“A new automated tool for the spectral classification of OB stars”

Kyritsis et al.,2021, accepted to A&A

Stellar spectroscopy and Astrophysical Parametrization from Gaia to Large Spectroscopic Surveys
21-23, September, 2021

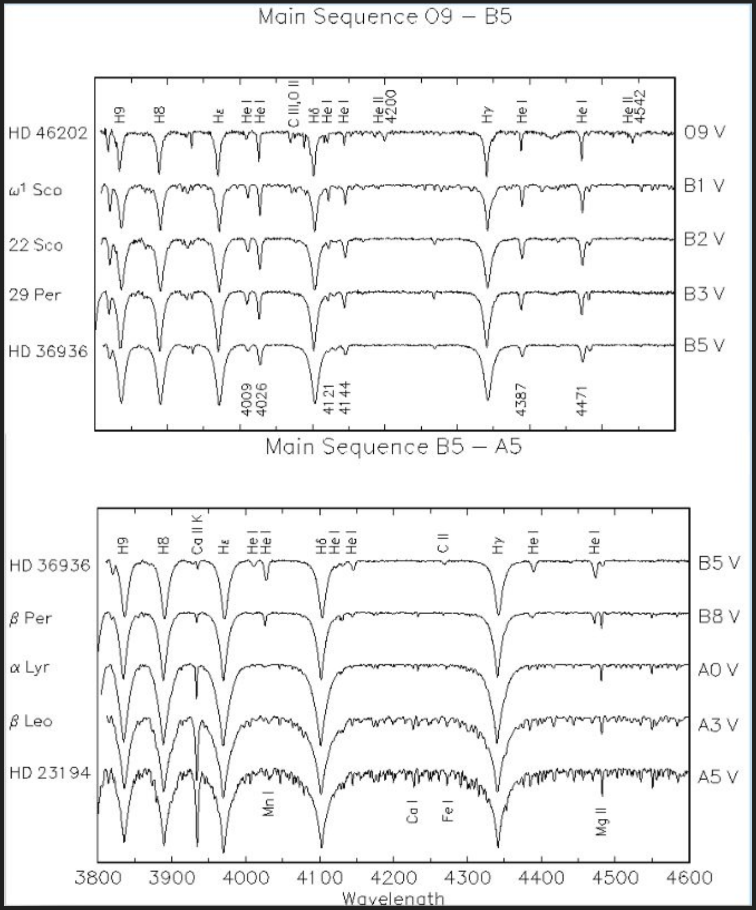
Elias Kyritsis - PhD student

ekyritsis@physics.uoc.gr

Collaborators: G. Maravelias, A. Zezas, P. Bonfini, K. Kovlakas, P. Reig

University of Crete & Institute of Astrophysics - FORTH

Optical spectroscopy as a tool for the study of OB stars



[NASA/IPAC]

The spectrum of a star reflects its physical parameters (e.g T, chemical abundances, gravity etc.)

Spectral Type → Proxy to T

Traditional way of spectral classification

Visual examination of the spectrum and recognition of characteristic spectral lines

2 Main Problems

- ❖ Subjectivity of human factor
- ❖ Time-consuming method

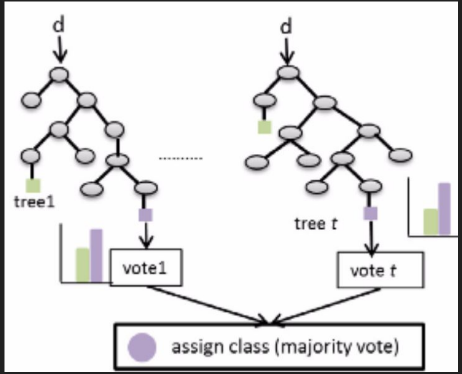
Solution

Usage of automated ways which are based on quantitative measurements of spectral features

