Multi-modal identification and classification of YSOs

NEMESIS — Novel Evolutionary Model for the Early Stages of Stars with Intelligent Systems

The NEMESIS General YSO Catalogue I: supervised image-based classification

G. Marton[1, 2], M. Madarász[1, 2, 3], J. Roquette[4, 5], M. Audard[4], I. Gezer[1, 2], D. Hernandez[6], and O. Dionatos[6]

marton.gabor@csfk.org

Abstract

- **Context: Young stellar objects are usually identified from infrared excess in their spectral energy distributions, but previous catalogues often rely on survey-specific colours and methods.**
- **Aim: We develop a deep-learning method that uses heterogeneous multi-wavelength data to identify YSOs more uniformly and accurately.**
- **Methods: We collect SEDs, light curves, and imaging data from VizieR and IRSA, convert them into image-based representations, and classify them with an ensemble of CNN models.**
- **Results: The method identifies YSO candidates with high accuracy while rejecting contaminants with very high reliability.**
- **Catalogue: Applying the classifier to literature YSO catalogues produces the NEMESIS General YSO catalogue, containing 275,023 unique, high-reliability YSO candidates.**
- **Validation: Gaia colour–magnitude diagrams and the spatial distribution of the sources relative to local Galactic structures support their young nature.**

Method in a glance

Literature candidates

VizieR + IRSA

SED / DTDM / WISE images

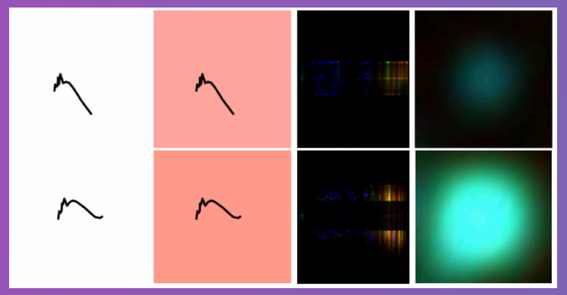
12 CNNs

6/12 vote threshold

NGYSO catalogue

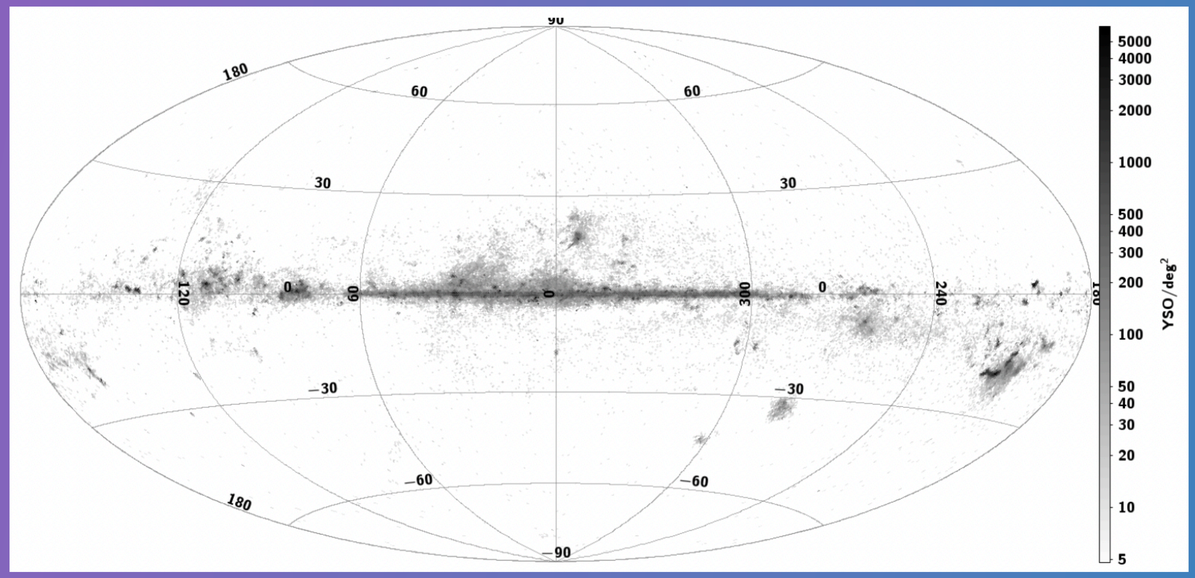
- **Multi-source data collection: We gathered heterogeneous photometric, variability, and imaging data from VizieR and IRSA for both known YSOs and common contaminant populations.**
- **Image-based representations: For each source, we created visual inputs such as SED plots, reddening-coloured SEDs, ZTF DTDM variability maps, and AllWISE image cutouts.**
- **CNN classification: These image representations were used to train multiple convolutional neural networks to distinguish YSOs from non-YSOs without manual feature engineering.**
- **Ensemble voting: The final classification combines the votes of 12 CNN models; a source is accepted as a YSO candidate when at least half of the models classify it as YSO.**

