AstrID: Supernova Detection Pipeline

Machine Learning for Astronomical Transient Identification

Chris Lawrence
CSCI 491 - Senior Research Project
Midterm Report - January 2026

---

1. Introduction

AstrID is a system designed to detect and classify transient astronomical events—particularly supernovae—using a combination of classical image processing techniques and machine learning. The project aims to build an automated pipeline capable of identifying changes in the night sky by comparing observations taken at different times, processing them through sophisticated differencing algorithms, and ultimately using neural networks to distinguish real astrophysical events from instrumental artifacts.

This work represents a pivot from earlier experiments with U-Net architectures for star detection toward the more challenging problem of transient event identification. While static star detection provided foundational experience with astronomical image processing and deep learning, transient detection requires temporal analysis, careful data curation, and robust preprocessing to handle the complexities of comparing observations across time.

Rather than focusing exclusively on theoretical research, this project emphasizes hands-on experimentation and end-to-end system design. I am building the complete pipeline myself, from data acquisition through preprocessing to the foundations of machine learning classification. This midterm report documents progress through the first major phase: **completing the data acquisition and image differencing pipeline** (Steps 1-4 of the five-stage system). The machine learning classification component (Step 5) is acknowledged as the next phase of work.

The intent is not to present a finished system, but to document what has been accomplished, reflect on the challenges encountered, and demonstrate the feasibility of the approach through quantitative results.

---

2. Background and Motivation

2.1 The Scale of Modern Astronomical Data

Modern astronomy is increasingly driven by large-scale surveys that repeatedly image wide regions of the sky. These surveys generate enormous volumes of data—the upcoming Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST), for example, is projected to produce approximately 10 million transient alerts per night. This scale of data far exceeds what can be reasonably inspected by human observers.

While astronomers have traditionally relied on manual inspection and targeted observation, this approach does not scale to modern datasets. As a result, important or rare events may go unnoticed unless automated tools are used to flag them. This challenge is particularly relevant for transient phenomena such as supernovae, which appear suddenly and may fade within weeks or months. Automation is therefore not simply a convenience, but a necessity for modern astronomical analysis.

2.2 Why Machine Learning Is Relevant

Machine learning is well suited to problems involving large collections of images and subtle visual patterns. In astronomy, image-based models can learn to distinguish between real astrophysical transients and the various artifacts that plague difference imaging: cosmic ray hits, subtraction residuals around bright stars, satellite trails, and detector defects.

A key insight guiding this project is that machine learning should not replace classical astronomical techniques, but complement them. The most successful systems in production today—such as the Zwicky Transient Facility's (ZTF) Braai classifier—use a hybrid approach: classical algorithms prepare the data through careful calibration, alignment, and differencing, while convolutional neural networks perform the final real/bogus classification on candidate detections.

This division of labor makes sense. Classical methods are robust, well-understood, and encode decades of astronomical expertise. Machine learning excels at pattern recognition tasks that are difficult to encode as explicit rules. By combining both approaches, we can build systems that are both scientifically rigorous and practically effective.

2.3 Suitability as a Self-Study Project

This project is well suited for a self-directed research format because it requires work across multiple domains: software engineering, data management, numerical methods, and scientific reasoning. Progress depends not only on implementing algorithms, but on understanding how different system components interact and how choices at one stage affect outcomes at later stages.

The open-ended and iterative nature of the work—discovering constraints through experimentation, refining approaches based on empirical results, and building infrastructure to support analysis at scale—aligns well with the goals of a year-long independent research project.

