Detection of Open Cluster Morphology Using Machine Learning Methods

Mohammad Noormohammadi * and Atefeh Javadi†

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

Abstract

Detecting cluster morphology is crucial for studying the formation and evolution of clusters and the interaction between galaxies and clusters. Achieving this requires accurate data and a reliable method to identify cluster members both inside and beyond the tidal radius. Machine learning techniques, combined with Gaia data releases, offer a valuable opportunity for determining cluster membership beyond the tidal radius, providing a detailed view of cluster morphology. In this study, we employed three machine learning algorithms to identify reliable cluster members using Gaia Data Release 3. First, we applied DBSCAN to detect cluster member candidates based on parallax and proper motion. Next, we used Gaussian Mixture Models to refine the selection within the tidal radius by incorporating positional, proper motion, and parallax data. To identify members beyond the tidal radius, we trained a Random Forest model using members detected by DBSCAN and GMM. The Random Forest classifier was applied based on proper motion, parallax, G magnitude, and the BP-RP color index. By integrating these three machine learning algorithms, we successfully identified both cluster members and escapees within and beyond the tidal radius. This approach enabled us to analyze cluster morphology over a wide field of view extending from the cluster center.

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

1. Introduction

Star clusters are fundamental tracers for investigating stellar formation and evolution. Open clusters, in particular, provide invaluable insights into the chemical and dynamical evolution of the Milky Way because they reside primarily within the galactic disk. Nevertheless, accurately distinguishing true cluster members from the dense background of field stars is a formidable task, especially in low-galactic-longitude fields characterized by high stellar crowding and interstellar extinction.

The first efforts for detecting cluster members were established by Sanders (1971). They constructed a method based on the Gaussian distribution of cluster members' proper motions. However, this traditional method suffers from several significant limitations:

1. It is not suitable for handling massive datasets.
2. It relies on only two parameters: the proper motion components in right ascension and declination.
3. It assumes that the proper motion distribution of cluster members forms a circular shape, while field stars follow an elliptical distribution—an assumption that does not hold true for all clusters.

Several subsequent methods were developed to address the limitations of Sanders' approach; however, most of them fail to yield satisfactory results as the volume and dimensionality of modern datasets dramatically increase.

Machine learning algorithms offer powerful capabilities for identifying patterns and grouping similar data within massive observational datasets. Consequently, the application of machine learning techniques for reliable member selection in open clusters has been explored in several recent studies (Agarwal et al., 2021, Cantat-Gaudin et al., 2018, Gao, 2018, Noormohammadi & Javadi, 2025, Noormohammadi et al., 2023, 2024, 2025). Furthermore, the third data release of the Gaia mission (Gaia DR3) (Gaia Collaboration et al., 2023) provides unprecedentedly complete and accurate astrometric and photometric information about the Milky Way, particularly for open cluster populations. Combining advanced machine learning algorithms

*monoorastro@ipm.ir, Corresponding author
†atefeh@ipm.ir

doi: https://doi.org/10.52526/25792776-26.73.2-334

with Gaia DR3 data presents a unique opportunity to investigate cluster membership and morphology with high precision.

In this work, we employ a combination of two unsupervised (DBSCAN and GMM) and one supervised (Random Forest) machine learning algorithms to identify reliable cluster members and map their morphology over a wide field of view. Utilizing Gaia DR3 astrometric data, this approach is applied to three open clusters: NGC 1582, NGC 5460, and NGC 7762.

The remainder of this paper is organized as follows: Section 2 describes the observational data and sample selection. Section 3 details the methodology and machine learning algorithms used. Section 4 presents the results and discussion, and finally, Section 5 provides a summary and conclusions.

2. Data

In this work, we utilize astrometric and photometric data from the third data release of the Gaia mission (Gaia DR3). For each of the three open clusters—NGC 1582, NGC 5460, and NGC 7762—we retrieved sources within a circular region with a 150 arcmin radius centered on the cluster coordinates. The extraction is based on five astrometric parameters (RA, DEC, pmRA, pmDEC , parallax) and two photometric bands (G magnitude and BP-RP color index).

