Membership Analysis of Open Clusters Using Machine Learning on Gaia Data Release 3

M.Noormohammadi *and A.Javadi†

School of Astronomy, Institute for Research in Fundamental Sciences (IPM), P.O. Box 1956836613, Tehran, IRAN

Abstract

The first and most important stage in studying open clusters is the detection of reliable members. Since open clusters form and evolve within the inner disk of the galaxy, they are surrounded by numerous field stars, making membership determination challenging. Because cluster members originate from the same molecular clouds, they exhibit similar physical parameters—such as proper motion and parallax—and align along a single main sequence in the color-magnitude diagram. For this reason, machine learning algorithms can identify cluster members as familiar data among field stars. In this work, we used a combination of unsupervised machine learning algorithms—DBSCAN and GMM—based on astrometric parameters, proper motion, parallax, and position from the latest Gaia data release (GDR3). After selecting reliable members within the tidal radius, we applied the Random Forest algorithm to detect members beyond the tidal radius, utilizing proper motion, parallax, G-band magnitude, and BP-RP color index as classification features. By leveraging accurate data and a suitable method capable of handling large datasets, we identified members both inside and beyond the tidal radius of clusters. We observed clusters with a comprehensive field of view and analyzed their morphology. All members outside the tidal radius fall within the range of proper motion, parallax, and the main sequence of members inside the tidal radius.

Keywords: *open cluster-data analysis-machine learning algorithms-Gaia data release 3*

1. Introduction:

Open clusters originate from a single interstellar cloud, resulting in member stars that share a common chemical composition and exhibit similar astrometric properties, such as position, proper motion, and parallax (Lada & Lada, 2003). A critical first step in the study of open clusters is the reliable identification of member stars using high-precision data. Over the past decade, the combination of Gaia data releases and machine learning techniques has significantly enhanced this process. Numerous studies have employed Gaia astrometric data alongside both supervised and unsupervised machine learning algorithms to determine cluster membership (Cantat-Gaudin et al., 2018, Gao, 2018, Hunt & Reffert, 2024, Noormohammadi et al., 2023, 2024).

The Pleiades is a well-known young open cluster, with an estimated age between 110 and 160 Myr and located at a distance of approximately 134 parsecs (Gossage et al., 2018). In this work, we apply a hybrid approach combining two unsupervised algorithms—DBSCAN and Gaussian Mixture Models (GMM)—with one supervised algorithm, Random Forest, to identify reliable members of the Pleiades cluster using data from Gaia Data Release 3 (GDR3) (Gaia Collaboration et al., 2023). In Section 2, we describe our data selection and method; in Section 3, we present our results and discuss them; and finally, we summarize our study in Section 4.

2. Data and Method:

To identify the full membership of the Pleiades cluster, we adopted a spatial selection radius of 6.5 degrees centered on the cluster, consistent with the approach used by Gao (2019). In this study, we applied

