Mass and Age Determination of the LAMOST Data with Different Machine-learning Methods

Qi-Da Li, Hai-Feng Wang, Yang-Ping Luo, Qing Li, Li-Cai Deng, Yuan-Sen Ting

The Astrophysical Journal Supplement Series, 262:20 (17pp), 2022 September https://doi.org/10.3847/1538-4365/ac81be

Mass and Age Determination of the LAMOST Data with Different Machine-learning Methods

Qi-Da Li[1], Hai-Feng Wang[1,2,3,5], Yang-Ping Luo[1], Qing Li[1], Li-Cai Deng[1], and Yuan-Sen Ting[4]
[1] Department of Astronomy, China West Normal University, Nanchong, 637002, People's Republic of China; hfwang@bao.ac.cn
[2] GEPI, Observatoire de Paris, Université PSL, CNRS, Place Jules Janssen, F-92195, Meudon, France
[3] CREF, Centro Ricerche Enrico Fermi, Via Panisperna 89A, I-00184, Roma, Italy
[4] Research School of Computer Science, Australian National University, Acton, ACT 2601, Australia
Received 2022 January 27; revised 2022 July 15; accepted 2022 July 15; published 2022 August 26

Abstract

We present a catalog of 948,216 stars with mass labels and a catalog of 163,105 red clump (RC) stars with mass and age labels simultaneously. The training data set is crossmatched from the Large Sky Area Multi-Object Fiber Spectroscopic Telescope DR5, and high-resolution asteroseismology data, mass, and age are predicted by the random forest (RF) method or a convex-hull algorithm. The stellar parameters with a high correlation with mass and age are extracted and the test data set shows that the median relative error of the prediction model for the mass of the large sample is 3%, and for the mass and age of RC stars is 4% and 7%. We also compare the predicted age of RC stars with recent works and find that the final uncertainty of the RC sample could reach 18% for age and 9% for mass; meanwhile, the final precision of the mass for the large sample with different types of stars could reach 13% without considering systematics. All of this implies that this method could be widely used in the future. Moreover, we explore the performance of different machine-learning methods for our sample, including Bayesian linear regression and the gradient-boosting decision tree (GBDT), multilayer perceptron, multiple linear regression, RF, and support vector regression methods. Finally, we find that the performance of a nonlinear model is generally better than that of a linear model, and the GBDT and RF methods are relatively better.

Unified Astronomy Thesaurus concepts: Stellar ages (1581); Stellar masses (1614); Support vector machine (1936); Random Forests (1935)

1. Introduction

To describe the current structure, evolution, and formation history of the Milky Way, it is necessary to accurately estimate the mass and age of a large number of stars distributed throughout our home galaxy. Through the spectra of stars, astronomers can acquire many stellar parameters (Mathur et al. 2017; Wu et al. 2019; Huang et al. 2020; Zhang et al. 2020, 2021). However, to date it is still not easy to acquire the age of stars accurately and precisely. The indirect isochrones method can obtain the age of clusters with relatively high precision by matching the observed data based on the stellar evolution model (Soderblom 2010; Xiang et al. 2017), but for field stars the precision of this method might not be perfect due to the highly accurate stellar parameters that are needed.

For a long time, due to the limitation of observations and data analysis, we can only estimate the ages of a small number of stars in the solar neighborhood (Edvardsson et al. 1993; Nordström et al. 2004; Takeda et al. 2007; Haywood et al. 2013; Bergemann et al. 2014). With large sky surveys, such as the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST), Xiang et al. (2015, 2017) estimated the ages of a large number of stars. Following this, it has been found that there is a relation between carbon and nitrogen abundances and the ages of giant stars, which has already been used to predict the ages of red giant branch (RGB) stars (Martig et al. 2016; Ness et al. 2016; Ho et al. 2017).

There are also other surveys that could provide the age of large sample. The Galactic Archaeology with HERMES (GALAH) survey, a high-resolution spectroscopic survey, aims toward a chemical tagging experiment (Freeman & Bland-Hawthorn 2002; Bland-Hawthorn et al. 2010); for very bright stars, more than 30 different elements can be measured. The age and kinematic inventory of the solar neighborhood was provided in Buder et al. (2019). Bright giant stars become the primary targets for the Apache Point Observatory Galactic Evolution Experiment (APOGEE) survey, a high-resolution spectroscopic survey that has lead to some works on the mass and age of stars (Zasowski et al. 2013; Martig et al. 2016; Majewski et al. 2017). Recently, a precision of ∼5% for mass and ∼20% for age was acquired by Silva Aguirre et al. (2020) with Transiting Exoplanet Survey Satellite (TESS) data. Meanwhile, there also are many other results with similar precisions such as ∼6% for mass and ∼20% for age in Stello et al. (2022) by using TESS asteroseismology of the Kepler red giants, and ∼10% for mass and ∼30% for age in Mackereth et al. (2021) with asteroseismology of giant stars in the TESS continuous viewing zones and beyond.

It has been found that there is a correlation between the age of solar-like stars and their surface rotation, and detailed studies have been carried out with asteroseismology data (García et al. 2014; McQuillan et al. 2014; Ceillier et al. 2016; van Saders et al. 2016). At the present time, it is known that asteroseismology is an effective method to estimate the mass and age of stars (Gai et al. 2011; Chaplin et al. 2014), however, it needs high-precision, long-duration, and high-resolution photometric observation so, unfortunately, we still do not have a large-enough asteroseismological sample.

