Identifying massive stars in nearby galaxies, in a smart way

Grigoris Maravelias[1]*, Alceste Z. Bonanos[1], Ming Yang[1], Frank Tramper[1], Stephan A. S. de Wit[1], Paolo Bonfini[1,2]

[1] IAASARS – National Observatory of Athens, Greece, [2] University of Crete, Greece

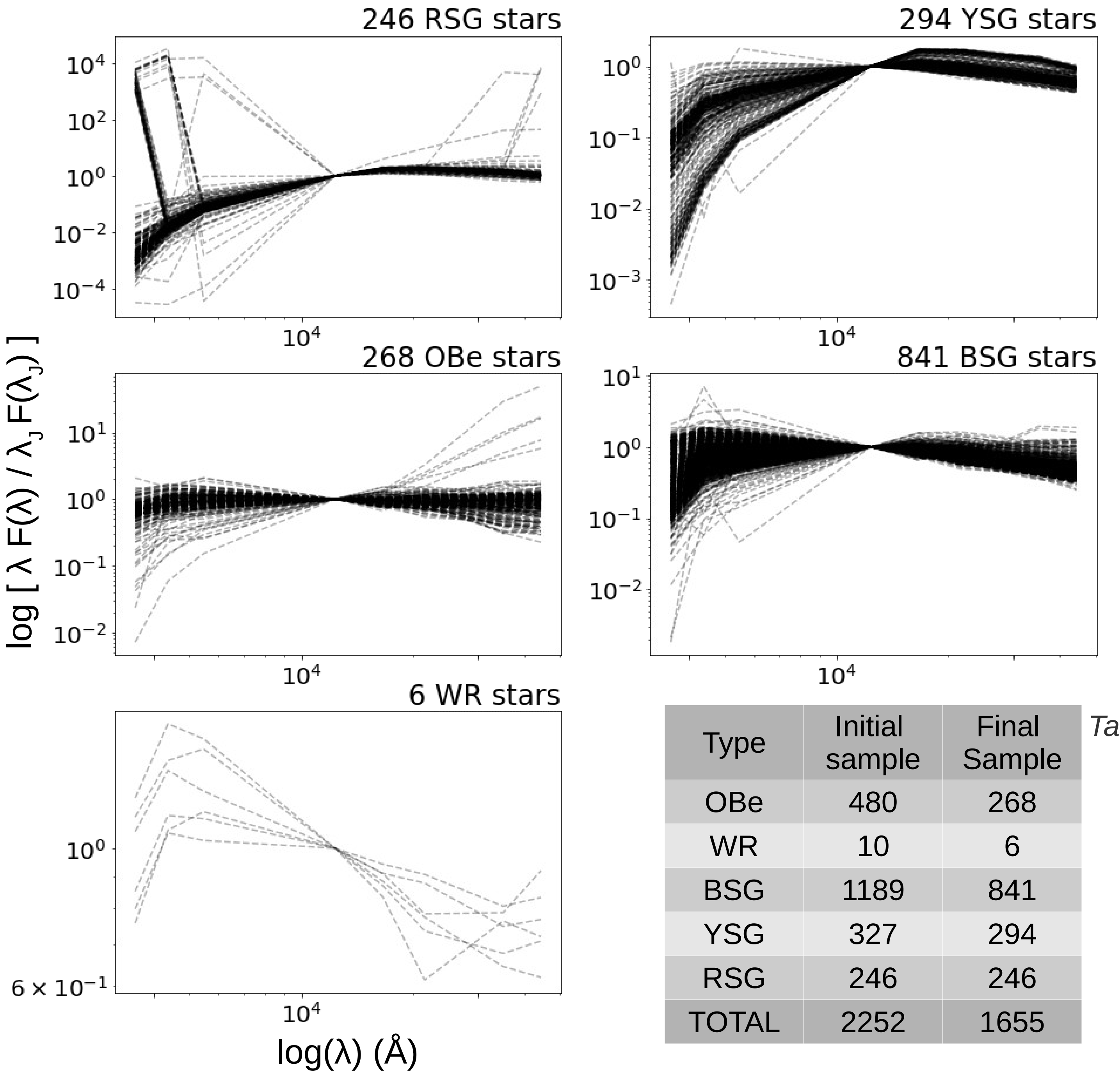
*contact:
maravelias@noa.gr

ABSTRACT

To better understand supernova remnants it is vital to establish a good perception of their progenitor stars. Important insights can only be acquired through the systematic study of these populations in different host galaxies. However, luminous massive stars may remain undetected, as they could be embedded in thick circumstellar environments due to their strong and sometimes eruptive mass-loss. To address this we have used the largest optical (e.g. Pan-STARRS, OGLE) and IR (e.g. 2MASS, Spitzer) photometric datasets to compile the most complete samples of massive stars for a number of nearby galaxies (e.g. the Magellanic Clouds, M31, M33). By taking advantage of multiple machine-learning techniques (i.e. Support Vector Machines, Random Forests, Convolutional Neural Networks) we have developed an algorithm that classifies supergiant stars with a success ratio of ~90-94% for the Magellanic Clouds. By applying this to the available photometric datasets, we can uncover previously unclassified sources, which will become our prime candidates for spectroscopic follow-up aiming to confirm both their nature and our approach.