*Monoorastro@ipm.ir, Corresponding author
†Atefeh@ipm.ir

doi:https://doi.org/10.52526/25792776-25.72.2-437

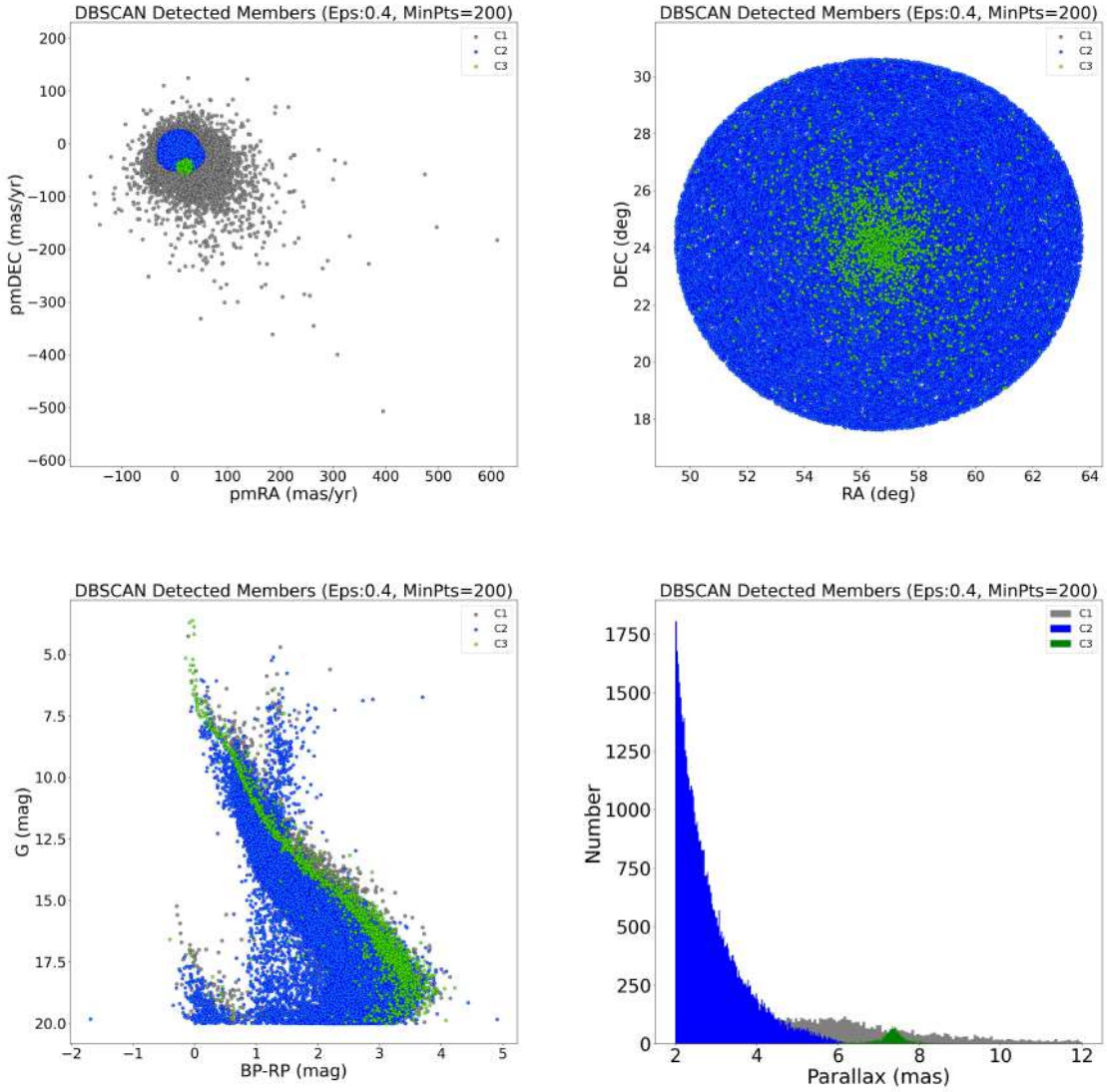

Figure 1. DBSCAN-detected members. The upper left panel shows position space, the upper right panel shows proper motion, the bottom left panel shows the Color-Magnitude Diagram, and the bottom right panel shows parallax. As can be seen, the Pleiades candidate members belong to cluster 3, which is illustrated in green.

.

initial filtering based on parallax and G-band magnitude. Specifically, we selected stars with parallaxes in the range [2, 12] mas and G magnitudes brighter than 20 mag. Additionally, we required completeness in all key astrometric and photometric parameters: right ascension (RA), declination (Dec), parallax, proper motions (pmRA and pmDEC), G magnitude, and BP–RP color index. After applying these criteria, approximately 63,609 sources were retained for further analysis using the DBSCAN clustering algorithm.

