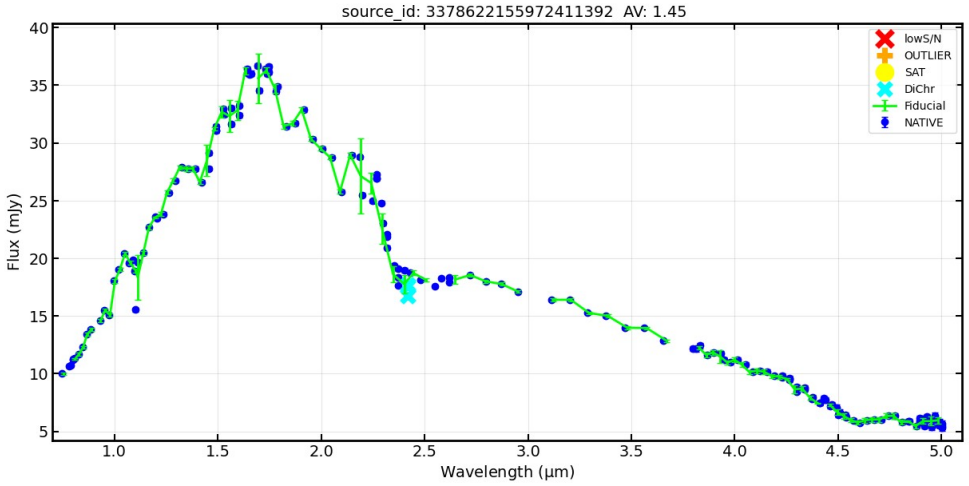
Representative Spectra Used for Classification II



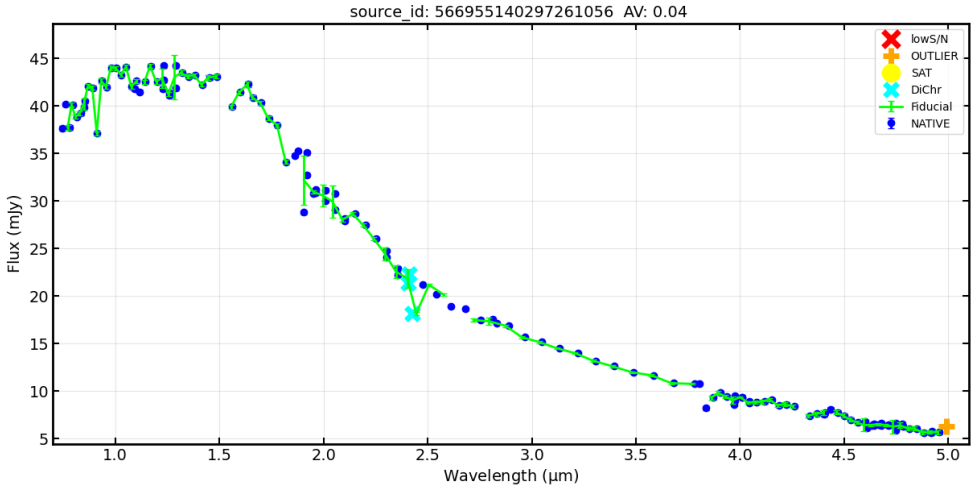
CAGB Spectrum



OAGB Spectrum



ms_t3500 Spectrum



ms_t5500 Spectrum

Classification problem

The Challenge

- Too many spectra
- Checking by hand is slow
- Young stars can look like other objects

The Goal

- Train a machine learning model
- Use spectra we already labeled
- Let the model classify new objects
- Sort objects into the right groups

Dataset

SCIENCE SPECTRA IN CURRENT RUN

5,000

LABELED TEMPLATES

4,611

NEW UNLABELED FILES

93,686

Inputs Used by the Pipeline

- FITS files
- Flux values and Uncertainty
- Wavelength data
- Extra ice features when available:

H2O, CO2, CO

Class Labels in the Templates

YSO groups: C1YSO, C2YSO, C3YSO

Contaminants: AGN, CAGB, OAGB

Stars: ms_t3500 to ms_t6500

Principal Component Analysis (PCA) and ice-depth features

PCA COMPONENTS

83

Components retained by the run

ICE FEATURES

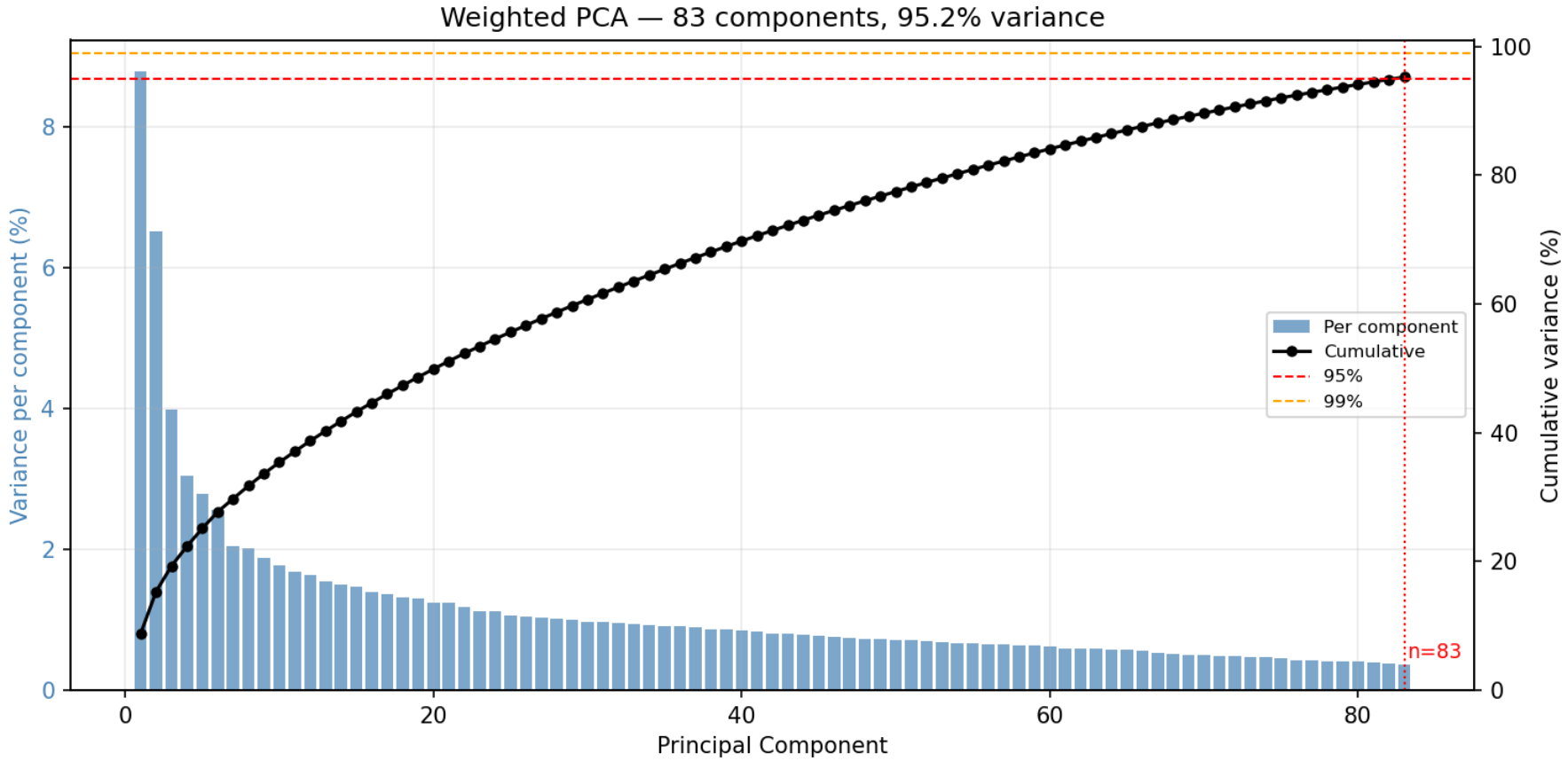
3

H2O, CO2, and CO optical depths

FINAL FEATURE VECTOR

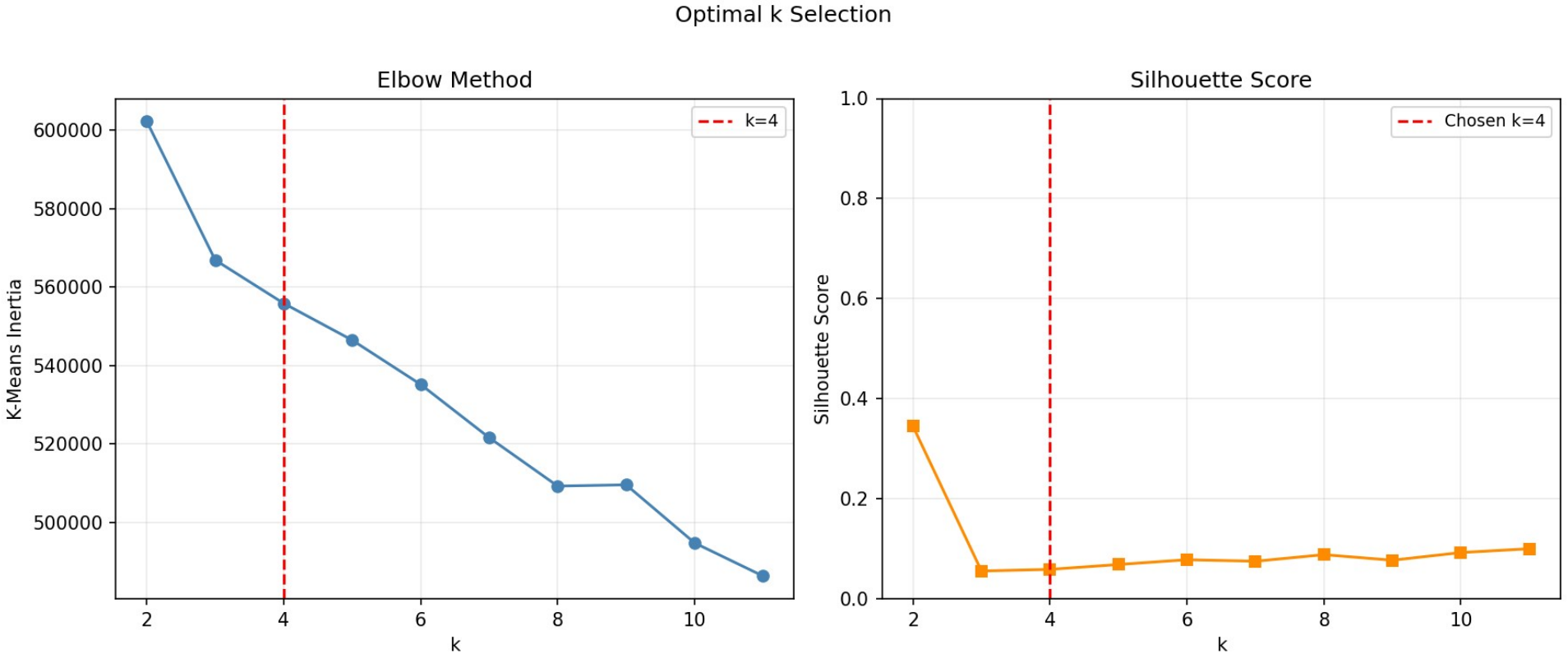
86

83 PCA components + 3 ice depths



EXPLORATION

Unsupervised learning



- Found 4 clusters
