Star Cluster Membership Identification By Supervised Machine Learning Models Applied To N-Body Simulations

by

Abylay Bissekenov

Submitted to the Department of Physics
in partial fulfillment of the requirements for the degree of

Master of Science in Physics

at the

NAZARBAYEV UNIVERSITY

Apr 2023

Author .....
Department of Physics
Apr 21, 2023

Certified by .....
Ernazar Abdikamalov
Associate Professor
Thesis Supervisor

Certified by .....
Bekdaulet Shukirgaliyev
Postdoctoral Researcher
Thesis Supervisor

Accepted by .....
Gonzalo Hortelano
Dean, School of Sciences and Humanities

Star Cluster Membership Identification By Supervised Machine Learning Models Applied To N-Body Simulations

by

Abylay Bissekenov

Submitted to the Department of Physics
on Apr 21, 2023, in partial fulfillment of the
requirements for the degree of
Master of Science in Physics

Abstract

This thesis investigates possible ways to apply supervised machine learning algorithms on N-body simulations. Because of the limitations of observational data, there is a motivation to research star clusters by the N-body simulations. The simulations used for the study are based on the Plummer model, and each has its star formation efficiency (SFE) and several random realizations. A random forest model was trained on the simulation with 15% star formation efficiency on a 20-100 Myr timeframe. The model was tested on the other N-body simulations with 17-25% SFEs and showed high classification accuracy throughout the whole dynamic evolution of tested simulations. The majority of mistakes of the model were the false positives (FP) that turned out to be within a 2 Jacobi radius, indicating that they might be gravitationally bounded to center of cluster. Framework and learning strategy can be considered effective and further applied for the mock observations of N-body simulations.

Keywords: Star clusters, N-body simulation, Machine Learning, Supervised Learning.

Thesis Supervisor: Ernazar Abdikamalov
Title: Associate Professor

Thesis Supervisor: Bekdaulet Shukirgaliyev
Title: Postdoctoral Researcher

Acknowledgments

I'd like to thank my supervisors professor Ernazar Abdikamalov and Dr. Bekdaulet Shukirgaliyev for guidance and support that they provided for me during my Master's. Special thanks to the Dr. Shukirgaliyev who was the proposer of the idea of this paper and this work would not have been possible without him. Additionally, I would like to thank Mukhagali Kalambay for lots of useful tips and support that he provided. I also would like to thank Energetic Cosmic laboratory for the support. Last but not least, I'd like to thank my family that always encourage and support me when I needed it the most.

Contents

1 Introduction **13**

1.1 Star clusters and N-body simulations 13

1.2 Previously applied machine learning methods on Star cluster membership analysis 15

1.3 Motivation for the use of Machine learning algorithms on N-body simulations 16

2 Methods **19**

2.1 Data analysis 19

2.2 Learning strategy and datasets 21

2.3 Testing method 24

2.4 Machine learning algorithm 25

2.5 Evaluation metrics 26

3 Results and discussions **29**

3.1 Performance based on evaluation metrics 29

3.2 Analysis of errors 35

4 Conclusion **45**

List of Figures

1-1 Pleiades OSC (left picture) and Omega Centauri GSC (right picture) captured with the WFI camera from ESO's La Silla Observatory. . . 14

2-1 Exploratory data analysis of OSC simulation with 15% SFE at 100 Myr. Diagonal plots are histograms, and other plots are features with respect to each other 22

2-2 Color/magnitude map of OSC simulation with 15% SFE at 100 Myr . 23

2-3 SFE=15% Cluster at 20 Myr, cut from 3 Jacobi radius and without cutting. 24

3-1 Performance evaluation of RF model on simulations with 17% SFE on violet models from Table 2.3 in terms of Accuracy (top panel), False Positive Rate (bottom left panel), and False Negative Rate (bottom right panel) throughout the dynamic evolution. 30

3-2 Performance evaluation of RF model on simulations with 17% SFE on violet models from Table 2.3 in terms of Accuracy (top panel), False Positive Rate (bottom left panel), and False Negative Rate (bottom right panel) throughout the dynamic evolution. 31

3-3 Performance evaluation of RF model on simulations with 20% SFE on pink models from Table 2.3 in terms of Accuracy (top panel), False Positive Rate (bottom left panel), and False Negative Rate (bottom right panel) throughout the dynamic evolution. 32

3-4 Performance evaluation of RF model on simulations with 20% SFE on violet models from Table 2.3 in terms of Accuracy (top panel), False Positive Rate (bottom left panel), and False Negative Rate (bottom right panel) throughout the dynamic evolution. 33
3-5 Performance evaluation of RF model on simulations with 25% SFE on pink models from Table 2.3 in terms of Accuracy (top panel), False Positive Rate (bottom left panel), and False Negative Rate (bottom right panel) throughout the dynamic evolution. 34
3-6 Performance evaluation of RF model on simulations with 25% SFE on violet models from Table 2.3 in terms of Accuracy (top panel), False Positive Rate (bottom left panel), and False Negative Rate (bottom right panel) throughout the dynamic evolution. 35
3-7 Classification result on 17% SFE cluster at 20 Myrs. Crosses are the negatives that are either true or false. Triangles are the positives both true or false. 37
3-8 Classification result on 17% SFE cluster at 100 Myrs. Crosses are the negatives either true or false. Triangles are the positives 38
3-9 Classification result on 17% SFE cluster at 500 Myrs 39
3-10 Classification result on 17% SFE cluster at 1 Gyrs 40
3-11 Number of FPs throughout the test that was within 2 Jacobi radius (blue) and between 2 and 3 Jacobi radius (green) on simulation with 17% SFE and 22 random realizations. No FPs beyond these distances. 41
3-12 Number of FPs throughout the test that were within 2 Jacobi radius (blue) and between 2 and 3 Jacobi radius (green) on simulation with 20% SFE and 22 random realizations. No FPs beyond these distances 42
3-13 Number of FPs throughout the test that was within 2 Jacobi radius (blue) and between 2 and 3 Jacobi radius (green) on simulation with 25% SFE and 22 random realizations. No FPs beyond these distances. 43

List of Tables

2.1 All random realizations of position and mass 20

2.2 All combinations of random realizations. 20

2.3 Types of simulations used for testing. Gray is the simulation that was not tested, but used for training. Pink is the simulations that are similar to training datasets either by position or mass. Violet are the simulations that are completely different from learning datasets by position and mass. All this is for 3 SFEs each having 4 similar and 4 different simulations. The total tested is 24 simulations. 25

2.4 Confusion matrix representing performance of the prediction compared to the actual value. 27

Chapter 1

Introduction

1.1 Star clusters and N-body simulations

This section provides general information about star clusters and N-body simulations used for these purposes.

In recent decades, observational data has been growing intensively with the development of technology. Analyzing these data gives us new concepts about the evolution of stellar systems, Galaxies, and the Universe [1]. And the study of star formation reveals more about the evolution of galaxies and the universe as a whole [1, 2]. Stars can form from interstellar dust and gas. Since the volume of such gases is very large, stars appear in groups, so-called star clusters. In this regard, we can assume that all the stars of the same cluster are the same age, and their number reaches 10^3 to 10^7 stars in the cluster. The cluster's size, mass, and dynamic state play an important role since stars can live in the parent cluster for up to several billion years, depending on it [3]. As we have already said, all stars of the same group are formed at approximately the same time, and it is easier to determine the age of stars in a cluster than for single stars in Galactic fields. Additionally, it is easier to observe star clusters than singular stars because it is almost impossible to observe singular stars in other galaxies. Therefore, studying the population of star clusters and their age distribution allows us to learn the history of star formation in galaxies, including the vicinity of the Sun in the Milky Way [4].

There are 2 types of clusters: open star clusters (OSCs) and globular star clusters (GSC). OSCs are generally young clusters (<100 Myr) mostly found on the plane of the galaxies, such as our galaxy Milky Way in the shape of spiral arms. GSCs are massive old clusters in the galactic halo and nearby center of Galaxy [3]. Stars in GSCs are strongly gravitationally bounded, and their numbers contain tens of thousands to millions of stars. In contrast, OSC's stars are weakly gravitationally bounded and have numbers of particles that could be hundreds to a few thousand. Examples of the types of clusters can be seen in Figure 1-1.

Figure 1-1: Pleiades OSC (left picture) and Omega Centauri GSC (right picture) captured with the WFI camera from ESO's La Silla Observatory.

For correct interpretations when observing clusters in the galaxy, we need to understand the formation and evolution of star clusters. Additionally, it is important to understand the mechanism of their decay since decay greatly changes the population of star clusters, and a wrong understanding of these processes can distort information. To address this problem dynamic evolution of star clusters needs to be studied.

The rapid rise of the computational capabilities of the hardware allowed us to solve N-body problems and simulate the entire cluster dynamic evolution with thousands of stars. As a result, throughout the years, scientists used and still use N-body simulations for various analyses of OSCs [5]. These simulations help to understand how clusters evolve from star formation and gas expulsion until violent relaxation

and further dissolution. These kinds of simulations are done with the gravitational N-body simulation code that runs on GPUs with the support of CUDA [6].

The main research concerns various parameters of star clusters, such as age, metallicity, etc. However, initial part of most of the research is on membership analysis of star clusters. This is especially important for the open star clusters that tend to be younger and scattered compared to far and older globular star clusters, as shown in figure 1-1.

1.2 Previously applied machine learning methods on Star cluster membership analysis

Considering the advances in observational means of research,, various tools such as machine learning (ML) were also applied. This section will provide information about previously applied machine learning methods. It was found that various machine learning methods were used as the random forest, k-nearest neighbors(KNN), and unsupervised learning methods such as StarGo, UPMASK, and Gaussian mixture model (GMM) on databases such as Gaia DR2 or Gaia DR3.

Unsupervised methods are used mostly as clustering algorithms, and those methods can also contribute to the increasing number of stars with membership probabilities. However, most of these studies were done with clustering algorithms that need 3D data for proper usage. Among those methods, a self-organized map algorithm named StarGo used a 5D topology map with 5 features and could identify membership information as in papers [7] and [8]. Another algorithm is the UPMASK, based on k-means clustering that enriched the Gaia DR2 database with new member star objects as seen in [9] and [10]. Additionally, a density-based algorithm called HDNSCAN was used on the M67 open star cluster from Gaia EDR3, which determined membership probabilities for further analysis of parameters of clusters [11]. Overall, clustering algorithms are the main ML algorithms used for membership identification of the OSCs, but they are nowadays used for labeling the data for training with supervised

ML models.