To ensure data quality, we applied standard quality filters: we retained stars with a Renormalised Unit Weight Error (RUWE) of RUWE<1.4 and a parallax over error>5, while discarding sources with missing or incomplete astrometric and photometric entries. After applying these filters, the final cleaned sample sizes for NGC 1582, NGC 7762, and NGC 5460 consist of 114,089, 113,816, and 156,846 stars, respectively.

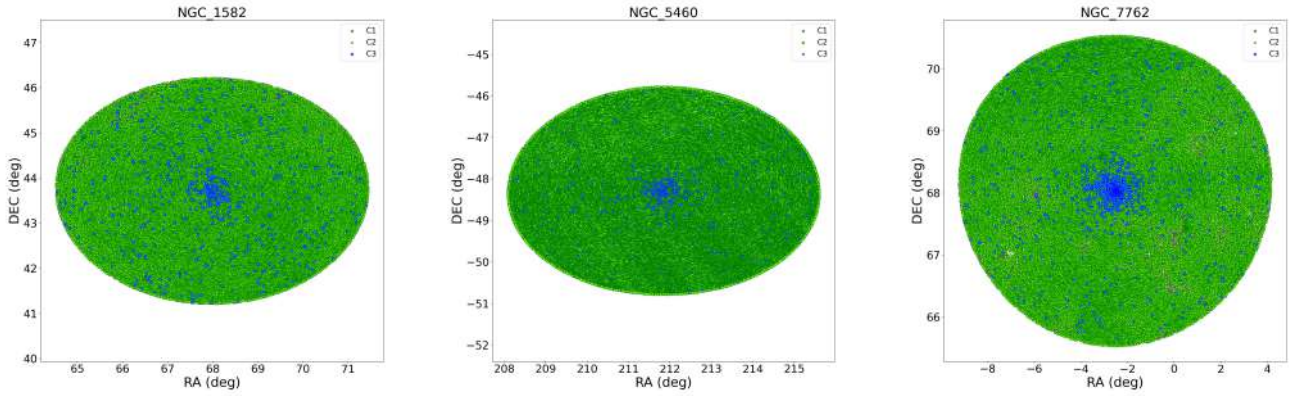

Figure 1. Spatial distribution of the DBSCAN-detected members. As can be seen, the cluster candidate members belong to cluster 3, which is illustrated in blue.

.

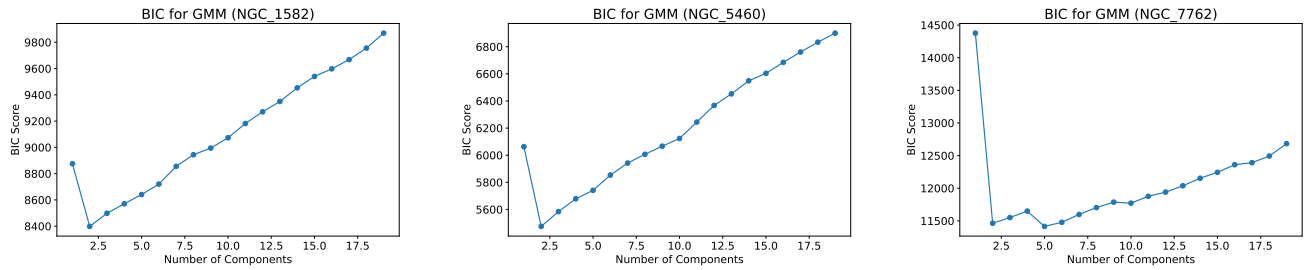

Figure 2. The BIC score for the Gaussian Mixture Model (GMM). Based on the BIC values, the optimal number of components for the GMM is 2.

.