- Silhouette score was low: 0.059
- The groups are not clearly separated

MODEL COMPARISON

Classical models

Models tested during the project

Random Forest Classifier

- Uses many decision trees
- Each tree votes for a class
- Easy to compare and understand

Gradient Boosted Trees

- Builds many decision trees one after the other
- Each new tree fixes past mistakes
- Strong for table-style data

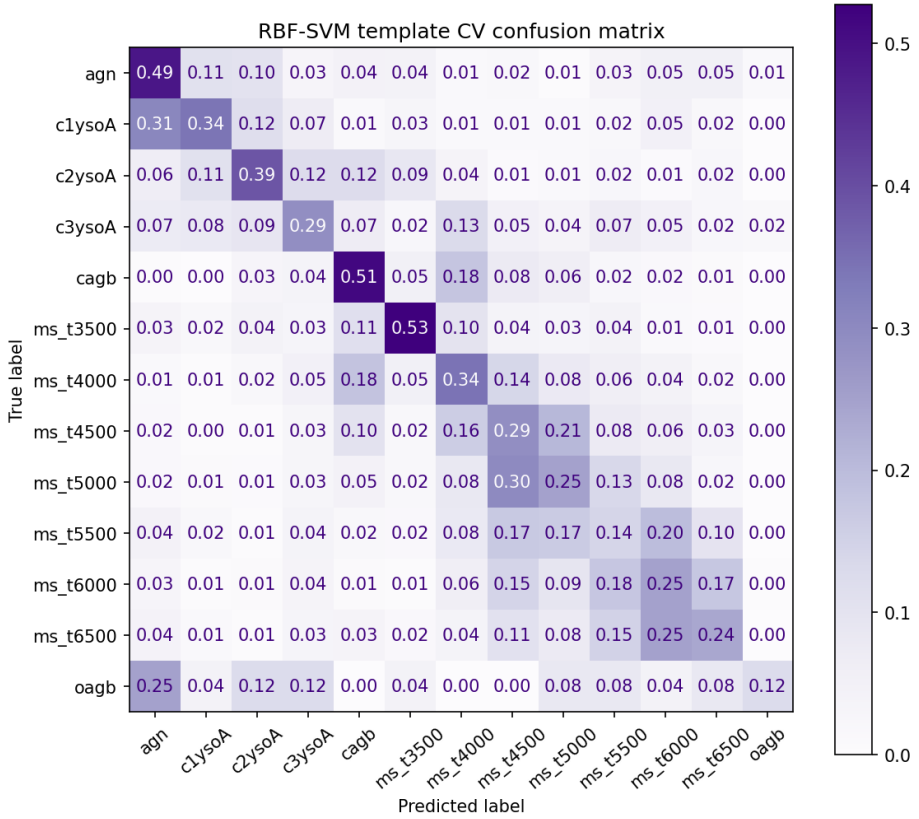
Radial Basis Function Support Vector Machine (RBF-SVM)