[5] Corresponding author.

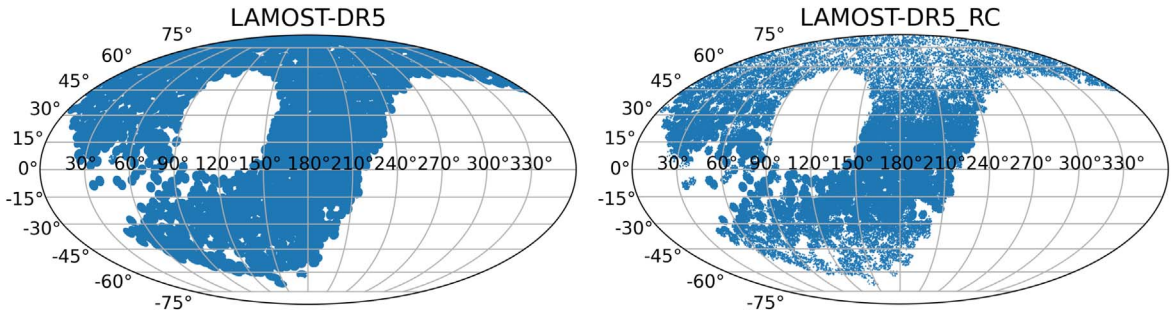

Figure 1. Distribution of the LAMOST data for mass estimation we use (left) and the RC distribution we use for age and mass (right).

Up to now, although there are many methods to predict the mass and age of stars, their precision and efficiency are still not perfect. We desperately need to make full use of big data to obtain more samples and try more methods to improve the precision of predictions, thus we can then explore both the history of the Galaxy's assembly more effectively, and more properties of the Milky Way such as mass distribution, population structure, and dynamical evolution (e.g., Wang et al. 2018a, 2018b, 2019, 2020a, 2020b, 2020c, 2022a, 2022b; Bland-Hawthorn et al. 2019; Yu et al. 2021; Yang et al. 2022) and references therein).

Machine learning is a branch of artificial intelligence and we could make full use of high-quality data for training through algorithms. By combining machine learning with high-quality asteroseismology data, we could predict the relationship between stellar mass (age) and stellar parameters, and thus we could then get these two parameters of a large sample with high confidence.

In this paper, we use a novel machine-learning method to estimate the mass of a larger sample and a smaller sample for the age and mass of red clump (RC) stars in LAMOST. Furthermore, we quantitatively compare the different machine-learning methods for the first time.

The paper is structured as follows: Section 2 presents the data we adopt, Section 3 is the method introduction we use, Section 4 shows our results, Section 5 is for discussion, and finally Section 6 gives a brief summary of our work.

2. Data

2.1. Catalogs

Xiang et al. (2019) have provided 8,162,566 stars from the LAMOST survey and the chemical abundances are derived from the DD-Payne model, which is inherited from both the Payne (Ting et al. 2019) and the Cannon (Ness et al. 2015). In this work, we use this catalog to obtain the chemical abundances of stars.

Ting et al. (2018) provided us with 175,202 RC stars in LAMOST with 3% contamination, and this work also includes two asteroseismology parameters ΔP and $\Delta\nu$. We use this catalog to obtain the RC stellar label and notice that the ΔP and $\Delta\nu$ are also obtained from stellar spectra, therefore the frequency separation ($\Delta\nu$) between adjacent acoustic p-modes and the period spacing (ΔP) of the mixed gravity g-modes and acoustic p-modes could be used for the separation of RC stars and RGB stars (Ting et al. 2018; Hawkins et al. 2018). The precision for LAMOST ΔP and $\Delta\nu$ is 50 s and 1 μHz, respectively, which is enough for the age/mass determination according to the previous results (Ting et al. 2018, 2019). In this work, we determine the final age and mass using new a training data set and new methods we chose, then we compare these with other catalogs in order to test the robustness of the different methods.

Pinsonneault et al. (2018) have provided ages of 6676 stars in APOKASC-2, which are derived from their model using mass, radius, [Fe/H], and [α/Fe]. We train our model for mass and age by this high-quality high-resolution asteroseismology catalog. To be more specific, this catalog contains stellar properties for a large sample of evolved stars with APOGEE spectroscopic parameters and Kepler asteroseismic parameters. With the help of five independent techniques, the median random mass uncertainties for RGB stars could reach 4%, for RC stars it could reach 9%, with the age precision being within 8%, which is suitable for the training sample.

In short, thanks to the works above we use the chemical abundance from Xiang et al. (2019), the precise mass and age from Pinsonneault et al. (2018), and the RC label, ΔP, and $\Delta\nu$ from Ting et al. (2019). Then we use the new machine-learning methods and new high-quality asteroseismic age and mass to estimate the mass of the large sample for Xiang et al. (2019) and the age and mass of RC stars for Ting et al. (2019).

After crossmatching the above catalogs, we first get 4479 stars to predict the large-sample mass (LS-mass) and 1806 stars for RC mass (RC-mass) and RC age (RC-age); notice that these are not the final data sets as shown in the next part. The distribution of the sample needed to be predicted in the Galactic longitude and latitude in celestial coordinates is shown in Figure 1.