Input data become images



The different image-based representations used for classification. Upper row shows data for the same source located at RA=98.804, Dec=12.166 degrees, bottom row for the source at RA=0.55529, Dec=64.90703 degrees. From left to right: the simple SED plot, the SED plot with background colour encoding the CSFD interstellar extinction seen in the direction of the source, the DTDM variability image, where the RGB layers represent the ZTF i, r and g band light curve data, and the AllWISE cutout stamp image of the source, respectively.

Result: a homogeneous all-sky YSO catalogue



275,023

unique high-reliability YSO candidates

F1 = 96.90

SED ensemble classification performance

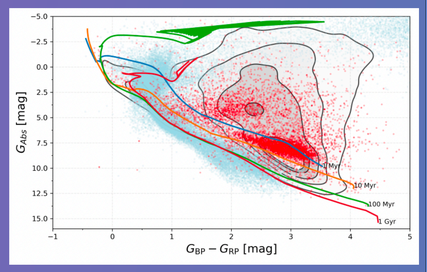
0.99%

external non-YSO sample misclassified

- **Nearly 90 literature catalogues are reassessed under the same standard.**
- **Full SED morphology separates real infrared excess from reddened contaminants.**
- **SED quality cut: Only sources with usable SEDs, containing at least 10 photometric points after cleaning, were classified.**
- **Voting criterion: A source was accepted as a YSO candidate if at least 6 of 12 SED classifiers voted for the YSO class.**
- **Final catalogue size: After merging duplicates within 2″, the NGYSO catalogue contains 275,023 unique high-reliability YSO candidates.**
- **Best data product: The SED ensemble was adopted as the catalogue backbone because it achieved the strongest performance: F1 = 96.90.**
- **All-sky but not complete: NGYSO combines region-focused surveys with large all-sky catalogues. It is broad and homogeneous but not fully complete.**
- **Flexible selection: Individual classifier probabilities are included, allowing to build stricter or more inclusive subsamples for their own science cases.**

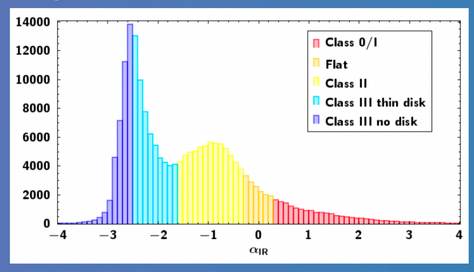
Gaia CMD validation

Distribution of the YSO training sample (red dots), the non-YSO training sample (light blue dots) and the sources classified as YSOs (contour lines) on the Gaia BP-RP – absolute magnitude diagram.



Contour levels are at the 0.8, 0.9, 0.95, 0.99 and 0.995 density quantiles. PARSEC evolutionary isochrones of 1, 10, 100 Myr and 1 Gyr are shown with blue, orange, green and red lines, respectively. Most sources occupy the expected 1-10 Myr PMS region.