Supervised methods such as random forest and KNN were used alongside unsupervised ML methods that labeled the observational data for the training. Proficient use of random forest with GMM was conducted on M67 and M45 clusters of Gaia DR2 data which gave membership probability results of 0.8 and 0.96 for respective clusters [12], [13]. Additionally, the random forest was used with spectral clustering (unsupervised ML method) and random forest on NGC 188 star cluster with 3780 sample stars which showed a high membership probability (>75%) of 645 stars [14]. A study on 15-star clusters of Gaia DR-2 with combined learning of KNN and GMM showed better results than using special software such as UPMASK [15]. Overall, the supervised approach gives results that help to enrich the existing catalogs or calculate probabilities of star memberships but are trained on data that was labeled with an unsupervised ML algorithm.

Those may seem considerable results, and these methods enrich the various catalogs, but comparison of these kind of studies show results that partially do not co-inside with each other. For example, 3 different groups (CG18 [9], KC19 [16], M21 [17]) studied membership of NGC 2516 cluster. All of them found different number of member stars, but only 25% of KC19 stars, 41% of M21 and 68% of CG18 stars are co-inside with each other[18]. This shows that different methods and studies may have errors and may consider non-members as members, or vice versa.

1.3 Motivation for the use of Machine learning algorithms on N-body simulations

As you saw in the previous section, all the previous machine learning methods applied for membership identification were mostly based on the observational data from various catalogs and served mostly as means of adding additional stars for the observed clusters. As discussed in the previous section, it was indicated that star memberships within catalogs might be far from the actual membership. Considering those factors

and observational limitations, there was an idea to use N-body simulations for the membership analysis. Stars in N-body simulations are already labeled as members or non-members for the whole simulated lifespan and have all the necessary features for applying and testing ML approaches. This kind of framework should allow us to test existing methods of membership analysis and possibly even create more effective methods.

This study aims to explore the possibilities of using ML models on N-body simulations for the membership analysis of OSC. Would we be able to use N-body simulations for supervised ML so that it can be further used on observational data? What kind of features should we use for training? How should the ML model be trained on many datasets containing the state of all stars during different timeframes of stellar evolution?

An attempt to answer these questions will be given in this paper.

Chapter 2

Methods

In this section there would be discussed all the used data, machine learning algorithms, learning strategies and evaluation criteria.

2.1 Data analysis

The data used for this study is the various N-body simulations of star clusters with different positions, mass and etc. Simulations differ by the number of stars, random realization and star formation efficiency (SFE) which the percentage of mass of the cluster that was used for the creation of the stars (100-SFE equals to the percentage of the mass that was blown away by gas expulsion) [19]. Each of these kinds of simulations contains various numerical data about OSC stars for each time step of cluster evolution in N-body time. This study's simulations are based on the Plummer model [19].

Random realizations are the probabilistic values assigned at the star of the simulation for the positions and mass. Simulations with different random realizations have either different mass or position. There are 3 random realizations of position and mass which can be seen in 2.1. For each SFE there are 9 simulations based on all combinations of random realizations of position and mass 2.2. There are 4 SFEs meaning that total number of simulations is 36.

Position (P)	Mass (M)
1	1
2	2
3	3

Table 2.1: All random realizations of position and mass

SFE-PM

11	12	13
21	22	23
31	32	33

Table 2.2: All combinations of random realizations.

In this study, simulations were used with SFEs of 15-25% with 10 000 stars because star clusters tend to be less stable and dissolve faster compared to star clusters with high SFEs and a higher number of stars. Among those 10 000 stars, around 20% of faint stars would be excluded from the training. Faint have ambiguously high color features and basically cannot be seen and thus should be excluded. From Gaia databases, we also know that stars with apparent magnitudes higher than 21 would be invisible to us, so they were cut out from the training and testing as other faint stars. So what's left is roughly 6000 stars. However, it is possible to cut it even more because some stars would be too far away from the cluster after a certain timeframe, but the cutting method will be explained in the next sections.

For learning and testing, features should be accessible and obtainable by observations, requiring a choice of specific features. Those features are the 2D galactocentric coordinates, velocities in the respective direction, color index, and apparent magnitude. The main reason is that those features are mostly available in various catalogs, and learning about them can be a viable choice. This allowed us to exclude background stars from testing and training because background stars do not affect data much when looked at from the 150 pc higher galactic plane. We used 2D coordinates because we tried to simulate a situation where we looked at the cluster from above the galactic plain along 150 parsecs higher Z axis. Also, the Galaxy has its other stars, which couldn't be stars of the cluster; it's called field stars. However, looking from the top side allowed us to exclude field stars from the consideration because they do not have a considerable effect from this perspective. The color index $(B-V)$ and apparent magnitude (m) are additional and secondary features. m is not in simulations

and instead can be calculated by the following formula:

$$m_v = M_v + 5(lgd - 1) \tag{2.1}$$

where $d = 150 - Z$. The exploratory data analysis of the features can be seen in 2-1, which shows the features' relationship and histogram of the features (diagonals). From this, we can see that by looking at all the features with respect to each other, we can see whether a star is a member and a non-member because both members and non-members are distinguishable. However, one ambiguous relationship is with $B-V$ and m because members and non-members from their side do not seem distinguishable enough. Nonetheless, upon closer look, we can see that members and non-members are distinguishable in 2-2.

Further, discussion about the data would continue in a sense of how ML model was trained and tested on given data.

2.2 Learning strategy and datasets

Selection and general strategy for training the model on star cluster simulations were the most crucial and the most challenging part of the study. Due to the nature of the open star cluster that tends to start dissolving in the early stages of the evolution and considering overall similarity on small time-frames and certain stages of cluster evolution, it is generally hard to decide on which snapshot model should be trained. Additionally, there would be a problem of imbalance when the cluster starts dissolving and having fewer members and more non-members.

The imbalance problem was solved by cutting out the stars that are 3 Jacobi radius away from the cluster so that they would not affect the learning process keeping the dataset fairly balanced. This ensured good training and testing because stars away from the cluster are easily classified as non-members because of their considerable distance from the cluster's actual domain. In 2-3, you can see the count-plots of both cut and full clusters for comparison. This allowed to development of new

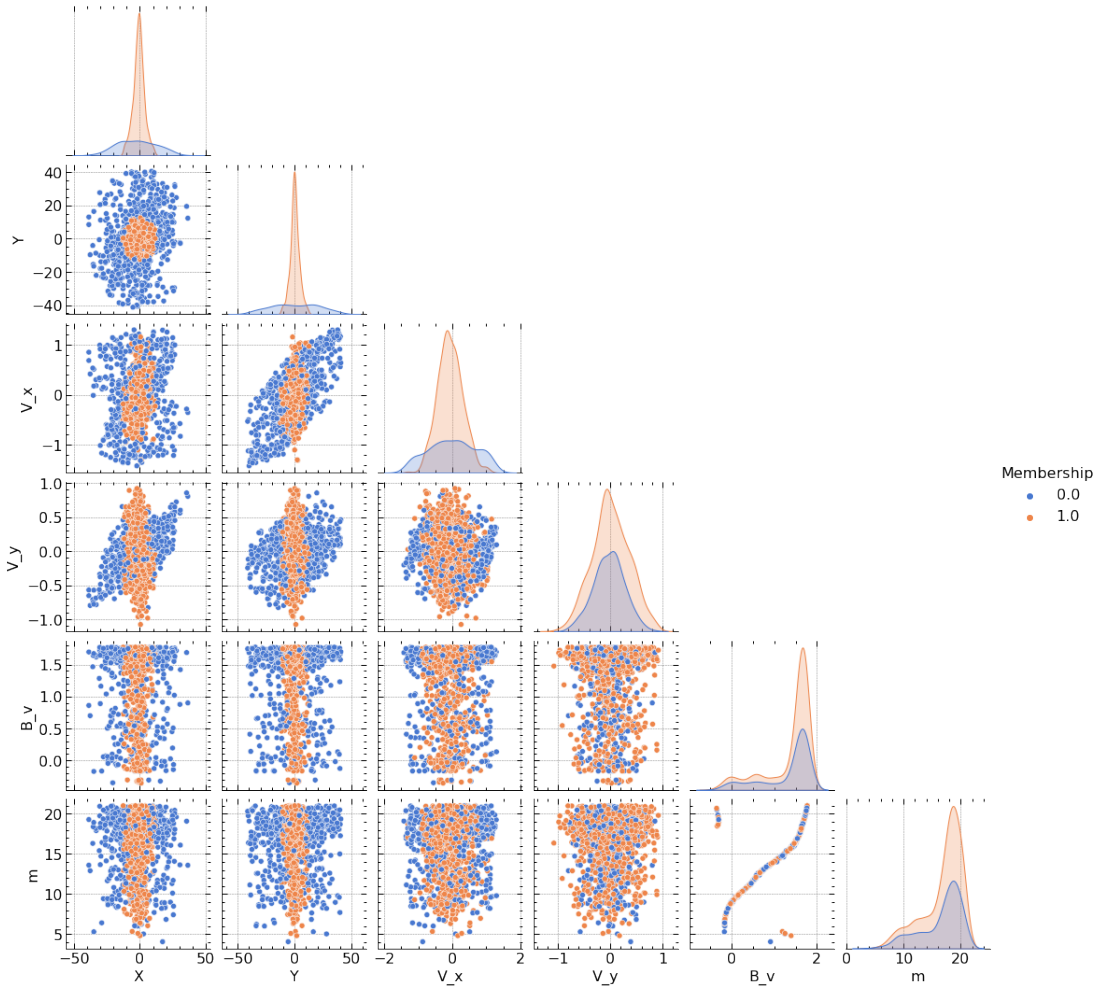

Figure 2-1: Exploratory data analysis of OSC simulation with 15% SFE at 100 Myr. Diagonal plots are histograms, and other plots are features with respect to each other

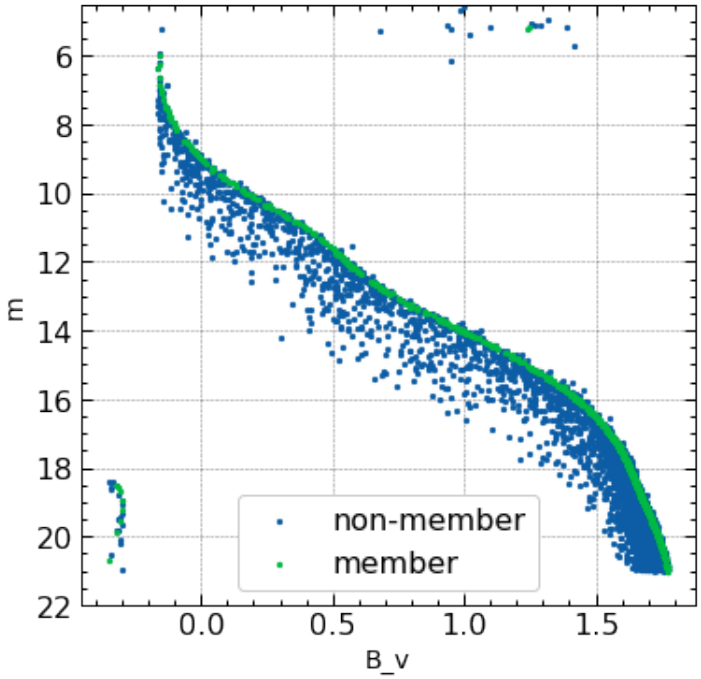

Figure 2-2: Color/magnitude map of OSC simulation with 15% SFE at 100 Myr

learning strategies that used further timeframes for training. This further decreases the number of stars in datasets, but this decrease is not fixed and would be different for all the timeframes.