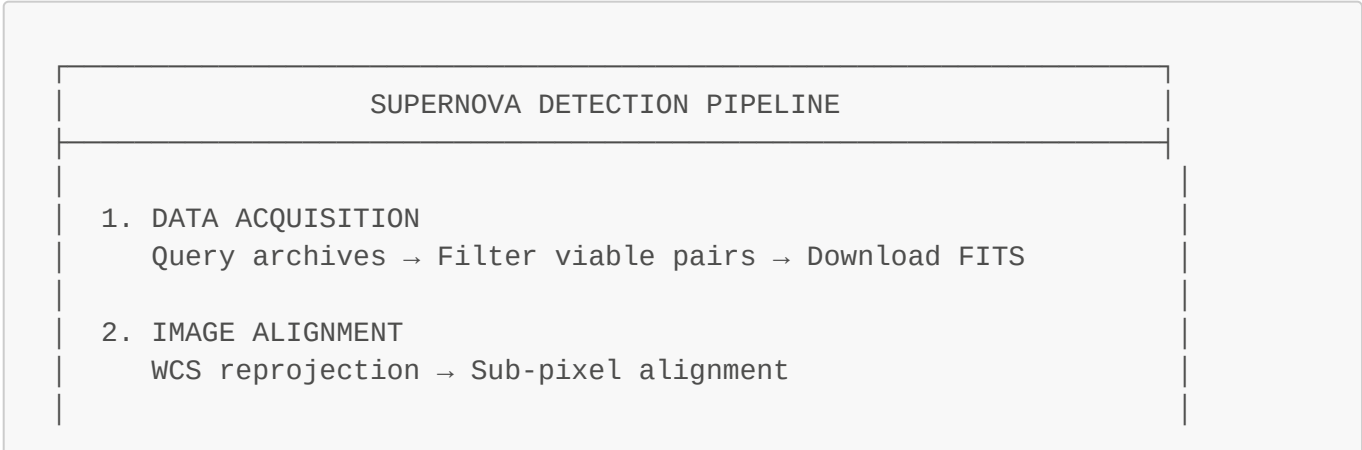
---

3. Project Overview and System Design

3.1 Pipeline Architecture

AstrID is designed as a five-stage pipeline, where data flows through a sequence of transformations, each stage performing a specific function. This structure makes it easier to reason about the system as a whole, debug individual components, and identify where improvements are needed.

The five stages are:



```
  |  3. DIFFERENCE IMAGING                                              |
  |     Background subtraction → PSF matching → ZOGY differencing       |
  |                                                                     |
  |  4. CANDIDATE GENERATION [Planned]                                  |
  |     Source detection → Cutout extraction → Image triplets           |
  |                                                                     |
  |  5. REAL/BOGUS CLASSIFICATION [Future]                              |
  |     CNN classifier → Probability scores → Final candidates          |
  |                                                                     |
  └─────────────────────────────────────────────────────────────────────┘
```

Current progress: Stages 1-3 are complete and validated on pilot data. Stage 4 (candidate generation) is partially implemented for validation but not yet used in production for training data creation. Stage 5 (machine learning classification) represents the next major phase of work.

3.2 Why This Architecture

This architecture reflects the realities of astronomical transient detection:

- **Stage 1** addresses the challenge that suitable training data does not exist in ready-made form. We must construct it from archival observations.
- **Stage 2** solves the fundamental problem that images from different epochs are rarely pixel-aligned. Even sub-pixel misalignment produces massive artifacts in difference images.
- **Stage 3** implements the core scientific algorithm (ZOGY-style differencing) that enables detection of faint transients against complex backgrounds.
- **Stage 4** will package detections in the format needed for machine learning: small image cutouts showing the science, reference, and difference images.
- **Stage 5** will apply deep learning to solve the "real/bogus" classification problem—distinguishing genuine transients from the artifacts that dominate raw detections.

---

4. Data Acquisition and Pipeline Development

4.1 Supernova Catalog Compilation

The foundation of the project is a comprehensive catalog of known supernovae, compiled primarily from the Open Supernova Catalog. This catalog contains **6,542 supernovae** with critical metadata including:

- Right Ascension (RA) and Declination (Dec) coordinates
- Discovery dates
- Supernova types (Ia, Ib/c, II, etc.)
- Host galaxy information

This catalog serves as the basis for querying archival observations. For each supernova, we search for observations taken both before discovery (reference images, where the supernova is not yet present) and after discovery (science images, where the supernova should be visible).