- Finds curved boundaries between classes
- Good when groups overlap
- Helps separate objects that look similar

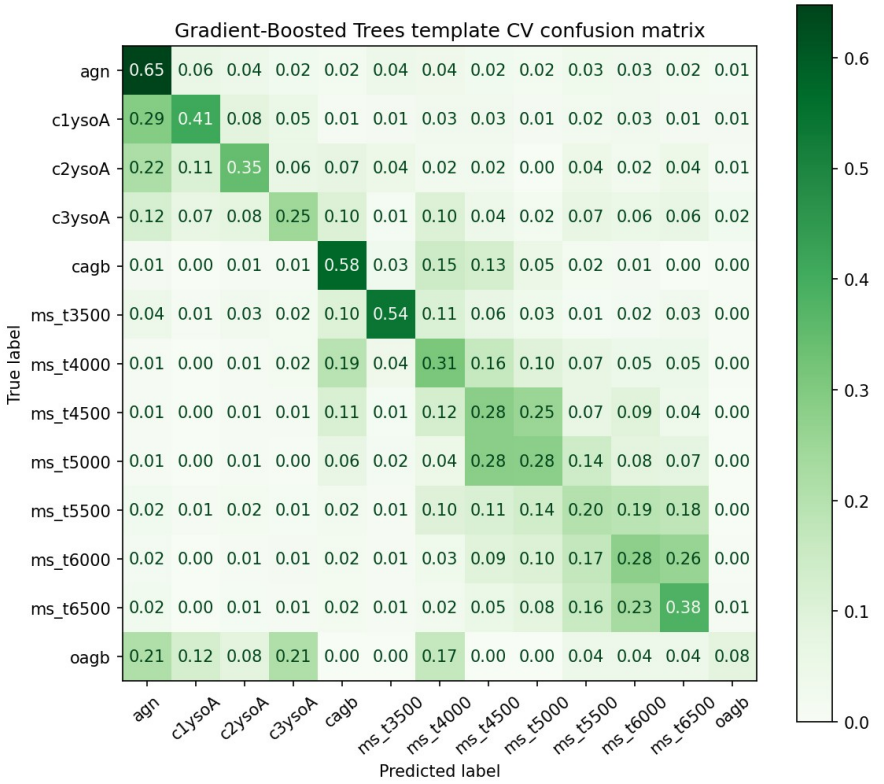
Convolutional Neural Network (CNN)

- Looks directly at the spectrum shape
- Learns patterns by itself
- Finds small and broad spectral features

RBF-SVM and Gradient-Boosted Trees template CV confusion matrices show the performance gap