The main learning strategy was to train on timeframe after violent relaxation because the cluster at that stage reaches equilibrium and would save the dynamic that would continue throughout the whole life-cycle of the cluster. This would cover the period from 20 Myr to 100 Myr, but only randomly chosen 20 time-frame datasets would be used for the training to avoid overtraining. If the ML model trains too much on the dataset, it will remember and fail when encountering a previously unseen test sample. This training strategy can be used for simulation with 15-17% SFE and the 1st type position and 1st type mass. Further types would be given as simple numbers after SFE number and "-" like "15-11". However, in this paper, there would be only the model trained on 15-11 because the model trained on both 15-11 and 17-11 were similar in performance and were excluded for easier readability of the paper.

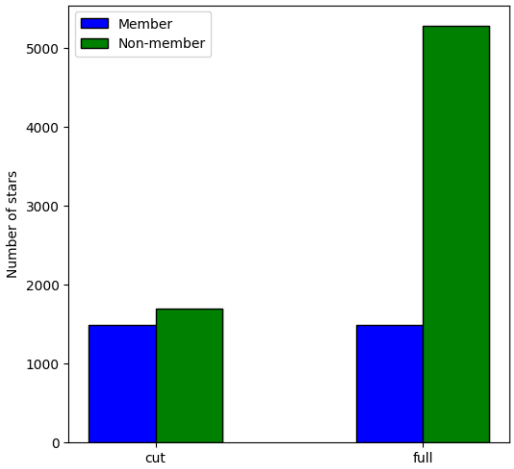

Figure 2-3: SFE=15% Cluster at 20 Myr, cut from 3 Jacobi radius and without cutting.

2.3 Testing method

As for the testing, the trained model was used to predict membership of OSC simulations with different SFEs, mass, and positions from 20 Myr (end of violent relaxation) until the time-step when the cluster would have a mass less than 100 solar Jacobi mass because it would not be considered as cluster anymore. Despite the absence of a mass feature in learning and testing, the state of the cluster with 100 solar mass can be calculated by the corresponding Jacobi radius formula:

$$R_J = \frac{GM_J}{(4-\beta^2)\omega^2} \tag{2.2}$$

where G is gravitational constant, M_J is Jacobi mass, β is normalized epicyclic frequency and ω is angular speed of star cluster [19]. Another indication on when the testing should stop is the number of member stars and OSC that have less than 50 stars would not be considered as OSC.

Almost all the available simulations would be tested and divided into 2 groups: simulations that have either the same position or mass and simulations without similarities. This can be seen in

Testing would be done on datasets already cut to a 3 Jacobi radius to reduce the redundant stars that are not members anymore.

SFE - 17, 20, 25

11	12	13
21	22	23
31	32	33

Table 2.3: Types of simulations used for testing. Gray is the simulation that was not tested, but used for training. Pink is the simulations that are similar to training datasets either by position or mass. Violet are the simulations that are completely different from learning datasets by position and mass. All this is for 3 SFEs each having 4 similar and 4 different simulations. The total tested is 24 simulations.

2.4 Machine learning algorithm

Multiple ML methods were tested for the classification of OSC simulations, but only 2 were the most successful in this: Random Forest (RF) and Feed-Forward Neural Networks. However, exactly in this study there would be presented only RF algorithms because performance compared to neural networks is similar, but neural networks are computationally more complex and thus considered less desirable for this study. The complexity mostly comes from the hardware capabilities available right now which do not really support CUDA as it should.

The random forest algorithm is the decision tree algorithm used for various high-dimensional data with a small number of samples. It works by bootstrapping (randomly creating new datasets from input datasets) and building a predictive tree. After that, each sample either for training or testing goes through these trees and the average output of those trees would be selected by the algorithm [20]. Despite being an ML model for datasets with multiple features, it proved to be very good at predicting membership of the OSCs [12], [14], [13]. Thus, testing and application of this algorithm were in high priority of the paper.

In the scikit-learn library for ML in Python, there are a lot of hyperparameters of RF classifier and major hyperparameters are number of trees, max tree depth, max number of features [21]. Number of trees or estimators depend on the dataset and larger number of trees can lead to more accurate classification. Max depth accounts for how many depth levels should the tree have when running the algorithm, making

the trees deeper affects the computational complexity of the learning and prediction. Max number of features are the number of features used for the node split of the tree and it can be either square root type or $\log_2$ type. As for the criterion, Gini index were used for measure the impurity in the values of datasets. Also, it is possible to use random states for the learning to reproduce the previous learning results. In this study, the random forest had 100 trees, a max depth of 10, square root type of max features (6 features $\approx$ 2.49), and a random state of 42. It was possible to use higher number of trees that would have also been deeper, but it was performing good as it is and there was no real need in increasing the complexity of both learning and testing which would have lead to overfitting. Other hyperparametes were left as default.

2.5 Evaluation metrics

The main evaluation metric is the accuracy of the prediction of the above machine learning models on other N-body simulations. It will show a percentage of how well the prediction was done in a broad sense.

Despite having general metrics, for binary classification to have a clearer picture of performance we need an analysis of True Negatives (TN) and True Positives (TP). The performance of the binary classification can be summarized with the so-called confusion matrix shown in 2.4. Diagonal represents the number of correctly classified samples and other values represent the number of samples classified incorrectly either as False Negative(FN) or False Positive(FP) [22]. Considering a significant proportion of the testing samples would be either TP or TN, it is wise to consider only the rate of FN and FP because the proportion of this metric would actually show how well the model performs. Thus, in this study, 2 metrics of FP and FN were used called False Positive Rate (FPR) and False Negative Rate (FNR). It is simply the rate of how many real positives or negatives are classified wrongly on the test. Formulas can be seen below:

$$FPR = \frac{FP}{TN + FP} \tag{2.3}$$

		Prediction	
		Non-Member	Member
Actual value	Non-Member	TN	FP
	Member	FN	TP

Table 2.4: Confusion matrix representing performance of the prediction compared to the actual value.

$$FNR = \frac{FN}{TP + FN} \tag{2.4}$$

Chapter 3

Results and discussions

In this chapter, I would like to present the results of this work and show the performance of models in accordance with evaluation metrics and further analysis of results.

3.1 Performance based on evaluation metrics

All the test was done on 2 groups of simulations for all available SFEs. All plots are given with the logarithmic scale because for each Myr until 175 Myr has 6 datasets per 1 Myr. This is the reason why the plot is thicker until 200 Myr.

The main evaluation metric is the accuracy or precision of the classification. Here you might notice several trends in all the tests for all SFEs. Firstly, it is evident that performance is very high that exceeds 90% which is an indication of very high performance. However, this kind of trend is not universal in all timeframes of the classification. Mostly it is due to the fact that OSC dissolves and as time goes there is less and less number of testing samples.

Classification on simulations with 17% SFE stays high on performance, but the actual drop happens on different timeframes on different simulations. This is because OSCs with lesser SFE tend to be less stable and they dissolve faster. Those simulations are the SFE 17-13 and 17-33. These random realizations have similar mass and they start dissolving at a similar time. Other simulations were more stable and the

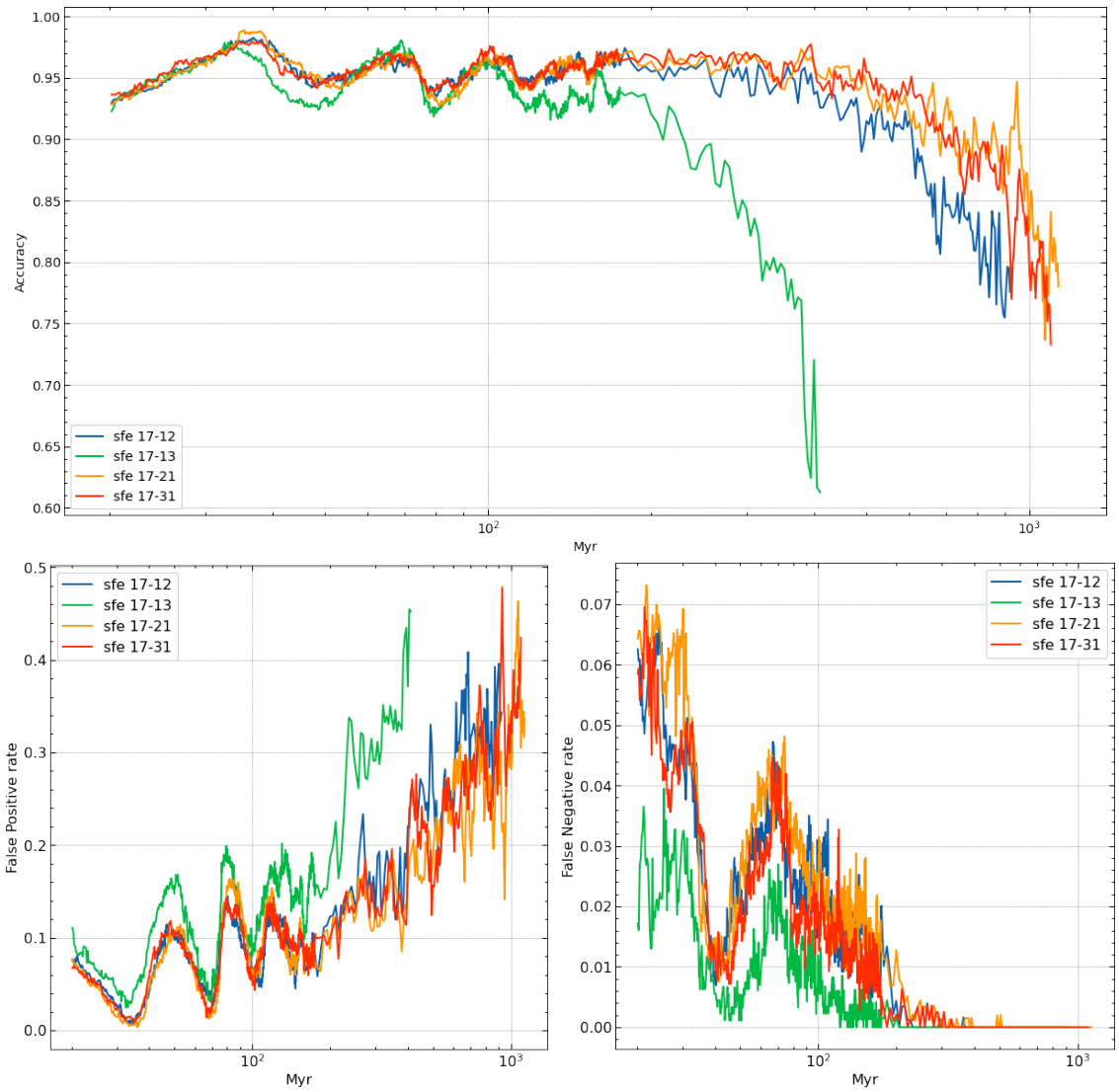

Figure 3-1: Performance evaluation of RF model on simulations with 17% SFE on violet models from Table 2.3 in terms of Accuracy (top panel), False Positive Rate (bottom left panel), and False Negative Rate (bottom right panel) throughout the dynamic evolution.

actual drop of performance less than 90% happened after 600 Myrs. However, even until the end they kept high performance closer to 80%. These results can be seen on the top panel of Figures 3-1 and 3-2.

