Analysis of Machine Learning Methods for Membership Association in Star Clusters Using GAIA DR3 Data

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE

OF

BACHELOR OF TECHNOLOGY

IN

Engineering Physics

Submitted by:

Anish Kalsi (2K19/EP/014)

Dhruv Tyagi (2K19/EP/032)

Harshit Choudhary (2K19/EP/038)

Under the supervision of

Dr. Ajeet Kumar

Department of Applied Physics

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

December 2022

DEPARTMENT OF APPLIED PHYSICS
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

We, Anish Kalsi (2K19/EP/014), Dhruv Tyagi (2K19/EP/032), and Harshit Choudhary (2K19/EP/038) students of **B.Tech Engineering Physics**, hereby declare that the Project titled "Analysis of Machine Learning Methods for Membership Association in Star Clusters Using GAIA DR3 Data" which is submitted by us to the Department of Applied Physics, DTU, Delhi in fulfillment of the requirement for awarding of the Bachelor of Technology degree is not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Fellowship, or other similar title or recognition.

Place: New Delhi	**Anish Kalsi**	**Dhruv Tyagi**	**Harshit Choudhary**
Date: 8/12/2022	**(2K19/EP/014)**	**(2K19/EP/032)**	**(2K19/EP/038)**

DEPARTMENT OF APPLIED PHYSICS

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project titled "Analysis of Machine Learning Methods for Membership Association in Star Clusters Using GAIA DR3 Data" which is submitted by Anish Kalsi (2K19/EP/014), Dhruv Tyagi (2K19/EP/032), and Harshit Choudhary (2K19/EP/038) for the fulfillment of the requirements for awarding of the degree of Bachelor of Technology (B.Tech) is a record of the project work carried out by the students under my guidance & supervision. To the best of my knowledge, this work has not been submitted in any part or fulfillment for any Degree or Diploma to this University or elsewhere.

Place : New Delhi

Date : 8/12/2022

Dr. Ajeet Kumar

(SUPERVISOR)

Assistant Professor

Department of Applied physics

Delhi Technological University

ABSTRACT

Keywords - machine learning, astrometry, membership classification, GAIA, open clusters, photometry, catalogs

The different machine learning methods in the membership association of stars in open clusters have been studied. By using the data from GAIA Data Release 3, high-precision astrometric parameters can be used to train models for enhanced prediction and identification of new members for existing clusters. By reducing both random and systematic errors introduced by the subjective nature of human classifications, machine learning not only expedites the classification process, which has been a limiting step in the creation of star cluster catalogs, but also improves the consistency of the classifications. The extended goals of this project are to use machine learning models to continue the astrometric and photometric examination of the selected clusters and determine the efficiency of each method over a range of cluster types.

ACKNOWLEDGEMENT

The successful completion of any task is incomplete and meaningless without giving any due credit to the people who made it possible without which the project would not have been successful and would have existed in theory.

First and foremost, we are grateful to **Prof. A.S Rao**, HOD, Department of Applied Physics, Delhi Technological University, and all other faculty members of our department for their constant guidance and support, constant motivation, and sincere support and gratitude for this project work. We owe a lot of thanks to our supervisor, **Dr. Ajeet Kumar**, Assistant Professor, Department of Applied Physics, Delhi Technological University for igniting and constantly motivating us and guiding us in the idea of a creatively and amazingly performed Major Project in undertaking this endeavor and challenge and also for being there whenever we needed his guidance or assistance. A sincere thanks to our faculty coordinator **Dr. M. Jayasimhadri**, Assistant Professor, Department of Applied Physics, Delhi Technological University for smooth and efficient management.

We would also like to take this moment to show our thanks and gratitude to one and all, who indirectly or directly have given us their hand in this challenging task. We feel happy and joyful and content in expressing our vote of thanks to all those who have helped us and guided us in presenting this project work for our Major Project. Last, but never least, we thank our well-wishers and parents for always being with us, in every sense, and constantly supporting us in every possible sense whenever possible.

Anish Kalsi	**Dhruv Tyagi**	**Harshit Choudhary**
(2K19/EP/014)	**(2K19/EP/032)**	**(2K19/EP/038)**

Contents

Candidate's Declaration **i**

Certificate **ii**

Abstract **iii**

Acknowledgement **iv**

List of Figures **vii**

List of Tables **viii**

CHAPTER 1: INTRODUCTION **1**

1.1 Overview 1

1.2 Objectives 2

1.3 Motivation 2

CHAPTER 2: BACKGROUND **3**

2.1 Literature Review 3

CHAPTER 3: DATA SAMPLE **5**

3.1 GAIA Data Release 5

3.2 Data Selection 6

3.3 Data Cleaning 7

3.4 Analysis 7

3.4.1 Hertzsprung-Russell Diagram 7

3.4.2 Color-Magnitude Diagram 8

CHAPTER 4: METHODOLOGY **12**

4.1 DBSCAN 12

4.2 GMM 13
4.3 UPMASK 14

CHAPTER 5: RESULTS AND DISCUSSION 16
5.1 Silhouette Score 16
5.2 Cluster Identification 17

CHAPTER 6: CONCLUSION 20

CHAPTER 7: FUTURE WORK 21
7.1 Expected Results 22

APPENDICES 23

REFERENCES 23

List of Tables

Table 3.1 : Table of Open Clusters 6

Table 5.1 : Clustering performance of DBSCAN algorithm for different open clusters 17

Table 5.2 : Clustering performance of GMM algorithm for different open clusters 17

List of Figures

Figure 3.1 : GAIA (E)DR3 Passbands for G, G_{BP} AND G_{RP} (Ref. [10]) 6

Figure 3.2 : Hertzsprung-Russell diagram (Ref: ESO Press Release eso0728) 8

Figure 3.3 : An example for a typical color-magnitude diagram. CMD for the globular star cluster M55. (Ref: B.J. Mochejska, J. Kaluzny (CAMK), 1m Swope Telescope) 9

Figure 3.4 : Color magnitude diagram of star clusters data obtained from GAIA DR3. 10

Figure 3.5 : Radial velocity diagram of star cluster data obtained from GAIA DR3. Blue points are blue-shifted sources and red points are red-shifted. 11

Figure 4.1 : DBSCAN algorithm that identifies clusters for Minimum Points(N) = 4. Core points are denoted by 'A', Boundary Points by 'B & C', and Noise by 'N'. [12] 13