2.2. Final Training Data Sets

In order to improve the precision of the machine-learning prediction, we do the following experiment for the three catalogs mentioned above.

The three data sets after the first crossmatch mentioned above are equally separated as the test and training samples, then we first use the random forest (RF) method to train and make mass and age predictions for the test data set. For large-sample stars (LS-mass), we select stars whose absolute error of mass prediction is less than 1 $M_\odot$ and relative error is less than 0.3, and for RC stars, we select stars whose absolute error of mass (age) prediction is less than 1 $M_\odot$ (3 Gyr) and relative error is less than 0.4. Notice that here we only use 200 decision trees and make full use of all stellar parameters shown in Figure 2 as inputs in the method to finish this step. After this,

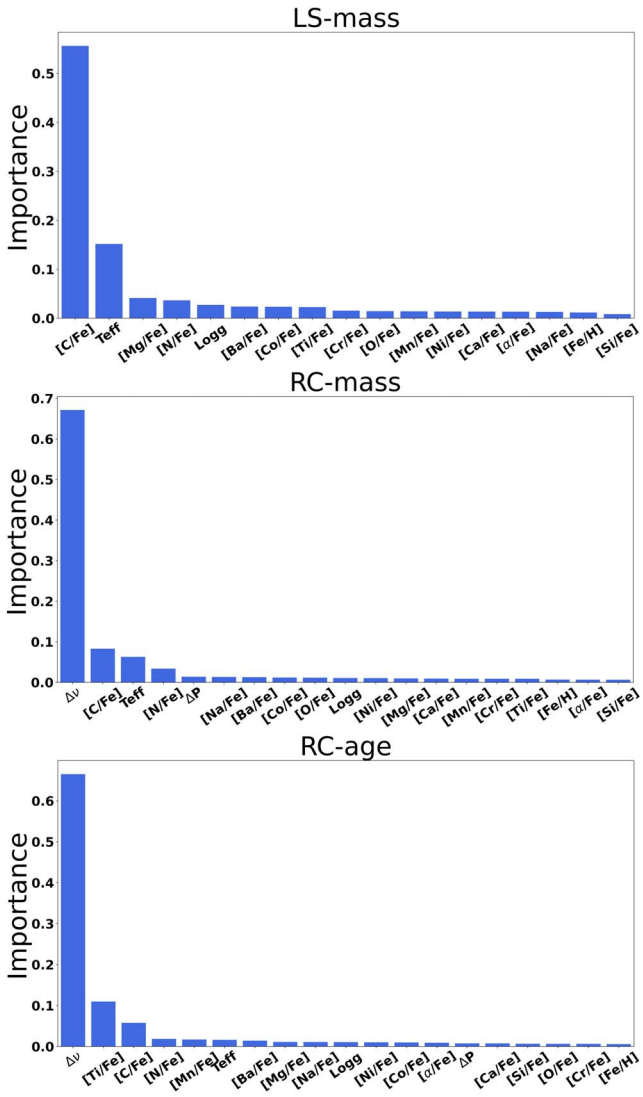

Figure 2. The results of feature extraction using random forest. The different panels are three different training samples that we have selected, the top one is for the large sample containing a different type of stars, for which we only estimate mass, and the middle and bottom ones are for RC stars, for which we could estimate mass and age. The importance represents the contribution of the stellar parameter to our prediction model, and it is actually the relative importance.

we finally get an LS-mass set of 4246 stars, an RC-mass set of 1751 stars, and an RC-age set of 1384 stars for training and predicting, as detailed in Figure 3, which shows the final training mass and age distribution on the $T_{\rm eff}$–log g plane.

3. Method

3.1. Feature Importance

The machine-learning methods used in this paper are mainly from Scikit-learn (sklearn; Anghel et al. 2019; Mediratta & Oswal 2019; Florescu & England 2020), and can be divided into six categories: classification, regression, clustering, dimensionality reduction, model selection, and preprocessing.

First, we explore the feature importance distribution of the stellar parameters for the mass/age of the three selected training samples with the RF method shown in Figure 2. In order to avoid the severe impact of one feature on the prediction due to the unexpected dimension problems, we choose to do the standardization that can accelerate the convergence of weight parameters. Standardization or Z-score normalization is the transformation of features by subtracting from the mean and dividing by the standard deviation.

The RF method adopted here is based on decision trees and the final prediction result is also dependent on these trees. The correlation between different parameters can be easily identified with the help of information gain used to train the model so that this method has good robustness and overfitting can be avoided.

The importance here is relative or not absolute, and we have a test that finds that the importance of many stellar parameters is highly correlated, so it is therefore reasonable that we choose to use the first six or nine parameters to estimate the mass and age. As shown in Equation (3) in Pinsonneault et al. (2018), the mass is very sensitive to the $\Delta\nu$ and it is known that the age is also sensitive to mass, so it is not strange to see the $\Delta\nu$ is the most important factor for the RC-age and mass.

3.2. Features Choice

The relation between the prediction precision and the number of features in the training data set, based on the relative error distribution versus feature numbers, is clearly shown in Figure 4. The mean relative error of the test data set decreases with the increase of the number of training features (orange line) until a stable pattern is reached. Based on this pattern, we choose the first six stellar parameters to train the model for LS-mass, the first nine features for the mass of RC stars, and the first six features for the age of RC stars.

