ML-MOC: Machine Learning based Membership Determination for Open Clusters

Manan Agarwal,1 Khushboo K. Rao,1 Kaushar Vaidya,1 and Souradeep Bhattacharya2
1Department of Physics, Birla Institute of Technology and Science – Pilani, 333031 Rajasthan, India
2Inter University Centre for Astronomy and Astrophysics, Ganeshkhind, Post Bag 4, Pune 411007, India

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT
The existing open cluster membership determination algorithms are either prior dependent on some known parameters of clusters or are not automatable to large samples of clusters. In this paper, we present, ML-MOC, a new machine learning based approach to identify likely members of open clusters using the Gaia DR2 data, and no a priori information about the clusters. We use the k-Nearest Neighbours algorithm and the Gaussian Mixture Model on the high-precision proper motions and parallax measurements from Gaia DR2 data to determine the membership probabilities of individual sources down to $G_{\text{mag}} \approx 20$ mag. To validate the developed method, we apply it on thirteen open clusters: M67, NGC 2099, NGC 2141, NGC 2243, NGC 2539, NGC 6253, NGC 6405, NGC 6791, NGC 7044, NGC 7142, NGC 752, Berkeley 18, and IC 4651. These clusters differ in terms of their ages, distances, metallicities, extinctions and cover a wide parameter space in proper motions and parallaxes with respect to the field population. The extracted members produce clean colour-magnitude diagrams and our astrometric parameters of the clusters are in good agreement with the values derived by the previous works. The degree of contamination in the extracted members range between 2% and 12%. The results show that ML-MOC is a reliable and scalable approach to segregate the open cluster members from the field stars.

Key words: methods: data analysis – open clusters and associations: general – methods: statistical – astrometry

1 INTRODUCTION
Galactic or Open Clusters (OCs) are the ideal laboratories to study the formation and evolution of stars (Krumholz et al. 2019), as they provide us chemically homogeneous groups of stars that are of the same age, share the same kinematics (proper motion and radial velocity), and are located at approximately the same given distance from us. The vast majority of them are located close to the Galactic plane and thus serve as excellent tracers of the formation history of the Galactic disk (e.g. Friel 1995; Chen et al. 2003; Jacobson et al. 2016). Accurate determination of cluster membership is essential for the studies of open clusters as it directly influences the estimation of the fundamental astrophysical parameters of clusters, e.g. age, photometric distance, reddening, and metallicity, among other things. The well-known open cluster catalogues, Dias et al. (2002) and Kharchenko et al. (2013, K13 hereafter), list about 3000 open clusters.

The ongoing European Space Agency (ESA) mission Gaia (Perryman et al. 2001), in particular the second Gaia Data Release, Gaia DR2 (Gaia Collaboration et al. 2018), has revolutionized the studies of open clusters by providing astrometric measurements with unprecedented precision. Cantat-Gaudin et al. (2018, CG18 hereafter) computed membership probabilities for 1229 OCs (including 60 previously unknown OCs) using Gaia DR2. They later added more clusters to get a catalogue of a total 1481 OCs (Cantat-Gaudin & Anders 2020). Gaia DR2 also facilitated the studies aimed at finding new OCs. Sim et al. (2019) reported 207 OCs by visually inspecting proper motion diagrams; Liu & Pang (2019) further found 76 OCs that were previously unknown; Castro-Ginard et al. (2019) identified 53 new OCs near the Galactic anticentre, and recently reported 582 more OCs in the Galactic disk (Castro-Ginard et al. 2020).

Various methods have been used for membership determination based on the analysis of the positions, proper motions, parallaxes, radial velocities, photometry, and their combinations (Vasilevsks et al. 1958; Sanders 1971; Cabrera-Cano & Alfaro 1990; Zhao & He 1990; Galadi-Enriquez et al. 1998). In the last few years, machine learning algorithms such as, DBSCAN (Gao 2014; Bhattacharya et al. 2017a), KMeans (El Aziz et al. 2016), GMM (Gao 2020), Random Forest (Gao 2018a), UPMASK (Cantat-Gaudin et al. 2018), and Artificial Neural Network (Castro-Ginard et al. 2018), have been put to the task of separating the true members of open clusters from the field stars. Most of these previous studies have only been applied to a
few old open clusters (e.g. Gao 2014 studied NGC 188; El Aziz et al. 2016 studied NGC 188 and NGC 2266) or are highly sensitive to the initial sample selection, making them uneasy for being scalable (e.g. Gao 2018a and Gao 2020). Methods developed by Cantat-Gaudin et al. (2018) and Castro-Ginard et al. (2018) were applied on a large number of open clusters, but they limited their membership analysis to sources brighter than $G \sim 18$ mag and $G \sim 17$ mag, respectively. Furthermore, Cantat-Gaudin et al. (2018) needs a priori information (distance and radius) about the cluster. Castro-Ginard et al. (2018), on the other hand, does not obtain the membership probability for individual stars.

In this paper, we propose, ML-MOC, a new probabilistic membership determination algorithm for open cluster members down to $G \sim 20$ mag using only Gaia DR2. Our algorithm is based on k-Nearest Neighbours algorithm (kNN, Cover & Hart 1967) and Gaussian Mixture Model (GMM, McLachlan & Peel 2000). We apply it on the high-precision Gaia DR2 proper motions and parallaxes to determine the membership probability of individual sources. We aim to facilitate homogeneous analysis of open cluster populations by developing a robust algorithm that works reliably on a large number of open clusters. The method is applied to thirteen open clusters: M67, NGC 2099, NGC 2141, NGC 2243, NGC 2539, NGC 6253, NGC 6405, NGC 6791, NGC 7044, NGC 7142, NGC 752, Berkeley 18 and IC 4651, which cover a range of ages, distances, metallicities and extinction.