Machine learning for the spectral classification of OB stars

Random Forest

- Why this algorithm ?
  - Imitates the human way of assessing the presence of a spectral line when visually examining a spectrum
- How does it work ?
  - RF is a collection of a large number of decision trees
  - Randomly-selected data subsets of the initial dataset
  - Randomly-selected subsets of the features in each node of each decision tree



[Belgiu & Draguit,2016]

Each tree in the forest suggests a class

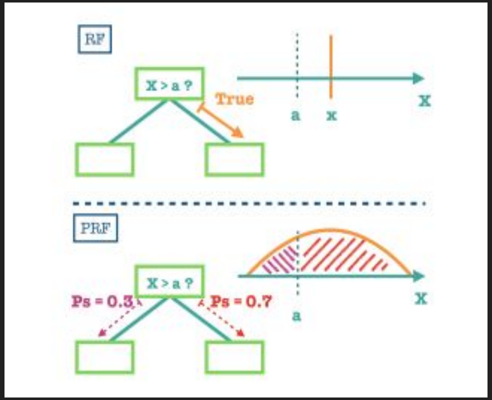
⬇

Majority vote

⬇

Final prediction

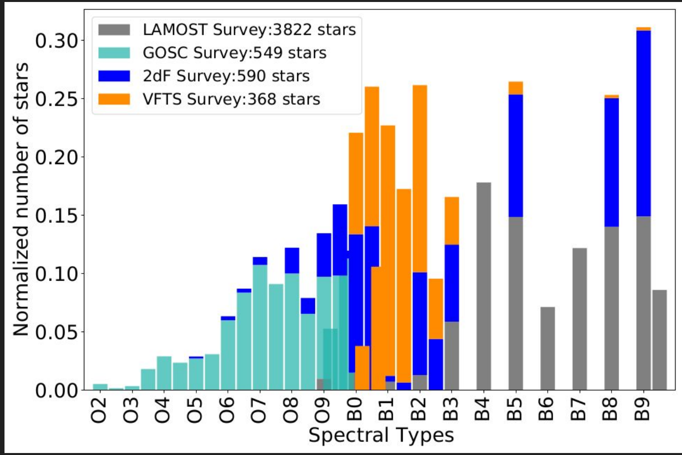
Probabilistic Random Forest



[Reis et al.,2019]

- Why this algorithm ?
  - It is a modified RF algorithm which accounts for the measurement errors during the training
- How does it work?
  - Similarly to RF but it handles the features values as PDFs instead of deterministic quantities

Spectroscopic data



Aiming to a **LC** and **metallicity** independent model

➢ 5329 available labeled spectra from large Galactic & extragalactic surveys within the range O2-B9.5 and LC I-V

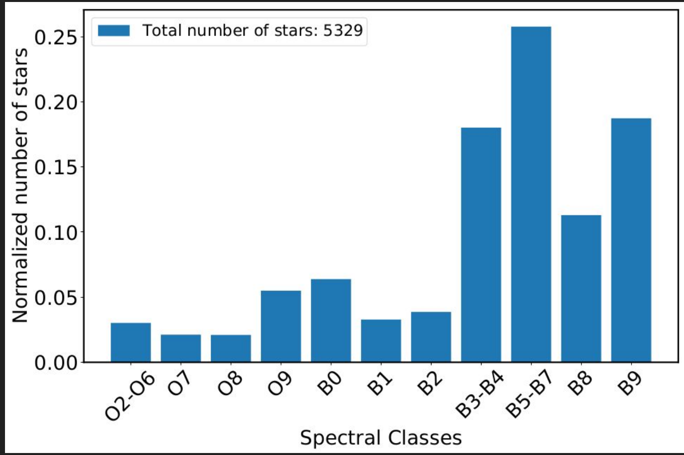
Limitation

➔ Early-mid O types are underrepresented. Non-optimal distribution for the training of the RF & PRF algorithms