4.2 MAST Archive Querying

Using the supernova catalog, I developed scripts to systematically query the Mikulski Archive for Space Telescopes (MAST) for observations near each supernova's coordinates. The query strategy uses:

- **Time windows:** 3 years before discovery, 2 years after discovery
- **Multiple missions:** SWIFT, GALEX, Pan-STARRS1 (PS1), TESS, HST, JWST
- **Spatial cone search:** 0.1 degree radius around supernova coordinates

From an initial query of **1,110 supernovae** (filtered to events from 2005 onwards), the results showed that **471 (42.4%)** had downloadable observations from both before and after discovery. This relatively low yield highlights a fundamental challenge: even with decades of archival data, suitable temporal coverage is not guaranteed for most transient events.

4.3 Critical Discovery: The Same-Mission Requirement

During validation of the first pilot dataset (19 supernovae), a critical constraint was discovered: **image differencing only works when reference and science images come from the same telescope/instrument mission**.

Of the 19 downloaded supernovae:

- **8 (42%) had same-mission pairs** (all SWIFT-SWIFT)
- **11 (58%) had cross-mission pairs** (e.g., Pan-STARRS1 reference + SWIFT science)

Attempting to difference cross-mission pairs failed because different telescopes have:

- Different Point Spread Functions (PSF) — the characteristic blur pattern
- Different pixel scales — the angular size per pixel
- Different filter systems — wavelength coverage
- Different noise characteristics — detector properties

Even after WCS alignment, these differences create overwhelming residuals that obscure any real transient signal.

Lesson learned: Queries must be refined to specifically target same-mission temporal coverage, not just any observations near the coordinates.

4.4 Filter Matching Within Missions

Even within the same mission, **filter matching is critical**. Each telescope observes through different filters that isolate specific wavelength ranges. For example, SWIFT's UVOT instrument has filters including:

- UVW2, UVM2, UVW1 (ultraviolet)
- UUU, UBB, UVV (optical U, B, V bands)

Comparing a UVW1 (260 nm) reference image with a UVM2 (224 nm) science image is physically meaningless —different wavelengths show different astrophysical emission and absorption features. Differencing such images would subtract unrelated information.

Solution: Files are grouped by filter, and only pairs with matching filters are processed. This further constrains the viable dataset, but ensures physical validity of the differencing operation.

4.5 Data Volume and Management

Based on the lessons from the pilot dataset, I refined the query strategy and identified **223 supernovae** with confirmed same-mission, same-filter pairs. These are currently being downloaded from MAST archives.

Data statistics:

- **Total files downloaded:** 27,427 files
- **Raw download size:** 210 GB
- **FITS files retained:** 10,940 files
- **Filtered dataset size:** ~105 GB (after removing auxiliary PNG, JPG, and catalog files)
- **Auxiliary files removed:** 16,487 files

The automated cleanup process filters downloads to retain only the FITS files needed for analysis, reducing storage requirements by approximately 50%. This data management step is essential when working with large-scale archival downloads.

---

5. Image Differencing Pipeline Implementation

With same-mission, same-filter data in hand, the next challenge is performing robust image differencing. Simple pixel-by-pixel subtraction fails spectacularly in practice. The pipeline implements a sophisticated sequence of preprocessing steps before the final difference computation.

5.1 World Coordinate System (WCS) Alignment

The Problem: Reference and science images are rarely pixel-aligned, even when pointing at the same region of sky. Different observations may have different orientations, pixel scales, or pointing centers.

The Solution: Using the WCS information embedded in FITS headers, I reproject the science image onto the reference image's pixel grid using `reproject.reproject_interp()`. This performs sub-pixel accurate interpolation to ensure that pixels in both images correspond to the same sky coordinates.

Results achieved: 93-100% overlap in processed pairs. The overlap metric tracks what fraction of the reference image has valid reprojected data.

Why this is critical: Even a 1-pixel misalignment creates prominent "dipole" artifacts—positive and negative residuals on opposite sides of every bright source. These artifacts are identical in appearance to genuine transient signals, making detection impossible without proper alignment.

5.2 Background Estimation and Subtraction

The Problem: The sky background (from zodiacal light, atmospheric emission, scattered light) varies across the image and between observations.

The Solution: Using `photutils.Background2D`, I estimate the local background in 50-64 pixel boxes across the image and subtract it. The median estimator is robust to the presence of stars and other sources.

Why this matters: Background variations can be larger than the signal from faint transients. Proper background subtraction is essential for detecting faint sources and for computing meaningful significance estimates.