The remainder of this paper is organized as follows. Section 2 describes the Gaia DR2 data and the cluster sample studied in this work, Section 3 describes the methodology of sample selection and the membership determination algorithm using M67 as the example cluster, in Section 4, we present our estimates of the degree of contamination, and make a comparison between our extracted members and the previously identified members of the clusters in the literature. Section 5 summarizes the main conclusions of this work.

2 DATA AND CLUSTER SAMPLE

We use Gaia DR2 (Gaia Collaboration et al. 2018) which catalogues more than 1.3 billion sources with unprecedented astrometric precision and accuracy. It provides a five parameter astrometric solution ($\alpha$, $\delta$, $\mu_{\alpha}$, $\mu_{\delta}$, $\omega$): stellar positions RA ($\alpha$) and DEC ($\delta$), proper motions in RA ($\mu_{\alpha}$) and in DEC ($\mu_{\delta}$), and parallaxes ($\omega$). It provides photometry for three broad bands ($G$, $G_{BP}$, and $G_{RP}$), containing sources up to a limiting magnitude of $G \sim 21$ mag. The large magnitude range leads to significant differences in the precision of various parameters of the bright and the faint sources. In parallax measurements, the uncertainties reach a precision of 0.02 milliarcsecond (mas hereafter) for $G < 14$ mag sources, and 2 mas for sources near $G \sim 21$ mag. In proper motions measurements, the uncertainties range between 0.05 mas yr$^{-1}$ for $G < 14$ mag sources, and 5 mas yr$^{-1}$ for sources with $G \sim 21$ mag. Among the other data products, Gaia DR2 also contains radial velocities for most sources brighter than $G \sim 13$ mag.

In this work, we determine the membership probabilities of thirteen open clusters: M67, NGC 2099, NGC 2141, NGC 2243, NGC 2539, NGC 6253, NGC 6405, NGC 6791, NGC 7044, NGC 7142, NGC 752, Berkeley 18 and IC 4651, that are located at different latitudes and cover a wide parameter space in proper motions and parallaxes relative to the foreground and the background contamination. They vary in terms of their ages, 0.53 Gyr to 8 Gyr, distances, ~450 pc to ~5500 pc, metallicities [Fe/H], ~0.54 to 0.43, and suffer from little extinction to as much as $A_v = 1.7$ mag. We use the cluster

![Figure 1](image)

Figure 1. The correlation of errors (photometry, parallax, and proper motions) with the $G$-mag of sources. The grey points are all Gaia DR2 sources for M67 within a radius of 150 arcmin from the cluster centre. The sources in blue are those with $G_{err} < 0.005$ mag.

M67 (NGC 2682) to demonstrate our methodology, as it is a well studied open cluster and its members are publicly available (Sanders 1977; Sandquist 2004; Sarajedini et al. 2009; Geller et al. 2015).

3 METHODOLOGY

Our approach to membership determination only uses astrometric measurements from Gaia DR2 and does not require any prior information about the cluster. The extraction of the cluster members is done in three main stages: extracting the sample sources, identifying high probability member sources and extending the list of members by identifying likely moderate probability members.

In the first stage, we use the kNN algorithm and extract the appropriate sample sources by removing the obvious field noise. In the second stage, we normalise proper motions and parallaxes and apply a three-dimensional GMM on the sample sources to identify member stars. Lastly, in the third stage, we include member stars with moderate probability to our selection of cluster members. Following is a detailed description of each stage.

3.1 First Stage: Extract the Sample sources

The aim of this stage is to remove a large number of obvious field stars. First, we download sources from Gaia DR2 in a cone around the cluster centre for a value of radius that is greater than the tidal radius of the cluster, as reported in K13. Though our algorithm is quite robust to the choice of this initial radius, as a rule of thumb, we generally use the value of radius that is 1.5 times the tidal radius. For M67, we download sources within a radius of 150 arcmin from the cluster centre. Next, we select the sources which satisfy the following criteria:

(i) each source must have the five astrometric parameters positions, proper motions and parallax as well as the three photometric parameters $G$-mag, $BP$-mag, and $RP$-mag in the Gaia DR2 catalog,
clusters. The selected range of parallax has a width between 0.4 and 2.5 mas. For example, as NGC 6405 has an estimated distance of just ~450 pc, we get a 2.5 mas wide parallax range whereas NGC 2243 is estimated at ~4400 pc and we get the parallax range width of 0.4 mas. The sources that are selected by applying the proper motion and parallax range on All sources are referred as "Sample sources" hereafter. For M67, we retrieve 2427 Sample sources. The proper motions and parallaxes of the Sample sources in M67 are shown in Figure 3 along with All sources of the cluster.

