MVSem Report on "Membership of Stars in Open Clusters using Random Forest with GAIA Data by Mahmudunnobe et al. [6]"

B. Rajpoot[1, *]

[1] *Department of Physics and Astronomy, Universität Heidelberg, 69117 Heidelberg, Germany*

This is a summary report of a talk given at the University of Heidelberg for the MVSem course "Dynamics of Galaxies, Star Clusters & Planetary Systems" in the Winter Semester 2022-23 on 11th Jan, 2023 under the supervision of Dr. Andreas Just, Dr. Rainer Spurzem, & Dr. Francesco Flammini Dotti. The seminar aimed at introducing the current research ongoing in the field, and the talk was aimed to introduce how Machine Learning (ML) techniques can be used to analyze big datasets from large survey missions such as GAIA, SDSS, etc., for studying specific astrophysical bodies, in this case, Open Clusters (OCs).

The paper presented in the seminar tests the performance of an ensemble-based ML method, Random Forest (RF), to estimate the membership probabilities of stars of 9 clusters based on the unprecedented 6-D data from GAIA's 2^{nd} data release (DR2) and a catalogue of OC population of Milky Way from Cantat-Gaudin et al. [2]. The RF method successfully identified new members of the clusters, which revealed further characteristics of OC, such as sub-structures beyond the previously reported cluster radius and non-main-sequence members of the clusters.

In this summary, we cover the basics of Supervised ML techniques and discuss how a RF classifier can solve the membership problem of OCs. Finally, we discuss how these data-driven methods can be further validated using the astrophysical properties of stars and clusters to improve the efficiency of these methods and what future prospects in this area lie.

Keywords: open clusters and associations: general - methods: data analysis - methods: statistical

CONTENTS

* bhavesh.rajpoot@stud.uni-heidelberg.de

I. INTRODUCTION

A. Open Clusters

Open clusters are an ensemble of $\approx 10^2 - 10^4$ stars, irregularly distributed and loosely bound to each other by mutual force of gravitation.

Anatomically speaking, a cluster should have a tidal boundary separating the cluster members from the field stars, as shown in fig. 1, and having a few heavily massed stars in the cluster can make clumps or sub-clusters inside the main cluster.

OCs form from molecular clouds, as seen in fig. 2; therefore, the members have the same ages but different masses, follow a common velocity and are younger than the later population stars. So, one can sample out the stars that follow such characteristics from a dataset by analyzing their behaviour in astrometric and photometric

