ISSN: 2836-8495

Research Article

Journal of Current Trends in Computer Science Research

A Novel Preparation Approach for Supervised ML Membership Determination of Open Clusters

Omid Rahimpour*, Mohsen Salek and Mehdi Khakian

Department of Physics, Energy Engineering, Amirkabir University, Tehran, Iran

***Corresponding Author**
Omid Rahimpour, Department of Physics, Energy Engineering, Amirkabir University, Tehran, Iran.

Submitted: 2024, Jun 10; **Accepted**: 2024, Jul 04; **Published**: 2024, Jul 18

Citation: Rahimpour, O., Salek, M., Khakian, M. (2024). A Novel Preparation Approach for Supervised ML Membership Determination of Open Clusters. *J Curr Trends Comp Sci Res, 3*(4), 01-10.

Abstract

Machine Learning methods have emerged as powerful tools for analyzing stellar clusters, which pose significant challenges. techniques such as DBSCAN and GMM have advanced remarkably in this domain. However, these clustering techniques exhibit imperfections and limitations, highlighting the need for careful data tuning and consideration of data characteristics to ensure meaningful result.

The utilization of supervised Machine Learning techniques for membership determination of the stellar clusters, especially open clusters, can lead to more accurate results. However, the absence of dataset for training on an open cluster presents a significant hurdle. To address the problem, we've introduced a novel approach to generate a labeled dataset for training the super- vised Machine Learning models. Our approach leverages data from Gaia DR3 Catalog, which provides precise astrometric and photometric measurements for millions stars in Milky Way, to construct a comprehensive dataset.

Our findings have significant implications for future astronomical research. By using Supervised machine learning techniques, we can achieve more accurate and efficient membership determination for stellar clusters, which can lead to a better understanding of the formation and evolution of galaxies. Our method not only enhances the accuracy of membership determination but also provides insights into the underlying data characteristics that influence cluster analysis.

Keywords: Gaia, Star Cluster, Stellar Characteristics, Stellar Classification, Astronomy Data Analysis

1. Introduction

Stellar clusters, as crucial components of stellar astrophysics, play a pivotal role in understanding the formation and evolution of galaxies. member determination is a fundamental task that aids in unraveling the dynamics and properties of these stellar systems. However, the limitations of existing algorithms such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Gaus- sian Mixture Models (GMM) have spurred the need for innovative approaches to enhance accuracy and efficiency in this domain.

1.1 Imperfections of DBSCAN Algorithm

DBSCAN, a popular density-based clustering algorithm, excels in identifying clusters of varying shapes and sizes. However, its performance can be hindered by its sensitivity to parameters such as epsilon and min Points. determining these parameters accurately can be challenging, especially in datasets with varying densities or noise. Moreover, DBSCAN struggles with clusters of varying densities, often leading to under-segmentation or over-segmentation issues.

1.2 Imperfections of GMM Algorithm

Gaussian Mixture Models offer a probabilistic approach to clustering by modeling data as a mixture of Gaussian distributions. While GMM is effective in capturing complex data distributions, it assumes that clusters are spherical and have equal variance, which

may not hold true for open cluster datasets
characterized by non-spherical and varying density distributions.

Moreover, GMM is sensitive to the number of components specified a priori, making it challenging to determine the optimal number of clusters in an unsupervised manner. These limitations underscore the necessity for advanced clustering methods based on the unique characteristics of open cluster data.

1.3 Novel Approach: Generate Labeled Dataset for Training Supervised ML

in light of the imperfections of DBSCAN and GMM in the context of open cluster membership de- termination, this paper introduces a novel approach harnessing the inherent physical principles and integrating domain knowledge with the idea of clustering algorithms, and then generating a training dataset, which can be used for training the supervised ML.

At this paper we'll implement the introduced method on M67 open cluster and will use Gaia DR3 Catalog. the approach used in this paper is generally include five steps:1- considering the Parallax parameter or equally distance and then restrict it to an appropriate range of it 2- map the frequency histograms and then identify the data having least frequency 3- calculate the probability density and then extract the data having the most membership probability 4- providing reference data 5- eventually the final analysis and do membership determination.

Cluster	Age	distance(pc)	Angular Diameter(arc-minutes)	Number of Stars
M67	3.2 to 5 billion-yr	850	30	500

Table 1: Properties of M67 Open Cluster

2. Method

After obtaining the initial raw data of the open cluster in CSV format, from the European Space Agency website, we encounter an extremely noisy file. in order to prepare a suitable initial input for the subsequent processing steps, we need to perform a series of initial tasks on the raw file related to the cluster.