Prior to applying DBSCAN, the input features—proper motion and parallax—were normalized. The algorithm was then executed with parameters MinPts = 200 and Eps = 0.4, using pmRA, pmDEC, and parallax as input dimensions. Fig 1 displays the DBSCAN cluster detections. As shown, candidate members of the Pleiades cluster are located in Cluster 3. This initial clustering identified approximately 1,857 candidate members potentially associated with the Pleiades.

In the second step, we applied the Gaussian Mixture Model (GMM) algorithm to the DBSCAN-selected members using RA, Dec, pmRA, pmDEC, and parallax. Based on the Bayesian Information Criterion (BIC), shown in Fig 2, we selected five clusters for the GMM model. To refine membership reliability, we focused on GMM clustering in position, proper motion, and parallax space. However, some stars—despite lying

doi:https://doi.org/10.52526/25792776-25.72.2-437

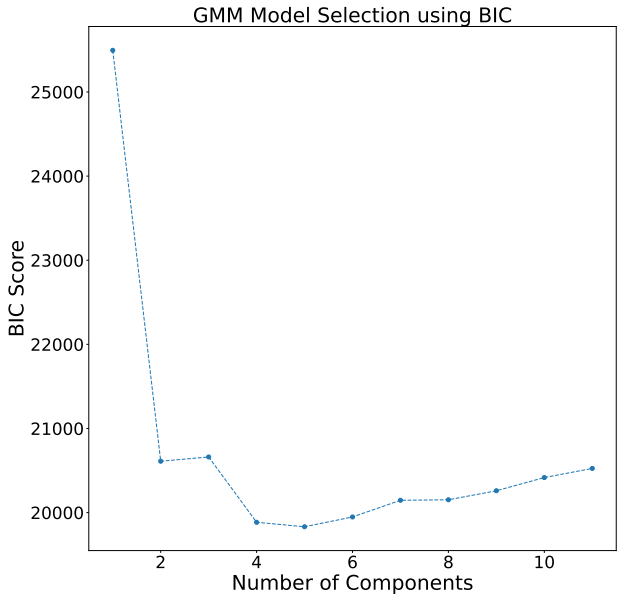

Figure 2. The BIC score for the Gaussian Mixture Model (GMM). Based on the BIC values, the optimal number of components for the GMM is 5.

outside the GMM-defined clusters—exhibited similar proper motion and parallax values and were located along the main sequence of the GMM-detected members.

To recover these potential members, we trained a Random Forest classifier using GMM-detected members as cluster stars and DBSCAN-rejected stars as field stars. We then applied the trained model to stars identified by DBSCAN but not retained by GMM. This approach allowed us to improve membership classification, including stars both within and beyond the tidal radius of the Pleiades cluster.

3. Results and Discussion:

Fig 3 shows GMM detection members with probability higher than 0.5. As could be seen Pleiades members are in C 5 that shows by blue color. Finally GMM detected 843 members with probability higher than 0.5. As could be seen in Fig 3 some members in other clusters have same value of proper motion and parallax to C 5 and also lied in same distribution of CMD. To detect these members, we applied the Random Forest algorithm. First, we trained the algorithm using the following data conditions:
1. Data not selected by DBSCAN and lying within the parallax range [5.5 , 9] mas were considered field stars (3925 sources).
2. Data identified by GMM as Pleiades members were considered cluster members.
After training, we applied the Random Forest algorithm with hyperparameters (n_estimators=500, max_depth=30, criterion=gini) to the data selected by DBSCAN but not classified as Pleiades members by GMM. The classification was based on five parameters: pmRA, pmDEC, parallax, BP-RP, and G magnitude. Random Forest detected 390 members with probability higher than 0.5.

Fig. 4 shows the members detected by the Random Forest algorithm in red, alongside the GMM-detected members shown in blue, across position, proper motion, parallax, and CMD diagrams. As can be seen, all members identified by Random Forest lie within the range of GMM-detected members in proper motion, parallax, and CMD, although they differ in positional space.