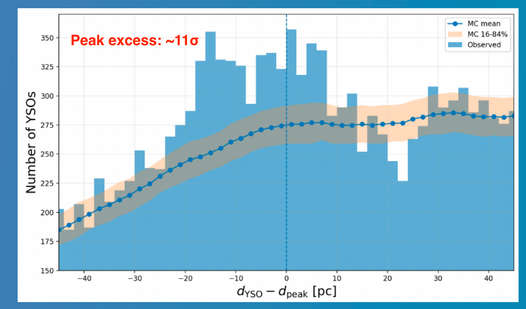
Evolutionary classes



Distribution of the α_{IR} values for 195 444 NGYSO sources. Different observational classes are presented with the different colours. Class boundaries are adopted from Grossschedl et al. (2019).

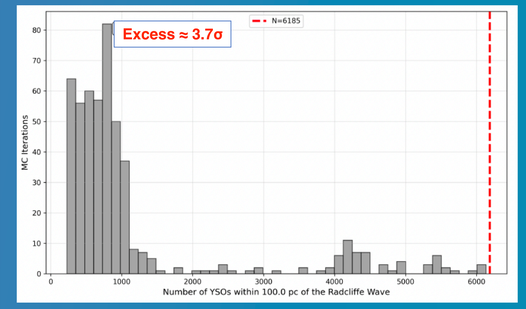
Galactic structure test

The Local Bubble



Distribution of YSO distances relative to the Local Bubble wall. The horizontal axis shows the signed offset Δ = d_{YSO} − d_{peak}, where d_{peak} is the distance of the shell density maximum along each line of sight. Negative values correspond to positions interior to the present day shell, while positive values indicate locations exterior to it. The blue histogram shows the observed distribution of YSOs, while the solid line indicates the mean of the Monte Carlo (MC) realisations in which YSO distances are preserved but their angular association with the wall is randomised. The shaded region marks the 16th–84th percentile range of the MC distribution. The vertical dashed line denotes Δ = 0, i.e. the location of the present-day shell peak.

Radcliffe Wave



Results of the Monte Carlo test between the NGYSO catalogue and the Radcliffe Wave. The grey histogram shows the expected background distribution of YSOs generated from 500 randomised rotations of the dataset around the Galactic Z-axis. The red vertical dashed line indicates the actual number of YSOs located within a 100 pc radius of the wave's 3D spatial spline.

Conclusion

- **We present a deep-learning framework that identifies YSO candidates from image-based representations of heterogeneous astronomical data.**
- **The SED-based ensemble performs best, reaching F1 = 96.90 with a very low false-positive rate.**
- **The final NGYSO catalogue contains 275,023 unique YSO candidates, making it the largest homogeneous and reproducible all-sky YSO sample to date.**
- **Independent validation confirms low contamination, with only 0.99% of external non-YSO sources misclassified as YSOs.**
- **Astrophysical tests support the young nature of the sample, including Gaia CMD positions, evolutionary class fractions, and spatial correlations with the Local Bubble and Radcliffe Wave.**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101004141

Take-home message

Image-based deep learning turns messy multi-survey catalogues into a reproducible map of young stars.

1 Konkoly Observatory, Research Centre for Astronomy and Earth Sciences, Hungarian Research Network (HUN-REN), H-1121 Budapest, Konkoly Thege Miklós út 15-17., Hungary
e-mail: marton.gabor@csfk.org
2 CSFK, MTA Centre of Excellence, Budapest, Konkoly Thege Miklós út 15-17., H-1121, Hungary
3 Department of Experimental Physics, Institute of Physics, University of Szeged, D.m t.r 9, 6720 Szeged, Hungary
4 Department of Astronomy, University of Geneva, Chemin Pegasi 51, 1290 Versoix, Switzerland
5 Department of Basic Neuroscience, University of Geneva, 1211 Geneva, Switzerland
6 Department of Astrophysics, University of Vienna, Türkenschanzstrasse 17, 1180 Vienna, Austria