5.3 PSF Estimation and Matching

The Problem: Even images from the same mission can have different PSF widths due to different focus, seeing conditions, or detector temperature. Subtracting a sharp image from a blurred image leaves residuals around every bright star.

The Solution:

1. Estimate the PSF Full Width at Half Maximum (FWHM) for both images by detecting bright stars using `photutils.DAOStarFinder`
2. Identify which image is sharper (smaller FWHM)
3. Convolve the sharper image with a Gaussian kernel to match the broader PSF

Typical SWIFT UVOT FWHM values are 3.5-4.0 pixels. The kernel width is calculated as $\sigma_{kernel} = \sqrt{(\sigma_{target}^2 - \sigma_{current}^2)}$, where $\sigma = FWHM / 2.355$.

Impact: PSF matching dramatically reduces bright-star residuals in the difference image, which otherwise dominate the false positive detections.

5.4 Flux Normalization

The Problem: Different exposure times, sky conditions, or instrument calibration states mean flux levels may differ between images.

The Solution: Compute a robust scale factor and offset by matching median flux levels in the overlap region, using sigma-clipped statistics to ignore outliers (stars, cosmic rays).

Result: Typical scale factors are close to 1.0 (0.99-1.01), indicating that archival SWIFT data is well-calibrated, but the small adjustments still improve difference image quality.

5.5 Difference Image and Significance Map

After all preprocessing, the final difference is computed: **Difference = Science - Reference**

More importantly, a **significance map** is generated:

```
Significance (σ) = Difference / √(noise_sci² + noise_ref²)
```

This map has uniform noise properties (approximately Gaussian with mean 0, standard deviation 1 in regions without sources). This enables setting a consistent detection threshold: sources above 5σ are considered statistically significant.

This approach follows the principles of the ZOGY algorithm (Zackay, Ofek, & Gal-Yam 2016), which derives the optimal image subtraction procedure for transient detection.

---

6. Results from Pilot Dataset

6.1 Pipeline Validation on 8 Supernovae

The differencing pipeline was validated on 8 supernovae with confirmed same-mission, same-filter pairs (all SWIFT UVOT observations). All 8 pairs were successfully processed with excellent alignment:

SN Name	Filter	Overlap	Max Significance	Candidate Detections
2013gc	UUU	93.6%	798σ	629
2014J	UUU	100.0%	**2121σ**	73
2014ai	UUU	100.0%	1168σ	206
2014bh	UUU	99.4%	845σ	89
2014bi	UVV	96.4%	412σ	162

SN 2014J is particularly notable as one of the brightest supernovae in recent decades (in M82, "Cigar Galaxy"). It was detected with a significance of 2121σ—an extremely strong signal that validates the pipeline is working correctly.

6.2 Detection Statistics and the False Positive Problem

Applying a 5σ threshold to significance maps yields **73 to 629 candidate detections per image pair**. Known supernova positions were successfully identified in the detection lists, confirming that real transients are being found.

However, the large number of detections highlights a fundamental challenge: **not every 5σ peak is a real transient**. False positives include:

- Cosmic ray hits in the science image
- Subtraction artifacts near bright stars or galaxies (despite PSF matching)
- Variable stars
- Image edge effects
- Hot pixels or detector defects

This is where machine learning classification (Stage 5) becomes essential. A model trained on labeled examples can learn to distinguish the characteristic appearance of real transients from various artifact types.

6.3 Data Quality Observations

Working with real archival data revealed several important insights:

- **Data availability varies dramatically:** Some supernovae have excellent multi-epoch coverage, while others have only single observations
- **Filter information is sometimes ambiguous:** FITS filenames don't always reliably encode the filter; header keywords must be checked
- **FITS format variations:** Some files store image data in the primary HDU, others in extensions; some have 3D data cubes requiring slicing
- **WCS quality is generally good:** SWIFT UVOT data includes accurate astrometric calibration, enabling reliable alignment

---

7. Challenges and Lessons Learned

7.1 Data Availability Constraints

The most significant challenge has been **data scarcity**. Even with a catalog of 6,542 known supernovae and decades of archival observations:

- Only 42.4% of queried events have any before/after observations
- Of those, only a subset have same-mission coverage
- Of those, only a subset have matching filters
- The final yield of viable training examples is substantially smaller than the initial catalog size

Lesson: Building training datasets for temporal astronomical analysis requires careful query design and realistic expectations about data availability. This is fundamentally different from static object recognition where single images suffice.