However, this kind of analysis would not be full without analyzing what kind of mistakes were made by the ML model. Due to the fact that it is a binary classification, there can be only 2 types of mistakes FP and FN (please look at Table 2.4). In the case of simulations with 17% SFE in the early timeframes rate of FN stays highest and

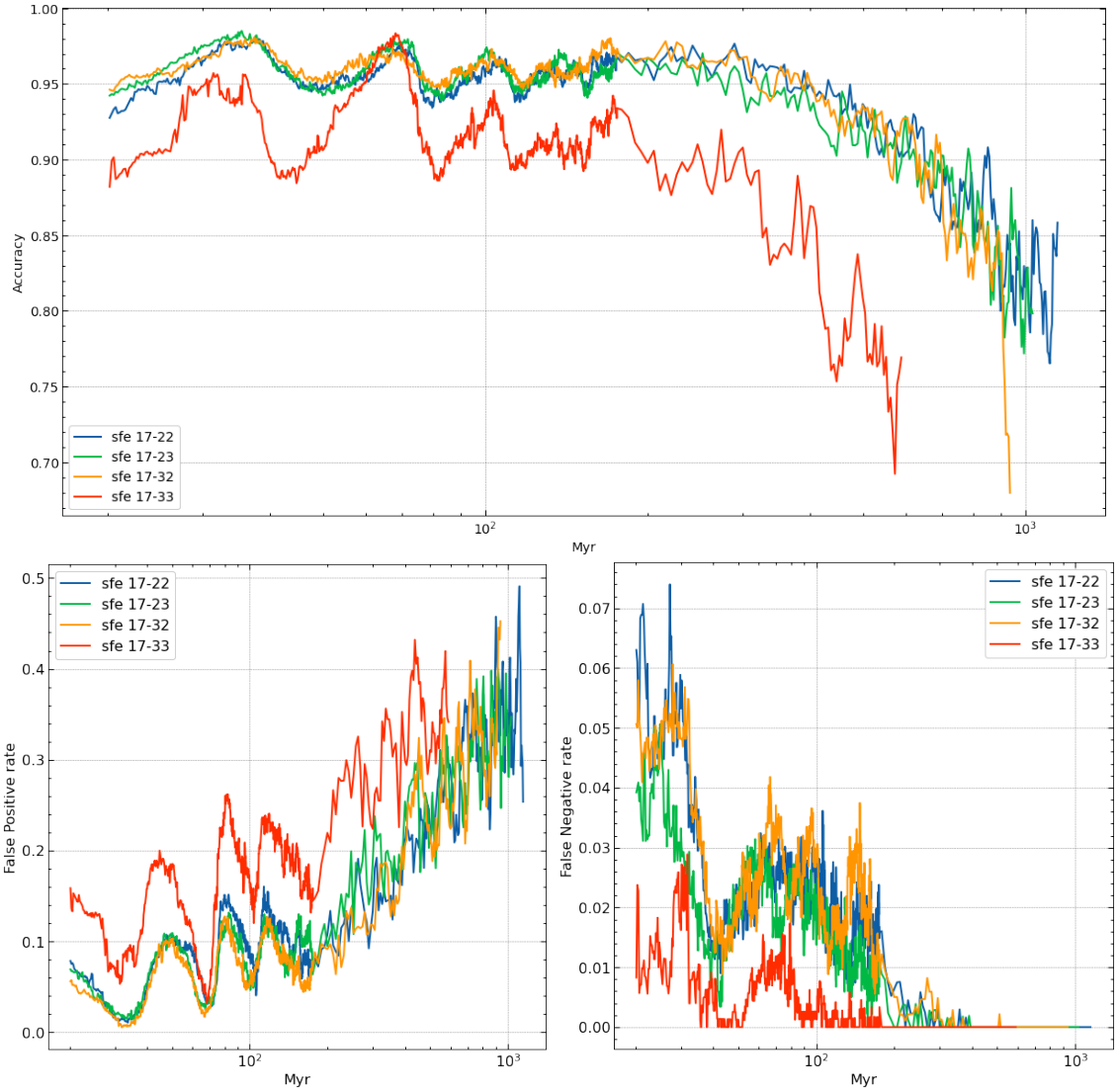

Figure 3-2: Performance evaluation of RF model on simulations with 17% SFE on violet models from Table 2.3 in terms of Accuracy (top panel), False Positive Rate (bottom left panel), and False Negative Rate (bottom right panel) throughout the dynamic evolution.

over time decreases eventually reaching 0. This means that model makes a marginal number of mistakes with stars that are actually members of the cluster and we mostly do not lose member stars during the classification. On the other amount of FP stays low at 20 Myr, but on further timeframes, it only increases and almost reaches 50% at the end. This means that the model mistake 0 to 50% of the non-members as members. Considering that it's the cluster that was cut to a 3 Jacobi radius, this is expected result.

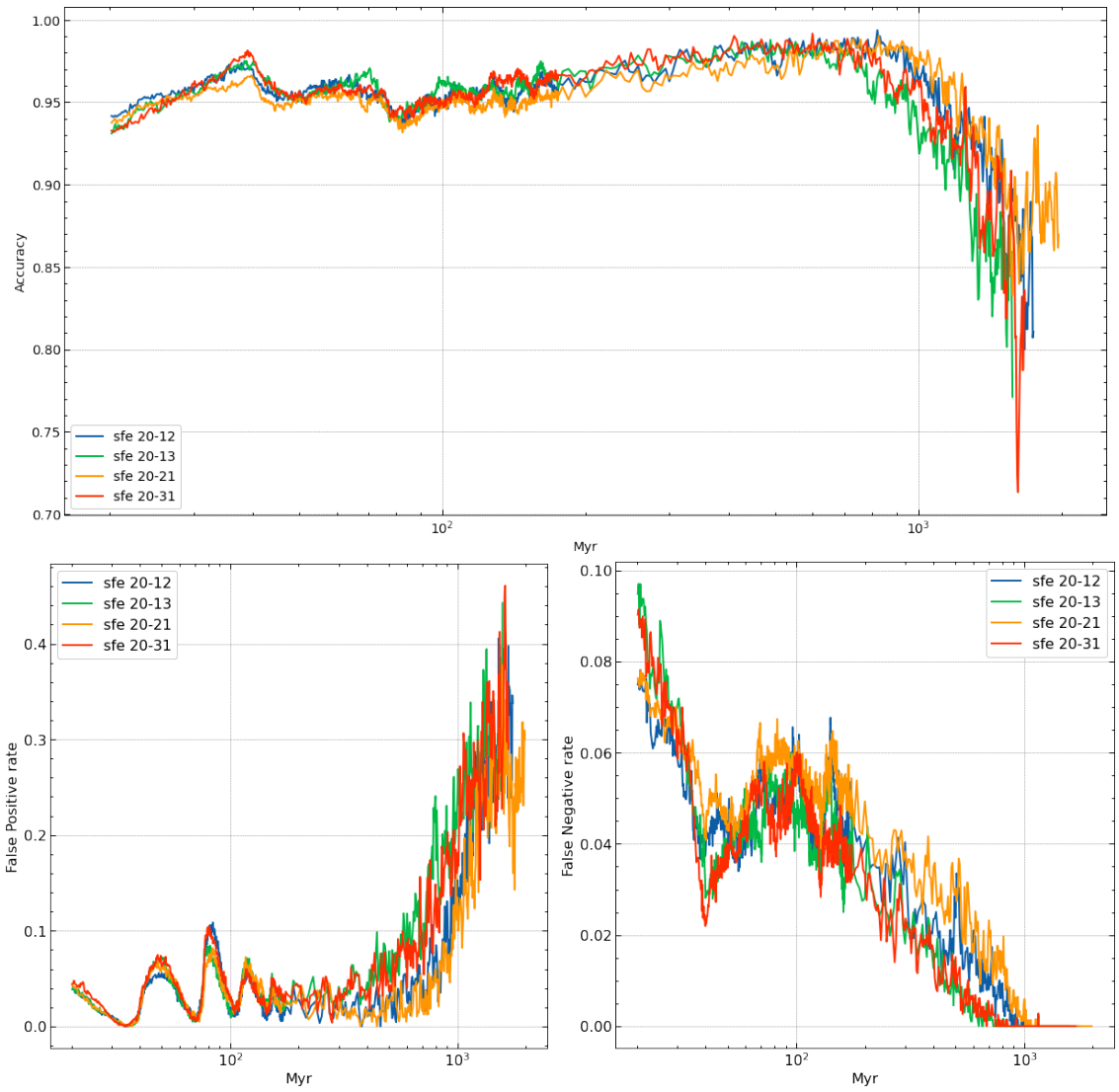

Figure 3-3: Performance evaluation of RF model on simulations with 20% SFE on pink models from Table 2.3 in terms of Accuracy (top panel), False Positive Rate (bottom left panel), and False Negative Rate (bottom right panel) throughout the dynamic evolution.

If we look at the results of the membership identification on simulations with 20% SFE, we can see that they perform similarly, but because of the higher stability of the OSC with 20% SFE, it shows higher results that you can see in Figures 3-3 and 3-4. The main difference is that accuracy doesn't drop after 600 Myr as in tests on simulations with 17% SFE. The decline of testing on these simulations mostly happens at the timeframe closer to 1 Gyr. Another interesting feature is that unlike on testing with 17% SFe rate of FP stays low for a longer period of testing, but still

increases closer to the OSC dissolution. However, the rate of FP is still lower. This shows that it works very well on the simulations with an SFE of 20%.