### 3.2 Second Stage: Identify the Member sources

To separate the likely cluster members from the field stars, we apply the GMM (McLachlan & Peel 2000), an unsupervised clustering algorithm, on the Sample sources. Prior to applying the GMM, we perform one more step, i.e., normalize the data $(\mu_\alpha, \mu_\delta, \omega)$ since the similarity measurements are sensitive to the differences in the magnitudes and the scale of the parameters. We perform the standard normalization for the proper motions and parallaxes following the process detailed below. Given N Sample sources, each with m parameters $[x^1, x^2, ..., x^m]$, we define the normalized parameter in the $j$th dimension $X_i^j$ as,

$$X_i^j = \frac{x_i^j - \mu_j}{\sigma_j} \quad (i = 1, 2, ..., N; j = 1, 2, ..., m) \tag{1}$$

where $x_i^j$ is the original parameter, $\mu_j$ being the median of $x^j$ distribution, and $\sigma_j$ its standard deviation.

The GMM is a probabilistic model that assumes that all the data points are drawn from a mixture of a finite number of Gaussian distributions with unknown parameters. It can be thought of as an extension of the K-Means clustering (MacQueen et al. 1967; Lloyd 1982) to incorporate information about not only the mean ($\mu$), but also the covariance ($\Sigma$), that describes the ellipsoidal shape of a distribution. The model is fitted by maximizing the likelihood estimates of the distribution parameters using the expectation maximization (EM) algorithm (Dempster et al. 1977). Given N data points $x = \{x_1, x_2, x_3, ..., x_N\}$ in a $M$-dimensional parameter space, the $K$-component GMM is defined as,

$$P(x) = \sum_{i=1}^{K} w_i G(x \mid \mu_i, \Sigma_i), \text{ such that } \sum_{i=1}^{K} w_i = 1 \tag{2}$$

---

1 Feature space refers to the vector space defined by the collection of numerical features used to characterize the data. A data point with n features is represented in an n-dimensional feature space.

2 For clusters which have an overlap in parameter $(\mu_\alpha, \mu_\delta, \omega)$ peak of field sources and cluster members (refer Section 4.5), we manually tweak the range.
Table 1. Comparison of our initial estimates of mean proper motions, mean parallaxes, and mean distances (using Bailer-Jones et al. 2018 determined distances), made by using the kNN algorithm, with the corresponding values for all the clusters from CG18.

| Clusters | Stage 1 (kNN estimated) | CG18 |
|----------|-------------------------|------|
|          | \( \mu_\alpha \) (mas/yr) | \( \mu_\delta \) (mas/yr) | \( \omega \) (mas) | \( \mu_\alpha \) (mas/yr) | \( \mu_\delta \) (mas/yr) | \( \omega \) (mas) | \( \text{dist} \) (pc) | \( \text{dist} \) (pc) |
| M67      | -10.913                 | -2.801              | 1.131              | 862.3              | -10.986                 | -2.964              | 1.135              | 859.1              |                          |
| NGC2099  | 1.820                   | -5.694              | 0.688              | 1396.4             | 1.924                   | -5.648              | 0.666              | 1438.1             |                          |
| NGC2141  | -0.045                  | -0.725              | 0.141              | 5296.3             | -0.028                  | -0.767              | 0.196              | 4441.3             |                          |
| NGC2243  | -1.314                  | 5.522               | 0.226              | 3758.3             | -1.279                  | 5.488               | 0.211              | 4167.8             |                          |
| NGC2539  | -2.313                  | -0.546              | 0.765              | 1260.0             | -2.331                  | -0.584              | 0.754              | 1277.4             |                          |
| NGC6253  | -4.450                  | -5.226              | 0.579              | 1698.5             | -4.537                  | -5.280              | 0.563              | 1689.7             |                          |
| NGC6405  | -1.216                  | -5.784              | 2.128              | 463.9              | -1.306                  | -5.847              | 2.172              | 454.3              |                          |
| NGC6791  | -0.417                  | -2.239              | 0.201              | 4279.3             | -0.421                  | -2.269              | 0.192              | 4530.8             |                          |
| NGC7044  | -5.015                  | -5.508              | 0.297              | 2872.6             | -4.976                  | -5.526              | 0.273              | 3315.6             |                          |
| NGC7142  | -2.698                  | -1.260              | 0.394              | 2368.0             | -2.747                  | -1.288              | 0.392              | 2376.4             |                          |
| NGC752   | 9.738                   | -11.865             | 2.240              | 441.9              | 9.810                   | -11.713             | 2.239              | 441.0              |                          |
| Be18     | 0.848                   | -0.044              | 0.203              | 3757.0             | 0.850                   | -0.057              | 0.152              | 5523.5             |                          |
| IC4651   | -2.402                  | -4.901              | 1.082              | 901.3              | -2.410                  | -5.064              | 1.056              | 921.3              |                          |

Figure 4. The frequency distributions of normalized proper motions and parallaxes of the Sample sources (blue). The sources with membership probability greater than 0.6, as determined by the GMM, are shown in red bars.

where \( P(x) \) denotes the probability distribution of data points \( x \) and \( w_i \) is the mixture weight of the \( i \)th Gaussian component, \( G(x \mid \mu, \Sigma) \), defined as

\[
G(x \mid \mu, \Sigma) = \frac{\exp\left[-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right]}{(2\pi)^{M/2} |\Sigma|^{1/2}}
\]

In eq. (3), \( \mu_i \) and \( \Sigma_i \) are the mean vector and the full covariance matrix of the \( i \)th Gaussian component, respectively. The GMM assigns each data point a soft membership probability to each cluster, i.e. how likely the data point is to be described by each cluster. The GMM has been previously used to determine the membership probabilities of the sources of open clusters (M67 by Uribe et al. 2006 using proper motion; NGC6791 by Gao 2020 using 5D astrometry data i.e. position, proper motions and parallaxes).