We notice that the LS-mass is mixed with different types of stars that might not belong to the training data set, so we use the first six stellar parameters of [C/Fe], $T_{\rm eff}$, [Mg/Fe], [N/Fe], log g, and [Ba/Fe] to construct a convex hull in order to determine which stellar types our training model are suitable for, as displayed in Figure 5. We can see our sample mainly consists of K-giant stars including RC and RGB stars; there are also very few possible other types of stars (e.g., G type stars), which is consistent with the result that APOKASC mainly consists of RGB and RC stars. Our large sample almost entirely consists of K giants and we find that LAMOST DR5 contains around 1 million K-giant stars; in this work we also use convex hulls to select 948,216 stars, which are self-consistent. Notice that in the future our method could be used for different types of stars if the quality and quantity of the training data set is good enough, and in order to avoid the mixing effects of RGB and RC stars, we choose not to estimate the age of all of the large sample here. The age of the RGB estimation will be shown in subsequent work. Algorithms that construct convex hulls of various objects have been used in astrophysics, mathematics, and computer science. We have 948,216 stars for LS-mass suitable for the training model based on Pinsonneault et al. (2018). Notice that we have also removed some vacancy values for the RC catalog before mass and age determination, and we finally get the 163,105 stars to be predicted without a convex-hull algorithm.

4. Results

4.1. Final Age and Mass Distribution

The final predicted mass of 948,216 stars (using [C/Fe], $T_{\rm eff}$, [Mg/Fe], [N/Fe], log g, and [Ba/Fe]), and mass (using $\Delta\nu$,

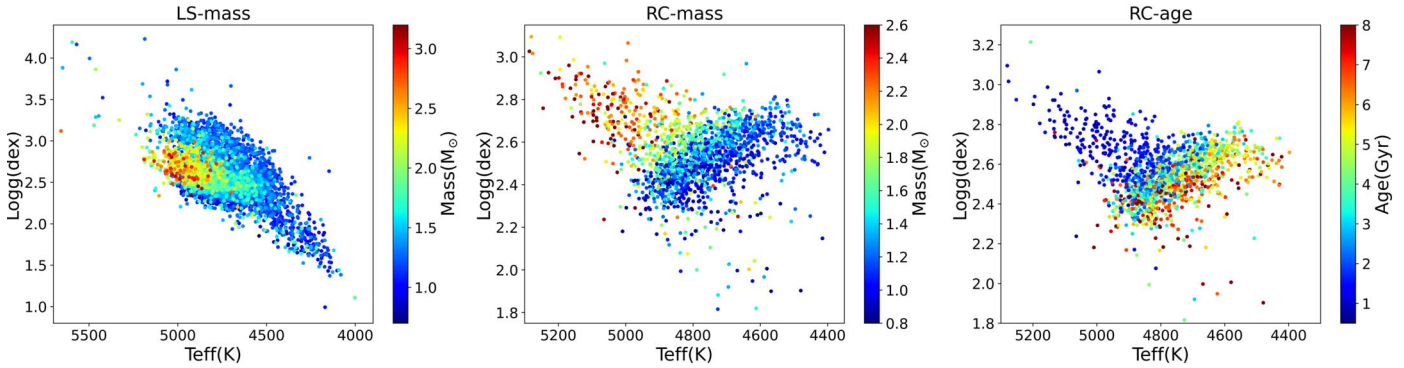

Figure 3. The final training sample distribution for mass and age on the $T_{\rm eff}$–log g plane.

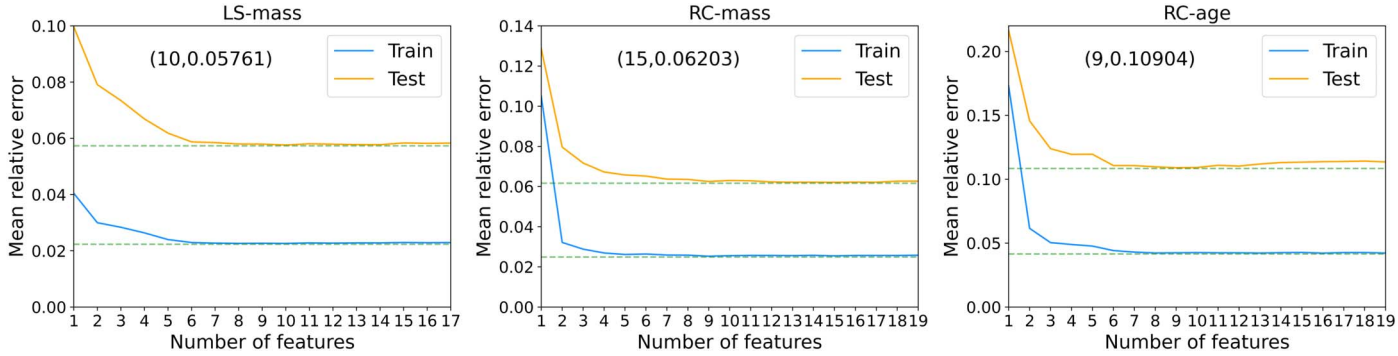

Figure 4. The relations between the number of training features and the mean relative error of the test data set in our prediction model. Different panels represent different samples. The blue line represents the training data set, the orange line represents the test data set, and the green dotted lines guide the eyes to the minimum value used for the stable pattern. The minimum values are also labeled at the top of each panel. Notice that it is the minimum value, but not the final feature, that we adopt; we choose final features empirically and accordingly.