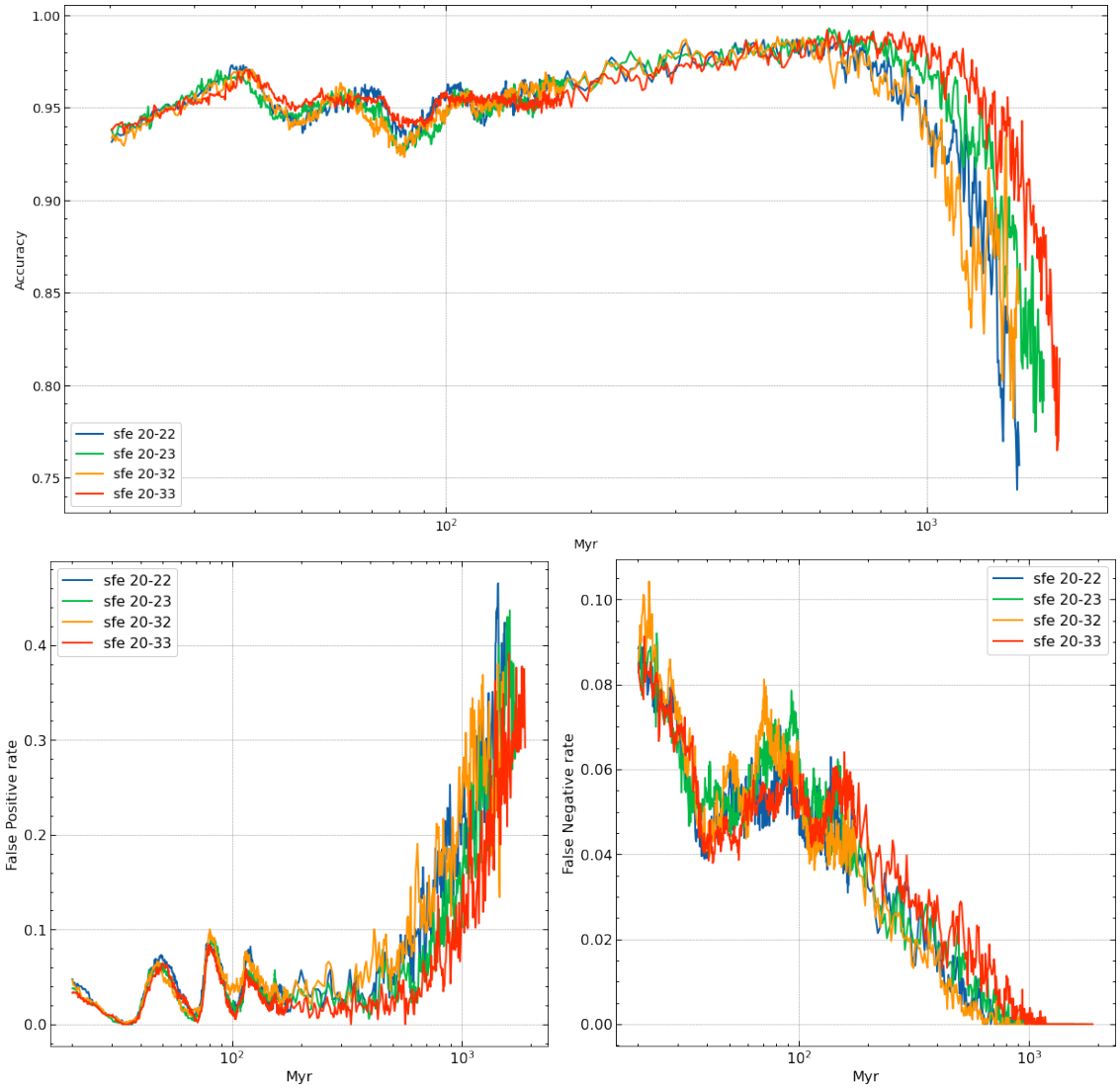

Figure 3-4: Performance evaluation of RF model on simulations with 20% SFE on violet models from Table 2.3 in terms of Accuracy (top panel), False Positive Rate (bottom left panel), and False Negative Rate (bottom right panel) throughout the dynamic evolution.

Last, but not least the tests on simulations with 25% SFE show very similar results to the tests on simulations with 20% SFE. The main difference is that unlike on 20% simulations, performance stays higher than 90% for a much longer period of time until 2.5 Gyr. Also, performance does not significantly drop after this period. Throughout the classification rate of FPs stays low on a level below 5% until reaching

2 Gyr where it becomes 20% and escalated to almost 50% at the end of the lifespan of the cluster. The rate of FNs starts from 8% and drops as time passes on reaching 0 at 1.5 Gyr.

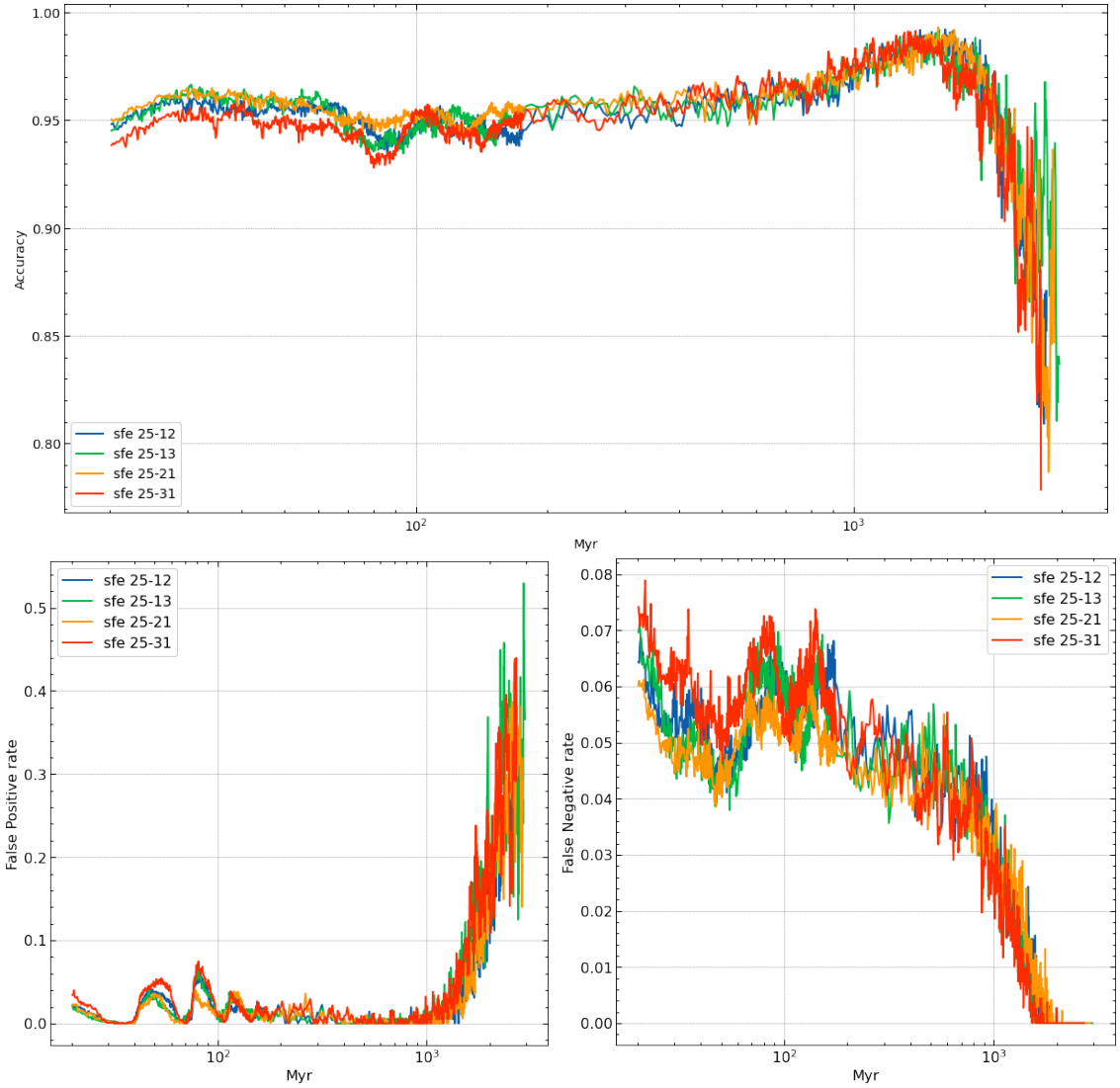

Figure 3-5: Performance evaluation of RF model on simulations with 25% SFE on pink models from Table 2.3 in terms of Accuracy (top panel), False Positive Rate (bottom left panel), and False Negative Rate (bottom right panel) throughout the dynamic evolution.