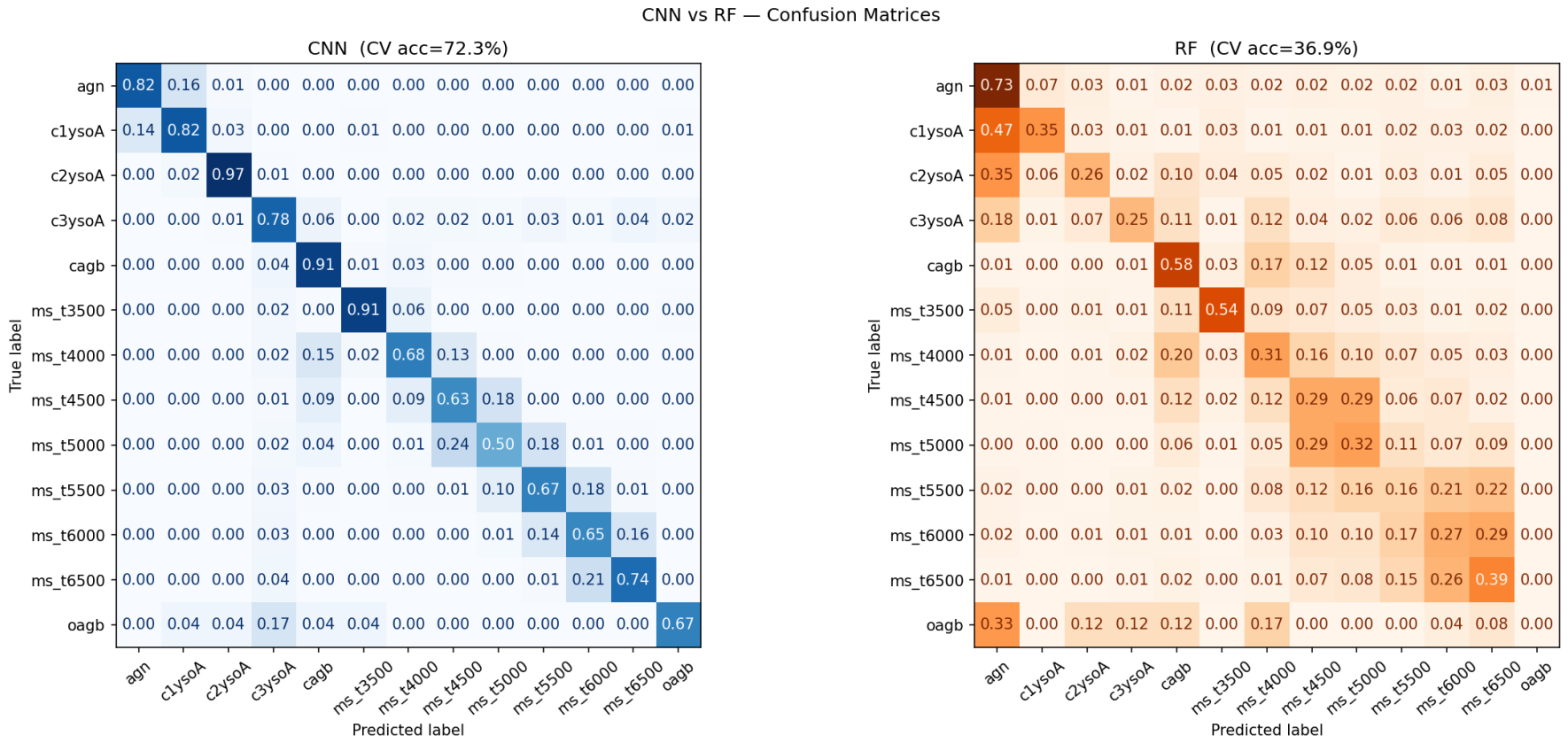
RBF-SVM confusion Matrix



Gradient-Boosted confusion Matrix

The stronger diagonal shows more correct class recovery.

CNN and Random Forest confusion matrices show the performance gap



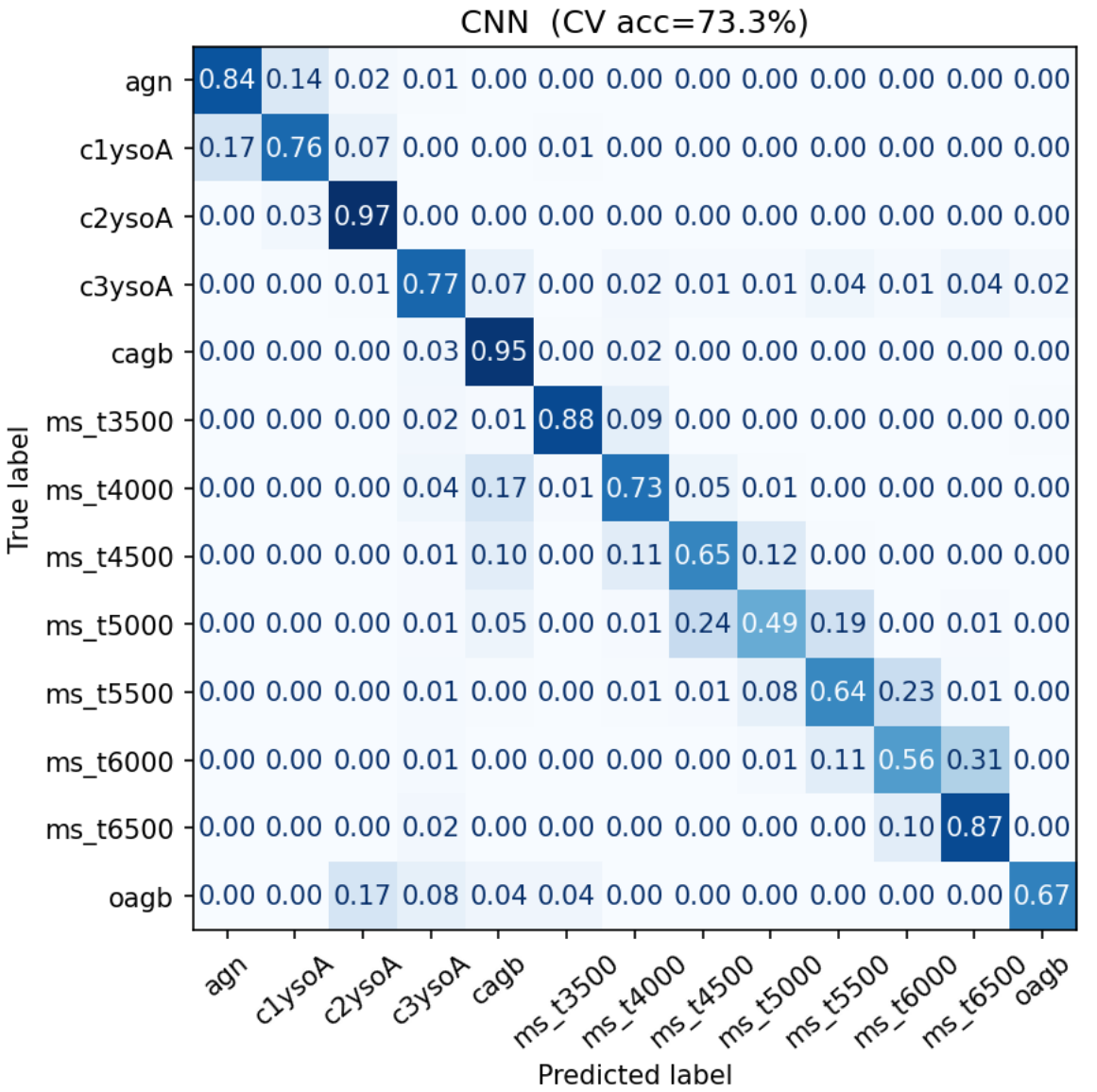
CNN CV accuracy = 72.3%, compared with RF CV accuracy = 36.9%. The stronger diagonal shows more correct class recovery.