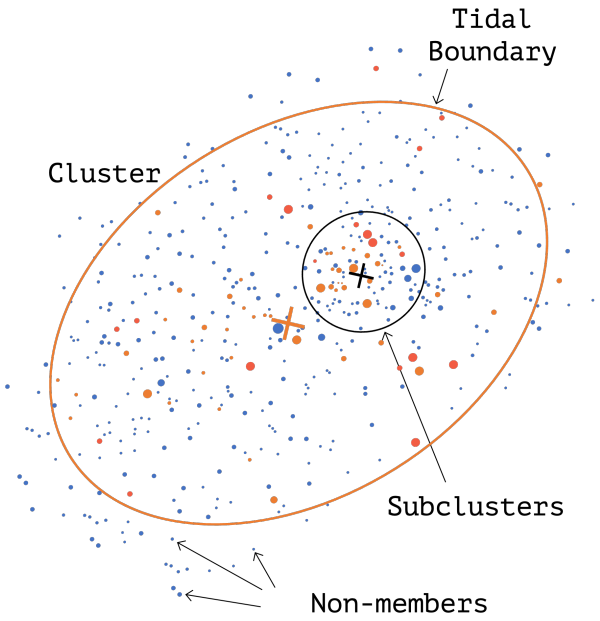

Figure 1: Anatomy of an Open Cluster

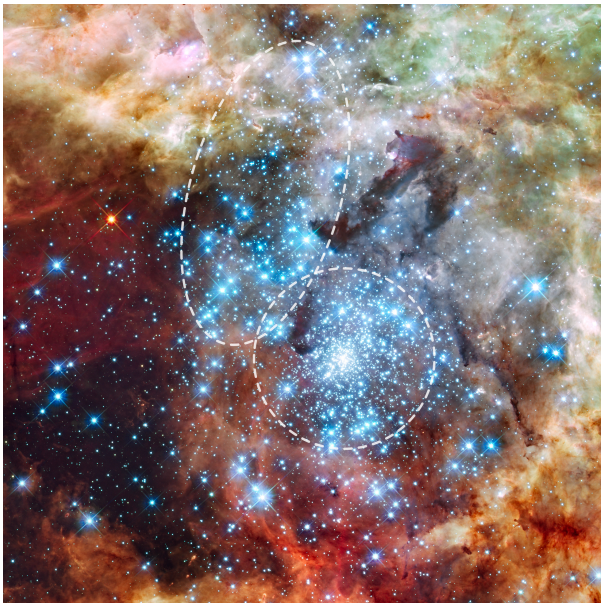

Figure 2: Two merging OCs in the Tarantula Nebula

parameter space and term them members. Table I lists the 11 parameters that can be used to quantify membership. But these samples are often contaminated by field stars, which could have similar characteristics like velocity or parallax.

Therefore, in OC studies, estimating membership probabilities, the probability of a star being a cluster member, is critical and demands both a good model and method.

Table I: Table of parameters used to estimate Membership Probability

Astrometric			Photometric	
Position	Proper Motion	Parallax	Magnitude	Reddening
δ,α	μ_δ,μ_α	ω	B, V, R, I	$E(B-V)$

B. Methods

Models can depend on the type of method we choose to use. Methods can be classified into two types, Model-Bound and Model-Free.

- Model-bound methods use purely theoretical models, such as stellar isochrones, that generate the grid of synthetic data given some input.
- Model-free methods use data-driven models where we don't imply any physical model but try to identify patterns in the dataset based on statistical analysis.

The problem is, given the nature of OCs, theoretically modelling them is very difficult as they don't have a uniform distribution like Globular Clusters, and observationally, we get limited by our instruments. The previous attempts of solving the membership problems by using both parametric [7] and non-parametric [e.g. 1, 4, 8, and references within] approaches worked till a limit but failed to completely solve the problem in a variety of scenarios as the solutions were model-dependent and therefore limited. [5]

The problem demanded a purely data-driven way to remove the limitations but keeping astrophysics in mind. So, we needed something in-between, the best of both worlds. And that was solved by a physics-based data-driven model. These models are very powerful, as shown in fig. 3, where a Neural Network fails to fit perfectly to the time evolution curve of the amplitude of a damped simple harmonic oscillator, but after incorporating some physics in the neural network, the model can reproduce the curve very precisely.

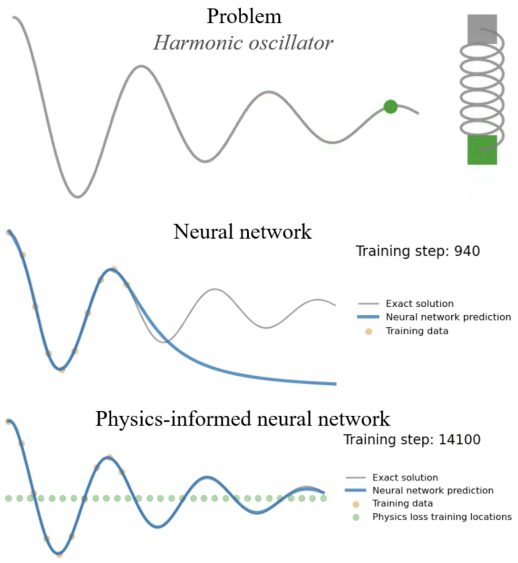

Figure 3: A Physics-informed Neural Network fitting a time evolution curve of a damped simple harmonic oscillator. Credits: Ben Mosely