The tidal radius was calculated by fitting a King profile (King, 1962) to the radial distribution. We calculated 10.58 pc for tidal radius and we found 55 members of Pleiades outside this radius. Fig 5 upper panel shows the King profile fitting (left panel) and the distribution of members inside and outside the tidal

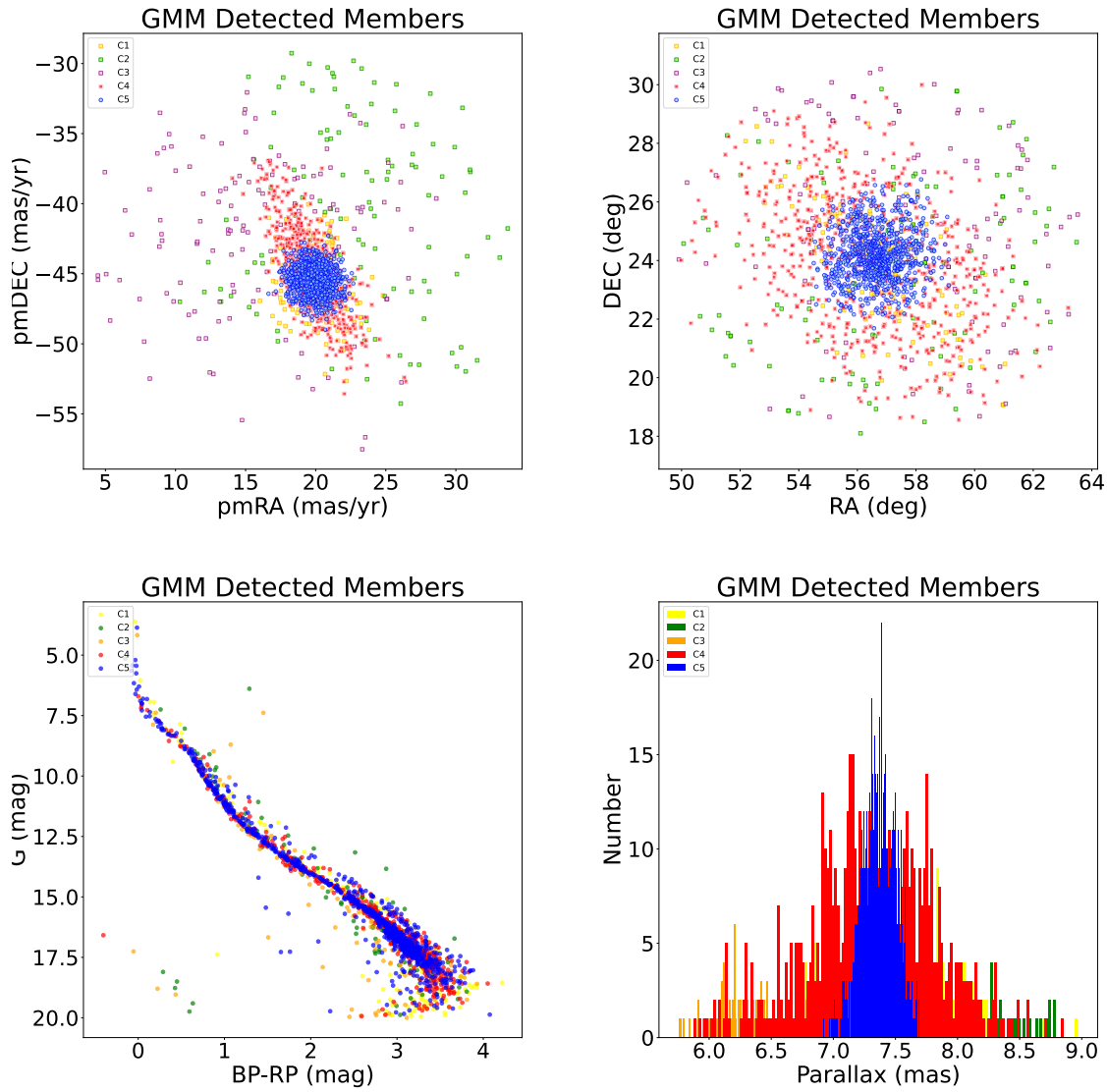

Figure 3. GMM-detected members. The upper left panel shows position space, the upper right panel shows proper motion, the bottom left panel shows the Color-Magnitude Diagram, and the bottom right panel shows parallax. As can be seen, the Pleiades candidate members belong to cluster5, which is illustrated in blue.