2.1 Restricting the Parallax

The stars in a star cluster are at approximately the same distance from the observer. as a result, by restricting the distance and choosing an appropriate distance range or equally a limited cone-shaped field of view, a large number of nearer and more distant stars can be excluded.

Although the distance parameter isn't directly available for the raw data, the parallax and Parallax Error parameters can be used to restrict the distance of stars. The appropriate range for parallax is calculated by the following formula:

$$ParallaxRange = \frac{1}{Distance} \pm average(ParallaxErrorColumn) \quad (1)$$

M67	Raw Data	After Parallax Restriction
Parallax in marcsec	-0.001653 to 9.8948	0.075 to 2.273
Number of Data	6210	4300

Table 2: Output After the Restriction of Parallax

the number of Data after Parallax Restriction has been shown in the table 2. and a considerable number of background and foreground stars have been eliminated.

2.2 Frequency Histogram

Star clusters are groups of stars that are bound together by gravity and have a common origin. The stars in a cluster have similar properties, such as age, metallicity, and motion. One way to identify and analyze a star cluster is by examining its motion components, Pm RA and Pm Dec, which represent the proper motion in right ascension and declination, respectively.

In this step of the approach, the goal is to clean more irrelevant data and significantly increase the number of stars belonging to the cluster compared to the number of field stars with the following steps:

- initially, we draw the Frequency Histogram for the motion component Pm RA and then identify the velocities having the least frequency.
- therefore, in order to establish a scale for recognition of the least frequency values, we fit a Gaussian curve to the drown Frequency Histogram and calculate its Standard Deviation.

$$\sigma = \sqrt{\frac{\sum_i^n (x_i - \mu)^2}{N}} \quad (2)$$

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \quad (3)$$

• eventually, we define Pm RA values that are located more than 2σ away from the center or average of the Gaussian function as noise and eliminate them. the mentioned steps are illustrated in the figures 1 and 2.

$$the\ \ appropriate\ \ Range\ \ of\ \ PmRA = Average\ of\ PmRA\ column\ \pm\ 2\sigma \qquad (4)$$

• repeat the above steps for Pm Dec motion component, and as a result, our data are analyzed as the same time by Pm RA and Pm Dec. as it can be seen in the figures 3 and 4.