Figure 4.2 : Typical GMM mixture model implementation which illustrates formation of gaussian probability density components. 14

Figure 4.3 : Simulated data for UPMASK 15

Figure 4.4 : UPMASK results for simulated data 15

Figure 5.1 : Cluster identification using Density-based spatial clustering algorithm for the following different open clusters (a) NGC188, (b) NGC225, (c) NGC6031, (d) NGC6756 18

Figure 5.2 : Cluster identification using Gaussian Mixture Model-based clustering algorithm for the following open clusters (a) NGC188, (b) NGC225, (c) NGC6031, (d) NGC6756 19

Chapter 1

INTRODUCTION

1.1 Overview

Cluster membership classification is one of the most important characteristics in the study of star clusters. We look for stars in the same area of the sky, at comparable distances, and with comparable velocities to identify members. Such samples might be tainted, though, because field stars with comparable velocities or distances do exist. Also, not all of the stars in a cluster often have photometric and spectroscopic (kinematic) data accessible. We calculate a membership probability, which is a challenging problem, for the majority of stars. Variable stars, outliers, etc. are difficult to identify as members when using the usual method, which uses a photometric envelope in the color-magnitude diagram when there is just photometric data available.

Various core Machine Learning techniques used for density-based clustering have been used in cluster identification of stars and galaxies. Some of the widely used techniques are Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Model (GMM), Unsupervised Photometric Membership Assignment in Stellar Clusters (UPMASK), and Random Forest (RF) method.

We aim to use a variety of machine learning methods used in previous literature and compile a comparative study of these methods on open clusters. We select a group of open clusters as our testing models. The astronomical data used for the purpose of this study is retrieved from the GAIA DR3 repository made available by the ESA. Data Re-

lease 3 (DR3) is a fairly recent and new dataset, hence it can give us a lot of insight into these clusters and the amount of information available about them is also pretty abundant compared to older Data Releases.

The goal of this study would be to carry forth the astrometric and photometric analysis of the chosen clusters using the machine learning models and see not only how effective each method is but also how effective it is for different types of clusters.

1.2 Objectives

- To obtain raw color-magnitude diagrams (CMD) and other astronomical plots for the selected open clusters. These analyses contain a range of field stars along with the member stars of clusters, henceforth introducing a lot of "noise" and making a further astrometric analysis like creating isochrones impossible.
- To determine the age of the star cluster, CMD and isochrone fitting are used. Therefore, it is essential to be able to identify members of the cluster as a preliminary step. To achieve this, a variety of machine learning techniques are used on the chosen data set.
- The obtained members will then be subjected to the analyses in the first step to obtain refined CMDs and can be ready for further astrometric study.
- This would be a comparative study of different machine learning methods currently in use for astronomical data analysis.

1.3 Motivation

It is very evident that a hoard of machine learning techniques is being deployed for astronomical data analysis these days including cluster membership identification. A thorough survey of astronomical literature on this topic reveals that although the work done on the viability of individual techniques is surplus, there is not a comparative study of the same. The motivation of this study is to be able to do such a comparative study and see how different algorithms vary for clusters of varying sizes and types.

Chapter 2

BACKGROUND

We had some prior experience and interest in astronomy. To conduct research from scratch, we first brushed up our knowledge and did a thorough literature review. With the goal to utilize data from GAIA DR3, we studied various published articles with GAIA DR2/EDR3 data. This helped us to select a sub-field and formulate our research question. In this process, we read a bunch of research papers, out of which a few relevant papers have been discussed below.

2.1 Literature Review

Star clusters are gravitationally bound star structures. Such structures are very helpful in studying the development of galaxies and our universe. [1] Rediscussion for membership classification in such clusters is a continuous process that betters our understanding of these clusters which in turn is a great tool for photometric analysis.[2]

Determination of cluster members involves either photometric [3] or astrometric [4] analysis of the astronomical data. The inhomogeneity of the data makes it difficult to conduct a systematic analysis of open clusters' type, size, number of members, and age, despite the fact that they are typical representatives of the Galactic disc population. On the one hand, there are generic bibliographic catalogs that were created from reviews of the literature. The parameters of uniform lists of a few hundred clusters, on the other hand, were obtained from cogent photometric or kinematic studies.

We are interested in the kinematic and photometric aspects used for the determination of the membership of stars in a particular cluster. There have been various successful implementations of supervised and unsupervised methods such as DBSCAN and Random Forest on the star cluster data.[6] [7] [8] [9]

The GAIA DR3 [10] provides an even wider dataset to train existing models and further enhance the state of determined members of star clusters. The work done previously has centered around GAIA DR2 [11] dataset. Therefore, a study based on the third release would be novel and more insightful, introducing new and never-before detected objects and improved resolution for the previously detected ones.

Some important work previously done in membership classification has shown improved results when compared to existing catalogs (e.g [2]) Random Forest was used effectively in the calculation of the membership probability of stars in cluster M67.[5] This methodology has been further implemented in [6] by training it on a set of 9-star clusters. This yielded a number of positive results with increased members detected and enriched the previously available CMDs. The number of detected stars almost doubled in certain instances compared to previously published results with a high precision of 90 percent.

Studies with the DBSCAN algorithm [8] over a set of different clusters have yielded positive results. An important aspect to be noted is that the model has been implemented for simulated data and not for real astronomical data. The astronomical data analysis of NGC 225 [9] reveals the contribution of binary stars in the total mass estimate. The proper estimation of binary members was made possible through the DBSCAN algorithm. Two lists of target stars were created, one for stars with probabilities $p > 50\%$ and the other for stars with $p > 90\%$, based on their probabilities. The high confidence set of $p > 90\%$ was then used for the photometric analysis of the cluster.

As shown in [7], the DBSCAN clustering algorithm can effectively segregate cluster members without using mathematical models or making assumptions about the distribution of stars. It is insensitive to the shape of clusters. The input parameters (MinPts, Eps) can affect the outcomes because the DBSCAN clustering method is sensitive to density.

Chapter 3

DATA SAMPLE

3.1 GAIA Data Release

The European Space Agency (ESA) launched Gaia in 2013 and anticipates that it will remain operational until 2025. The spacecraft is made for astrometry, which involves measuring the locations, separations, and motions of stars with incredibly precise measurements.