BEST MODEL

CNN Model Improvement

Improved template data before training

- Regrouped C1YSO, C2YSO, and C3YSO into smaller subgroups
- Selected only the cleanest and best-looking spectra
- Used these high-quality spectra as the model templates



APPLYING THE CLASSIFIER

Model Outputs

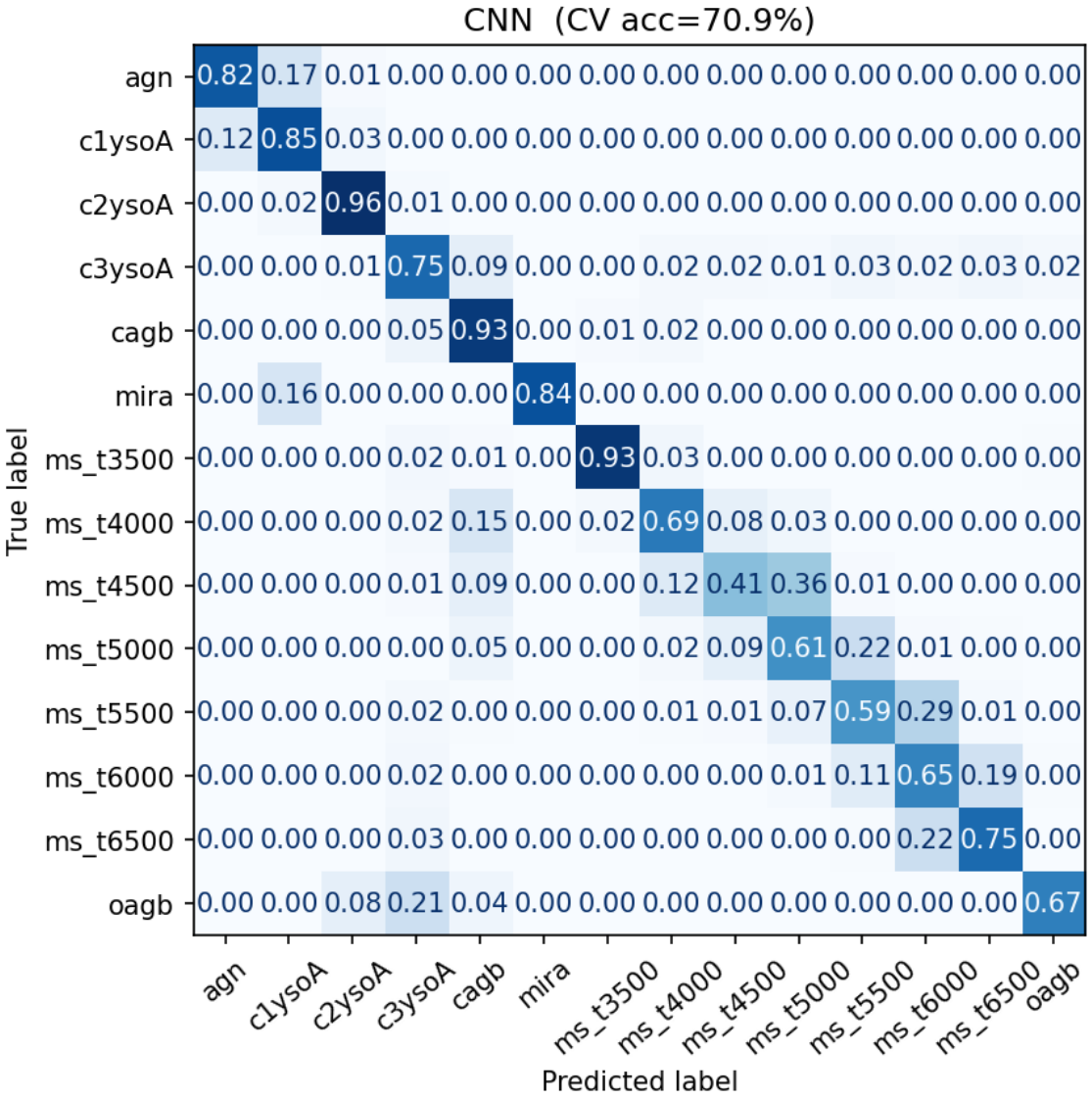
- Takes unlabeled FITS files as input
- Processes spectra in safe batches
- Outputs predictions as a CSV and grouped class list