In order to get reliable membership probability estimates for cluster members using the GMM, the following conditions should be met: (1) there is a high precision data set, (2) there is a high proportion of cluster members i.e. a high cluster sources to field sources ratio, and (3) there is a difference in the location of the peaks of the distribution of the field star and cluster members (de Graeve 1979). If even one of these conditions is not met, the computed membership probabilities will lose significance. The unprecedented high-quality astrometry data of Gaia DR2 and the initial selection of sources with the \( G \)-mag error less than 0.005, ensures that we have a reliable data set to satisfy the first condition. To tackle the second of the above issues, namely the need of high proportion of cluster members in our sample on which we apply the GMM, we choose to apply the GMM on the Sample sources instead of All sources. The third condition is not in our hand, but it is unlikely that all the three peaks (\( \mu_\alpha, \mu_\delta, \omega \)) of the field stars coincide with those of the cluster members. We give examples of applying our algorithm to clusters which have overlapping peaks in one or more of the parameters: NGC 2539 has overlapping proper motion in RA (\( \mu_\alpha \)); IC 4651 has overlap in both proper motion RA and DEC (\( \mu_\alpha, \mu_\delta \)); NGC 2141 and Berkeley 18 (refer Appendix A) have overlap in proper motion in RA and parallax (\( \mu_\alpha, \omega \)) with the field sources. The membership determination for these clusters is discussed in detail in Section 4.5.

We use a two-component, i.e. the cluster and field, GMM on the normalised three-dimensional parameter space (\( \mu_\alpha, \mu_\delta, \omega \)). This assumes that the proper motions and parallaxes of the Sample sources follow a two-component Gaussian distribution. We have not used position in the parameter space as that would have restricted the cluster members within a definite ellipsoidal border, in turn, making it impossible to study the morphology of the cluster and to identify potential escapers. It must also be noted that since the
3.3 Third Stage: Including Moderate Probability Source

The high probability members (≥ 0.6) from the GMM give a definite border to the parallax and proper motion. The moderate membership probability sources lie either on the periphery of the proper motion or on the parallax border. We observed that the sources for which the membership was verified by the radial velocity had a probability greater than 0.8 assigned by the GMM. Thus we used the parallax range determined by our member sources with probability greater than 0.8 to filter the low probability members. We extend the member sources by including the Sample sources whose parallax values lie in this range and have a membership probability between 0.2 and 0.6. This allows us to selectively include likely cluster members while keeping the degree of contamination to the minimum. In case of M67, we add 71 sources with a membership probability between 0.2 and 0.6, to get a total of 1221 members. Figure 5 shows the vector point diagram of the proper motions, the parallax distribution and the colour-magnitude diagrams (CMD) for the Sample sources of M67. In Figure 6, we show the membership probabilities assigned by GMM to the Sample sources of M67. We provide the membership results for the rest of the open clusters in Section 4.

4 ANALYSIS AND RESULTS

4.1 Comparison with spectroscopically confirmed members

The distributions of field stars and cluster members are overlapped both in parallax and in proper motions, so it is expected that some field stars could have been erroneously extracted as members. In order to analyze the accuracy of our membership determination algorithm, we use the members identified by the WIYN Open Cluster Study (WOCS) as a proxy of ground truth. WOCS uses the WIYN 3.5m telescope to make definitive cluster membership measurements via precise (0.5 km/s) radial velocities for proper-motion candidate members. The survey extends down to $G$ 16.5 mag. Out of the thirteen OCs studied in this work, WOCS has observed M67 (Geller et al. 2015), NGC 6791 (Tofflemire et al. 2014) and NGC 6253 (Anthony-Twarog et al. 2010). For M67, it classified 549 sources as members and 558 sources as non-members. We only consider the WOCS sources for which we found a Gaia counterpart with defined astrometric (positions, proper motions and parallax) and photometric (G-mag, BP-mag, and RP-mag) parameters. We retrieve 486 (88.52%) common members with WOCS, while mislabeling 98 sources as members which are identified as non-members by WOCS. CG18 identified 429 (78.14%) common members while misclassifying 79 sources. Gao (2018b), who also uses Gaia DR2 to identify members for M67, finds 510 (92.89%) common members and misclassifies 101 sources. For NGC 6791, WOCS identifies 111 sources as members and 103 sources as non-members. Both, this work and CG18, retrieve 105 (94.59%) common members with WOCS.
Table 2. An estimate of the degree of contamination in the member sources extracted by our algorithm. Column 2 gives the cluster radius, columns 3 and 4 give core and tidal radii, respectively, as estimated from the King’s profile fitting, column 5 gives the radius that encloses twice the area as enclosed by the tidal radius, column 6 gives the number of member stars for each cluster, column 7 gives the number of sources identified by the algorithm between the tidal radius of the cluster and the radius that encloses twice the area as the tidal radius, column 8 gives the degree of contamination for each cluster.