The mission's goal is to create the largest and most accurate 3D space database ever created, with a total of about 1 billion celestial objects, mostly stars but also including planets, comets, asteroids, and quasars. The GAIA mission makes the data available for public use in the form of Data Releases (DR), the first being released on 14th September 2016.
On June 13, 2022, Gaia Data Release 3 (Gaia DR3) was made available. The Gaia Archive has the information available (and through the partner data centers). The Early Data Release 3 (issued on December 3, 2020) collection is the foundation for the Gaia DR3 catalog, which adds various new data products including extended objects and non-single stars for the same period of time and the same set of observations.

For this study, we accessed the dataset from GAIA DR3. Using the ADQL (Astronomical Data Query Language) interface for complex searches, we searched the archive for GAIA sources. The type of objects has been limited to open star clusters for this part of the study, however, we intend to extend it further to globular star clusters.

3.2 Data Selection

A total of 4 open star clusters have been studied henceforth, namely, NGC 188, NGC 225, NGC 6031, and NGC 6756. The various photometric and observational details of the clusters have been mentioned below.

Table 3.1: Table of Open Clusters

Object Name	Constellation	Magnitude B (Blue, 445nm)	Magnitude V (Visual, 551nm)	Angular Size	Right Ascension J2000	Declination J2000
NGC 188	Cepheus	8.91	8.1	17.7 arcmin	00h 47m 27s	+85° 16' 10"
NGC 225	Cassiopeia	7.43	7	4.2 arcmin	00h 43m 36s	+61° 46' 00"
NGC 6031	Norma	8.92	8.5	4.8 arcmin	16h 07m 35s	-54° 00' 54"
NGC 6756	Aquila	-	10.6	3.6 arcmin	19h 08m 43s	+04° 42' 20"

For uniformity, a radius of 25 arc mins around the center of each cluster has been chosen for data extraction from the GAIA archive. A query was run into the ADQL and the data was imported in CSV format.

To create the Colour Magnitude Diagrams (CMD) we used the Gaia DR3 passbands consisting of the G_{BP} and G_{RP} passband. The passbands FOR G, G_{BP} AND G_{RP} have the following transmissivity characteristics across different wavelengths -

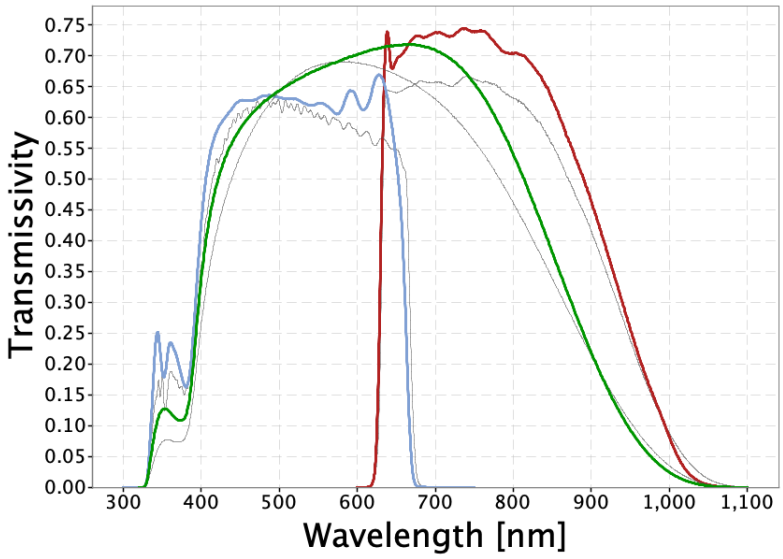


Figure 3.1: GAIA (E)DR3 Passbands for G, G_{BP} AND G_{RP} (Ref. [10])

3.3 Data Cleaning

The parameters necessary for cluster membership association are Proper Motion Right Ascension (pmra), Proper Motion Declination (pmdec), and the Parallax. These three features in conjugation assign the proper positions of stars in 3-D coordinates. This information is further utilized for machine learning computation. However, the data from the GAIA archive contains a lot of noise terms and negligible terms which need to be dropped and the data has to be cleaned before the algorithm code can be run on them.

We used the following code snippet to clean our data and carry forth the machine-learning analyses.

```
features = ["pmra", "pmdec", "parallax"]
X = ds[features]
X = X.replace([np.inf, -np.inf], np.nan).dropna(axis=0)
```

3.4 Analysis

3.4.1 Hertzsprung-Russell Diagram

One of the most crucial tools for studying stellar evolution is the HR diagram, the Hertzsprung-Russell diagram. It was independently developed by Ejnar Hertzsprung and Henry Norris Russell in the early 1900s. It graphs either the color of stars against their absolute magnitude or the temperature of stars against their brightness. It graphs either the color of stars against their absolute magnitude or the temperature of stars against their brightness.

Every star goes through several developmental stages that are determined by its internal structure and method of energy production depending on its starting mass. The temperature and luminosity of the star fluctuate with each of these stages, and as the star develops, it may be seen moving to various areas on the HR diagram.

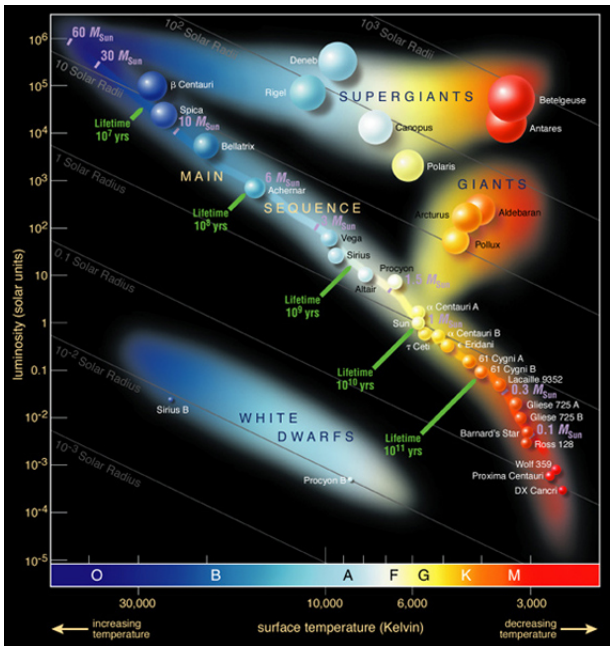


Figure 3.2: Hertzsprung-Russell diagram (Ref: ESO Press Release eso0728)

3.4.2 Color-Magnitude Diagram

The Colour Magnitude Diagram (CMD) is a representation of observational data that demonstrates how the brightness (luminosity) and color of a population of stars can be represented (surface temperature). The idea that stars can be thought of as black-body sources, allowing us to employ Wien's Law, provides the foundation for our ability to interpret a star's color as a measurement of its temperature. Generally, the star's spectral class is plotted on the x-axis (inverted) using this temperature.