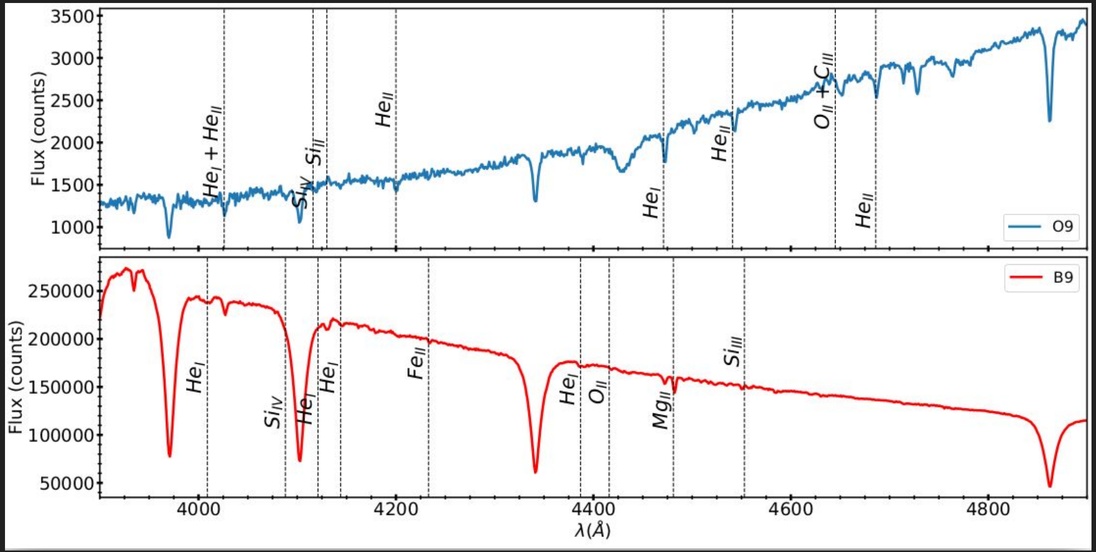
Solution

Adaptive binning of the **spectral types** to **spectral classes** which accounts for spectral similarities between adjacent bins

➔ Increase the number of objects per spectral class for the underrepresented classes



Classification scheme



17 characteristic spectral lines based on the criteria of Maravelias et al. 2014

- HeI, HeII spectral lines ⟶ Strong indicators for O-type and early B-type stars
- MgII, SiI,SiII spectral lines ⟶ Indicators for mid and late B-type stars

★ Balmer lines are not included in the feature scheme since their profile can be affected in the cases of Oe/Be stars

★ Our model can classify such stars which are companions in HMXBs

Quantification of spectral features - EW measurements

Quantification of spectral lines ⟶ Measurement of Equivalent Width (EW)

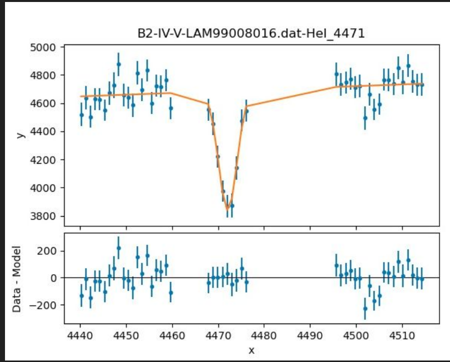
EW measurements

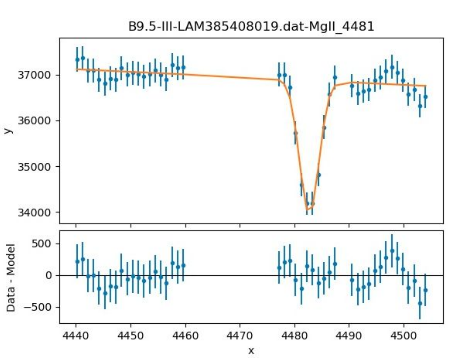
❖ Spectral line fitting using the *Sherpa* package v 4.13.0

➢ Fit each of the 17 lines per spectrum (1 Gaussian profile + 1 polynomial)

➢ Development of an automated pipeline for this purpose
17 x 5329 ~ 90000 fits !