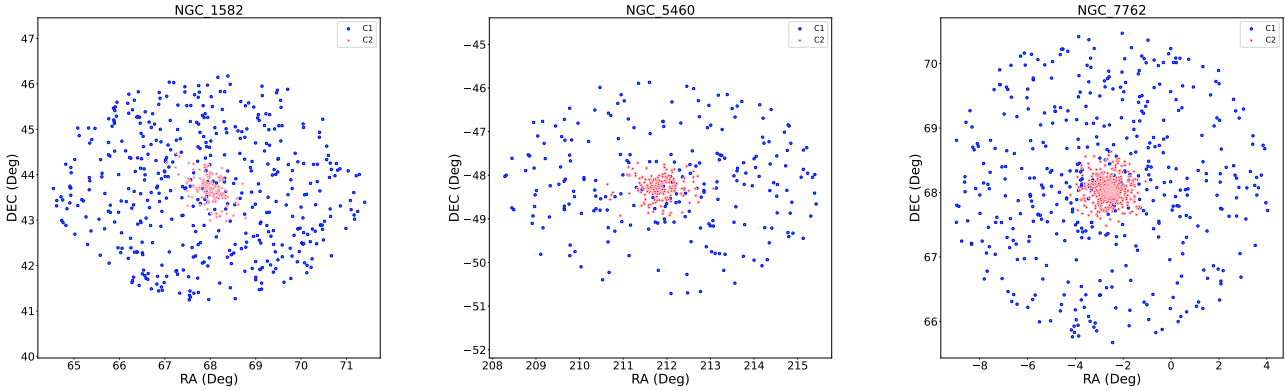

Figure 3. Spatial distribution of the GMM-detected members. As can be seen, the cluster candidate members belong to cluster 2, which is highlighted in pink.

.

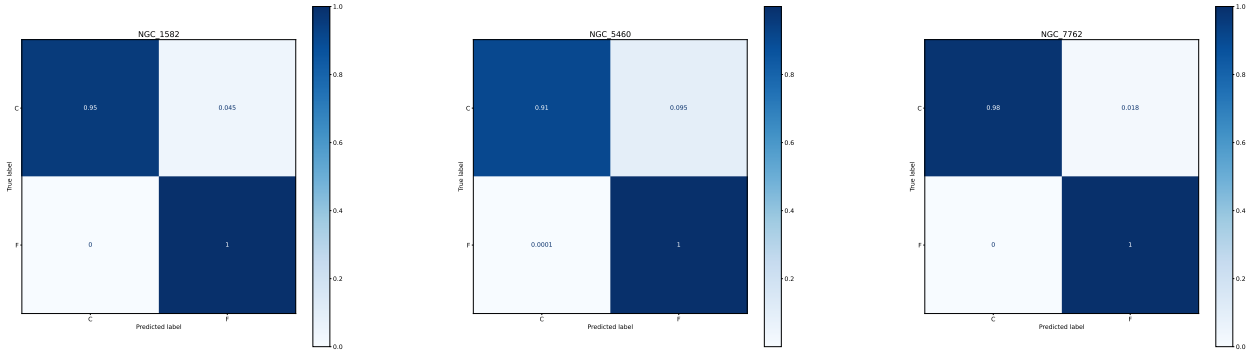

Figure 4. Confusion matrices obtained from the Random Forest classification for the three open clusters (NGC 1582, NGC 5460, and NGC 7762) using the optimized hyperparameters.

.

3. Method

In this work, we build upon our previous methodologies (Noormohammadi et al. (2023, 2024)), as outlined below:

In the first step, after data normalization, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm was applied to the sample sources using three parameters: proper motion Right Ascension (pmRA), proper motion Declination (pmDEC) and parallax. Optimal values for the hyperparameters minPts and eps were carefully selected for each cluster. The primary goal of applying DBSCAN at this initial stage is to isolate high-density core candidates and increase the signal-to-noise ratio, as Gaussian Mixture Models (GMM) can be sensitive to heavy field star contamination when the initial signal-to-noise ratio is low.

In the second step, GMM was applied on DBSCAN condidate members using five astrometric parameters Right Ascension (RA), Declination (DEC), proper motion Right Ascension (pmRA), proper motion Declination (pmDEC) and parallax. The optimal hyperparameters for GMM were determined based on the Bayesian Information Criterion (BIC) score.