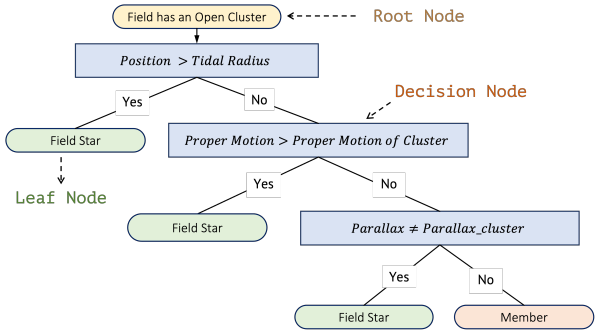

Figure 4: A labelled example Decision Tree for classifying Members from Field Stars

These physics-informed models come under the umbrella of Supervised ML techniques, where we use labelled datasets, meaning each data point contains features and a corresponding label, to train algorithms to classify data or predict outcomes accurately. Any kind of Supervised ML model is highly dependent on the training data we give it as that determines its performance and accuracy on unseen data.

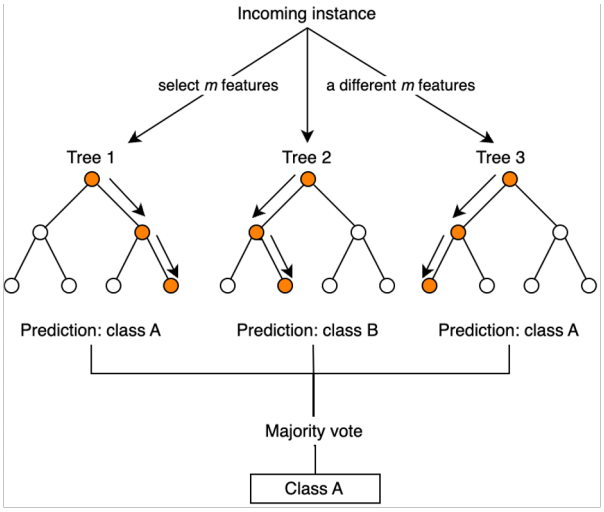

Figure 5: A labelled Random Forest Diagram

C. Crash Course on Random Forest

Among a no. of supervised classification methods, Random Forest is the most appropriate method for distinguishing between members and field stars based on their characteristics [6], as it comprises Decision Trees that mimic the human decision-making process.

- A Decision Tree classifies class labels into separate leaf nodes by making optimal splits on different features at various decision nodes, as shown in fig. 4.
- Training a tree means learning the optimal threshold value for each feature to correctly classify data points into their respective classes.
- A trained model is then used to classify unlabelled data.
- Decision Trees are very fast on tabular data and can handle large datasets but are prone to overfitting.

And as a solution to that problem, we use Random Forest.

- Random Forest classifiers use an ensemble of trees and the feature bagging technique to overcome overfitting and improve classification accuracy.

- In feature bagging, a no. of bootstrapped training data samples are independently trained in parallel to avoid correlation between features, and then an aggregate/majority of their votes are taken.
- They can also deal with large, high-dimensional datasets and has a short execution time compared to other ML methods.
- However, the model's strength is highly dependent on the quality of the training data. Messing up with training data means wreaking havoc on the model predictions.

II. SAMPLE SELECTION

Table 2 of Mahmudunnobe et al. [6] lists 9 Open Clusters (NGC 581, NGC 1893, IC 1805, NGC 6231, NGC 6823, NGC 3293, NGC 6913, NGC 2264 and NGC 2244) the authors used as a sample. Both the table and fig. 6 shows these 9 clusters are situated at different galactic latitudes and longitudes. The reason for such selection is that the more diverse the sample is, the more generalis ze the Random Forest model will become, and therefore the more powerful it will be.

III. RANDOM FOREST IN ACTION

Any ML model takes two types of parameters: model parameters and model hyperparameters.

Model Parameters: parameters the model uses to make a prediction

- Ex. Astrometric and Photometric parameters
- These are learned during the training process

Model Hyperparameters: parameters that control the learning, performance and efficiency of the ML model

- Ex. No. of decision trees, no. of features in each tree, min. samples in each leaf node, etc.
- These are not learned during the training process

As the hyperparameters control the performance, we tune these hyperparameters by performing Cross-Validation to generate a well-trained ML model.

A. Training, Testing and Validation

Initially, the complete labelled dataset is split into the Train and Test datasets, usually in a ratio of 70 : 30, respectively. The ML model learns on the Train dataset, and its performance, how accurate the model can make a prediction, is measured on the Test dataset.