Example Grouped Output

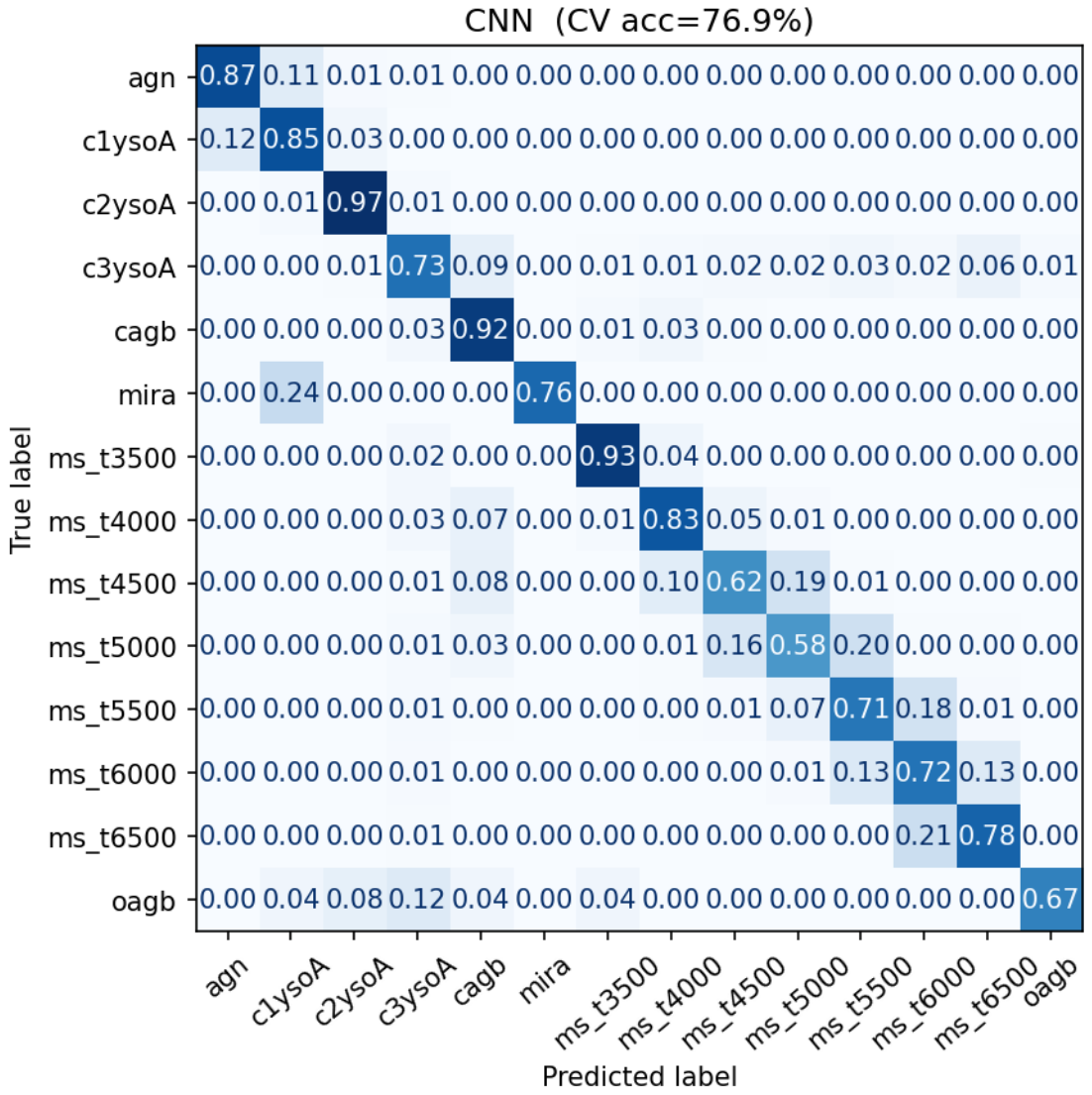
```
C1YSOA
  object_a.fits    confidence=0.921
  object_b.fits    confidence=0.884

C2YSOA
  object_c.fits    confidence=0.907

UNKNOWN
  object_d.fits    confidence=0.241
```