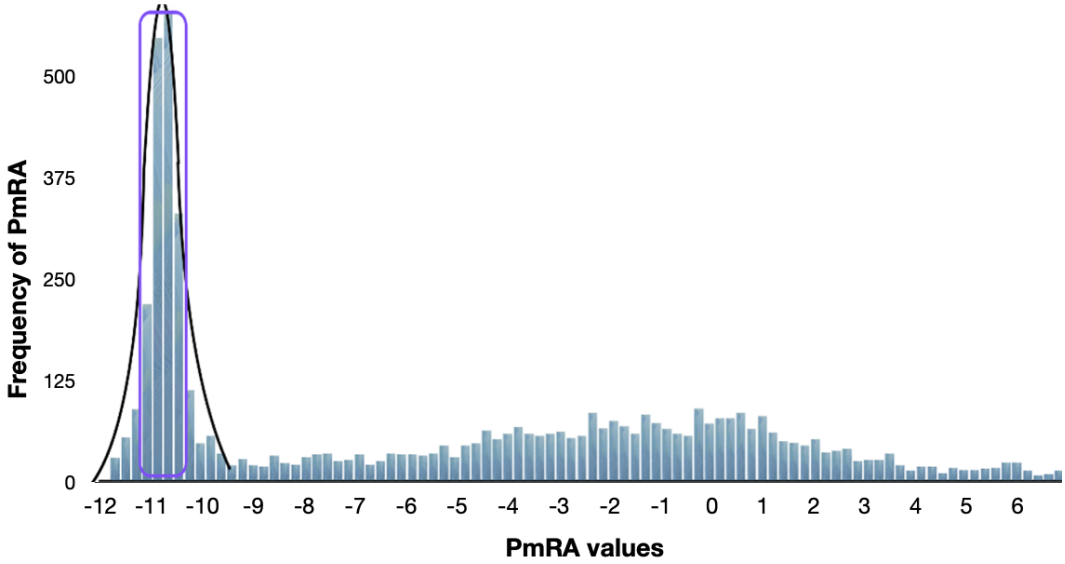


Figure 1: Frequency Histogram of Motion Component Pm RA

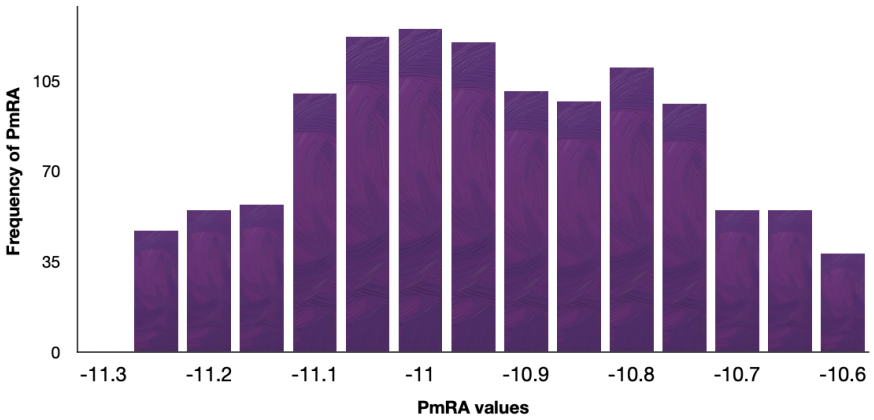


Figure 2: Frequency Histogram of Motion Component Pm RA in the Purple Rectangle

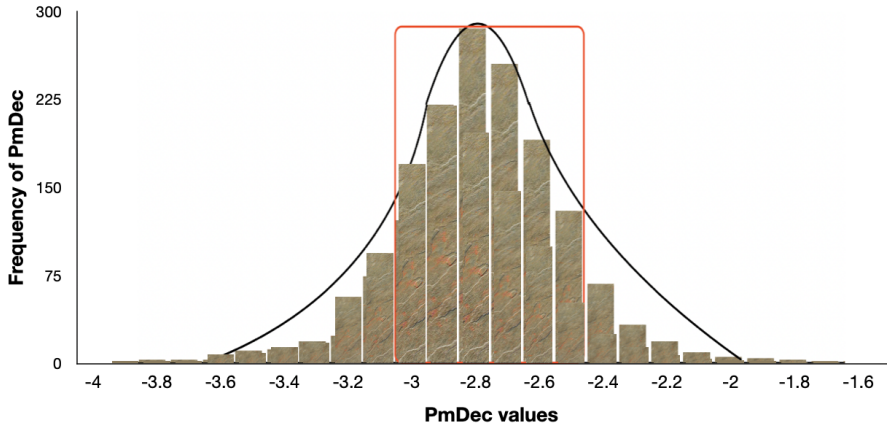


Figure 3: Frequency Histogram of Motion Component Pm Dec

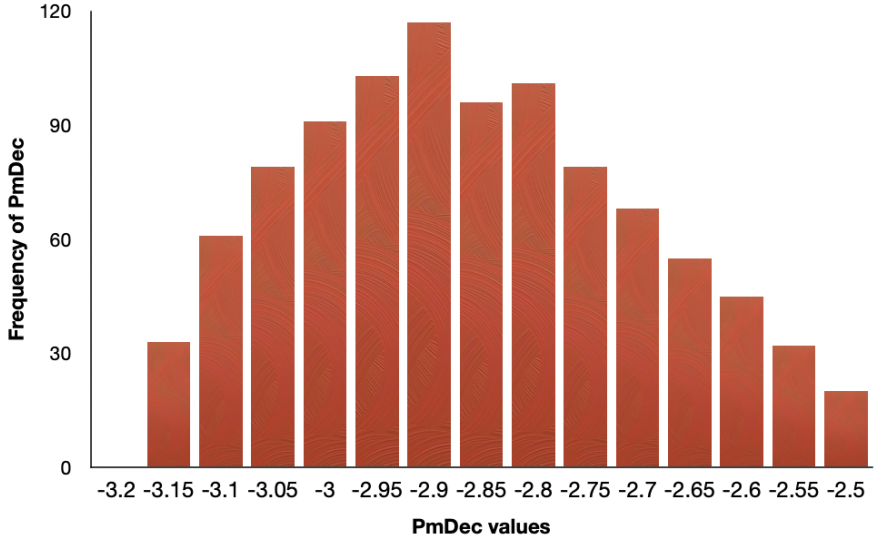


Figure 4: Frequency Histogram of Motion Component Pm Dec in the Orange Rectangle

as it's been shown in the table 3, it's blatantly obvious that how much effective has been the analysis of Frequency Histogram. however, in case we're going to draw the HR diagram.

M67	Raw Data	Restricted Parallax	Frequency Histogram
the number of stars	6210	4300	981

Table 3: Output After the Analyzing Frequency Histograms

2.3 Measuring the Accuracy of the Steps Taken

the Hertzsprung-Russell diagram is a fundamental tool for studying the photometry of stars and evaluating the accuracy and validity of the steps taken to analyze star clusters. by comparing the Magnitude and effective surface temperature of stars, astronomers can group stars according to their evolutionary state and gain insights into the properties and evolution of stars.

most of the stars belonging to a star cluster are expected to located on the main sequence where the sun is, so plotting the Hertzsprung-Russell diagram is a suitable way to measure the accuracy and validation of the output data so far. we plotted the Hertzsprung-Russell diagram for the processed data sofa in Figure 5.