To further improve model training by tuning the hyperparameters, we perform *k-fold Cross-Validation*, a method that gives the best model by performing hyperparameter tuning. Fig. 7 shows how the dataset is split.

B. Hyperparameter Tuning

Hyperparameter tuning is the process of finding the best set of values of hyperparameters that maximizes the model performance by minimizing the validation error. There are multiple algorithms to perform this search, namely *Grid Search*, where all the hyperparameter values are tested and hence becomes very exhaustive, and *Random Search*, where the algorithm begins with a few random hyperparameter values and finds the optimal ones by evaluating model accuracy.

The metric paper used for this task was *Precision*,

$$Precision = \frac{TP}{TP + FP}$$

where,

True Positive (TP): no. of times a true member is classified as a member

False Positive (FP): no. of times a true member is classified as a field star

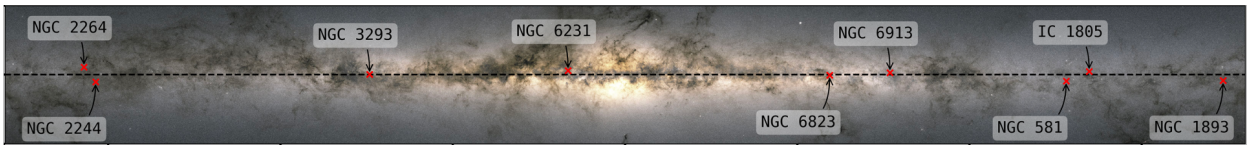

Figure 6: Position of sample clusters chosen for study with respect to the galactic centre and disk. Credits: Rajpoot, B./[6]/[3]

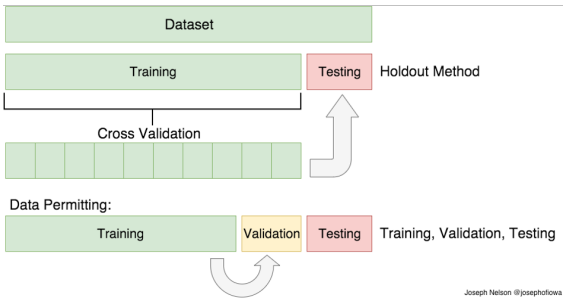

Figure 7: Visualization of how the dataset is spilt for model learning.

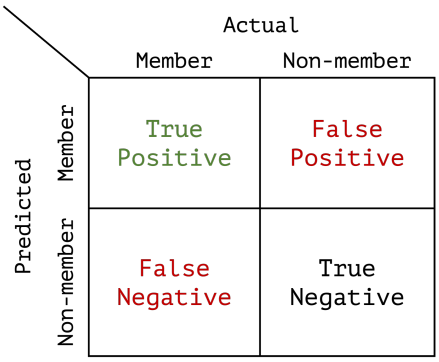

Figure 8: Confusion Matrix for Membership Problem

Both True and False Positives are calculated from a confusion matrix, fig. 8.

In [6], the authors performed a 5-fold Cross-Validation with Random Search on fig. 10 for 100 iterations and chose the model which maximized the Precision.

C. Training Data

As mentioned earlier, choosing the right training data is very important. Therefore, to balance the data and avoid overfitting, the authors took a random sample of an equal number of non-members and members for the cluster.

For Members:

- A GAIA[3] DR2 catalogue of the open cluster population in the Milky Way by Cantat-Gaudin et al. [2] was used.
- Physical Constraints:
  - Members $\in$ `search_radius` $= 2 \times$ `cluster_radius`
  - `cluster_radius` := distance of the farthest members from the center
  - Additional criteria to filter bad data:
    $parallax/parallax\_err > 3$
    $pmra\_err < 0.3$
    $pmdec\_err < 0.3$
  - Filtered members = `CG members`

By fixing the `search_radius` for each cluster, we reduced the dimensionality of data from 5D (Astrometric parameters) to 3D (Proper Motion and Parallax).