| Clusters     | $r$ (arcmin) | $r_c$ (arcmin) | $r_t$ (arcmin) | $r_{t2}$ (arcmin) | N till $r_t$ | N between $r_t$ and $r_{t2}$ | Degree of contamination (%) |
|--------------|--------------|----------------|----------------|-------------------|-------------|-----------------------------|---------------------------|
| M67          | 53           | 5.603          | 98.686         | 139.563           | 1194        | 23                          | 1.93                      |
| NGC2099      | 41           | 5.249          | 57.471         | 81.276            | 1640        | 103                         | 6.28                      |
| NGC2141*     | 11           | 3.811          | 16.883         | 23.876            | 828         | 102                         | 12.32                     |
| NGC2243*     | 14           | 1.207          | 32.911         | 46.543            | 583         | 12                          | 2.06                      |
| NGC2539*     | 23           | 5.808          | 39.508         | 55.873            | 466         | 38                          | 8.15                      |
| NGC6253*     | 12           | 3.364          | 23.651         | 33.448            | 743         | 47                          | 6.33                      |
| NGC6405      | 53           | 16.511         | 58.869         | 83.253            | 688         | 39                          | 5.67                      |
| NGC6791      | 14           | 3.041          | 20.507         | 29.001            | 2422        | 134                         | 5.53                      |
| NGC7044*     | 9            | 1.597          | 16.982         | 24.016            | 693         | 46                          | 6.64                      |
| NGC7142*     | 12           | 2.842          | 22.719         | 32.130            | 316         | 21                          | 6.65                      |

4.2 Calculation of cluster radial density parameters

WOCS data are only available for M67, NGC 6791 and NGC 6253, and for sources brighter than $G \sim 16.5$ mag. To get a more complete estimate of the degree of contamination, we compare the number of cluster members identified up to the tidal radius ($r_t$) and the radius ($r_{t2}$) that encloses twice the area than the tidal radius. The tidal radius indicates the distance at which the radial density profile reaches the theoretical zero level (King 1962), so all our member sources lying beyond this distance can be classified as noise. To find the tidal radii of the clusters we follow the process described below.

All the identified member sources are used to construct the cluster stellar density radial profiles. We calculated the mean stellar surface density in concentric rings centred on the cluster centre as

$$\rho_i = \frac{N_i}{\pi(r_{i+1}^2 - r_i^2)}$$

where $N_i$ is the number of stars in the $i$-th ring with inner and outer radius $r_i$ and $r_{i+1}$, respectively.

The radial distribution of number of sources, and of projected surface density for M67 are shown in Figure 7. The cluster radius is defined as the distance from the cluster centre, where the combined cluster plus background profile is no longer distinguished from the background alone. We then use all the member sources within the cluster radius to fit the King’s profile (King 1962) to derive the core ($r_c$) and and tidal radii ($r_t$) of the clusters. Figure 8 shows the best fit King’s profile for M67.

4.3 Degree of contamination

To calculate the degree of contamination, we download sources up to $r_{t2}$ i.e. the radius that encloses twice the area as enclosed by the $r_t$, and perform the algorithm as described in the Section 3 to identify sources satisfying the same criteria as members of the clusters but now between $r_t$ and $r_{t2}$. All the sources identified in the annular region of same area as the cluster area are considered as erroneously labelled members to calculate the degree of contamination. For M67, we find 1194 members stars within $r_t$ i.e. 98.686 arcmin and 23 additional sources up to $r_{t2}$ i.e. 139.563 arcmin. This gives a degree of contamination of 1.93% for M67. In Figure 9, we compare the member sources of M67 within the tidal radius with the sources in the annular region of same area as the cluster area. We observe that these 23 sources between $r_t$ and $r_{t2}$ are indistinguishable from

![Figure 7](image1.png)
![Figure 8](image2.png)
Figure 9. Comparing the member sources (in red) of M67 within $r_t$ with the sources (in black) in the annular region of same area as the area up to $r_t$ from the cluster centre. Out of the 1217 sources within $r_t^2$, we find 23 sources beyond $r_t$.

Figure 10. Upper Panel: The CMDs of members identified for the cluster M67 by the three algorithms. Lower Panel: The radial distribution, the proper motion distribution, and the parallax distribution of members by the three algorithms.

In Table 2, we show the estimated values of $r_c$ and $r_t$, and the degree of contamination for all the open clusters studied in this work except NGC 752, Be 18, and IC 4651. For a few clusters, marked by ‘*’ in the Table 2, on applying our method to the sources up to $r_t^2$, it fails to clearly segregate the cluster members from the field stars. This is because when considering sources up to $r_t^2$, we are adding a large number of field sources to the Sample sources, hence reducing the proportion of cluster to field sources, which makes the GMM
Figure 11. The spatial distribution, the proper motion distribution, the parallax distribution and the CMD of the Sample sources and the members identified by our algorithm.
Figure 11 (Continued). The spatial distribution, the proper motion distribution, the parallax distribution and the CMD of the Sample sources and the members identified by our algorithm.
unreliable. For these clusters, we take sources brighter than $G \sim 18$ mag to estimate the degree of contamination. For NGC 752, Be 18 and IC 4651, due to their sparse spatial distribution, we are unable to calculate the degree of contamination even with $G = 18$ mag limit.