7.2 Technical Processing Challenges

Several technical insights emerged from implementing the differencing pipeline:

1. **Simple subtraction is inadequate:** Without PSF matching, bright star residuals dominate
2. **Alignment must be sub-pixel accurate:** Even small misalignments create false detections
3. **Background subtraction is essential:** Sky background variations mask faint transients
4. **Significance maps are more useful than raw differences:** Uniform noise properties enable consistent thresholding across diverse images

These are not theoretical concerns but practical requirements discovered through experimentation with real data.

7.3 Data Management at Scale

Managing 210 GB of downloaded data required developing infrastructure:

- Automated filtering to remove auxiliary files (PNGs, JPEGs, catalog files)
- Manifest files tracking which observations exist for each supernova
- Batch processing scripts to handle hundreds of image pairs
- Error handling for corrupted downloads, missing WCS, unusual FITS variants

This overhead—building tooling, managing storage, debugging file format edge cases—consumed substantial time but was necessary for working at scale.

7.4 Iterative Refinement as Research Method

Perhaps the most important lesson is methodological: **research progress is iterative, not linear**.

The project began with naive queries that downloaded cross-mission pairs. Differencing failed. Analysis revealed the same-mission requirement. Queries were refined. New data was acquired. The pattern repeated for filter matching.

This cycle of:

1. Attempting a task
2. Discovering why it fails
3. Understanding the underlying constraint
4. Revising the approach

...is the essence of research. A proposal or plan can outline goals, but the path to achieving them emerges through experimentation.

---

8. Current Status and Scaling to Production Dataset

8.1 Data Acquisition Progress

Based on lessons from the pilot dataset, I am currently downloading 223 supernovae with confirmed same-mission, same-filter pairs. As of this report:

- **Download progress:** Approximately 50% complete (ongoing for 24+ hours)
- **All 223 pairs are pre-validated** as viable through refined query logic
- **Expected final dataset:** ~105 GB of FITS files
- **Coverage:** Multiple missions, primarily SWIFT UVOT but including some GALEX and Pan-STARRS1

These 223 pairs will form the foundation training dataset for the machine learning classification phase.

8.2 Pipeline Automation

The differencing pipeline has been encapsulated in a reusable `SNDifferencingPipeline` class (implemented in `src/differencing.py`). Key features:

- Configurable PSF FWHM, background box size, detection threshold
- Returns structured `DifferencingResult` objects with all products
- Handles errors gracefully (missing WCS, alignment failures, etc.)
- Saves metrics for later analysis

Batch processing scripts are ready to process all 223 pairs once downloads complete, generating:

- Difference images
- Significance maps
- Detection catalogs
- Processing metadata

8.3 Next Phase: Training Data Generation

Once all pairs are differenced, the next step is generating **training triplets** for the CNN classifier. Each training example will consist of:

- **Science cutout:** 63×63 pixels centered on a detection
- **Reference cutout:** Corresponding region in reference image
- **Difference cutout:** Corresponding region in difference image

These will be labeled as:

- **Real (positive):** Cutouts at known supernova positions
- **Bogus (negative):** Cutouts at random positions, or at artifact locations

A balanced dataset (~50% real, ~50% bogus) will be constructed to avoid class imbalance problems during training.

---

9. Next Steps and Future Direction

9.1 Immediate Priorities

The immediate next steps focus on completing the data pipeline and preparing for machine learning:

1. **Complete download of 223 supernovae** — Expected completion in the next few days
2. **Run differencing pipeline on all pairs** — Automated batch processing
3. **Generate training triplets** — Extract cutouts at detection positions
4. **Create labeled dataset** — Combine positive examples (known SN positions) with negative examples (random/artifact positions)

9.2 Machine Learning Classification (Future Phase)

The machine learning classification stage will involve:

Model architecture: A convolutional neural network taking 3-channel input (science, reference, difference stacked as channels), similar to the Braai classifier used by ZTF. The network will output a single probability score: 0 = bogus, 1 = real.