To detect extended or scattered members and achieve a comprehensive view of the cluster halo, we employed a Random Forest (RF) classifier. The training strategy for the Random Forest algorithm was constructed as follows: sources not selected by DBSCAN were considered field stars, stars confirmed as cluster members by GMM were designated as core members, and sources selected by DBSCAN but excluded by GMM were treated as ambiguous or unlabelled. The RF model was then trained using five features: proper motion Right Ascension (pmRA), proper motion Declination (pmDEC) , parallax, G magnitude and BP-RP color

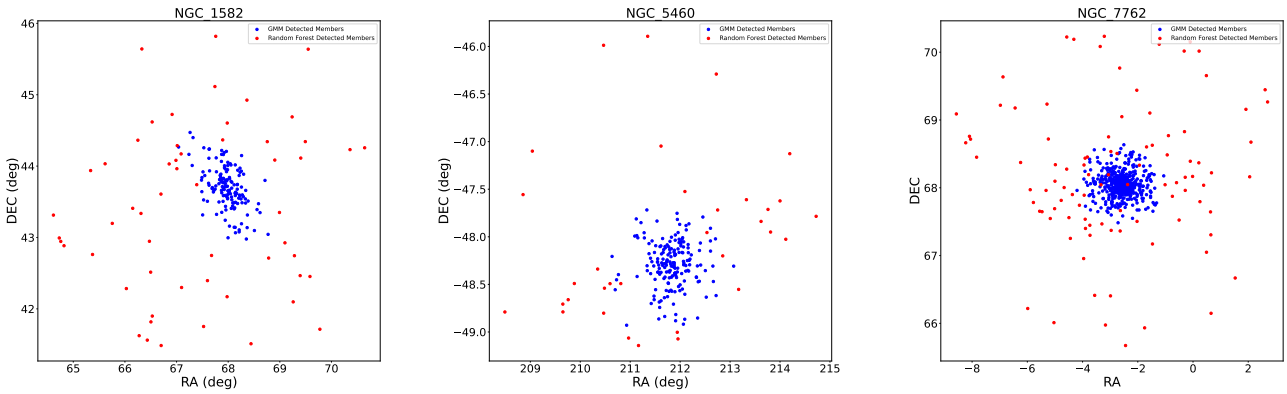

Figure 5. Spatial distribution of the cluster members, where blue dots represent GMM-detected core members and red dots indicate Random Forest-detected scattered members.

.

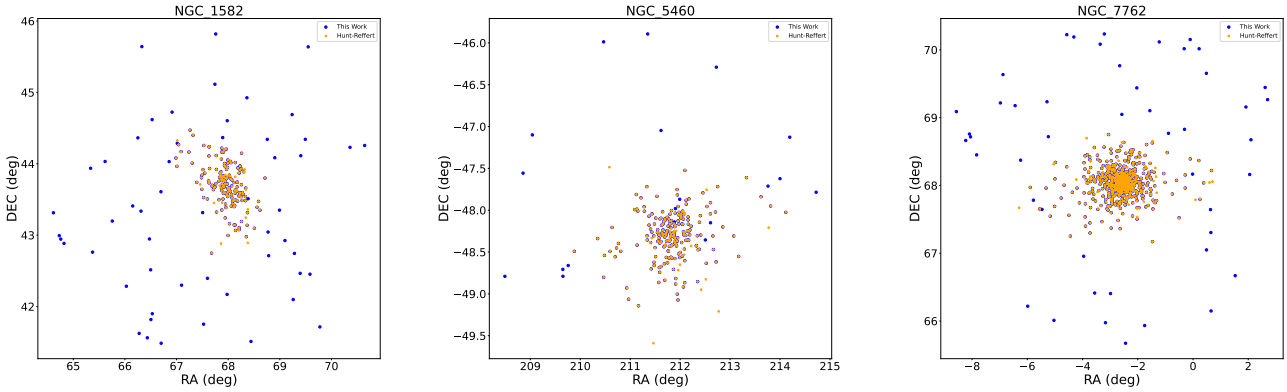

Figure 6. Comparison between our detected members (blue circles) and the catalog of Hunt-Reffert (orange stars) in the spatial distribution. As can be seen, our method successfully identifies extended scattered members across all clusters.