4.4 Comparison with other clustering algorithms

We compare the members extracted in this work with the members identified by CG18 and Gao (Gao 2018b and Gao 2018a, hereafter collectively referred as Gao18). CG18 employed the membership assignment code, Unsupervised Photometric Membership Assignment in Stellar Clusters (UPMASK), by Krone-Martins & Moitinho (2014) on the proper motions and parallaxes from Gaia DR2. They consider sources only with $G < 18$ mag and those that are located within a radius twice as large as the diameter reported by Dias et al. (2002). Gao18 use a combination of GMM and Random Forest method on the astrometric and photometric measurements from Gaia DR2 to investigate the membership of open clusters M67 and NGC 6405. In Figure 10, we show a comparison of the CMDs, proper motion and parallax distributions, of our member sources with those of CG18 and Gao18 for the cluster M67. The cluster members determined by all the three result in clean CMDs but this work and Gao18 find the member sources up to $G \sim 20$ mag, whereas CG18 only goes up to $G \sim 18$ mag.
Table 3. The comparison of the astrometric parameters estimated in this work with CG18. The RA and DEC are expressed in degree, the mean proper motions along with the uncertainties are expressed in mas/yr, the mean parallaxes and the corresponding uncertainties are expressed in mas, and the distances are expressed in pc.

| Clusters | RA  | DEC  | This work | CG18 |
|----------|-----|------|-----------|------|
|          | $\mu_\alpha$ | $\mu_\delta$ | $\omega$ | RA   | DEC  | $\mu_\alpha$ | $\mu_\delta$ | $\omega$ | distance |
| M67      | 132.852 | 11.836 | -10.981 | -2.949 | 1.135 | 860.7 | 132.846 | 11.814 | -10.986 | -2.964 | 1.135 | 859.1 |
| NGC2099  | 88.064 | 32.547 | 1.928  | -5.636 | 0.663 | 1466.9 | 88.074 | 32.545 | 1.924  | -5.648 | 0.666 | 1438.1 |
| NGC2141  | 90.742 | 10.455 | -0.025 | -0.750 | 0.197 | 3812.3 | 90.734 | 10.451 | -0.028 | -0.767 | 0.196 | 4441.3 |
| NGC2243  | 97.390 | -31.283 | -1.285 | 5.489  | 0.213 | 3606.0 | 97.395 | -31.282 | -1.279 | 5.488  | 0.211 | 4167.8 |
| NGC2539  | 122.670 | -12.845 | -2.337 | -0.584 | 0.757 | 1280.5 | 122.658 | -12.834 | -2.331 | -0.584 | 0.754 | 1277.4 |
| NGC6253  | 254.769 | -52.713 | -4.521 | -5.289 | 0.567 | 1718.6 | 254.778 | -52.712 | -4.537 | -5.280 | 0.563 | 1689.7 |
| NGC6405  | 265.088 | -32.285 | -1.360 | -5.816 | 2.157 | 460.6 | 265.069 | -32.242 | -1.306 | -5.847 | 2.172 | 454.3 |
| NGC6791  | 290.226 | 37.776 | -0.422 | -2.274 | 0.218 | 4086.6 | 290.221 | 37.778 | -0.421 | -2.269 | 0.192 | 4530.8 |
| NGC7044  | 318.287 | 42.494 | -4.976 | -5.523 | 0.273 | 3273.5 | 318.284 | 42.494 | -4.976 | -5.526 | 0.273 | 3315.6 |
| NGC7142  | 326.287 | 65.753 | -2.743 | -1.283 | 0.392 | 2401.3 | 326.290 | 65.782 | -2.747 | -1.288 | 0.392 | 2376.4 |
| NGC752   | 29.156 | 37.809 | 9.825  | -11.724 | 2.229 | 443.8 | 29.223 | 37.794 | 9.810  | -11.713 | 2.239 | 441.0 |
| Be18     | 80.472 | 45.396 | 0.843  | -0.078 | 0.139 | 4652.0 | 80.531 | 45.442 | 0.849  | -0.057 | 0.152 | 5523.5 |
| IC4651   | 261.230 | -49.873 | -2.424 | -5.039 | 1.055 | 933.0 | 261.212 | -49.917 | -2.410 | -5.064 | 1.056 | 921.3 |

4.5 Stellar parameters of the cluster sample

We applied our method to twelve more open clusters located between ~450 pc and ~5500 pc. In Figure 11, we show the Sample sources and the identified member sources for the rest of these open clusters, except for Berkeley 18 which is shown in Figure A2. As mentioned in the Section 3, it is difficult for the GMM to differentiate between field and cluster sources when there is a low concentration of cluster members and/or the parameter peaks coincide. In case of NGC 2539, our approach initially struggled at reliably identifying cluster members because the peak of $\mu_\alpha$, for field sources coincides with that of cluster sources and there is a low cluster to field sources ratio. This was easily overcome by only considering sources brighter than $G = 18$ mag, which improves the cluster to field sources ratio. Similarly for IC 4651, NGC 2141 and Berkeley 18, which all have overlapping peaks with the field sources, we consider sources brighter than $G \sim 19$ mag to reliably extract cluster members. The figures comparing our identified members with CG18 and Gao18 are shown in Appendix B.

To accurately compute the astrometric parameters of the cluster, we should consider only the most reliable member sources. Therefore, we use the sources that have a membership probability greater than 0.6 and lie within the estimated cluster radius. To reduce the uncertainty in the calculated parameter values, we only consider sources that are brighter than $G = 18$ mag. In this work, we compute the mean proper motion and mean parallax of the cluster by taking a simple average of the individual proper motions and parallaxes of the most reliable member stars, without any consideration about their errors and covariances. For M67, the mean proper motion is determined to be $(\mu_\alpha, \mu_\delta) = (10.981 \pm 0.006, -2.949 \pm 0.006)$ mas/yr and mean parallax $(\omega) = 1.135 \pm 0.002$ mas, using 965 bright and high probability members. Since reliable distances to the Gaia DR2 sources cannot be obtained by simply inverting the parallax, we take the mean of the individual source distances obtained by Bailey-Jones et al. (2018). For M67, the mean distance is determined to be 860.7 pc. The centre coordinates of the clusters are computed using the Mean shift algorithm (Comaniciu & Meer 2002). We compare the astrometric parameters determined by us with CG18 in Table 3.