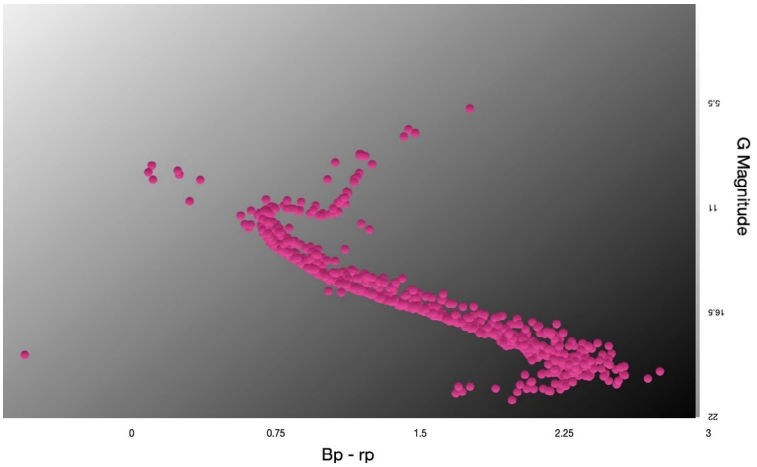


Figure 5: HR Diagram M67 After Initial Analyses

2.4 Calculation of Membership Probability

probability density function is a fundamental mathematical concept that is defined for a random variable and calculate the membership probability of the variety values of a variable. More over the values being near the middle of function have the most membership probability, therefore the values on the edge of function have the least membership probability.

in order to calculate the membership probability, we choose the motion components Pm RA and Pm Dec variables and obtain the probability of them by following the following steps:

• draw the Gaussian function curve for the motion component Pm RA and then split it into variety of equal intervals. the result has been shown in the figure 6.

• calculate the membership probability of the all the selected intervals of Pm RA. the function f(x) is defined as the probability density of variable x, so the membership probability can be acquired with integration. for instance, the probability that variable x is between a and b value is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad \sigma = \sqrt{\frac{(x-\mu)^2}{N}} \tag{5}$$

$$p(a < x < b) = \int_b^a f(x)dx = \frac{-\sigma}{\sqrt{2\pi}(x-\mu)} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{6}$$

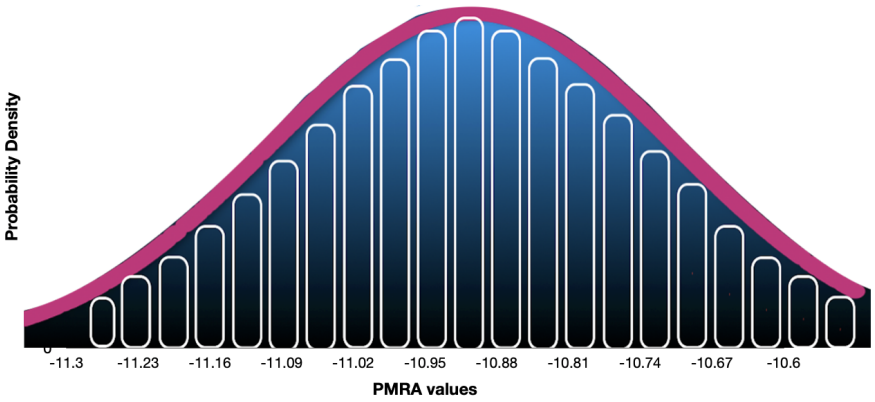
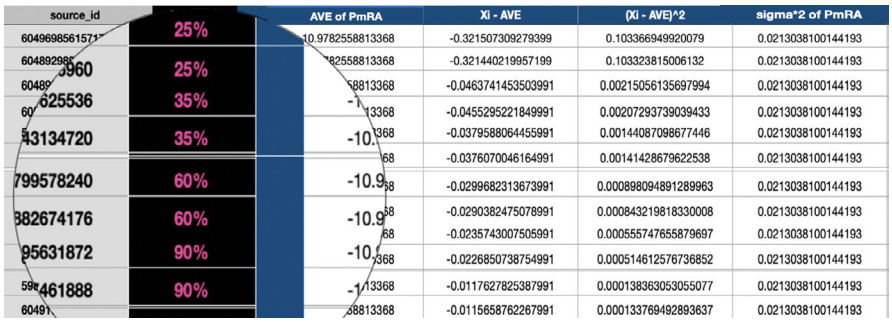