.

index. This classifier successfully identified the scattered members of the clusters. The results are presented below.

4. Results

A major challenge in implementing the DBSCAN algorithm is determining the optimal values for its hyperparameters, namely MinPts and eps. In our framework, DBSCAN serves to isolate reliable candidate members with a high signal-to-noise ratio, providing a robust initial subset for the subsequent GMM step. Through careful tuning, we selected MinPts = 150 and eps = 0.07 for NGC 1582, MinPts = 75 and eps = 0.085 for NGC 5460, and MinPts = 65 and eps = 0.09 for NGC 7762. Figure 1 illustrates the results of applying DBSCAN in the pmRA, pmDEC and parallax parameter space for our sample. As can be seen, the candidate members are in cluter 3 and clearly separated from the field population.

In the next step, a Gaussian Mixture Model (GMM) was applied to the DBSCAN candidate members using five parameters: RA, DEC, pmRA, pmDEC and parallax. The optimal number of components for GMM was determined based on the Bayesian Information Criterion (BIC) score. Figure 2 illustrates the BIC scores as a function of the number of components for each cluster. As can be seen, the minimum BIC score (indicating the optimal model complexity) is achieved at two components. Figure 3 displays the members identified by the GMM. As shown in the figure, the true cluster members are clearly separated

doi: https://doi.org/10.52526/25792776-26.73.2-334

within cluster 2.

In the next step, to detect scattered members and map the cluster halos, we applied the Random Forest (RF) classifier using five features: pmRA, pmDEC, parallax, G magnitude and BP-RP color index. Figure 4 illustrates the confusion matrices. Based on the evaluation metrics derived from the confusion matrices, the optimal hyperparameters for each cluster were determined as follows:

- **NGC 1582:** n_estimators = 100, max_depth = 20, criterion = 'gini'
- **NGC 5460:** n_estimators = 150, max_depth = 25, criterion = 'gini'
- **NGC 7762:** n_estimators = 100, max_depth = 30, criterion = 'gini'

Figure 5 displays the members identified by both the GMM and the Random Forest classifier. As can be seen, the Random Forest algorithm successfully detects the extended scattered members and maps the outer regions of the clusters beyond the dense core.

Figure 6 presents a comparison between our identified members and those from the catalog of Hunt & Reffert (2024). As can be seen, our method successfully uncovers a wider field of view and reveals more extended cluster halos compared to Hunt & Reffert (2024) for all three clusters.

5. Conclusions

In this work, we presented a hybrid machine learning pipeline—combining two unsupervised algorithms (DBSCAN and GMM) with a supervised classifier (Random Forest)—to accurately determine the membership and trace the morphology of three open clusters (NGC 1582, NGC 5460, and NGC 7762) using Gaia DR3 astrometric and photometric data.

Our approach sequentially leverages the strengths of each method: DBSCAN filters dense candidate regions to improve the signal-to-noise ratio, GMM reliably separates core cluster members in the astrometric space, and the Random Forest model successfully recovers outer, scattered members. The results demonstrate that our framework identifies both the centrally concentrated core populations and the extended halos over a wide field of view. A direct comparison with previous studies, such as hunt-reffert, indicates that our pipeline successfully detects extra halo members and traces the full spatial extent of open clusters with higher completeness, offering a robust tool for future wide-field morphological studies of the Milky Way cluster population.

References

Agarwal M., Rao K. K., Vaidya K., Bhattacharya S., 2021, Mon. Not. R. Astron. Soc. , 502, 2582

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

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

Noormohammadi M., Javadi A., 2025, Communications of the Byurakan Astrophysical Observatory, 72, 437

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

Noormohammadi M., Javadi A., Khakian Ghomi M., 2025, Mon. Not. R. Astron. Soc. , 544, 3696

Sanders W. L., 1971, Astron. Astrophys. , 14, 226