For Non-Members/Field Stars:

- Stars from a concentric ring outside the `search_radius` were taken as non-members
- GAIA DR2 catalogue was used with the same filter criterion as used for members

Lastly, in the training set, the membership probability of members was set to 1 and non-members to 0. The model is then trained on this dataset, and the resulting model is then used to classify the members and non-members of the unlabelled GAIA DR2 dataset of the search region of each cluster.

IV. RESULTS AND CONCLUSION

The model predictions, [as listed in 6, table 3, pg. 13], shows an increase of almost a factor of 2

Figure 9: The Correlation Matrix of the Prediction Table showing correlation trends in various parameters.

in the no. of members with an average precision of 92%.

A. Correlation Matrix

The correlation matrix of the prediction table 9 can be used to analyze the model's performance further. The matrix shows that:

- Precesion is
  - *mildly anti-correlated* to `search_radius, old_member & new_member`, and
  - *correlated* to `ratio of new to CG`
- Ratio of new to CG is:
  - *anti-correlated* to `old_member`
  - *correlated* to `search_radius`
- *Tight correlation* b/w `old_member & new_member`

The correlation of Precision here is astrophysically justified as we increase our `search_radius`, beyond the cluster's tidal radius, the probability of the star being a non-member would be higher, hence the Precision drops.

And the correlations between the no. of old and new members and its related measure, like ratio, speak about the dependency of our Random Forest classifier on the training set. The more the no. of old members it gets to train on, the better it performs on the unlabelled dataset.

Using the prediction table, we can also visualize the results in the form of Distribution plots in both Astrometric and Photometric parameter spaces.

B. Distribution Plots

Fig. 11 shows the distribution plots of only 2 out of 9 clusters from Mahmudunnobe et al. [6].

The similarity in proper motion and parallax tells that the results agree well with CG members and overall the new members add more information to the cluster studies. The CMDs of both clusters in fig. 11 shows that by not using Photometric information during the ML model building, the classifier was able to identify new non-Main-Sequence members, such as Pre-Main Sequence stars, Unresolved Binaries, etc., in the cluster region.

C. Kernel Density Estimate (KDE) Plots

As the OCs do not follow a specific radial profile, the authors plot Kernel Density Estimates (KDEs) of the Member and Proper Distribution, fig. 12. In doing so, the KDEs reveal bimodalities in both the Position and Proper Motion space of NGC 1893, NGC 2244, NGC 2264 and NGC 6913, which means the presence of substructures and dynamic interactions inside the clusters. KDE plots of some clusters like NGC 581 only showed bimodalities in Position Space, indicating asymmetric stellar distribution in the cluster.

D. Conclusion

- Results indicate that the ensemble-based machine-learning-based method is highly suitable for membership determination of open clusters in high dimensional feature space.

- Increment in the overall no. of members by almost 2–3 times. Therefore, the accuracy of determining various physical parameters of the clusters, such as distance, extinction and mass function, improved.
- Identification of sub-structures in NGC 1893, NGC 2244, NGC 2264 and NGC 6913.
- Identification of variables, pre-main sequence stars (in NGC 1893, NGC 3293), unresolved binary sequences (in NGC 6231), as well as all other possible non-main-sequence members of the cluster from its CMD.

V. FUTURE PROSPECTS

- A comparative study to benchmark other Supervised ML methods for membership determination
- Possible study with GAIA EDR3 dataset

VI. HANDOUT

In the spirit of the advent of Machine Learning techniques in Astronomy, a Jupyter Notebook was created to showcase the process of using such techniques. Here in the GitHub repository, https://github.com/Bhavesh012/oc_with_rf, one can find that Jupyter Notebook that recreates the results of the paper Mahmudunnobe et al. [6].

The complete process, from querying the datasets from GAIA servers to training the random forest model and generating the plots, has been explained in the notebook. The code is written in Python and can be easily run on Jupyter Notebook/Lab, VS Code or Google Colab. One can find complete instructions to create a conda environment for running this notebook in the README.md file in the repository.

For any problems, either in running the notebook or understanding a step, please get in touch with me. I would be happy to help.

ACKNOWLEDGMENTS

I want to thank my supervisor, Dr. F. Flammini Dotti, for listing this paper for the seminar and allowing us to appreciate the Machine Learning techniques as really useful tools in astrophysics research, and also for clearing my doubts.