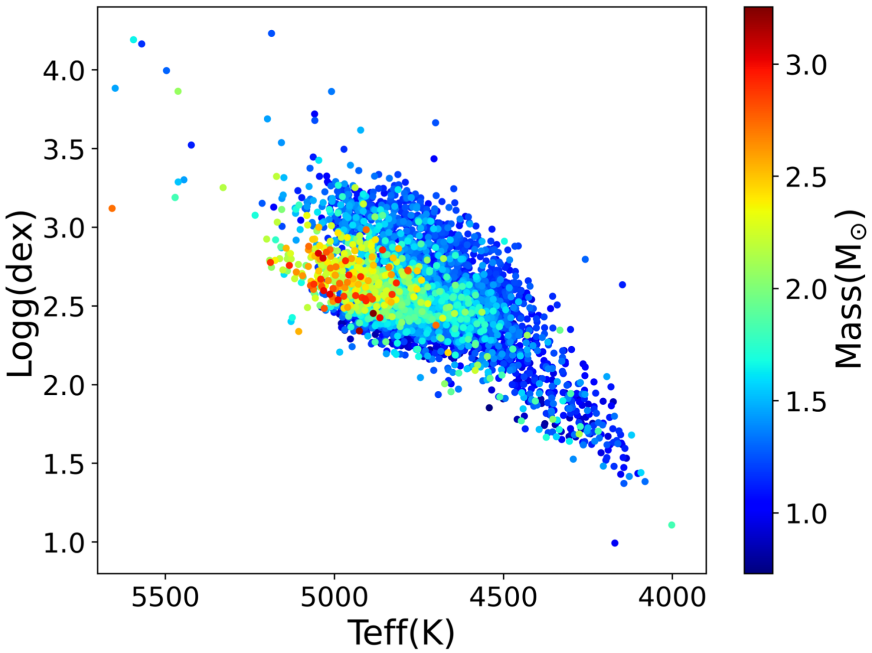

Figure 5. A mass distribution of the stars we use to create a convex hull and train prediction models on the $T_{\rm eff}$ and log g plane. Different colors represent different masses.

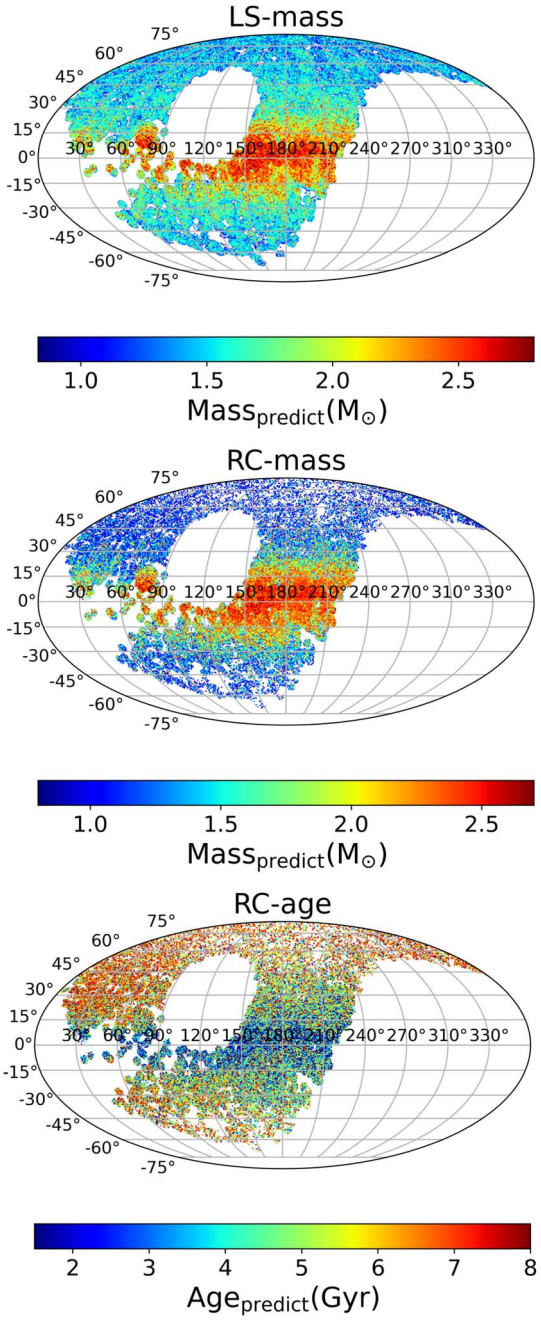

Figure 6. The distribution of predicted mass (age) in celestial sphere coordinates. The top panel is the mass distribution of the large sample, the middle panel is the mass distribution of RC stars, and the bottom one is the age distribution of RC stars.

[C/Fe], $T_{\rm eff}$, [N/Fe], ΔP, [Na/Fe], [Ba/Fe], [Co/Fe], and [O/Fe]) and age (using $\Delta\nu$, [Ti/Fe], [C/Fe], [N/Fe], [Mn/Fe], $T_{\rm eff}$) of 163,105 RC stars are vividly presented in Figure 6, colored by the mass or age on the Galactic longitude and latitude celestial sphere. For the mass distribution, we can see the more massive stars are located in the disk similarly to the mass pattern for the RC stars in the middle panel, and the age distribution of RC stars also shows that the younger stars are mainly located in the low latitudes. It could be naturally understood that there are more star-forming regions in the disk, so therefore more massive stars and younger stars are located in the disk and low latitudes.