➢ Derived quantities: EW, δEW for each line of 17 characteristics spectral lines

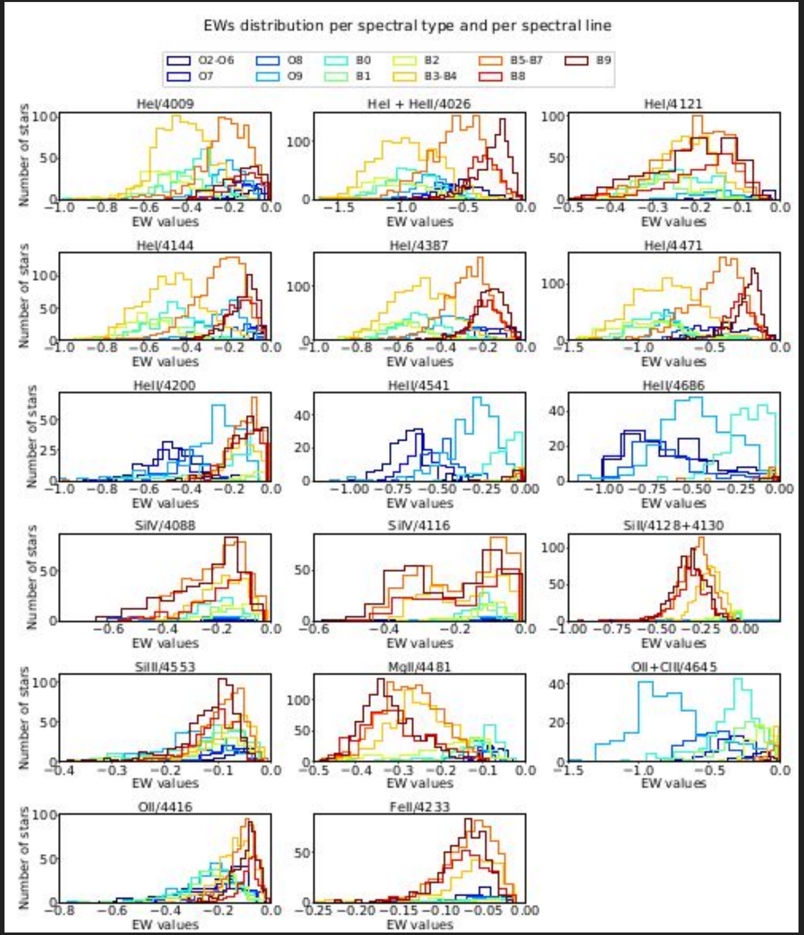




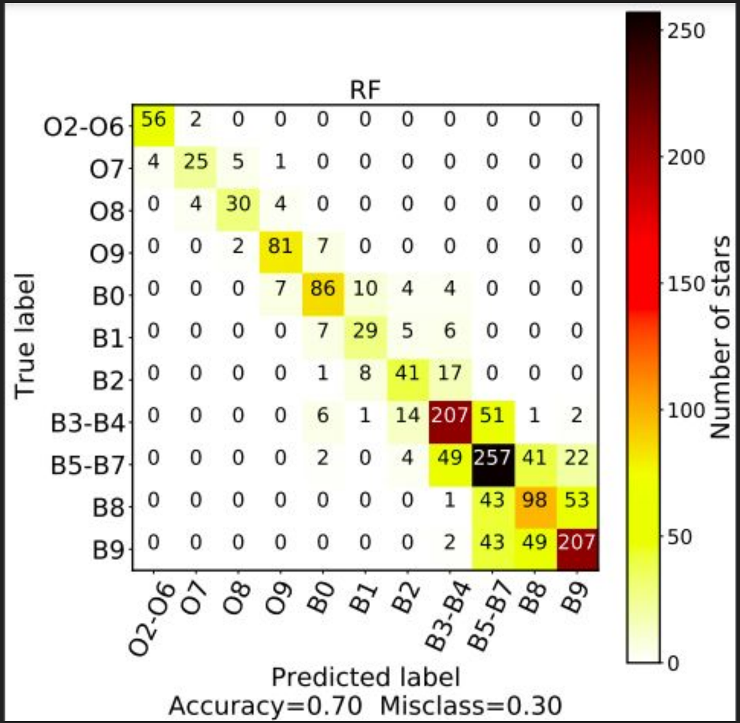
Data post-processing & Building the RF and PRF models