I would also like to thank Dr. F. Hamprecht for organizing the core course, Machine Learning and Physics, this winter semester. The course proved very useful not only in advancing understanding and developing the skill set for strengthening our research with such valuable tools but also in developing the personality and thought process as a whole to become a better researcher when it comes to data analysis.

I want to thank the organizers of this Master's seminar, Dr. A. Just, Dr. R. Spurzem and Dr. F. Flammini Dotti, for giving us a chance to know the field of Galactic Dynamics more closely with recent research going on in the area. Therefore, I would also like to thank the participants of this year's seminar for giving such wonderful talks throughout the semester.

Last but not least, I would like to thank this beautiful city of Heidelberg for inspiring me every day to explore and solve the mysteries of nature and for taking away my stress with its beautiful sunsets.

[1] J. Cabrera-Cano and E. J. Alfaro, Astronomy and Astrophysics **235**, 94 (1990). 2
[2] T. Cantat-Gaudin, C. Jordi, A. Vallenari, et al., Astron. and Astrophys. **618**, A93 (2018). 1, 5
[3] Gaia Collaboration, T. Prusti, J. H. J. de Bruijne, et al., Astron. and Astrophys. **595**, A1 (2016). 5
[4] G. Javakhishvili, V. Kukhianidze, M. Todua, and R. Inasaridze, Astronomy & Astrophysics **447** (3), 915 (2006). 2
[5] A. Krone-Martins and A. Moitinho, Astronomy & Astrophysics, Volume 561, id.A57, <NUMPAGES>12</NUMPAGES> pp. **561**, A57 (2014). 2
[6] M. Mahmudunnobe, P. Hasan, M. Raja, and

Table 2 The random selection grid with the chosen range of values for important RF model parameters

Model parameter	Chosen values to select from
bootstrap (Whether bootstrapping the samples)	True, False
ccp_alpha (Complexity parameter)	2^{-10}, 2^{-9}, 2^{-8}, 2^{-7}, 2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 0
max_depth (Maximum depth of the tree)	10, 20, 30, 40, 50, 60,70, 80, 90, 100, 110, None
max_features (Number of features in each tree)	'auto', 'sqrt'
min_samples_leaf (Minimum samples required for a leaf node)	1, 2, 4
min_samples_split (Minimum samples required to split a node)	2, 5, 10
n_estimators (Number of decision trees)	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000

Figure 10: The Hyperparameter table from Mahmudunnobe et al. [6]