The distribution of age in the R.A. and decl. plane is also shown in the left panel of Figure 7, the numbers and fractions for decl. beyond 20° or 30° are denoted at the top: they are 110,071 and 68%, and 82,739 and 51%, respectively. The middle panel of this figure is for density distribution in the longitude and latitude planes and star counts; fractions beyond 20° or 30° for latitude are labeled at the top of this panel: they are 47,926 and 29%, and 24,673 and 15%, respectively. The right panel in this figure is the R and Z planes in cylindrical Galactic coordinates colored by density/stellar number; fractions larger than 10 or 15 kpc for distance are also denoted at the top, they are 78,424 and 48%, and 2694 and 2% separately.

Figure 8 shows the results of our method for the test data sets of three groups. From the top left to the top right, the y-axis is the predicted mass, the absolute mass error, and the relative error, and the x-axis is the true mass from asteroseismology. As shown in the figure, the predicted dispersion of the large-sample mass is 0.13 $M_\odot$, the mean absolute error is 0.08 $M_\odot$, and the median is 0.05 $M_\odot$; the mean relative error is 6% and the median is 3%. Dispersion means the standard deviation of the predicted ages/mass minus the true values in the catalog we used, the absolute error is the predicted value minus the true value, and the relative error is the predicted value minus the true value divided by the true value; notice in this work we use the median relative error for the final precision uniformly.

Similarly, the middle row of Figure 8 is the RC stars' mass; as shown in the label the predicted dispersion of mass of RC stars is 0.14 $M_\odot$, the mean absolute error is 0.09 $M_\odot$, the median value is 0.05 $M_\odot$, the mean relative error is 6%, and the median value is 4%.

It can be found in Figure 8 for the prediction of mass that the precision of RC stars (4%) is slightly worse than that of large-sample stars (3%) for the test data set. The main reason is that the number of stars in the training samples is different. The larger the sample size, the more effectively the machine-learning method could find the rule. Moreover, the predicted dispersion of the age of RC stars is 0.68 Gyr, the mean absolute error is 0.42 Gyr, the median value is 0.21 Gyr, the mean relative error is 11% ,and the median relative value is 7%. We could speculate that the precision of the age of RC stars could be higher if we have a higher-quality catalog.

We then explore the relations between the predicted age and [C/N], as shown in Figure 9. We can see that in the region where age is less than or equal to 8 Gyr, the age and [C/N] show a good linear relationship, which is consistent with our expectation. In the region where age is older than 8 Gyr, it seems that there is no obvious pattern because the RC stars are inclined to the relatively younger group, and the number of old stars is very small in our sample, so it is impossible to make high-precision statistics.

4.2. More Comparisons

Figure 10 shows the comparison between the mass or age we predict and the reference values we use; the consistency provides verification for the robustness of our method. We also compare our predicted age with other works based on LAMOST, APOGEE, and Gaia data, which will also provide independent verification for the method. The comparison results are shown in Figure 11, where the top-left panel is a

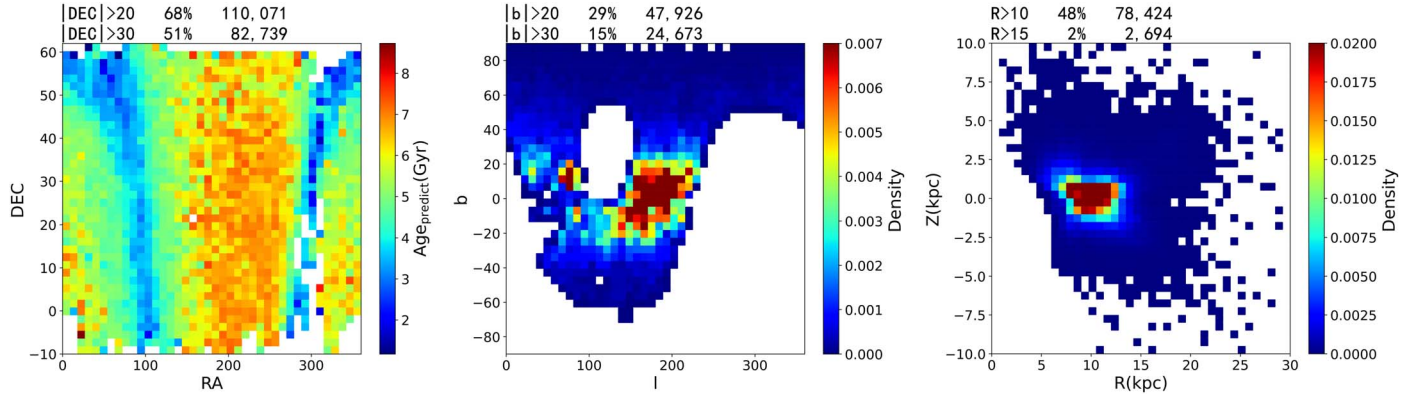

Figure 7. The distribution of RC-age on the R.A. vs. decl. plane is shown in the left panel colored by age, the number and fraction for decl. beyond 20° or 30° are denoted at the top of the panel. The middle panel is for density distribution in the longitude and latitude planes and star counts; fractions beyond 20° or 30° for latitude are labeled at the top of this panel. The right panel is the R and Z planes in cylindrical Galactic coordinates colored by density/star counts; fractions larger than 10 or 15 kpc for radial distance are also denoted at the top.