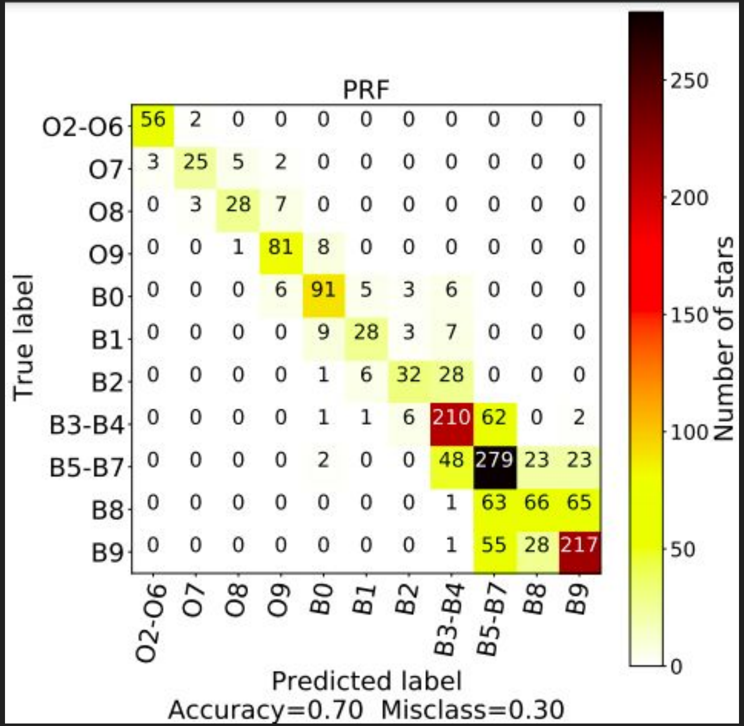
- Keeping only spectra with S/N > 50
- Detections
  - EW < 0 (absorption lines)
  - EW S/N > 3 (i.e. EW/δEW > 3)
- Non-Detections
  - Unconstrained fit parameters
  - EW > 0 (emission lines)
  - EW S/N < 3 (i.e. EW/δEW < 3)
  - Flagged and included in the analysis

- Splitting the initial sample: 70% Training - 30% Test
  - RF input: Spectral Types, EW of each line
  - PRF input: Spectral Types, EW, δEW of each line



Results

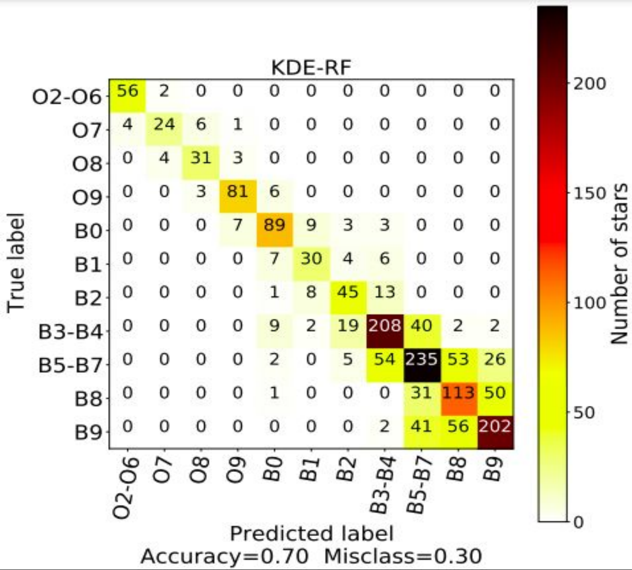




Score ~70 % !!!