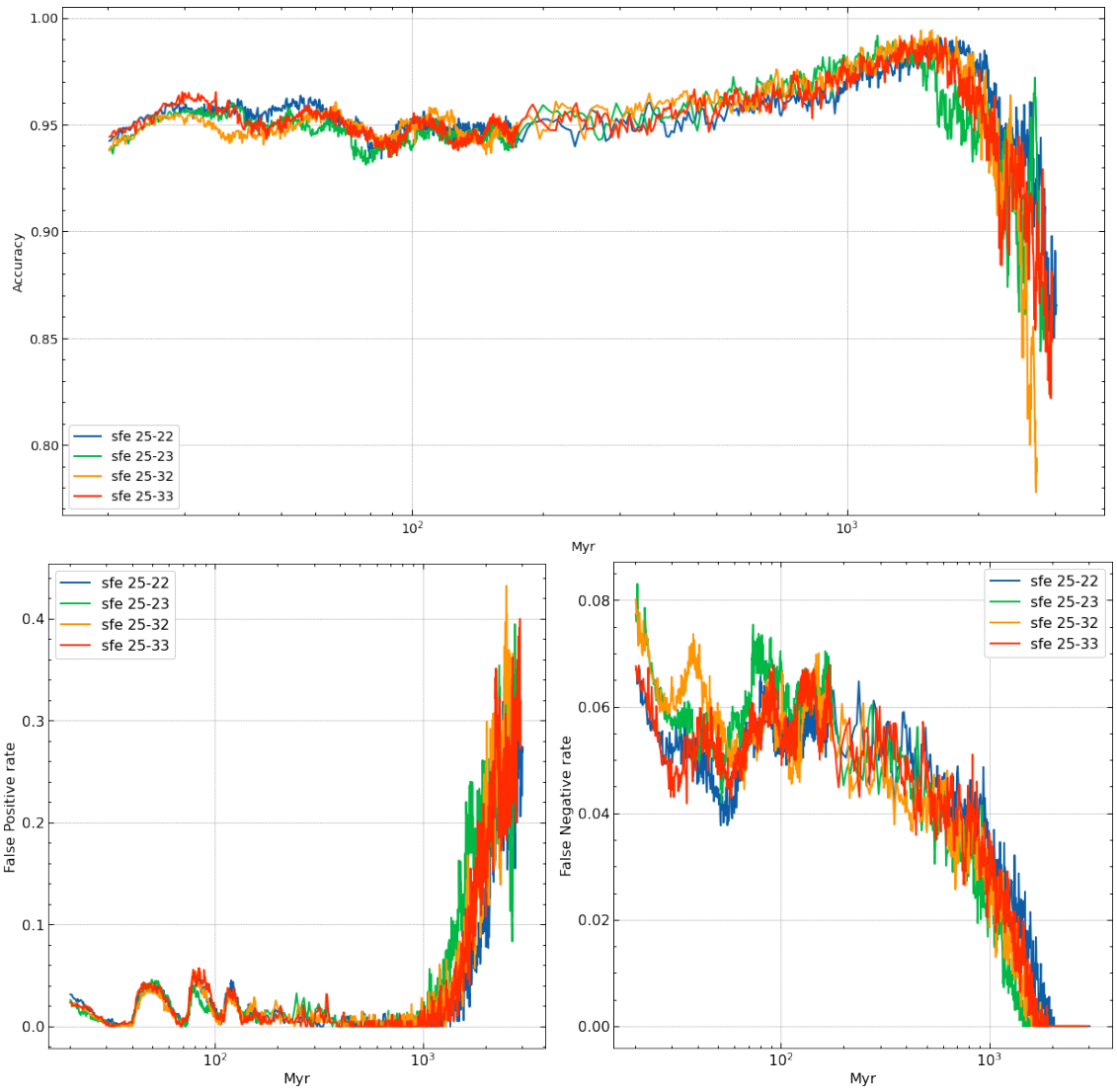

Figure 3-6: Performance evaluation of RF model on simulations with 25% SFE on violet models from Table 2.3 in terms of Accuracy (top panel), False Positive Rate (bottom left panel), and False Negative Rate (bottom right panel) throughout the dynamic evolution.

3.2 Analysis of errors

All of these show that performance overall is very good reaching and exceeding 90% accuracy, but still, there is a considerable amount of FPs throughout the whole testing. In order to understand these mistakes we need to look at how stars were classified from a 2D perspective.

In simulations violent relaxation end at 20 Myr and this is the starting period

of testing. In Figure 3-7, you can see how cluster members and non-members were classified as well as with the indication of whether they are TP, TN and etc. Generally at this stage classification accuracy was around 92.7% and there were 157 FPs and 83 FNs. From the picture you can see that shape of the cluster was generally captured and it indeed classified correctly both members and non-members for the most part. This is evident from the circular shape of the identified cluster and the location of the TPs and TNs. However, you may also see that all the FP and FN are located at the 2D plane or on the borders of the cluster. This means that mistakes mostly happen because of the absence of the Z plane because the majority of the FPs are located upper or lower to the X and Y plane on the Z axis. FN is located on borders meaning that they are marginal statistical errors of ML model. Despite these errors, these are very good results for young OSC with 17% SFE and 20 Myrs.

Similar results can be seen with the same cluster after 80 Myrs later in figure 3-8, but the accuracy is higher and reached 95%. The numbers of FP and FN are 39 and 33 respectively. Considering the lower number of non-members overall this can be considered as a very good result. Placements of the FPs and FNs are similar, but now mistakes are considerably lower than before.

Now let's look at the further timeframe of 500 Myrs that you can see in figure 3-9. The accuracy at that timeframe is 93% and there are no FNs and 42 FPs. FPs are both on borders and on the Z plane higher or lower than the cluster. Overall number of stars is decreasing and this kind of no FN will continue further for the whole dynamic evolution.

The performance further decreases, but it's mostly because of decreasing number of stars in the cluster and does not mean that model works worse than before. For example, you can see an almost dissolved cluster in Figure 3-10, which contained only 118 stars at the 1.1 Gyr, and only 27 of them were misclassified as non-members or FPs, which gives an accuracy of 77%.

From the analysis of the layout of the stars, we identified that most of the mistakes happen to stars on cluster borders and on stars that are higher or lower than the Z plane. Most mistakes are the FPs, and FNs are mostly statistical mistakes that are

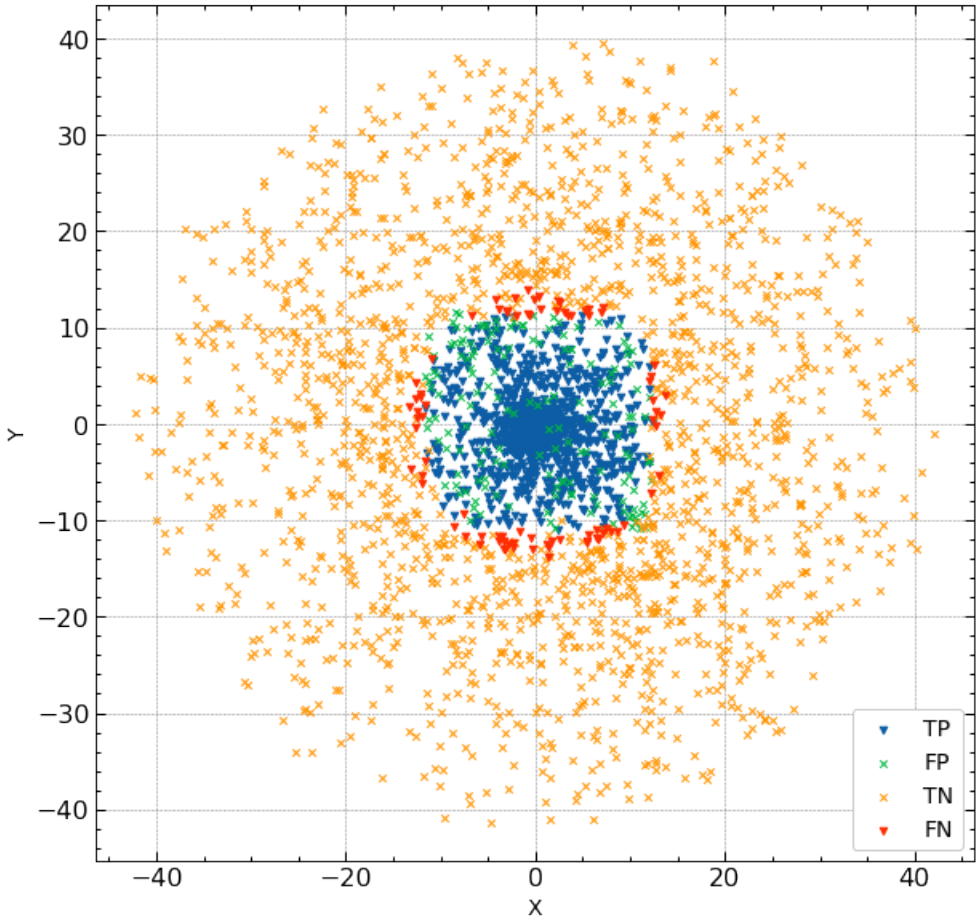

Figure 3-7: Classification result on 17% SFE cluster at 20 Myrs. Crosses are the negatives that are either true or false. Triangles are the positives both true or false.

not present in the further testing timeframe. This raises the question of whether this number of FPs is acceptable for the membership analysis. To explore this, we need to know how far these stars are from the cluster's center. The cluster border is found by the Jacobi radius, and FPs are not within the Jacobi radius. However, they might be within a 2 or 3-Jacobi radius, and at 2-Jacobi radius distance, stars might still be gravitationally bounded to the center of the cluster. Thus, there is a need to look at the numbers of FP stars that are within a 2 or 3-Jacobi radius. This can be seen in figures 3-11, 3-12, and 3-13.

From these results, we can see that the majority of the FPs were within 2 Jacobi radius in all the classifications with different SFEs, and it's possible that they are still gravitationally bound to the center of the cluster.

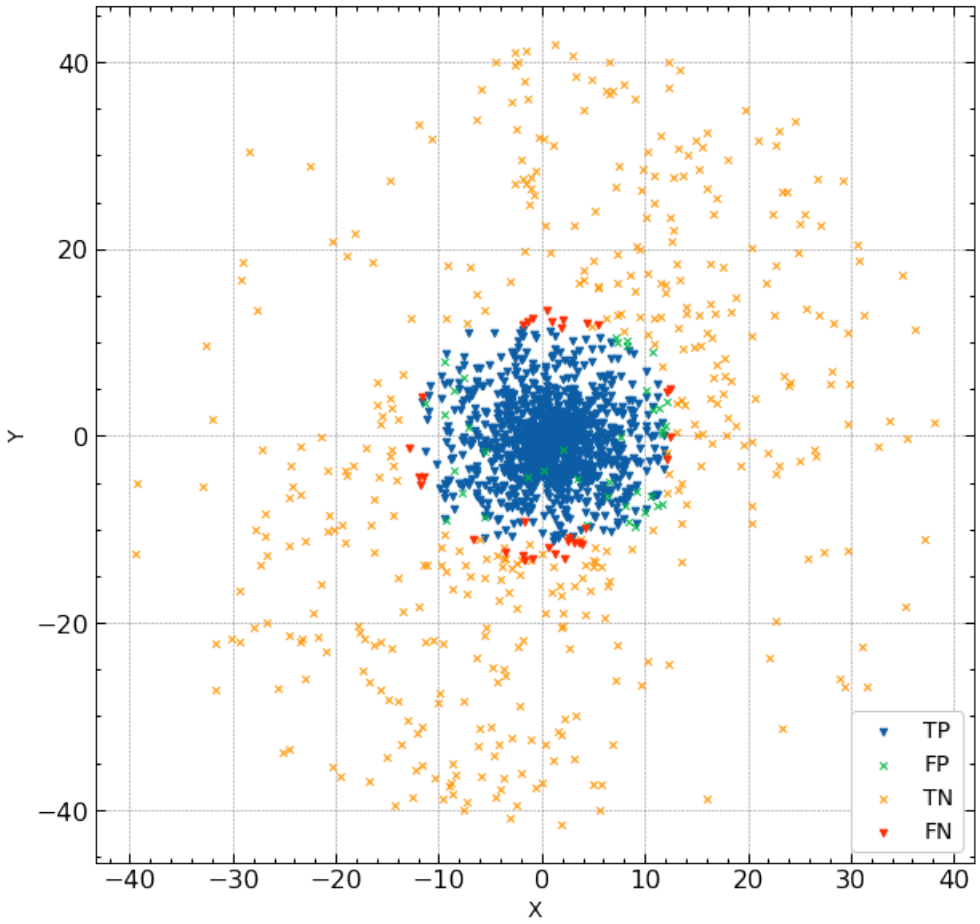

Figure 3-8: Classification result on 17% SFE cluster at 100 Myrs. Crosses are the negatives either true or false. Triangles are the positives