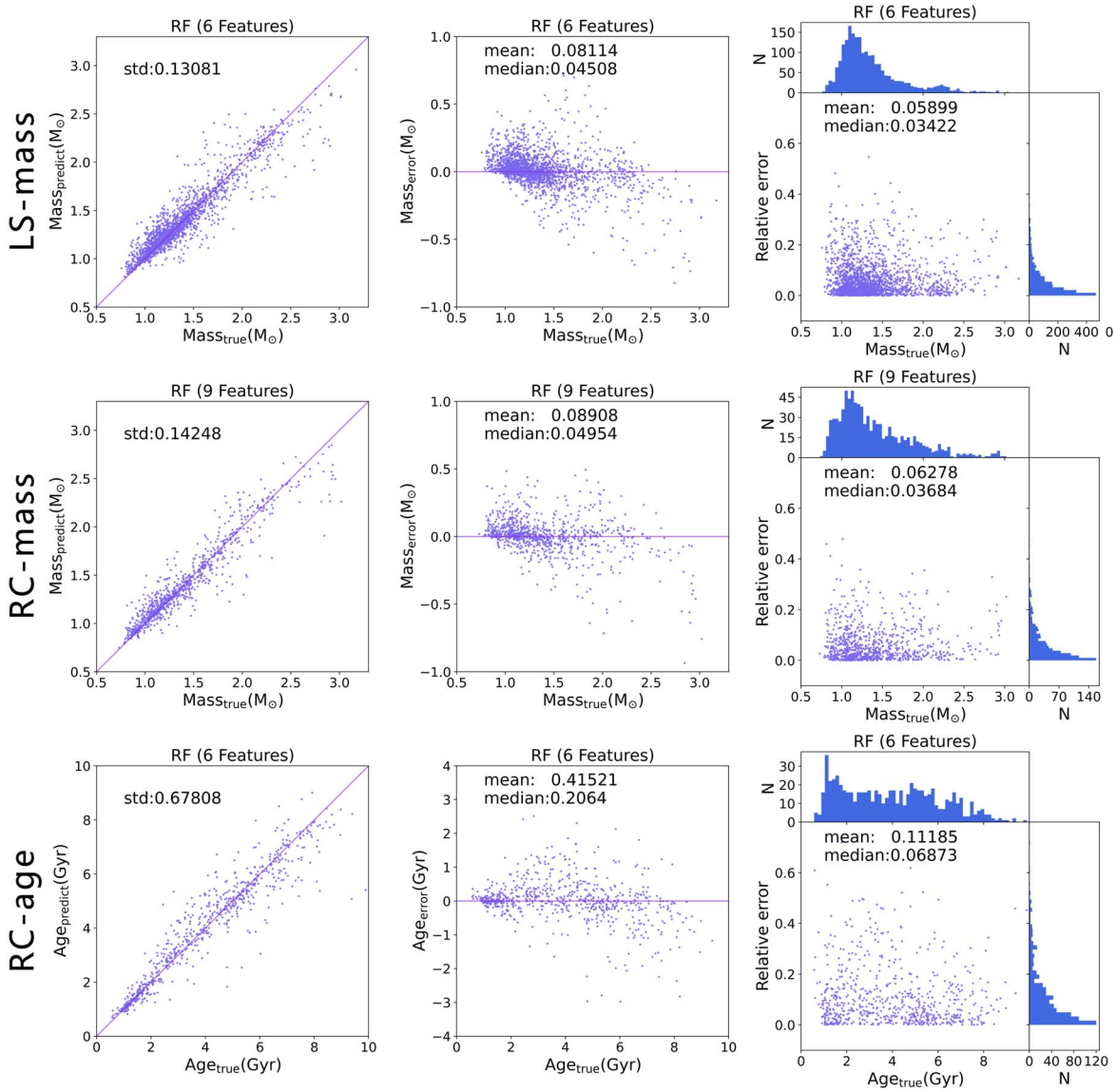

Figure 8. The predicted results of our test data sets using RF. Different rows represent different groups of samples, and different columns show the dispersion ($M_\odot$, Gyr), absolute error ($M_\odot$, Gyr), and relative error, respectively. Dispersion means the standard deviation of the predicted ages/mass minus the true values in the catalog we used, the absolute error is the predicted value minus the true value, and the relative error is the predicted value minus the true value divided by the true value; notice in this work we use the median relative error for the final precision uniformly The dispersion, mean, and median values of the data are marked in the upper-left corner of each figure, and the final number of features we adopt to train each model is marked at the top of each panel.