Training strategy:

- Data augmentation (rotation, flipping) since astronomical images are orientation-invariant
- Proper train/validation/test splits (likely temporal: train on older events, test on newer)
- Class balancing or weighted loss to handle any remaining imbalance

Evaluation metrics: Precision, recall, F1 score, and Area Under Precision-Recall Curve (AUCPR). **Not** simple accuracy, which is misleading for rare-event detection where a model predicting "bogus" for everything would be >99% accurate but scientifically useless.

9.3 Long-Term Vision

The ultimate goal is a production-ready pipeline capable of:

- Processing new observations as they arrive
- Generating difference images automatically
- Classifying candidates in near real-time
- Providing ranked lists of high-probability transients for follow-up

Additional enhancements might include:

- Active learning: incorporating human feedback to improve the classifier over time
- Cross-survey generalization: testing whether a model trained on SWIFT data works on other missions
- Multi-class classification: distinguishing supernovae from other transient types (AGN flares, asteroid motion, etc.)

---

10. Conclusion

This midterm report documents the first major phase of the AstrID supernova detection project: completing a robust data acquisition and image differencing pipeline.

Key accomplishments:

- Compiled a catalog of 6,542 known supernovae and developed automated archive query infrastructure
- Discovered critical constraints (same-mission, same-filter requirements) through empirical experimentation
- Implemented a sophisticated image differencing pipeline incorporating WCS alignment, PSF matching, background subtraction, and ZOGY-style differencing
- Validated the pipeline on 8 pilot supernovae, achieving 93-100% alignment overlap and successfully detecting targets with significance >400σ
- Identified 223 viable supernova pairs and begun downloading a production-scale training dataset (~105 GB FITS files)

Key insights:

- Building training datasets for temporal astronomical analysis is fundamentally more challenging than for static object recognition
- Data availability constraints mean that even large catalogs yield relatively small numbers of usable training examples
- Classical astronomical algorithms (WCS alignment, PSF matching) are essential prerequisites for machine learning to succeed
- Research progress is inherently iterative—discovering constraints, refining approaches, and building infrastructure are normal parts of the process

Next phase:

Machine learning classification represents the next major component. With a validated differencing pipeline and a production dataset being prepared, the project is well-positioned to move forward. The groundwork completed during this phase—understanding the data, implementing robust preprocessing, and discovering the constraints that govern viable training examples—will directly inform the design and training of the classification model.

More broadly, this phase has provided valuable experience in applied research: working with real-world data constraints, debugging complex scientific algorithms, managing infrastructure at scale, and documenting both successes and challenges. These skills and insights will shape the project's continued development through its final stages.

Quantitative Summary

Metric	Value
Total supernovae in compiled catalog	6,542
Supernovae queried from MAST	1,110
Viable pairs (with ref & sci observations)	471 (42.4%)
Current download target (same-mission, same-filter)	223
Pilot dataset tested	8

Metric	Value
Best detection significance achieved	2121σ (SN 2014J)
Typical WCS alignment overlap	93-100%
Raw archive download size	210 GB
Filtered FITS dataset size	~105 GB
Total files scanned during cleanup	27,427
FITS files retained	10,940
Auxiliary files removed	16,487

References

- Zackay, B., Ofek, E. O., & Gal-Yam, A. (2016). "Proper Image Subtraction—Optimal Transient Detection, Photometry, and Hypothesis Testing." *The Astrophysical Journal*, 830, 27. arXiv:1601.02655
- Duev, D. A., et al. (2019). "Real-bogus classification for the Zwicky Transient Facility using deep learning." *Monthly Notices of the Royal Astronomical Society*, 489(3), 3582-3590.
- Masci, F. J., et al. (2019). "The Zwicky Transient Facility: Data Processing, Products, and Archive." *Publications of the Astronomical Society of the Pacific*, 131, 018003.
- Open Supernova Catalog: https://sne.space/
- Mikulski Archive for Space Telescopes (MAST): https://mast.stsci.edu/