.

radius (right panel). Based on the young age of the cluster, it is expected that most members reside within the tidal radius.

We also investigated mass segregation by calculating the cumulative distribution function (CDF). Fig 5, bottom panel (left), shows the CDF for the Pleiades cluster. Stars with luminosity greater than 2 $\frac{L}{L_\odot}$ are considered high-mass stars, those with luminosity between 1 and 0.1 $\frac{L}{L_\odot}$ are classified as intermediate-mass stars, and stars with luminosity below 0.1 $\frac{L}{L_\odot}$ are considered low-mass stars. As can be seen, the cluster exhibits a degree of mass segregation. We also estimated an age of approximately 120 Myr for the Pleiades using Perren, G. I. et al. (2015) as shown in the bottom right panel of Fig 5.

4. Conclusion:

In this work, we employed a combination of machine learning techniques to identify reliable members of the young open cluster, the Pleiades. Member candidates were detected across a wide field of view. Additionally, we calculated the tidal radius and classified stars located both inside and outside this boundary.

doi:https://doi.org/10.52526/25792776-25.72.2-437

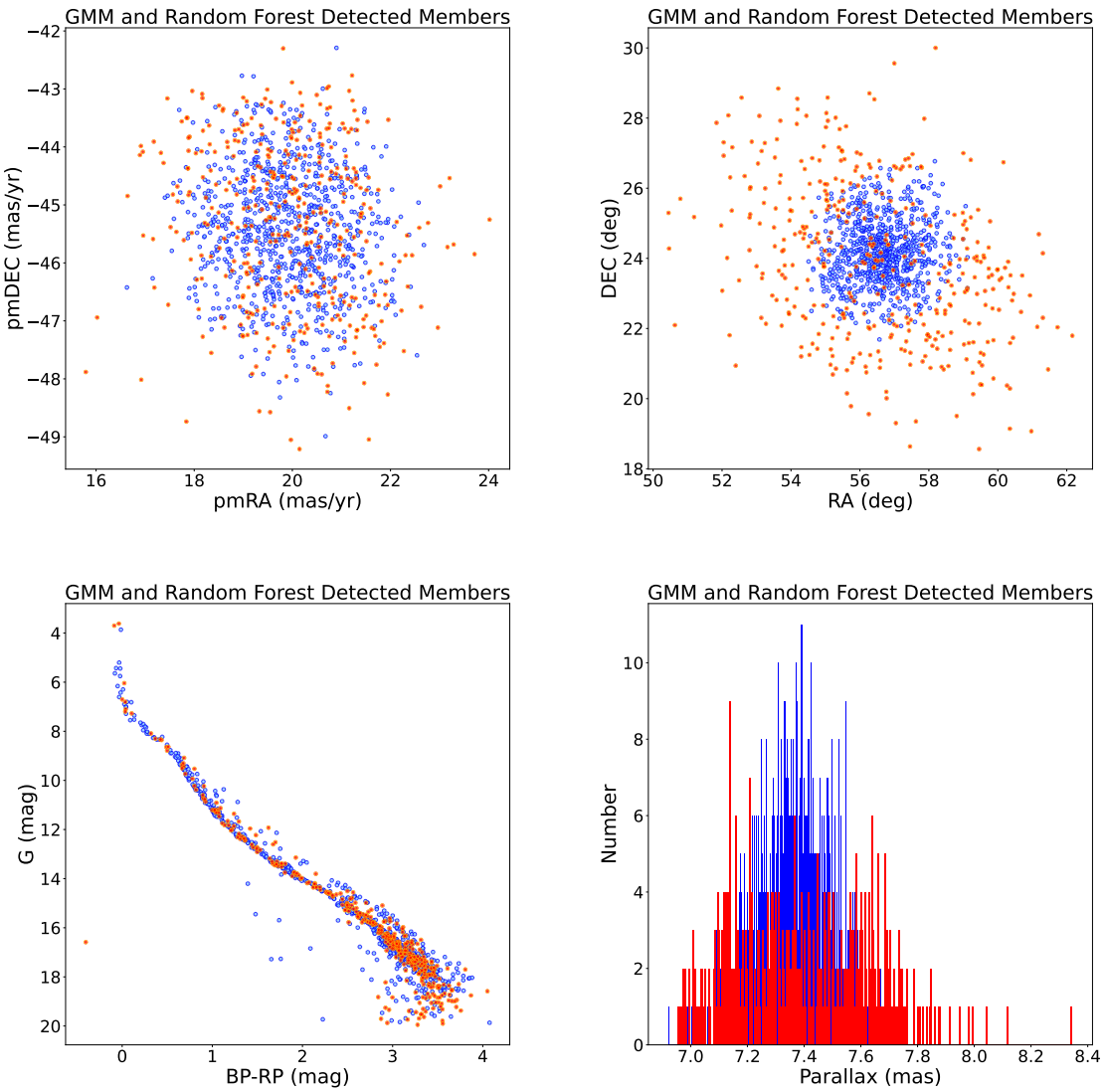

Figure 4. GMM detected (blue dots) and Random Forest-detected members (red dots). The upper left panel shows position space, the upper right panel shows proper motion, the bottom left panel shows the Color-Magnitude Diagram, and the bottom right panel shows parallax.

.

Owing to the cluster's young age, a large fraction of its members were found within the tidal radius. We computed the cumulative distribution function (CDF) to investigate mass segregation and found evidence supporting its presence in the Pleiades. Finally, we estimated the age of the Pleiades to be approximately 120 Myr by fitting an isochrone to the reliably detected members.