Typically CM diagrams are used rather than HR diagrams, due to their greater practicality. For example

- If the distances to all the stars have been determined, then this might be a plot of B - V versus absolute V band magnitude.
- The distances to all the stars have not been determined, such as in the case of a star cluster, then this might be a plot of B - V versus apparent V band magnitude.

These are both equivalent to plots of luminosity vs. temperature. Near- and mid-infrared measurements can be combined with optical data or used on their own to make CMDs.

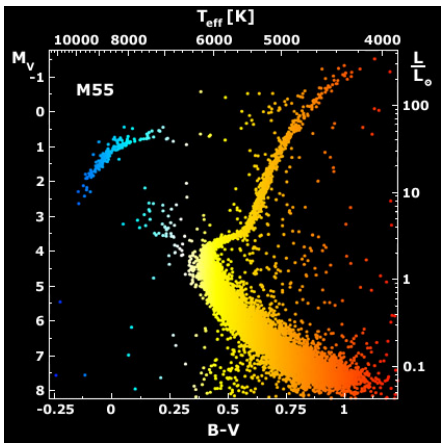
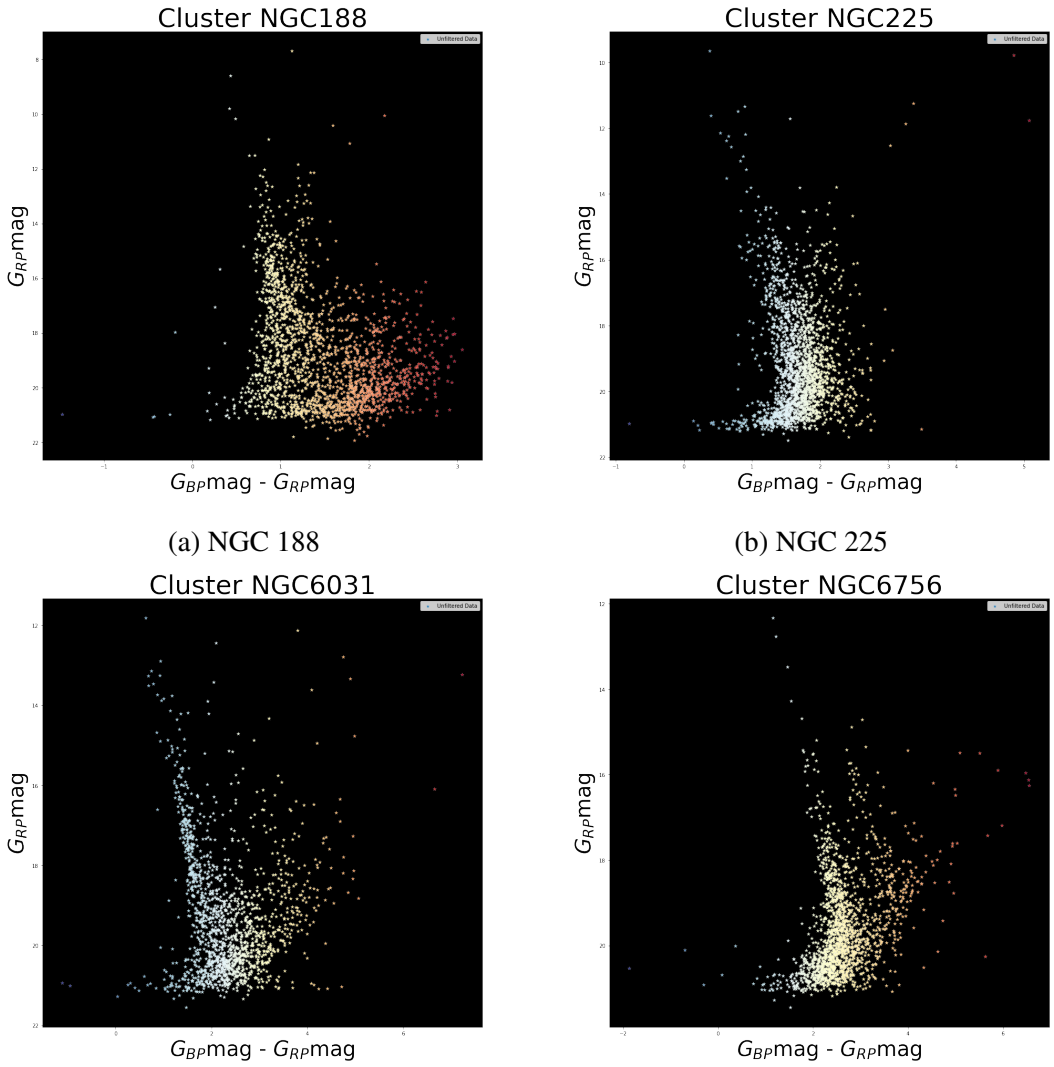


Figure 3.3: An example for a typical color-magnitude diagram. CMD for the globular star cluster M55. (Ref: B.J. Mochejska, J. Kaluzny (CAMK), 1m Swope Telescope)

We have plotted a color-magnitude diagram for our sample clusters (NGC188, NGC225, NGC6031, and NGC6756). As per GAIA data, on the x-axis, we have the difference of apparent magnitudes in B-band and R-band filters (G_{BP}mag - G_{RP}mag). Which signifies the relative temperature of the given stars. On the y-axis, the apparent magnitude of the R-band is plotted, which is equivalent to luminosity in the HR diagram.

In order to better understand the dynamics of a cluster, we analyzed its radial velocities as well. Positive values of radial velocities indicate the source is moving away from the observer and for a negative value of radial velocity, the source is moving towards the observer.

As observing a source to determine its radial velocity is complicated. GAIA doesn't have radial velocity values for all the observed sources. Due to limited data, the obtained plots have low density but it still conveys enough information about the rotational direction of the cluster and its viewing angle.