Figure 6: Gaussian Function Curve of M67

• determine the membership probability of motion component Pm RA values of data based on that which interval it places. some of the probabilities calculated for motion component Pm RA have been shown in the figure 7



source_id		AVE of PmRA	Xi - AVE	(Xi - AVE)^2	sigma^2 of PmRA
6049698561571	25%	10.9782558813368	-0.321507309279399	0.103366949920079	0.021303810014419
60489298 3960	25%	82558813368	-0.321440219957199	0.103323815006132	0.021303810014419
6048	35%	8813368	-0.0463741453503991	0.00215056135697994	0.021303810014419
60 625536		13368	-0.0455295221849991	0.00207293739039433	0.021303810014419
43134720	35%	-10. 368	-0.0379588064455991	0.00144087098677446	0.021303810014419
		68	-0.0376070046164991	0.00141428679622538	0.021303810014419
799578240	60%	-10.9 68	-0.0299682313673991	0.000898094891289963	0.021303810014419
382674176	60%	-10.9 68	-0.0290382475078991	0.000843219818330008	0.021303810014419
95631872	90%	-10. 368	-0.0235743007505991	0.000555747655879697	0.021303810014419
59 461888	90%	-1 13368	-0.0226850738754991	0.000514612576736852	0.021303810014419
60491		8813368	-0.0117627825387991	0.000138363053055077	0.021303810014419
			-0.0115658762267991	0.000133769492893637	0.021303810014419

Figure 7: Some of the Pm RA Probabilities Calculated for M67

• repeat the above steps for the motion component Pm Dec, so we have the probability membership of variables Pm RA and Pm Dec at the same time.

source_id	<10%	AVE of PmDec	Xi - AVE	(Xi - AVE)^2	Sigma^2 of PmDec
6049687910053212		-2.88456251839928	-0.30764026975396	0.0946425355742893	0.0264819094973113
60496926774626	40%	-2.88456251839928	-0.30449971720443	0.0927200777775778	0.0264819094973113
604996106997 3	40%	-2.88456251839928	-0.03772334402429	0.00142305068437494	0.0264819094973113
59890010529	63%	2.88456251839928	-0.03739501401369	0.00139838707308407	0.0264819094973113
6049081460 2	63%	.88456251839928	-0.03639046257133	0.00132426576615537	0.0264819094973113
6049747695 2		.88456251839928	-0.03624306945519	0.00131356008353373	0.0264819094973113
60490972661		2.88456251839928	0.02331453402874	0.000543567496977275	0.0264819094973113
604906118842	>63%	-2.88456251839928	0.02422826530445	0.000587008839662816	0.0264819094973113

Figure 8: Some of the Pm Dec Probabilities Calculated for M67

2.5 Provide high-precision Reference Data or Samples

So far, a significant number of field stars have been eliminated, so the disparate parameters associated to cluster stars and field stars are extremely close to each other, and as a result, it poses serious challenges to identify and separate cluster stars from field stars according the former diagrams. to address this issue, we need high-precision reference data or samples in order to enable us to accurately recognize the behaviors of cluster stars.

We separate the stars having the motion components Pm RA and Pm Dec membership probabilities higher than 60 percent at the same time as the reference data or samples. some of these samples have been shown in the figure 9.

source_id	P of PmRA	P of PmDec
604970096675364352		
604962331374512640	90%	63
604712158118974080	90%	>63%
6050030133046241		
6050010547995782	60%	>63%
604987139105593	> 90%	>63%
605001875138050944	> 90%	>63%
604913952863074688		
604969168962428672	0%	>6
604988376056169088	> 90%	>63%
604997172149101312		
604921202767809664	> 90%	>63%
604917835513458816	70%	>63%
5988836426850663		
604920030241180672	60%	>63%
604976315787962240	60%	>63%
604925875692160640		
604919725299047424	90%	>63%
604909726615246336		

Figure 9: Some of the Considered Reference Data for M67

2.6 Recognition of the Cluster Behavior

in this step, we're going to recognize the behavior of stars belonging to cluster such as physical properties of them. so, we've used Right Ascension, Declination and Parallax of the samples acquired in the previous subsection in order to derive a reference interval for the coordinate and velocity of stars belonging to the cluster. the reference interval is calculated by the following formula:

$$reference - interval(RA, Dec, Parallax) = samples - average(RA, Dec, Parallax) \pm 3\sigma \quad (7)$$

Min proper RA	Max proper RA
132.441696826901	133.214568640701

Figure 10: Obtained Reference RA Interval for M67

Min proper Dec	Max proper Dec
11.4644461154933	12.3621802251301

Figure 11: Obtained Reference Dec Interval for M67

Min proper Parallax	Max proper Parallax
0.80140751213055	1.52834973294857

Figure 12: Obtained Reference Parallax Interval for M67

2.7 Processing According to the Reference Interval

now we've recognized the behavior of the most crucial parameters such as Right Ascension, Decli- nation and Parallax and eventually it enables us to detect and separate the stars belonging to cluster from field stars accurately.