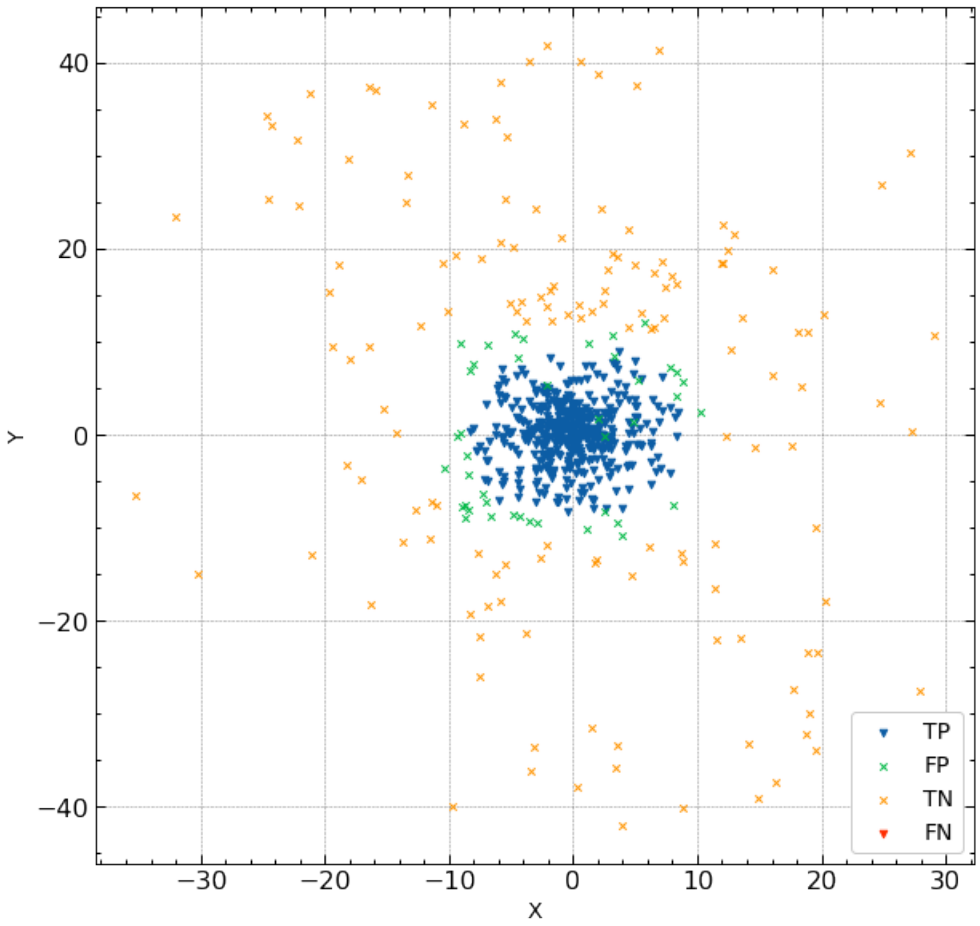

Figure 3-9: Classification result on 17% SFE cluster at 500 Myrs

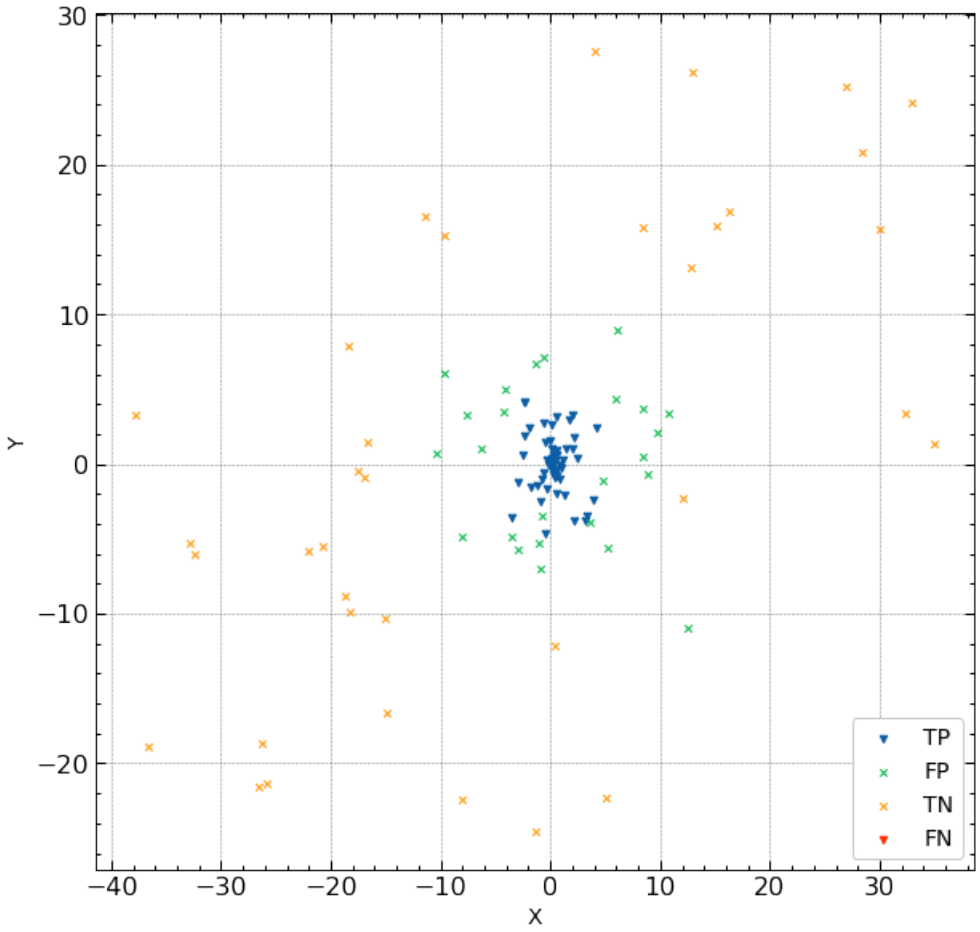

Figure 3-10: Classification result on 17% SFE cluster at 1 Gyrs

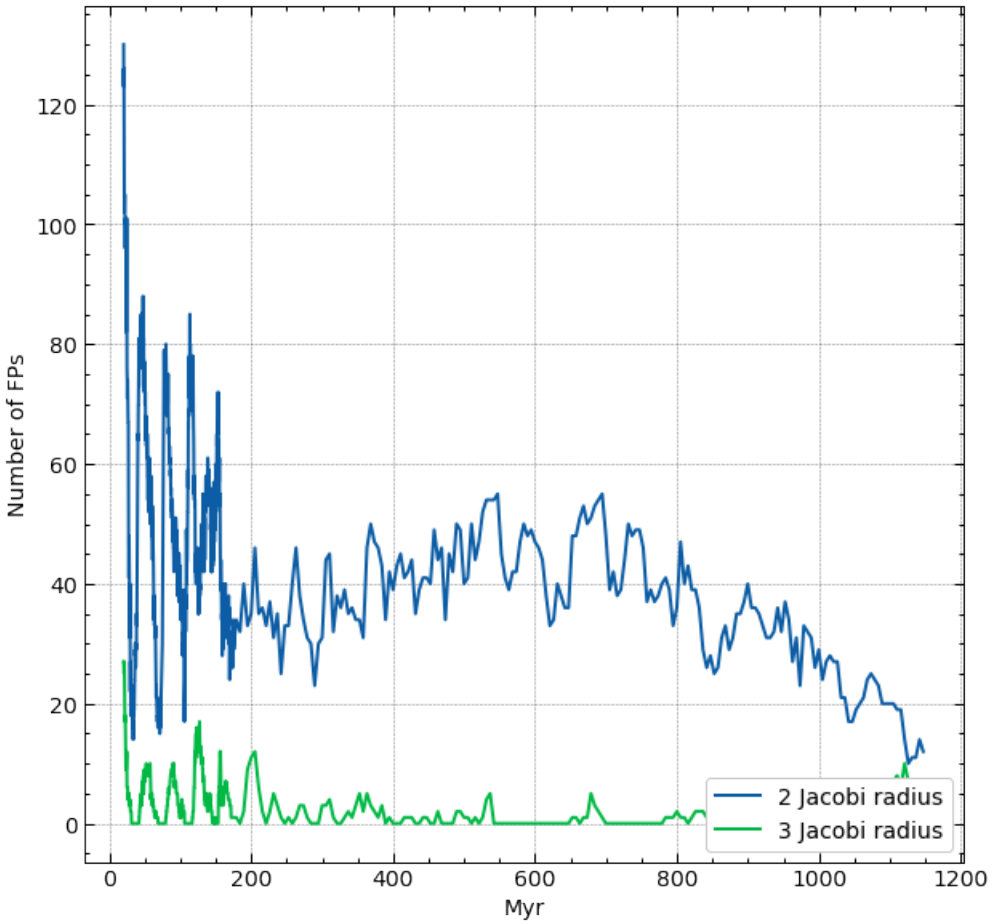

Figure 3-11: Number of FPs throughout the test that was within 2 Jacobi radius (blue) and between 2 and 3 Jacobi radius (green) on simulation with 17% SFE and 22 random realizations. No FPs beyond these distances.