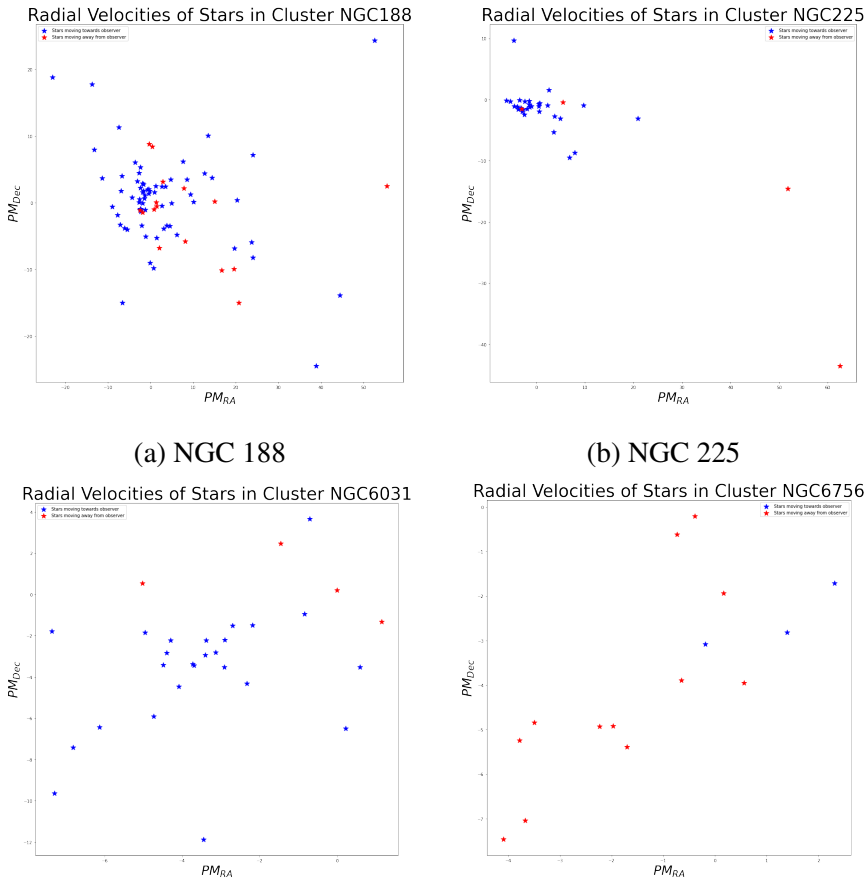
(a) NGC 188

(b) NGC 225

(c) NGC 6031

(d) NGC 6756

Figure 3.4: Color magnitude diagram of star clusters data obtained from GAIA DR3.



(a) NGC 188

(b) NGC 225

(c) NGC 6031

(d) NGC 6756

Figure 3.5: Radial velocity diagram of star cluster data obtained from GAIA DR3. Blue points are blue-shifted sources and red points are red-shifted.

Chapter 4

METHODOLOGY

In this section, the methodology adopted to classify & distinguish between members & field stars of distinct open clusters is discussed. A brief description of the various supervised & unsupervised algorithms implemented to identify members of open clusters has been provided. The distinct approaches employed by each algorithm in the identification of clusters leads to different performances of algorithms.

4.1 DBSCAN

Density-based spatial clustering is a clustering algorithm that assigns points that are packed tightly in close proximity to one another to the same cluster. DBSCAN assigns clusters to data points based on 2 user-defined parameters namely Epsilon (ε) & Minimum points (N). Epsilon (ε) corresponds to the maximum distance between 2 points to label that one is in the neighborhood of the other and Minimum points (N) is the number of points in a neighborhood of a point to label it as a core point.

Based on the selected values of Epsilon & Minimum points, the DBSCAN algorithm operates as follows: The nearest neighboring points in the Epsilon range around every point are located & consequently the points with neighbors greater than minimum points (N) are identified as ”core points”. Points that possess less than $'N'$ neighbors are identified as ”boundary points”, meanwhile points that possess 0 neighbors are labelled as noise. An example of DBSCAN algorithm cluster identification is shown in figure 4.1.

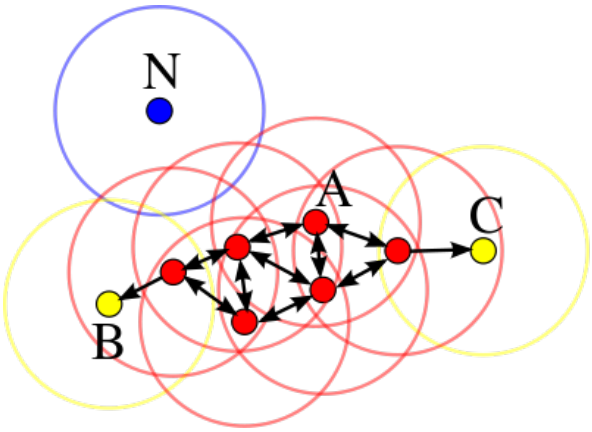


Figure 4.1: DBSCAN algorithm that identifies clusters for Minimum Points(N) = 4. Core points are denoted by 'A', Boundary Points by 'B & C', and Noise by 'N'. [12]

4.2 GMM

Gaussian Mixture Models (GMM) are a type of supervised clustering algorithm which assumes that clusters are distributed in a Gaussian or normal distribution & based on this assumption estimates the probability density of data points belonging to a particular cluster. Gaussian Mixture Model-based clustering assigns data points to clusters with some 'soft' probability & hence allows for the determination of membership probability of stars in open clusters.

The Gaussian mixture model forms different mixture components based on the different features of the dataset being used. It then assigns a set of cluster weights (denoted by π_k) to each Gaussian component. A distinct cluster is represented by each mixture component. This distinct cluster is specified by the following 3 cluster parameters:

$$\{\pi_k, \mu_k, \sigma_k\}$$

Where, π_k are the mixture weights, π_k is the mean vector, & σ_k is the covariance matrix. Thus the probability that the ith point in a dataset belongs to the kth cluster by GMM is given as:

$$p(z_i = k) = \pi_k$$

Where z_i is the cluster assignment for observation 'X_i' of the dataset.