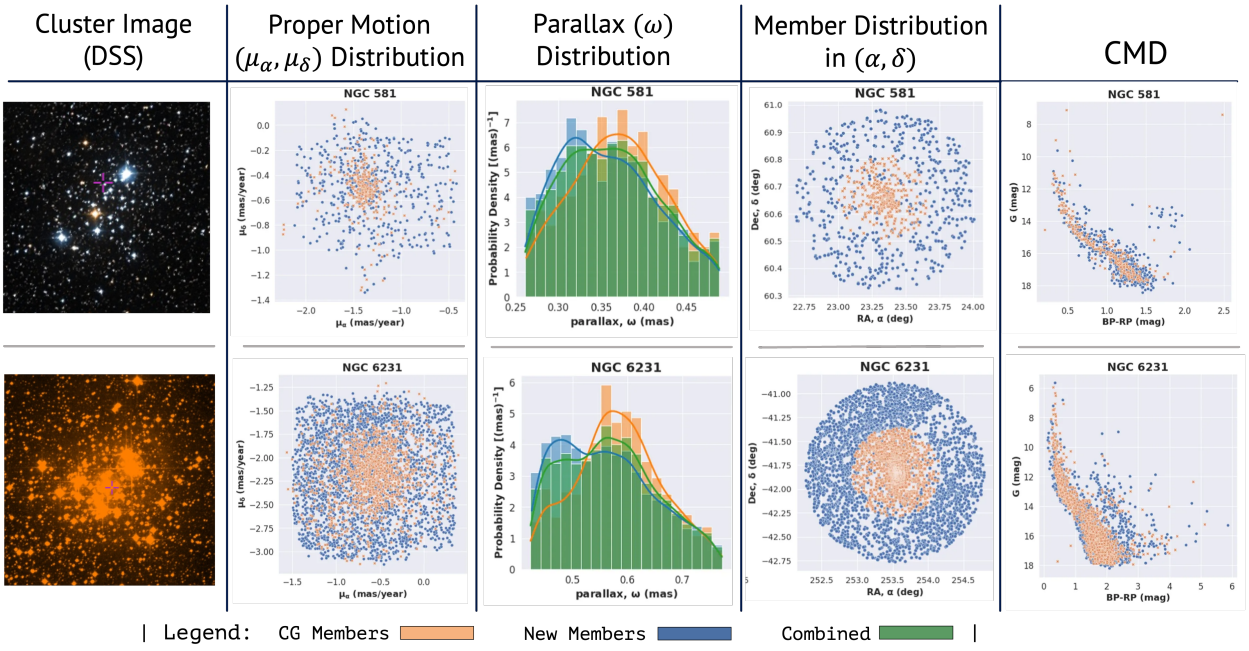

Figure 11: Distribution plots of NGC 581 and NGC 6231 for both old and new members in Astrometric and Photometric parameter spaces, from Mahmudunnobe et al. [6]

S. N. Hasan, The European Physical Journal Special Topics **230** (10), 2177 (2021). 1, 3, 4, 5, 6, 7, 8, 9

[7] W. L. Sanders, Astronomy and Astrophysics, Vol. 14, p. 226-232 (1971) **14**, 226 (1971). 2

[8] J. L. Zhao and Y. P. He, Astronomy and Astrophysics **237**, 54 (1990). 2

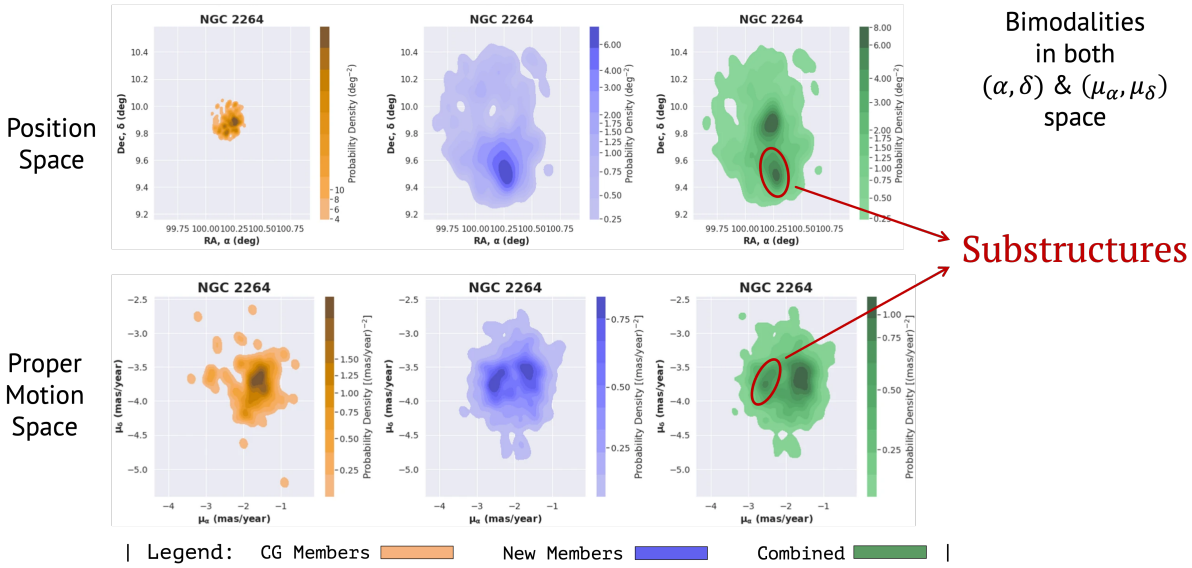

Figure 12: Position and Proper Motion KDE plot of NGC 2264 from Mahmudunnobe et al. [6]