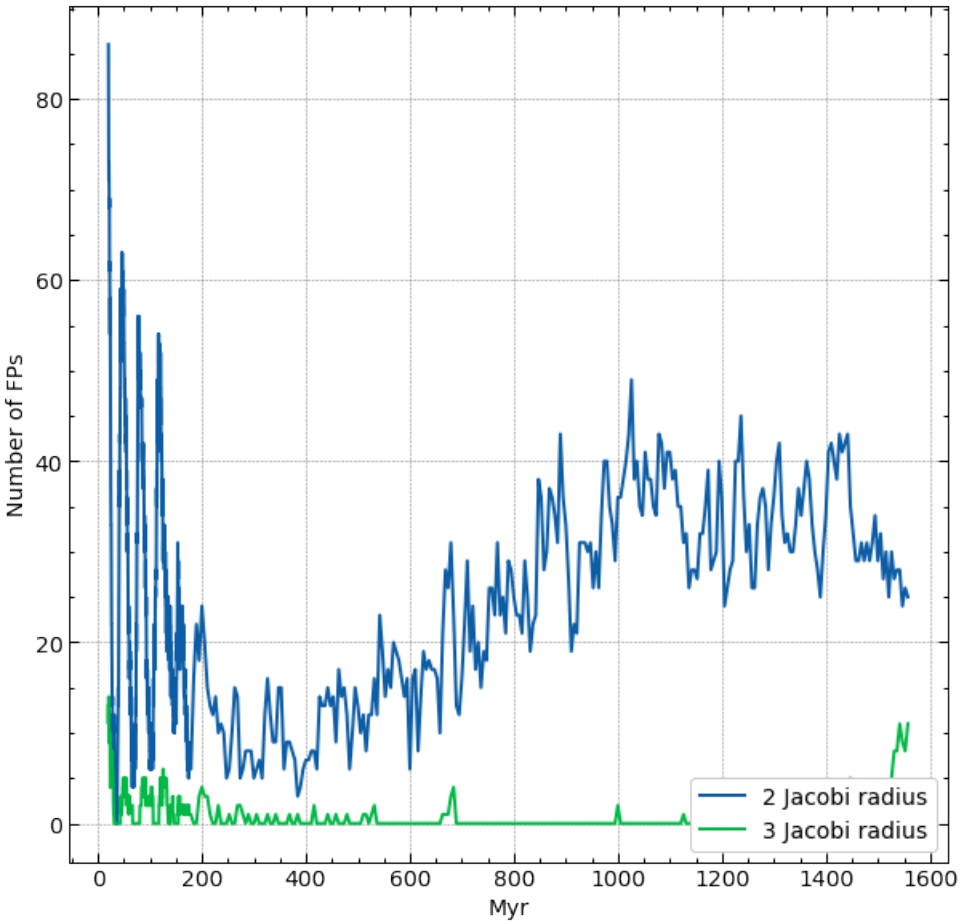

Figure 3-12: Number of FPs throughout the test that were within 2 Jacobi radius (blue) and between 2 and 3 Jacobi radius (green) on simulation with 20% SFE and 22 random realizations. No FPs beyond these distances

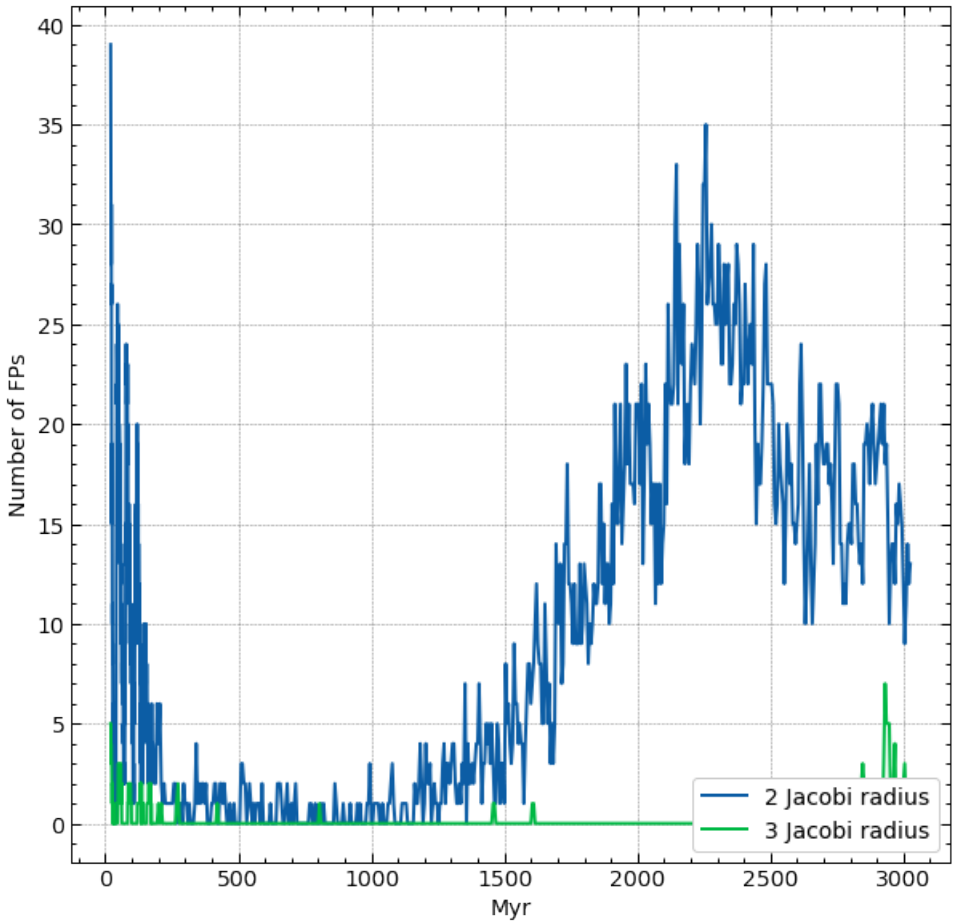

Figure 3-13: Number of FPs throughout the test that was within 2 Jacobi radius (blue) and between 2 and 3 Jacobi radius (green) on simulation with 25% SFE and 22 random realizations. No FPs beyond these distances.

Chapter 4

Conclusion

This work explored the possibilities of using ML on the N-body simulation for the membership analysis. The primary ML algorithm was the RF, and the model was trained on simulation with 15% SFE on the timeframe of 20-100 Myr (after violent relaxation) with a semi-random approach. Results were tested on the N-body simulations with different SFE and random realizations. In result, the model was able to predict the membership of all N-body simulations at the timeframes of 20 Myrs to their respective dissolution. Despite having high accuracies, it was identified that the model tends to mistake non-member stars as members that were called FPs, the amount of FPs only increased due to a lesser number of stars in testing and closer to the complete dissolution. However, it was found out that the majority of the FPs were located inside the 2 Jacobi radius, which might indicate that those stars are still gravitationally bound to the center of the cluster and may still be considered member stars in some sense. This framework shows that it is possible to train an ML model that could predict membership of similar and completely different N-body simulations with 6 easily obtainable features as 2D coordinates, their velocities, color index, and apparent magnitude. Nonetheless, this was just a framework, and conditions on which the approach was tested here cannot be made on Earth for a similar observation.

As for the future plans, it was a test of concept and strategy that can be considered effective, and the near future plan is to apply this strategy to the simulations that were

converted to mock observed simulations. Current simulations are in galactocentric coordinates, and mock observation versions of the simulations would have features in a format that can be seen in real observations. Also, an observer would be placed on the plane of the galaxy. This would make classification both harder and easier because, in this case, background stars would play a considerable role but would make stars more distinguishable by their apparent magnitude and color index (in current tests, stars are too far from the observer that both apparent magnitude and color index of stars looks similar). It is also planned to test out the unsupervised learning methods, such as various density-based scans and StarGO on N-body simulations presented here.

Bibliography

[1] Piero Madau and Mark Dickinson. Cosmic star-formation history. *Annual Review of Astronomy and Astrophysics*, 52:415–486, 2014.

[2] Robert C Kennicutt Jr and Neal J Evans. Star formation in the milky way and nearby galaxies. *Annual Review of Astronomy and Astrophysics*, 50:531–608, 2012.

[3] Mark R Krumholz, Christopher F McKee, and Joss Bland-Hawthorn. Star clusters across cosmic time. *Annual Review of Astronomy and Astrophysics*, 57:227–303, 2019.

[4] Rupali Chandar, S Michael Fall, and Bradley C Whitmore. New tests for disruption mechanisms of star clusters: the large and small magellanic clouds. *The Astrophysical Journal*, 711(2):1263, 2010.

[5] Roland Wielen. Dynamics of open star clusters. In *Symposium-International astronomical union*, volume 113, pages 449–462. Cambridge University Press, 1985.

[6] Jeroen Bédorf, Evghenii Gaburov, and Simon Portegies Zwart. A sparse octree gravitational n-body code that runs entirely on the gpu processor. *Journal of Computational Physics*, 231(7):2825–2839, 2012.

[7] Xiaoying Pang, Yuqian Li, Shih-Yun Tang, Mario Pasquato, and MBN Kouwenhoven. Different fates of young star clusters after gas expulsion. *The Astrophysical Journal Letters*, 900(1):L4, 2020.

[8] Yu Zhang, Shih-Yun Tang, WP Chen, Xiaoying Pang, and JZ Liu. Diagnosing the stellar population and tidal structure of the blanco 1 star cluster. *The Astrophysical Journal*, 889(2):99, 2020.

[9] Tristan Cantat-Gaudin, C Jordi, Antonella Vallenari, Angela Bragaglia, L Balaguer-Núñez, C Soubiran, D Bossini, A Moitinho, A Castro-Ginard, A Krone-Martins, et al. A gaia dr2 view of the open cluster population in the milky way. *Astronomy & Astrophysics*, 618:A93, 2018.

[10] T Cantat-Gaudin and F Anders. Clusters and mirages: cataloguing stellar aggregates in the milky way. *Astronomy & Astrophysics*, 633:A99, 2020.

[11] Esan Mouli Ghosh, Princess Tucio, Muhammad Fajrin, et al. Membership and age determination of m67 open cluster using gaia edr3 data. In *Journal of Physics: Conference Series*, volume 2214, page 012009. IOP Publishing, 2022.

[12] Xinhua Gao. A machine-learning-based investigation of the open cluster m67. *The Astrophysical Journal*, 869(1):9, 2018.

[13] Xin-hua Gao. An investigation of the pleiades cluster using machine learning. *Publications of the Astronomical Society of the Pacific*, 131(998):044101, 2019.

[14] Xin-Hua Gao. Memberships, distance and proper-motion of the open cluster ngc 188 based on a machine learning method. *Astrophysics and Space Science*, 363:1–8, 2018.

[15] Manan Agarwal, Khushboo K Rao, Kaushar Vaidya, and Souradeep Bhattacharya. Ml-moc: machine learning (knn and gmm) based membership determination for open clusters. *Monthly Notices of the Royal Astronomical Society*, 502(2):2582–2599, 2021.

[16] Marina Kounkel and Kevin Covey. Untangling the galaxy. i. local structure and star formation history of the milky way. *The Astronomical Journal*, 158(3):122, 2019.

[17] Stefan Meingast, João Alves, and Alena Rottensteiner. Extended stellar systems in the solar neighborhood-v. discovery of coronae of nearby star clusters. *Astronomy & Astrophysics*, 645:A84, 2021.

[18] LG Bouma, JL Curtis, JD Hartman, JN Winn, and GÁ Bakos. Rotation and lithium confirmation of a 500 pc halo for the open cluster ngc 2516. *The Astronomical Journal*, 162(5):197, 2021.

[19] Bekdaulet Shukirgaliyev. *The life of star clusters, from birth to dissolution: a new approach*. PhD thesis, 2018.

[20] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[22] Matthias Kohl. Performance measures in binary classification. *International Journal of Statistics in Medical Research*, 1(1):79, 2012.