DATA PRE-PROCESSING - THE SMC

The sample of blue stars (Blue supergiants; BSG, O/B supergiants with emission lines; OBe, Wolf-Rayet; WR) is taken from Bonanos+2010, while Yellow supergiants (YSG) and Red supergiants (RSG) are retrieved from Neugent+2010 and Davies+2018, respectively. Optical (*U,B,V,I* from Massey+2002) and IR (*J,H,Ks* from 2MASS; 3.6μm, 4.5μm, 5.8μm, 8.0μm, 24μm from Spitzer) is converted to flux ($\lambda F_\lambda = F_{0,\lambda} 10^{(-0.4 m)}$), thus removing band independency. For the model we actually use only a subset of bands (*U,B,V,J,H,K*,[3.6],[4.5]) which are normalized to the J band (see Fig. 1). Our targets need to have magnitudes in all bands, else they are removed from the training sample (see Table 1) We group into 5 major classes: WR, BSG, OBe, YSG, RSG.



Type	Initial sample	Final Sample
OBe	480	268
WR	10	6
BSG	1189	841
YSG	327	294
RSG	246	246
TOTAL	2252	1655

Table 1.

Fig. 1: Example SEDs for different classes of objects.

APPLYING ML METHODS

We split the sample we have built into training (1328 objects), and validation and test sets (166 objects each). For the SMC, we used three different machine-learning methods. For the validation sets the accuracy we achieved is:

82% for Support Vector Machines,
86% for Random Forests,
86% for Convolutional Neural Networks.

Combining the classification results from all three methods, i.e. taking the most frequent classification result (mode), we are able to achieve an accuracy of **94%** on the test sample (used for testing the algorithm).