- ❖ High performance for earlier spectral classes
- ❖ Difficulty to distinguish between later spectral classes where weaker spectral lines are important

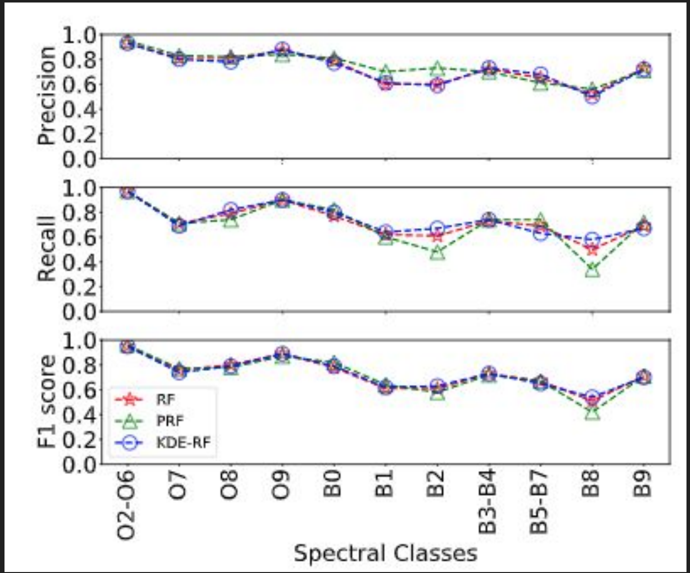
Results



Kernel -Density Estimation - RF

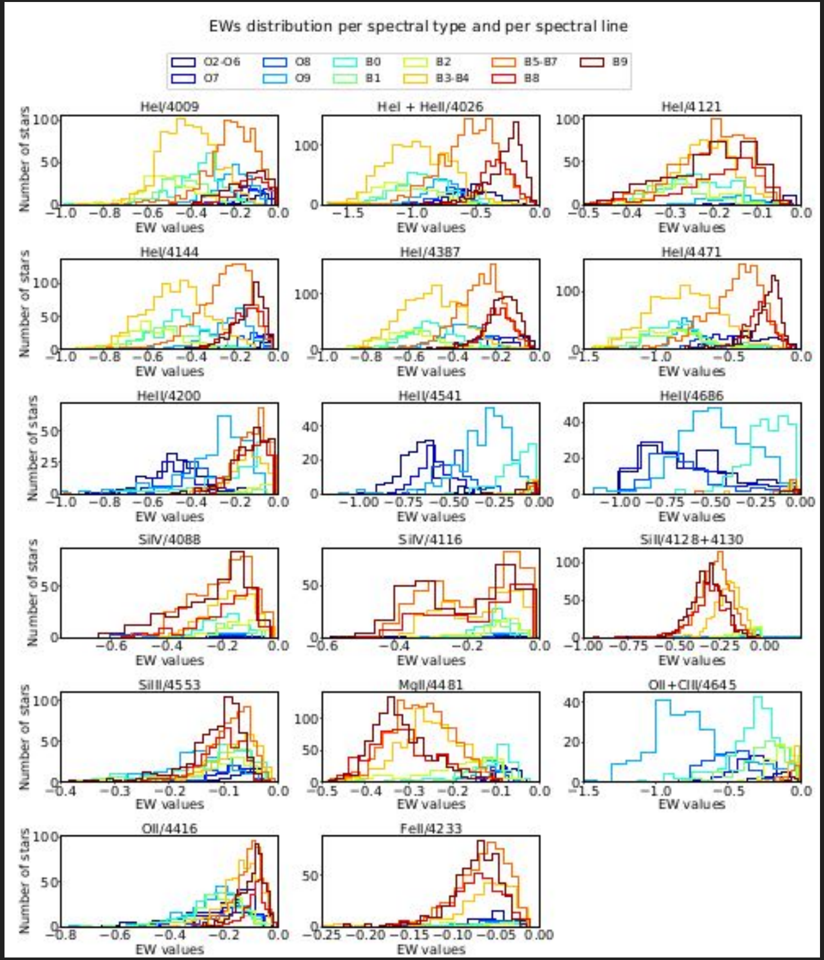
An alternative approach for defining training set

- Accounts for spectral feature correlations
- Corrects for imbalance among the classes
- Tests the stability of the algorithm



- ➢ Similar results across all approaches - **70 %** !
- ➢ O2-O6 and O9 > 90 % - The majority of spectral classes reaches scores > 65 %
- ➢ The similarity in the performance indicates the robustness and the reliability of RF algorithm

Results



Why not higher score per class ?