5 DISCUSSION

In this work we presented, ML-MOC, a new method for open cluster membership determination using only astrometric measurements from Gaia DR2 and no a priori information about the cluster. We employed a combination of well-known Machine Learning algorithms, i.e. k-Nearest Neighbours algorithm and Gaussian Mixture Model, to estimate the membership probability of individual sources down to $G \sim 20$ mag. Our approach does not rely on any strong physical assumption concerning the nature of the cluster (no assumptions on density profile modelling or on the structure in the photometric space). We applied ML-MOC to thirteen open clusters that cover a wide parameter space in terms of their distances, ~450 pc to ~5500 pc, ages, 0.53 Gyr to 8 Gyr, metallicities [Fe/H], -0.54 to 0.43, and extinctions, 0.2 mag to 1.7 mag. The cluster members identified by ML-MOC successfully produce clean CMDs. On considering the radial velocity verified members by WOCS for M67, NGC 6791 and NGC 6253 as ground truth, ML-MOC retrieves more members than CG18 while maintaining a similar number of false classifications.
The cluster centres, mean proper motions and mean parallaxes of the clusters measured with ML-MOC are in excellent agreement, as shown in Table 3, with those determined by CG18 who make use of the same input catalogue (Gaia DR2), but without any a priori information on cluster properties.

The open clusters whose proper motions and parallaxes merge with the field sources are challenging for ML-MOC. Most of these clusters can be resolved by considering the brighter sources and reducing the initial selection radius. However, we cannot entirely exclude the possibility that some of the sources identified as member stars in such clusters are field stars. To estimate the field stars erroneously labelled as cluster members we evaluate the degree of contamination. For M67 and NGC 2243, the contamination is estimated at just 2%. In six of studied clusters we determine the contamination between 5% and 7%. The highest degree of contamination was estimated in NGC 2539 and NGC 2141 at 8.15% and 12.32% respectively. The high contamination is likely due to the overlap in peak of field sources and cluster members in (\(\mu_\alpha, \mu_\delta\)) for NGC 2539 and (\(\mu_\alpha, \omega\)) for NGC 2141.

ML-MOC is a reliable and scalable approach to extract the member stars of open clusters without relying on any a priori information about the cluster. This makes us well equipped to perform an unbiased all-sky search for new open clusters. The non-reliance of ML-MOC on spatial information will help us to find non-circular morphology and study the internal dynamical processes including those that influence the formation of tidal tails (Bhattacharya et al. 2017b; Tang et al. 2019). The comprehensive information about the membership of a large number of open clusters will enable us to study a host of issues of relevance in Astrophysics such as, the dynamical evolution (Lee et al. 2013; Carrera et al. 2019), signatures of mass segregation (Allison et al. 2009; Maurya et al. 2020), evaporation of lower-mass cluster members (Lee et al. 2013), and dissolving of old open clusters (Lammers et al. 2010; Carrera et al. 2019), and studying the luminosity functions (Miret-Roig et al. 2018; Maurya et al. 2020), initial mass functions (de La Fuente Marcos 1997; Prisinzano et al. 2016), and estimating primordial binary population (de La Fuente Marcos 1997; Kouwenhoven et al. 2007), in young open clusters. The algorithm also has a great potential to allow the study of exotic stellar populations such as the blue straggler stars in open clusters (Bhattacharya et al. 2019; Vaidya et al. 2020; Rain et al. 2020), and their formation mechanisms. Moreover, with the expected release of Gaia Early Data Release 3 and later Gaia Data Release 3 which will provide more precise astrometric, photometric, and radial velocity information, our algorithm will improve the reliability of membership determination and classify even dimmer sources.

**ACKNOWLEDGEMENTS**

This work has made use of data from the European Space Agency (ESA) mission Gaia (https://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement. This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France. This research made use of Astropy, a community-developed core Python package for Astronomy (Astropy Collaboration et al. 2013; Price-Whelan et al. 2018), scikit-learn (Pedregosa et al. 2011) and Numpy (van der Walt et al. 2011). The figures in this paper were produced with Bokeh, a Python library for interactive visualization (Bokeh Development Team 2020) and Matplotlib (Hunter 2007). This research also made use of NASA’s Astrophysics Data System (ADS).

**DATA AVAILABILITY**

The data underlying this article are publicly available at https://archives.esac.esa.int/gaia. The derived data generated in this research will be shared on reasonable request to the corresponding author.