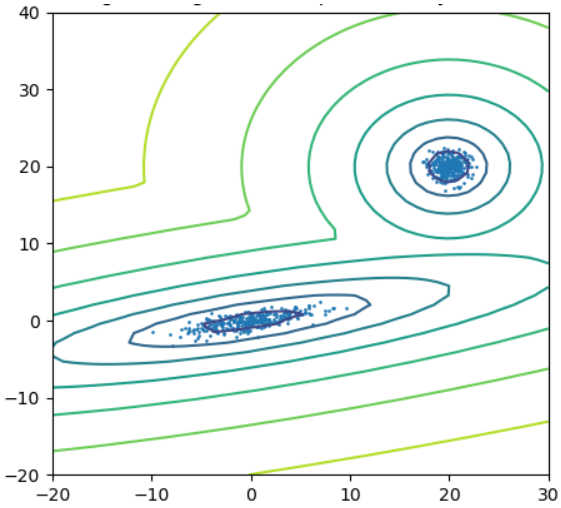


Figure 4.2: Typical GMM mixture model implementation which illustrates formation of gaussian probability density components.

4.3 UPMASK

UPMASK stands for Unsupervised Photometric Membership Assignment In Stellar Clusters. As the name suggests it is an unsupervised machine learning algorithm to determine the membership probabilities of stars in the given field of view. Principal component analysis, a clustering technique, and kernel density estimations are the foundations of the methodology used in this study for membership evaluation. Arbitrary error models can be taken into consideration using the approach, UPMASK. Its ability to successfully segregate cluster and field populations is demonstrated by running the algorithm on simulated data clusters. Under a wide range of circumstances, it is possible to reconstruct the general spatial structure and distribution of cluster member stars in the color-magnitude diagram.

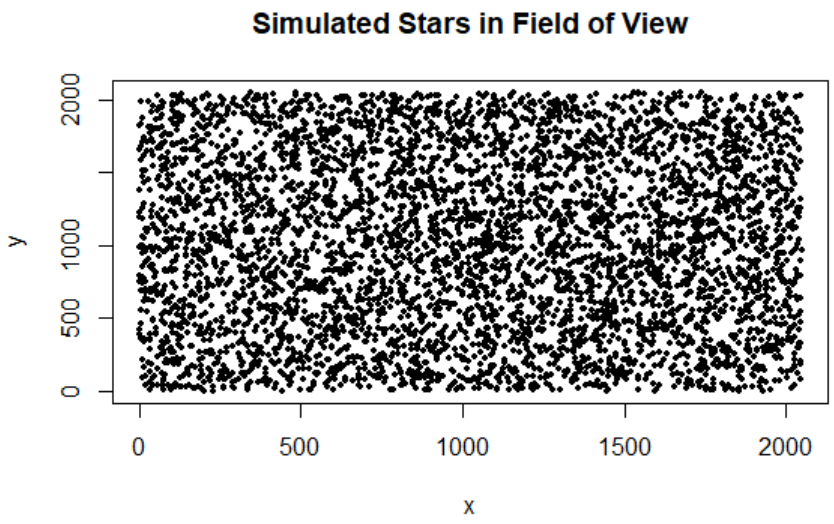


Figure 4.3: Simulated data for UPMASK

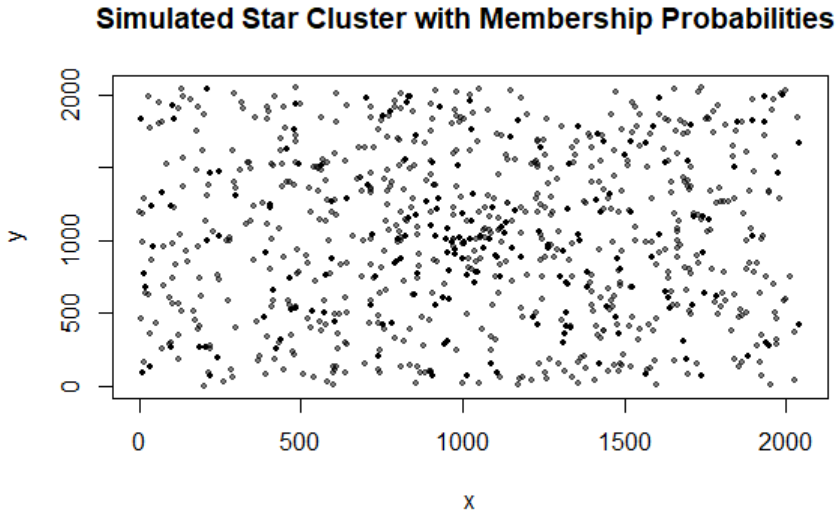


Figure 4.4: UPMASK results for simulated data

Here is a sample to understand the working of UPMASK. A simulated data set is analyzed and results have been shown. With darker the shade of the marker, the more the probability of that star being a member of a given cluster.

Chapter 5

RESULTS AND DISCUSSION

This section presents an overview of the results as well as the performance of the different supervised/unsupervised clustering algorithms implemented. The members of 4 distinct open clusters are determined through different machine learning-based clustering algorithms. The performance & accuracy of different clustering algorithms are compared through metrics such as silhouette scores as well as a graphical representation of the open clusters in 3D space. The open clusters are represented in 3D space through plots of proper motion components of the stars vs. their parallaxes.

5.1 Silhouette Score

Silhouette Score is a metric used to assess the performance of the fit provided by a clustering algorithm. Since DBSCAN provides labels corresponding to noise and labels corresponding to clusters separately, there are two types of silhouette scores that may be computed for DBSCAN algorithms. One includes noise points in the calculation of the silhouette score & the other considers only those data points which are assigned to a distinct cluster. The effectiveness of the algorithm in determining field stars around the open cluster is an important feature to consider when assessing the overall performance of the clustering algorithm. The silhouette scores (including noise points) of the DBSCAN Algorithm for 4 different open clusters are shown in table 5.1 below.

Table 5.1: Clustering performance of DBSCAN algorithm for different open clusters

Sr. No	Open Cluster ID	Epsilon (ε)	Minimum points (N)	Silhouette Score
1.	NGC 188	0.90	40	0.1594
2.	NGC 225	0.90	8	0.5885
3.	NGC 6031	0.90	12	0.3629
4.	NGC 6756	0.90	12	0.3739

The silhouette scores of Gaussian Mixture Model (GMM) clustering algorithms for 4 different open clusters is shown in table 5.2 below.