Finally, according to the reference interval obtained in the previous step, we eliminate the values outside the reference interval and keep the rest.

M67	Raw Data	Restricted Parallax	Frequency Histogram	reference data
the number of stars	6210	4300	980	458

Table 4: The Number of Data After Final Processing

3 Discussion and Conclusion

In the field of investigating stellar evolution, membership of stellar clusters is vital for understanding the formation and evolution of galaxies. we looked for a way to obtain a labeled dataset for the star cluster which can be used as training dataset for supervised machine learning methods, which are significantly more accurate than unsupervised Machine Learning methods due to the train-process being in them.

initially we cut the region where the cluster is located on the galaxy map of the European Agency Gaia and then received all the data available on the celestial bodies in that region. Then, by performing various processes on the received raw data, ,we managed to create a labeled dataset for the M67 open cluster. The number of members of cluster M67 is about 500 stars reported in various databases. However, it should be noted that the method presented in this research identified 456 members out of 6211 given Raw data. Also, about a hundred of stars have less possibility compared to other data, although they are still relatively valid candidates for studying cluster members.

at the end to evaluate the accuracy and validation of the method adopted in this paper, we'll compare the change in the Hertzsprung–Russell diagram at the initial stage the final stage and then check the Colour-Colour diagram to ensure the accuracy of the cluster stars extinction.

3.1 Hertzsprung–Russell

the Hertzsprung-Russell diagram is a fundamental tool for studying the photometry of stars and evaluating the accuracy and validity of the steps taken to analyze star clusters.

Most of the stars belonging to a star cluster are expected to located on the main sequence where the sun is, so plotting the Hertzsprung-Russell diagram is a suitable way to measure the accuracy and validation of the output data, which the horizontal axis and vertical axis are graded according to the colour spectrum BP-RP colour and G Magnitude, respectively.

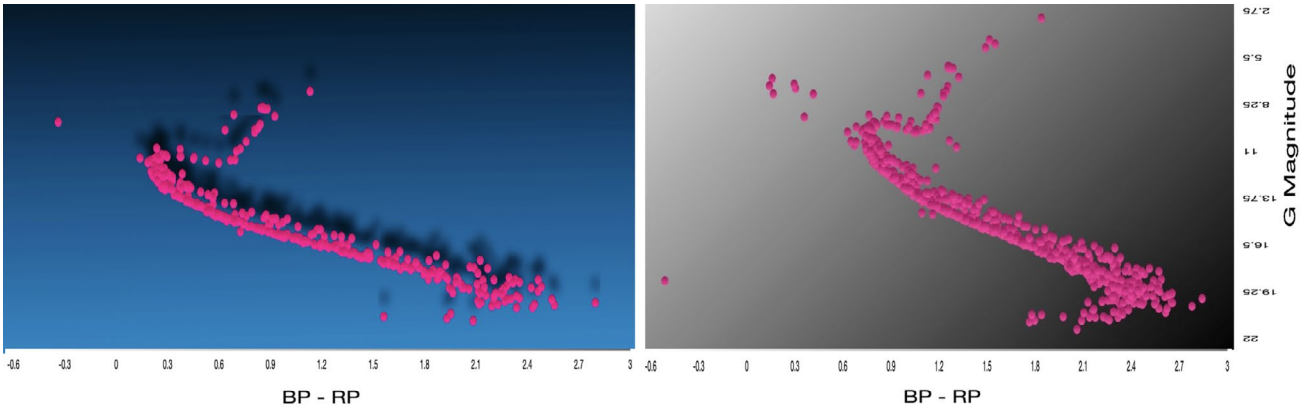


Figure 13: Compare HR Conclusion for M67

3.2 Colour-Colour Diagram

Color-color diagram is an important tool in the field of astronomy, in order to analyze the black body properties of stars and their extinction. In this scheme, two different colors (g-r, b-g, b-r) of a celestial body, which come from the physical magnitude difference, are usually displayed on two sets of coordinates.

Also, due to Rayleigh scattering, which is proportional to 1 , the extinction of the star is greater at smaller wavelengths. As a result, the factor that causes the star to move away from the black body curve is the extinction effect.

the Colour-Colour diagrams of output data have been shown in the figure 14.