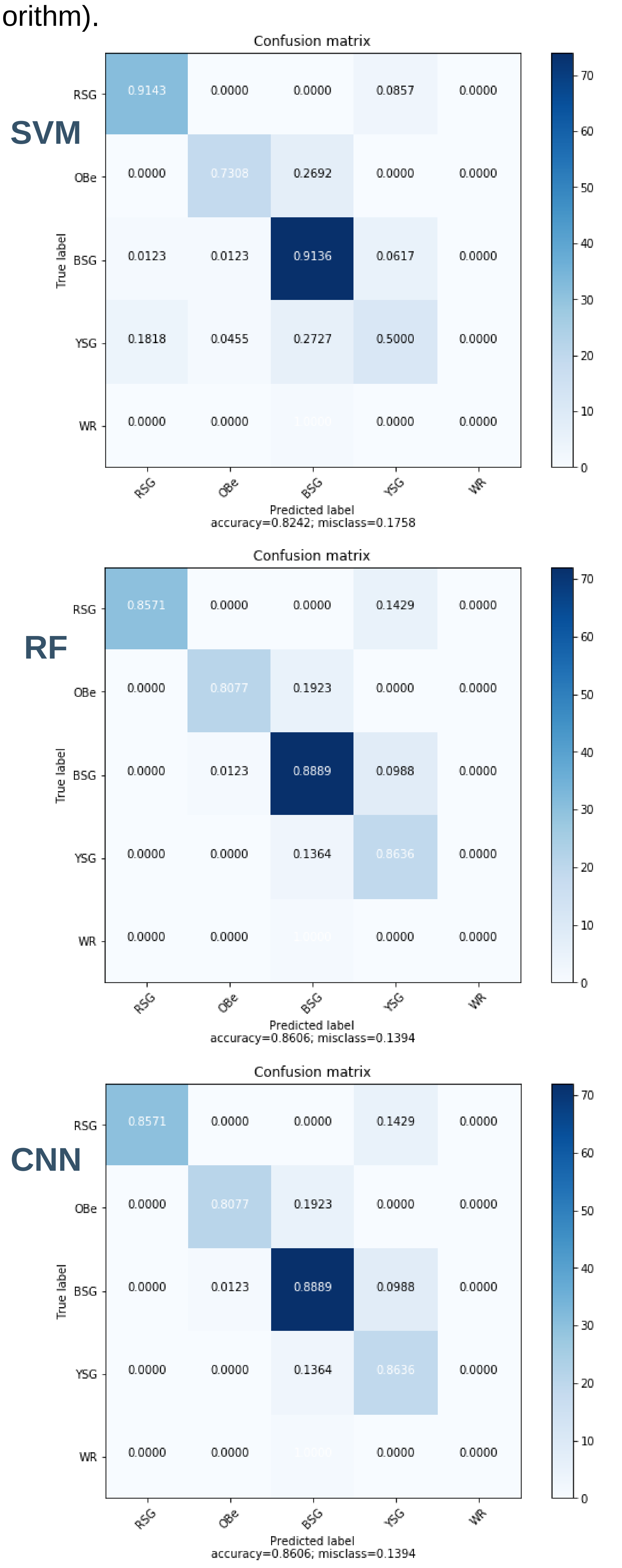


Fig. 2: Confusion matrices for the different methods used.

THE LMC CASE

Similar to the SMC we build the LMC sample of classified blue, yellow, and red supergiants from Bonanos+2009, Neugent+2012, and Davies+2018, respectively. Following the SMC approach, we convert the magnitudes of the same selected bands (*U,B,V,J,H,K*,[3.6],[4.5]) and we train the algorithm to detect the same 5 classes: WR, BSG, OBe, YSG, RSG (see Table 2). Splitting further the remaining 905 objects to training (724), validation (90), and test (91) sets, we get an accuracy of:

80% for Support Vector Machines,
82% for Random Forests,
82% for Convolutional Neural Networks,

and combining the results from all methods we get an accuracy of **90%** on the test sample.

Type	Initial sample	Final Sample
OBe	73	62
WR	91	72
BSG	370	266
YSG	213	203
RSG	306	302
TOTAL	1053	905

Table 2.

CONCLUSIONS

In this work we present our recent results from the application of a machine-learning algorithm to classify massive stars based on their multi-wavelength photometry. The datasets have been compiled by combining optical and IR surveys, and their corresponding magnitudes have been converted to fluxes before training. Taking into account the results from three different methods we are able to achieve an accuracy of more than 90% in the Magellanic Clouds. However, our preliminary results on M31 and M33 are not equally good. This is attributed to the (much) smaller sample of stars and the effect of the over-representation of BSG in this sample (that drives the training of the model). The algorithm will be used to identify new candidate massive stars for which we will perform follow-up spectroscopic observation to confirm their nature and our approach.

DATA PRE-PROCESSING FOR M31 AND M33

The source lists for M31 and M33 are not cleared from foreground sources. To account for these we use the astrometric information from GAIA DR2. To derive the necessary criteria to flag foreground sources we exclude all sources (e.g. ~146000 initial sources in M31) that satisfy any of the selected criteria (*pmra_error* ≥ 3.0 mas, *pmdec_error* ≥ 3.0 mas, *phot_g_mean_mag* ≥ 20.5, *parallax_error* ≥ 1.5 mas, *astrometric_excess_noise* ≥ 1.0) and can be considered as low-quality objects (~7000 in M31), as well as those without any estimate on the proper motion in either R.A. or Dec (e.g. due to crowding; ~61000 in M31). From the remaining sources (~78000 in M31) we produce histograms of their proper motions in RA and DEC (over their corresponding errors) and parralax/error (see Fig. 3). The optimal fit is achieved by fitting a spline to known foreground sources (i.e. all stars outside the indicated ellipse) and a Gaussian fit. From the standard deviation of the later, we may derive the 3-sigma criteria, out of which we flag stars as potential foreground stars.

Secure spectral classification for M31 and M33 is derived from Neugent +2019, Humphreys +2017, Massey +2016, Gordon +2016, Drout +2009. The lists include initially 574 (M31; shown as blue dots in Fig. 3) and 636 (M33) objects with photomery in optical and IR bands. From these we identify and remove 20 and 55 foreground stars, respectively.