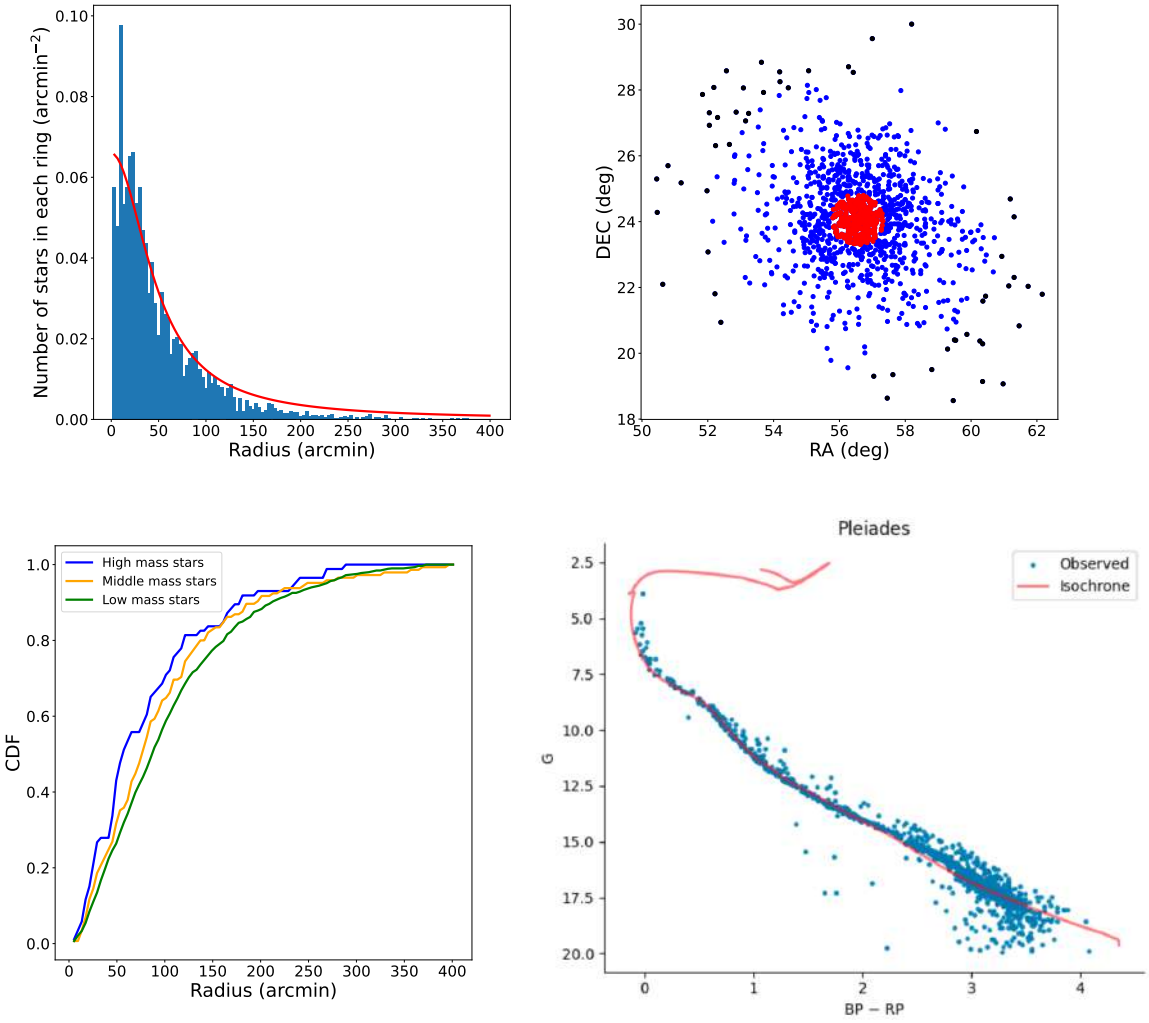

Figure 5. Physical parameters of the Pleiades based on our selected member data. The upper left panel shows the radial density profile with a fitted King model (red line). The upper right panel displays members located outside the tidal radius (black dots) and those inside the tidal radius (blue dots). The bottom left panel presents the cumulative distribution function (CDF), providing evidence of mass segregation. The bottom right panel shows the isochrone fitting, indicating an estimated age of 120 Myr for the Pleiades.

Acknowledgements

The data used in this work are from Gaia DR3, available at https://gea.esac.esa.int/archive/. We are also willing to share our dataset upon research request. Additionally, we would like to thank Perren, G. I. et al. (2015) for their isochrone fitting method used to estimate the age of the Pleiades.

References

Cantat-Gaudin T., et al., 2018, Astron. Astrophys. , 618, A93

Gaia Collaboration et al., 2023, Astron. Astrophys. , 674, A1

Gao X., 2018, Astrophys. J. , 869, 9

Gao X.-h., 2019, Publ. Astron. Soc. Pac. , 131, 044101

Gossage S., Conroy C., Dotter A., Choi J., Rosenfield P., Cargile P., Dolphin A., 2018, Astrophys. J. , 863, 67

Hunt E. L., Reffert S., 2024, Astron. Astrophys. , 686, A42

King I., 1962, Astron. J. , 67, 471

Lada C. J., Lada E. A., 2003, Ann. Rev. Astron. Astrophys. , 41, 57

Noormohammadi M., Khakian Ghomi M., Haghi H., 2023, Mon. Not. R. Astron. Soc. , 523, 3538

Noormohammadi M., Khakian Ghomi M., Javadi A., 2024, Mon. Not. R. Astron. Soc. , 532, 622

Perren, G. I. Vázquez, R. A. Piatti, A. E. 2015, A&A, 576, A6