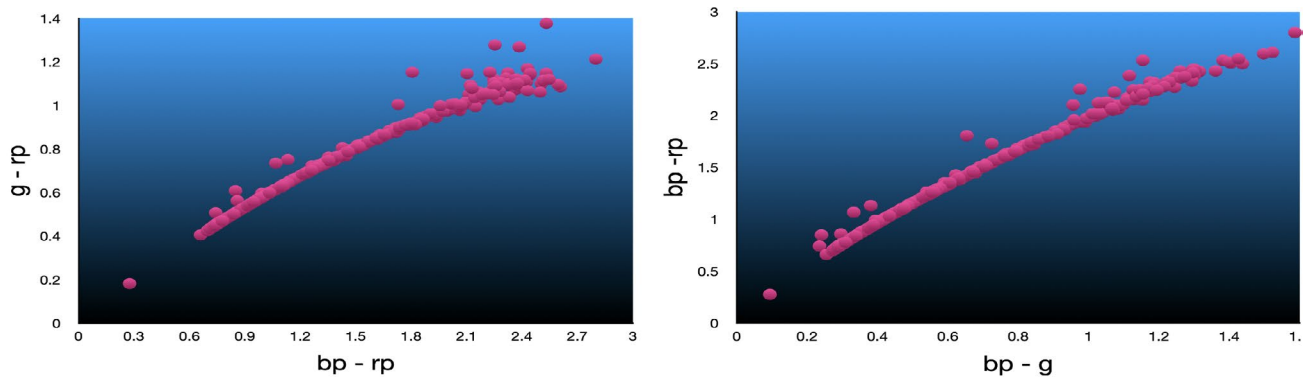


Figure 14: Colour-Colour Diagram of Final Output for M67

According to the concepts mentioned about the color-color diagram, we expect that most of the stars belonging to a cluster are located on the black body curve, in addition, the stars outside the curve have at least one of the following characteristics: 1) The stars are located at a further distance or equally have less Parallax. 2) The stars should be located in the dense area of the cluster or in other words close to the center of the cluster.

So, we plot the coordinates Right Ascension and Declination to see the position of the stars being out of black-body curve, also calculate the average Parallax of them to see if they're at a further distance or not.

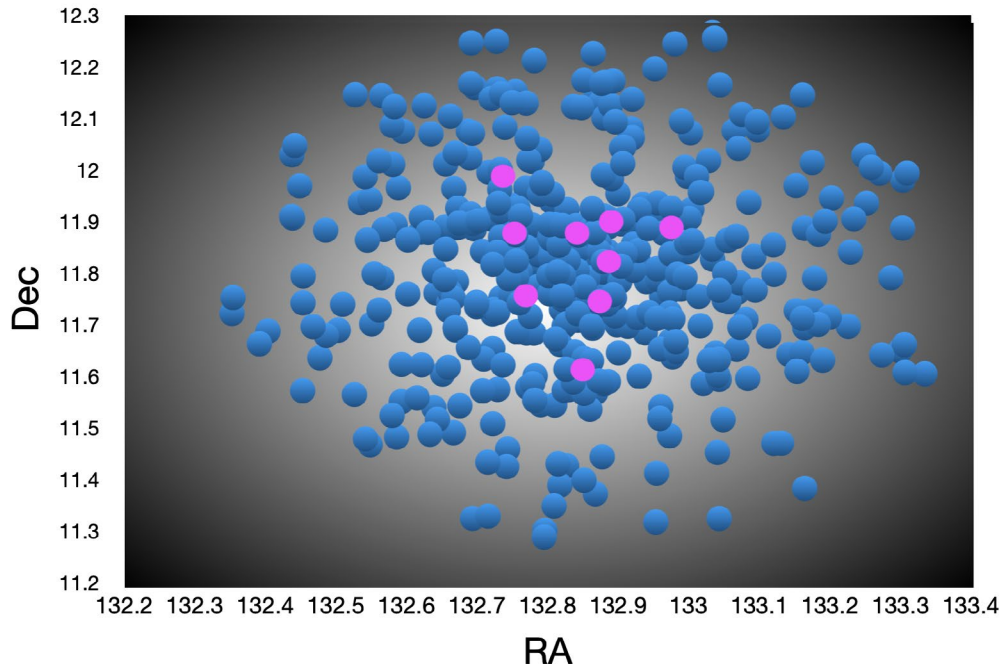


Figure 15: The Pink Points: Points Out of the Black-Body Curve, the Blue Points: Points on the Black- Body Curve

As it can be seen, the stars being out of the black-body curve are in the dense areas or close to the center of cluster, moreover, they have less Parallax or equally are at a further distance. So, it makes sense that the stars have more extinction and so be out of black-body curve.