**REFERENCES**

Allison R. J., Goodwin S. P., Parker R. J., de Grijs R., Zwart S. F. P., Kouwenhoven M. B. N., 2009, *The Astrophysical Journal*, 700, L99

Anthony-Twarog B. J., Deliyannis C. P., Twarog B. A., Cummings J. D., Maderak R. M., 2010, *AJ*, 139, 2034

Astropy Collaboration et al., 2013, *A&A*, 558, A33

Bailer-Jones C. A. L., Rybizki J., Fouesneau M., Mantelet G., Andrae R., 2018, *AJ*, 156, 58

Bhattacharya S., Mahulkar V., Pandaokar S., Singh P. K., 2017a, *Astronomy and Computing*, 18, 1

Bhattacharya S., Mishra I., Vaidya K., Chen W. P., 2017b, *The Astrophysical Journal*, 847, 138

Bhattacharya S., Vaidya K., Chen W. P., Beccari G., 2019, *A&A*, 624, A26

Bokeh Development Team 2020, Bokeh: Python library for interactive visualization. https://bokeh.org/

Cabrera-Cano J., Alfaro E. J., 1990, *A&A*, 235, 94

Cantat-Gaudin T., Anders F., 2020, *A&A*, 633, A99

Cantat-Gaudin T. et al., 2018, *A&A*, 618, A93

Carrera R., et al., 2019, *A&A*, 627, A119

Castro-Ginard A., Jordi C., Luri X., Julbe F., Morvan M., Balaguer-Núñez L., Cantat-Gaudin T., 2018, *A&A*, 618, A59

Castro-Ginard A., Jordi C., Luri X., Cantat-Gaudin T., Balaguer-Núñez L., 2019, *A&A*, 627, A35

Castro-Ginard A., et al., 2020, *A&A*, 635, A45

Chen L., Hou J. L., Wang J. J., 2003, *AJ*, 125, 1397

Comaniciu D., Meer P., 2002, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 603

Cover T., Hart P., 1967, *IEEE Transactions on Information Theory*, 13, 21

Dempster A. P., Laird N. M., Rubin D. B., 1977, *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1

Dias W. S., Alesi B. S., Moitinho A., Lépine J. R. D., 2002, *A&A*, 389, 871

El Aziz M. A., Selim I. M., Esmail A., 2016, *Experimental Astronomy*, 42, 49

Friel E. D., 1995, *ARA&A*, 33, 381

Gaia Collaboration et al., 2018, *A&A*, 616, A1

Galadi-Enriquez D., Jordi C., Trullols E., 1998, *A&A*, 337, 125

Gao X., 2020, *Ap&SS*, 365, 24

Geller A. M., Latham D. W., Mathieu R. D., 2015, *AJ*, 150, 97

Hunter J. D., 2007, *Computing in Science & Engineering*, 9, 90

Jacobian H. R., et al., 2016, *A&A*, 591, A37

Kaluzny J., 1997, *A&A*, 121, 455

Kharchenko N. V., Piskunov A. E., Schilbach E., Röser S., Scholz R. D., 2013, *A&A*, 558, A53

King I., 1962, *AJ*, 67, 471

Kouwenhoven M. B. N., Brown A. G. A., Portegies Zwart S. F., Kaper L., 2007, *A&A*, 474, 77

Krone-Martins A., Moitinho A., 2014, *A&A*, 561, A57

Krumholz M. R., McKee C. F., Bland-Hawthorn J., 2019, *ARA&A*, 57, 227

https://www.cosmos.esa.int/web/gaia/earlydr3
APPENDIX A: BERKELEY 18

Berkeley 18 is an old open cluster with the estimated distance of ~5.8 kpc (Kaluzny 1997). On directly applying the method described in this work we fail to segregate the likely cluster members of Be18 from the field stars. This is because of (1) very low cluster members to field sources ratio and (2) the overlapping mean values of $\mu_\alpha$ and $\omega$ for cluster and field sources (shown in Figure A1). These are clear violations of the criteria, as mention in second stage in section 3, required for reliable determination of cluster members by the Gaussian Mixture Model. To overcome this, we need to reduce the number of field sources while retaining the maximum cluster members. We achieve this by only considering sources within 22 arcmin (according to K13 the tidal radius, $r_t$, is 26 arcmin and cluster radius is 16.5 arcmin) and brighter than $G = 18.75$ mag. In Figure A2 we show the result after applying these limits and in Figure A3 we compare our extracted members with CG18. It is possible to reduce the radius at the expense of $G$-mag and vice-a-versa, so depending on the particular science case one can decide these limits.

APPENDIX B: ADDITIONAL FIGURES

This paper has been typeset from a \TeX/L\LaTeX file prepared by the author.
Figure A1. Distribution of Be18 All sources and Sample sources in $\mu_\alpha$, $\mu_\delta$, $\omega$. The peak of $\mu_\alpha$, and $\omega$ of field sources overlaps with the corresponding peaks of cluster members.

Figure A2. The grey points represent the Sample sources and the red points are the identified cluster members for Be18 that are within 22 arcmin and brighter than $G = 18.75$ mag.

Figure A3. Comparing the cluster members of Be18 determined in this work with CG18.
Figure B1. Upper Panel: The CMDs of members identified for the cluster NGC 6405 by our algorithm, CG18 and Gao18. Lower Panel: The radial distribution, the proper motion distribution, and the parallax distribution of members by the three algorithms.
Figure B2. Comparing the cluster member sources determined in this work with the members by CG18. The sources in red are extracted by the algorithm defined in this work. CG18 member sources (i.e. sources with membership probability greater than 0.5) are shown in green.
Figure B2 (Continued). Comparing the cluster member sources determined in this work with the members by CG18. The sources in red are extracted by the algorithm defined in this work. CG18 member sources (i.e. sources with membership probability greater than 0.5) are shown in green.
Figure B2 (Continued). Comparing the cluster member sources determined in this work with the members by CG18. The sources in red are extracted by the algorithm defined in this work. CG18 member sources (i.e. sources with membership probability greater than 0.5) are shown in green.