Table 5.2: Clustering performance of GMM algorithm for different open clusters

Sr. No	Open Cluster ID	N_components (No. of clusters)	Silhouette Score
1.	NGC 188	2	0.4749
2.	NGC 225	2	0.5885
3.	NGC 6031	2	0.3599
4.	NGC 6756	2	0.3749

5.2 Cluster Identification

The performance of clustering algorithms can also be discerned visually through 3D plots of pmra vs pmdec vs parallax. Where, pmra & pmdec are components of proper motion of stars and parallax is the angle of inclination of an object between two distinct lines of sight. A clear difference in the identified members for the same open clusters is observed between figures 5.1 & 5.2 which is also validated by the different silhouette scores observed in tables 5.1 & 5.2.

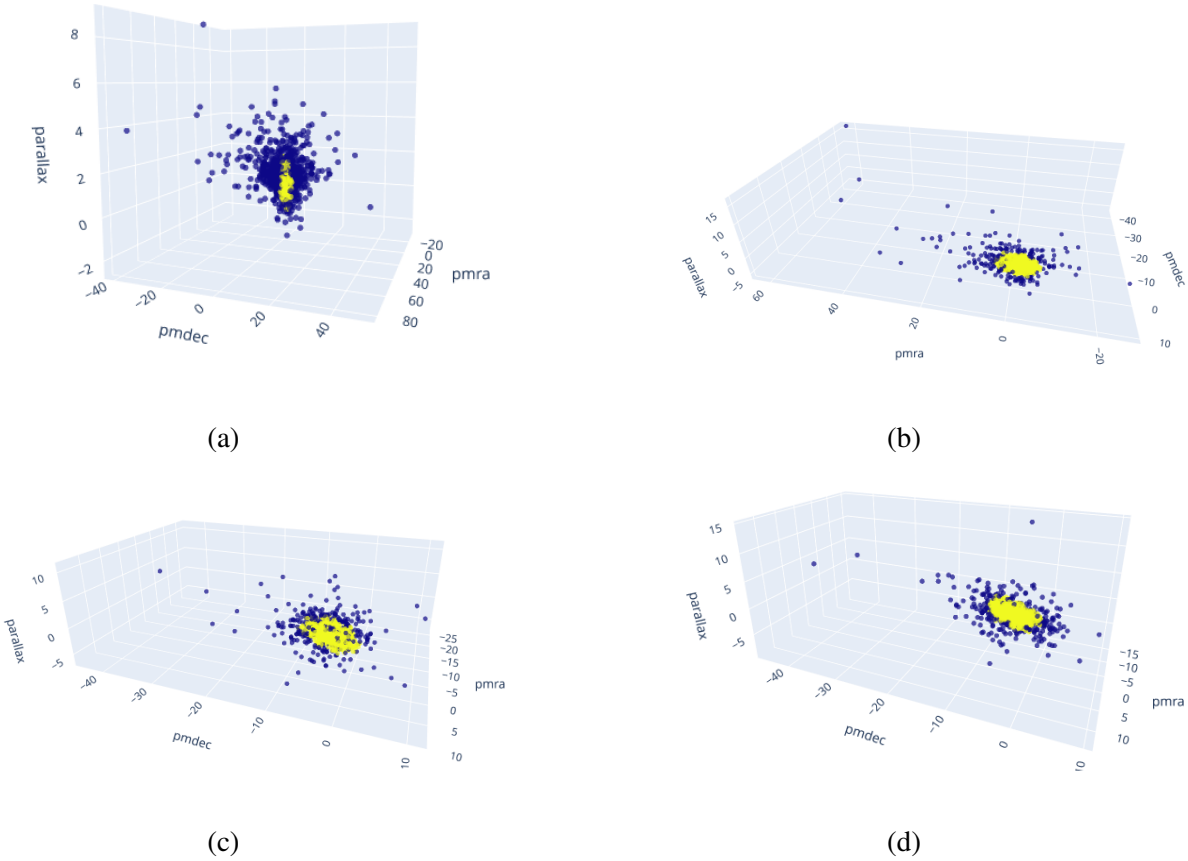
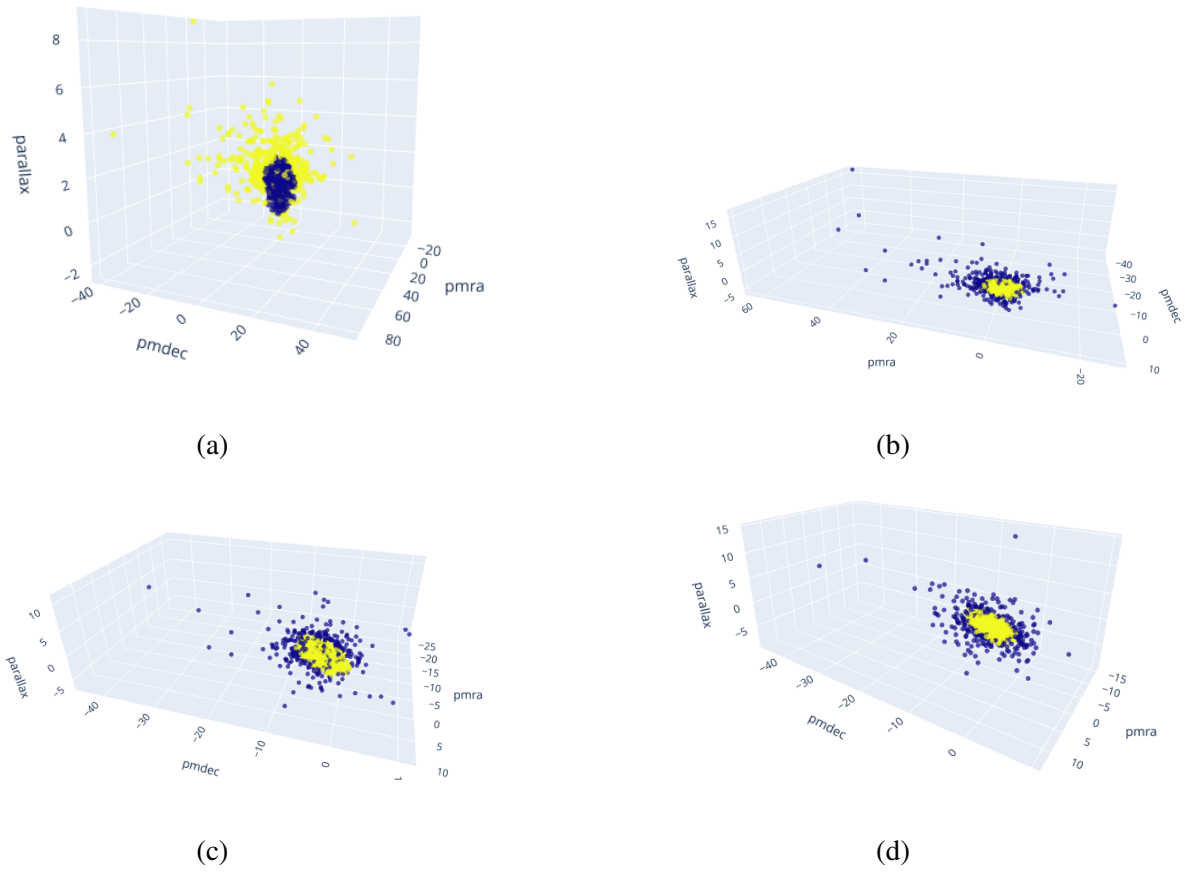