M67	Average Parallax	Standard Deviation of Parallax
all objects	1.14769918	0.09270235
objects out of black-body curve	1.13353892	0.1464153749

Table 5: Comparison of Parallax of Stars on the Black-Body Curve Verses Out of Black-Body

References

1. Collaboration, G., & Bono, G. (2016). Gaia Data Release 1. Summary of the astrometric, photometric, and survey properties. *Astronomy & Astrophysics, 595.*
2. Erik Høg. Astrometric Accuracy of Positions, 2024.
3. Reyes-Reyes, S. D., Stutz, A. M., Megeath, S. T., Xu, F., Álvarez-Gutiérrez, R. H., Sandoval-Garrido, N., & Liu, H. L. (2024). Benchmarking the IRDC G351. 77− 0.53: Gaia DR3 distance, mass distribution, and star formation content. *Monthly Notices of the Royal Astronomical Society, 529*(3), 2220-2233.
4. Belwal, K., Bisht, D., Bisht, M. S., Rangwal, G., Raj, A., Dattatrey, A. K., ... & Bhatt, B. C. (2024). Exploring NGC 2345: A Comprehensive Study of a Young Open Cluster through Photometric and Kinematic Analysis. *The Astronomical Journal, 167*(5), 188.
5. Mahmudunnobe, M., Hasan, P., Raja, M., Saifuddin, M., & Hasan, S. N. (2024). Using GMM in open cluster membership: An insight. *Astronomy and Computing, 46*, 100792.
6. Raja, M., Hasan, P., Mahmudunnobe, M., Saifuddin, M., & Hasan, S. N. (2024). Membership determination in open clusters using the DBSCAN Clustering Algorithm. *Astronomy and Computing, 47,* 100826.
7. Divakar, D. K., Saraf, P., Sivarani, T., & Doddamani, V. H. (2024). Possibilities of identifying members from Milky Way satellite galaxies using unsupervised machine learning algorithms. *Journal of Astrophysics and Astronomy, 45*(1), 5.
8. Jamal, S., & Bailer-Jones, C. A. (2024). Improved source classification and performance analysis using Gaia DR3. *arXiv preprint arXiv:2405.01340.*
9. Li, Z. M., & Mao, C. Y. (2024). BSEC method for unveiling open clusters and its application to Gaia DR3: 83 new clusters. *Research in Astronomy and Astrophysics, 24*(5), 055014.
10. Sariya, D. P., Jiang, G., Bisht, D., Yadav, R. K. S., & Rangwal, G. (2023). A Gaia based analysis of open cluster Berkeley 27. *New Astronomy, 98*, 101938.
11. He, Z., Li, C., Zhong, J., Liu, G., Bai, L., Qin, S., ... & Chen, L. (2022). New Open-cluster Candidates Found in the Galactic Disk Using Gaia DR2/EDR3 Data. *The Astrophysical Journal Supplement Series, 260*(1), 8.
12. Guilherme-Garcia, P., Krone-Martins, A., & Moitinho, A. (2023). Detection of open cluster rotation fields from Gaia EDR3 proper motions. *Astronomy & Astrophysics, 673,* A128.
13. Raja, M., Hasan, P., Mahmudunnobe, M., Saifuddin, M., & Hasan, S. N. (2024). Membership determination in open clusters using the DBSCAN Clustering Algorithm. *Astronomy and Computing, 47,* 100826.
14. Schmeja, S. (2011). Identifying star clusters in a field:

A comparison of different algorithms. *Astronomische Nachrichten, 332*(2), 172-184.
15. Lada, C. J. (2010). The physics and modes of star cluster formation: observations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 368*(1913), 713-731.
16. Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. *Wiley interdisciplinary reviews: data mining and knowledge discovery, 1*(3), 231-240.
17. Milone, A. P., & Marino, A. F. (2022). Multiple populations in star clusters. *Universe, 8*(7), 359.
18. Mahmudunnobe, M., Hasan, P., Raja, M., & Hasan, S. N. (2021). Membership of stars in open clusters using random forest with gaia data. *The European Physical Journal Special Topics, 230,* 2177-2191.
19. He, Z., Li, C., Zhong, J., Liu, G., Bai, L., Qin, S., ... & Chen, L. (2022). New Open-cluster Candidates Found in the Galactic Disk Using Gaia DR2/EDR3 Data. *The Astrophysical Journal Supplement Series, 260*(1), 8.
20. Bishop, C. M. (2006). Pattern Recognition and Machine Learning - Chapter 9. Microsoft Research. Available at: https://www.microsoft.com/en-us/research/people/cmbishop/prml-book/chapter9. pdf
21. He, Z., Li, C., Zhong, J., Liu, G., Bai, L., Qin, S., ... & Chen, L. (2022). New Open-cluster Candidates Found in the Galactic Disk Using Gaia DR2/EDR3 Data. *The Astrophysical Journal Supplement Series, 260*(1), 8.