EVALUATION

CNN Training Model With Mira



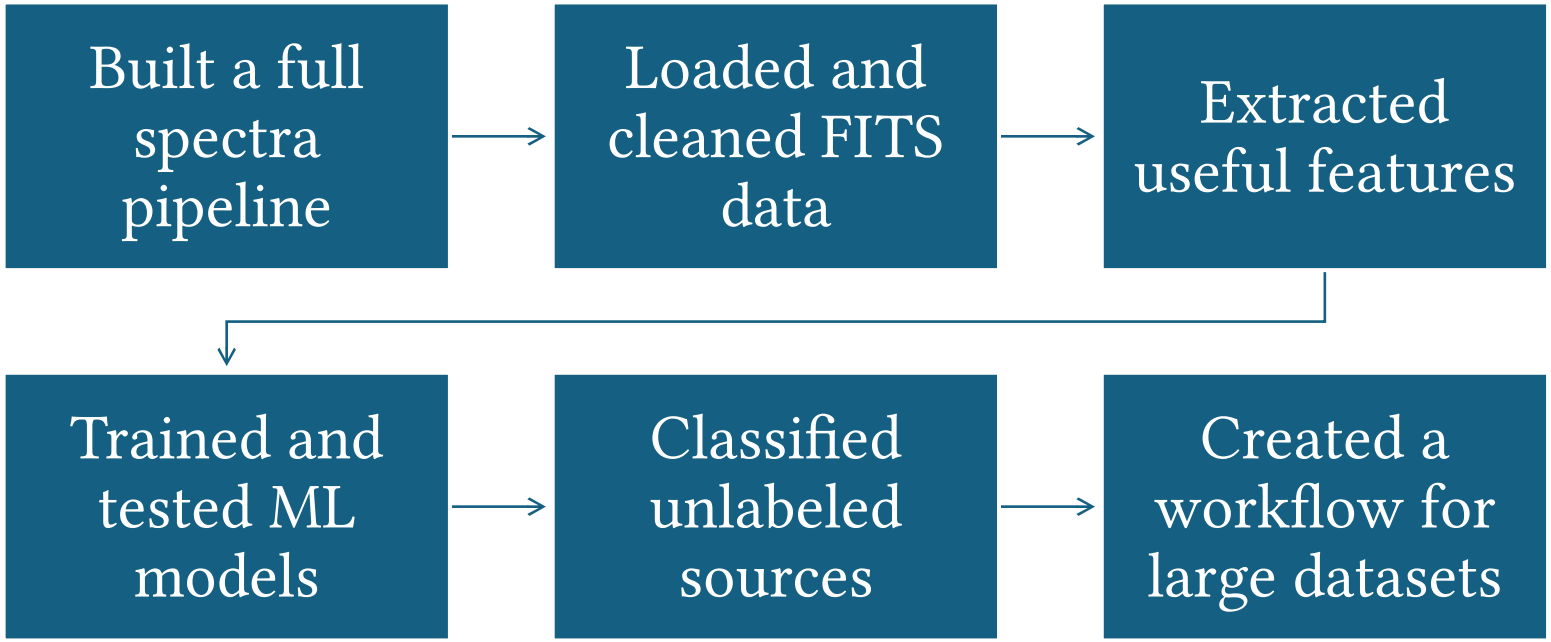
Extended CNN Training Increased Cross-Validation Accuracy



CONCLUSION

Key Takeaways

Accomplishments:



- CNN models can sort large source catalogs quickly
- The most useful result is a ranked candidate list
- Helps scientists focus on the best objects first

Questions?

Thank You!!!