Figure 5.1: Cluster identification using Density-based spatial clustering algorithm for the following different open clusters (a) NGC188, (b) NGC225, (c) NGC6031, (d) NGC6756

From figure 5.1 & it may be concluded that the DBSCAN algorithm performs relatively poorly on certain irregularly populated clusters such as NGC188, which is also reflected by the low silhouette score of 0.1549 as may be seen in table 5.1. In contrast, the GMM algorithm performs substantially better on the same cluster which again may be ascertained through its high silhouette score for the same shown in table 5.2. The better performance of the GMM algorithm for NGC188 might be attributed to the fact that gaussian mixture models utilize probability density based assignment for classification of cluster members whereas DBSCAN provides hard assignment to data based on the selected input parameters.



(a) (b)

(c) (d)

Figure 5.2: Cluster identification using Gaussian Mixture Model-based clustering algorithm for the following open clusters (a) NGC188, (b) NGC225, (c) NGC6031, (d) NGC6756

Chapter 6

CONCLUSION

The membership association of stars in open clusters using various machine learning techniques has been researched. By leveraging the data from GAIA Data Release 3, high-precision astrometric characteristics were used to train algorithms for increased prediction and identification of new members for existing clusters.

We were successfully able to implement and analyze two machine learning methods for membership classification in open cluster stars. DBSCAN and GMM provide a great tool for astronomical cluster identification. The CMD diagrams generated as a part of the analysis are much denser than the previous releases as the new data has a lot more information on the clusters.

Chapter 7

FUTURE WORK

Results so far have been as per our requirements but we have some more plans built upon this project and have this published. Some of the major goals for upcoming months have been listed below:

- Implement other widely used methods for determining or checking membership probability like Unsupervised Photometric Membership Assignment in Stellar Clusters (UPMASK) and Random Forest.
- Include more clusters in this analysis. By using a sample of less studied globular star clusters as well. We have shortlisted NGC 5986, NGC 6293, and NGC 6401 to include for now.
- Compare the results obtained from these methods and conclude, to understand which method is performing better or which method will be suitable for a given task.

These are a few tasks planned for now, which are subject to change as we proceed with the results obtained at each step. We are optimistic that, we will be able to complete our analysis with proper results by January 2023. After which we will prepare a paper and send it to a conference or a journal depending on the results obtained by February 2023.

7.1 Expected Results

We are comparing multiple widely used methods to determine the membership probabilities of stars in a given cluster. To have better inferences we are using different types of clusters. We expect to see a slight variation between the results from these methods such as the number of stars identified to be part of a cluster.

Further, we can study all the common stars that were identified by all the methods and which all stars were clustered by some methods but not others. It will be interesting to see how these algorithms perform with astronomical data.

Bibliography

[1] K. A. Janes and R. L. Phelps, 'The galactic system of old star clusters: The development of the galactic disk', Astron. J., vol. 108, p. 1773, Nov. 1994.

[2] N. V. Kharchenko, A. E. Piskunov, S. Röser, E. Schilbach, and R.-D. Scholz, 'Astrophysical parameters of Galactic open clusters', Astron. Astrophys., vol. 438, no. 3, pp. 1163–1173, Aug. 2005.

[3] A. L. Tadross, P. Werner, A. Osman, and M. Marie, 'Morphological analysis of open clusters' properties', New astron., vol. 7, no. 8, pp. 553–575, Dec. 2002.

[4] W. S. Dias, M. Assafin, V. Flório, B. S. Alessi, and V. Líbero, 'Proper motion determination of open clusters based on the UCAC2 catalogue', Astron. Astrophys., vol. 446, no. 3, pp. 949–953, Feb. 2006.

[5] X. Gao, 'A machine-learning-based investigation of the open cluster M67', Astrophys. J., vol. 869, no. 1, p. 9, Dec. 2018.

[6] M. Mahmudunnobe, P. Hasan, M. Raja, and S. N. Hasan, "Membership of stars in open clusters using random forest with gaia data," The European Physical Journal Special Topics, vol. 230, no. 10. Springer Science and Business Media LLC, pp. 2177–2191, Jul. 08, 2021.

[7] X.-H. Gao, 'Membership determination of open cluster NGC 188 based on the DBSCAN clustering algorithm', Res. Astron. Astrophys., vol. 14, no. 2, pp. 159–164, Feb. 2014.

[8] M. Zhang, 'Use density-based spatial clustering of applications with noise (DBSCAN) algorithm to identify galaxy cluster members', IOP Conf. Ser. Earth Environ. Sci., vol. 252, p. 042033, Jul. 2019.

[9] L. Yalyalieva, G. Carraro, E. Glushkova, U. Munari, and P. Ochner, 'The young Galactic cluster NGC 225: binary stars' content and total mass estimate', Mon. Not. R. Astron. Soc., Apr. 2022.

[10] Gaia Collaboration et al., 'Gaia Data Release 3: Summary of the content and survey properties', 2022.

[11] Gaia Collaboration et al., 'Gaia Data Release 2', Astron. Astrophys., vol. 616, p. A1, Aug. 2018.

[12] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, 'DBSCAN revisited, revisited: why and how you should (still) use DBSCAN', ACM Transactions on Database Systems (TODS), vol. 42, no. 3, pp. 1–21, 2017.