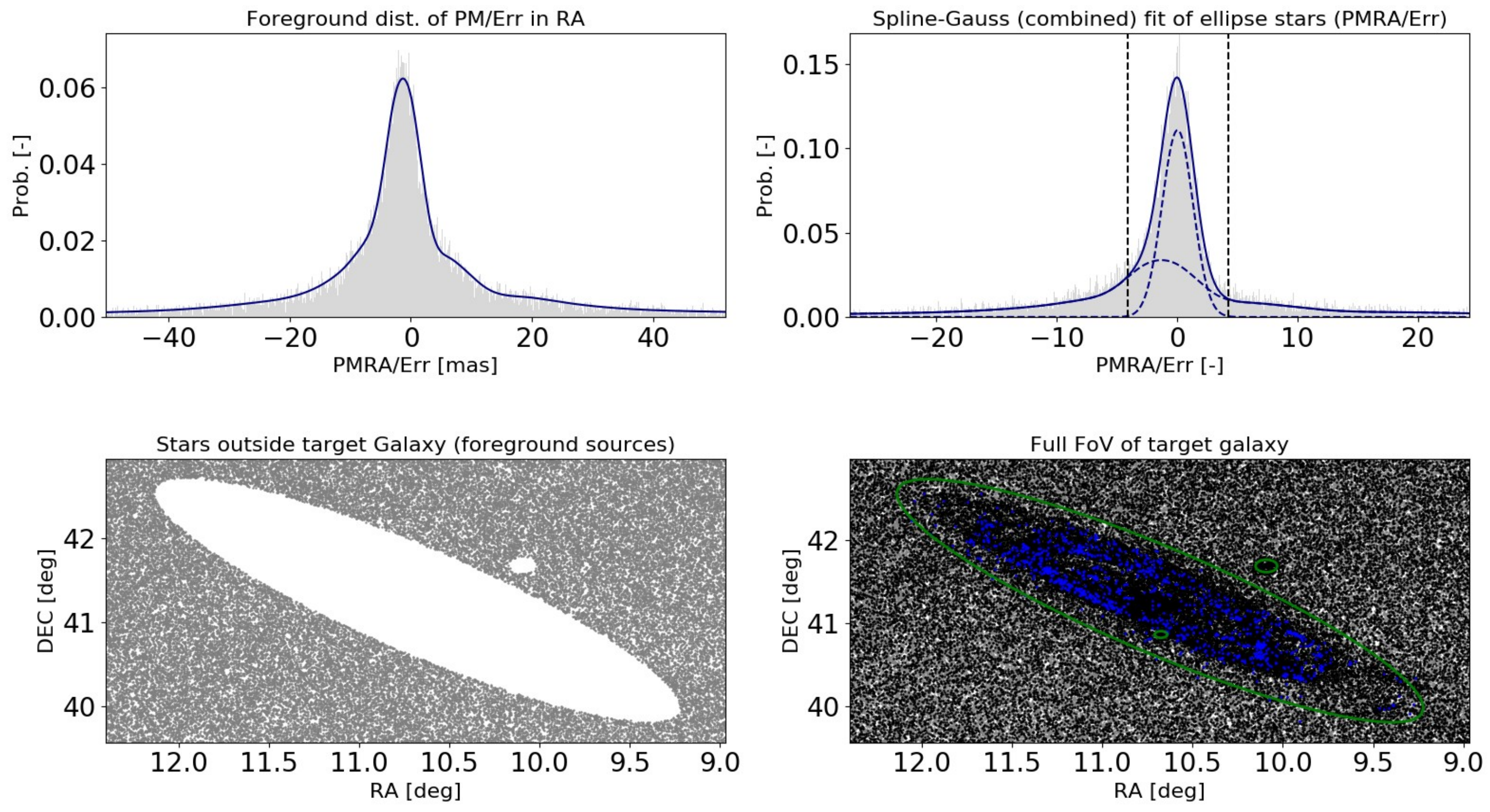


Fig. 3: Distribution and fits for the proper motion in RA (upper panels). The spatial distribution of the foreground stars outside M31 and of the massive stars with know spectral types indicated as blue dots (lower panels)

PRELIMINARY results

ML RESULTS ON M31+M33

We obtain photometry from Pan-STARRS DR2 (*g,r,i,z,y* bands) and Spitzer mission (3.6μm, 4.5μm, 5.8μm, 8.0μm, 24μm). Following the previous approach we convert all magnitudes to fluxes and we normalize all data to 4.5μm band (in this case we use all available bands; see Table 3). Due to the smaller number of objects we need to combine both galaxies. We group into 4 major classes: Luminous Blue Variables (LBVs), BSG, YSG, and RSG. However, our model is biased in favor of BSGs that dominate the sample (by ~85%). Thus, the accuracy is lower to ~65% across all methods, fitting actually the BSGs.

Table 3.

Type	Initital sample	Final sample
BSG	774	483
YSG	54	29
RSG	141	48
LBV	7	4
TOTAL	1041	612

# **References** #
Bonanos +2009, AJ, 138, 1003
Bonanos +2010, AJ, 140, 416
Davies +2018, arXiv:1804.06417
Drout +2009, AJ, 703, 441
Gordon +2016, ApJ, 825, 50
Humphreys +2017, ApJ, 844, 40
Massey +2016, AJ, 152, 82
Neugent +2010, ApJ, 719, 1784
Neugent +2012, ApJ, 749, 177
Neugent +2019, ApJ, 875, 124