- ❖ Strong overlap between the EW distributions especially for later type stars and weaker lines

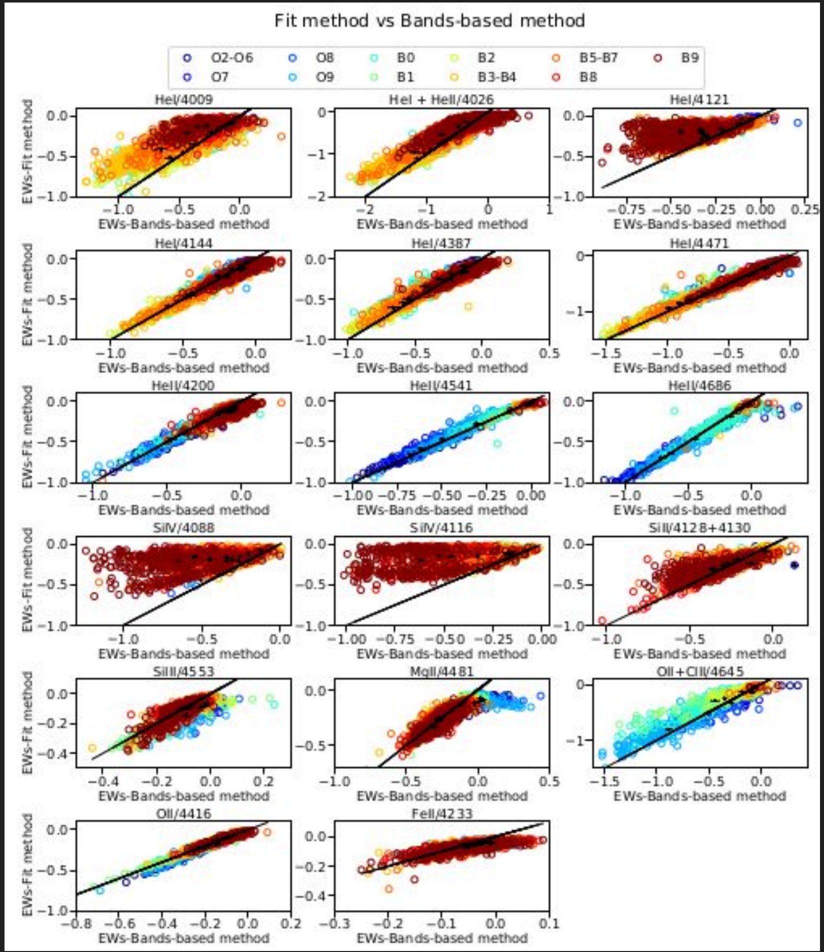
- ❖ More data is needed for better representation of the parent distributions of each spectral type

Results

Comparing 2 independent methods for EW measurement

- ❖ Consistent especially for the stronger lines

- ❖ The score of the algorithms is not driven by statistical scatter due to measurements errors. Instead it is driven by intrinsic scatter of the EW values within the examined spectral range



Take home message and future perspectives

- ➢ Development of a new automated tool for the spectral classification of OB stars
- ➢ Our model is Luminosity Class and Metallicity independent !
- ➢ It can be used also for the spectral classification of Oe/Be stars
- ➢ Reaches 70 % accuracy !
- ➢ Maximum performance limited due to intrinsic scatter of the EWs distributions

- ➢ Inclusion of extra features (i.e. multi-band colors, Luminosity)
- ➢ RF regression for stellar parameter determination


Thank you
ekyritsis@physics.uoc.gr