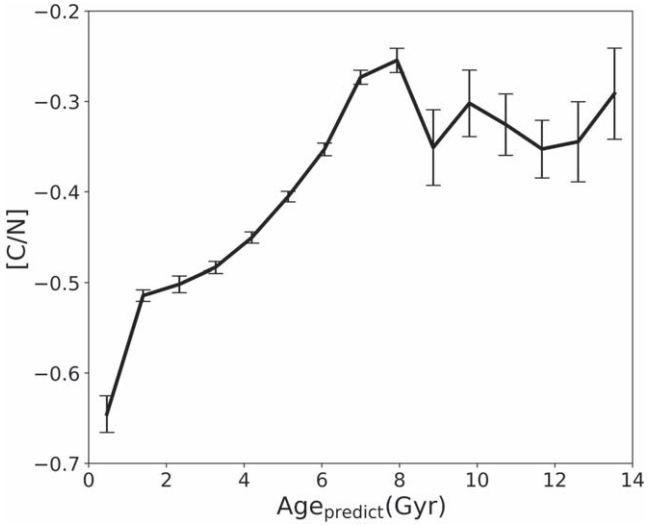

Figure 9. The relationship between the predicted age and [C/N]. The black line is the median value in each bin with Poisson error.

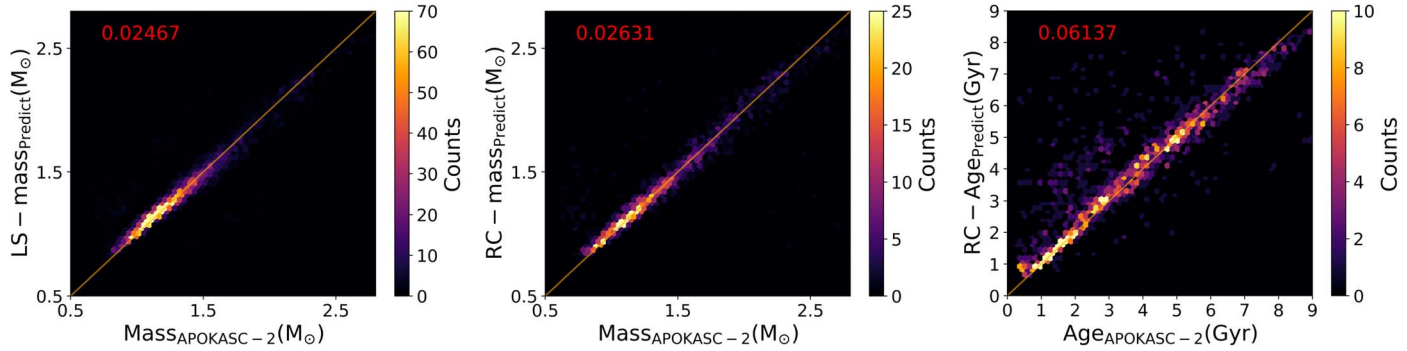

Figure 10. The comparison between the mass and age we predict and the reference values we use during this work, where the number marked on the figure represents the median value of relative error for our method. It consists of the common stars of LAMOST data we predict and APOKASC-2 in this work. The purpose here is method validation and since we use APOKASC-2 to predict our sample, the precision is naturally quite good for the data set since we use the APOKASC-2 data for training.

comparison for the common stars of APOGEE (Ting & Rix 2019), the top-right panel is for LAMOST (Ting et al. 2018),[6] the bottom-left panel is for Gaia (Sanders & Das 2018), and the bottom-right panel is for Ho et al. (2017). We can see that although there are some differences, for the overall trend the consistency is acceptable. Similarly, the first four panels of Figure 12 show the mass comparisons for other works. The left column is compared to Yu et al. (2018), the right column is compared to Ho et al. (2017), the top row is for LS-mass, and the bottom row is for RC-mass; all are matched well with some reasonable difference.

Compared with the APOGEE high-quality data we could claim for this work, the precision of RC-age could reach 18% (top left of Figure 11) and by matching with the high-precision Kepler asteroseismology data we can claim that our uncertainty of RC-mass could reach 9% (bottom left of Figure 12). Meanwhile, the precision of LS-mass could be 13% (top left of Figure 12). All these final precisions are based on the final relative error analysis using a high-precision asteroseismology data set and we frankly admit that the systematics might be ignored so more work on this is needed in the future.

[6] Age is not shown clearly in the paper but it is determined simultaneously.

Moreover, we also compare the open cluster (OC) age using our final sample; the OC is chosen by the spatial locations, kinematics (line-of-sight velocity, proper motions) and metallicity-clustering distributions. As we can see in Figure 13 the relative errors are NGC 6811: 9.1%, NGC 2420: 9.3%, NGC 6819: 23.4%, NGC 2682: 9.5%, NGC 6791: 2.7%, and Be 17: 33.5%. The final median relative error is 9.5%, which strongly supports our final conclusions. Notice that we use our final LAMOST RC catalog to select OC memberships and then compare these with literature values. In our final RC catalog, the stellar number of memberships for these open clusters mentioned above is NGC 6811: 2, NGC 2420: 1, NGC 6819: 6, NGC 2682: 2, NGC 6791: 2, and Be 17: 4.

We also explore the relationship between RC-age relative error and signal-to-noise ratio (S/N; the ratio of the intensity of a signal to the background noise detected by a measuring instrument for spectra used for estimation of LAMOST stellar parameters). As shown in Figure 14, the relative error tends to be stable with the increase of S/N. The distributions of the relative errors of mass and age for our test data set with stellar parameters $T_{\rm eff}$, log g, and [Fe/H] are also displayed in Figure 15, which shows the robustness of our method with a small dispersion.

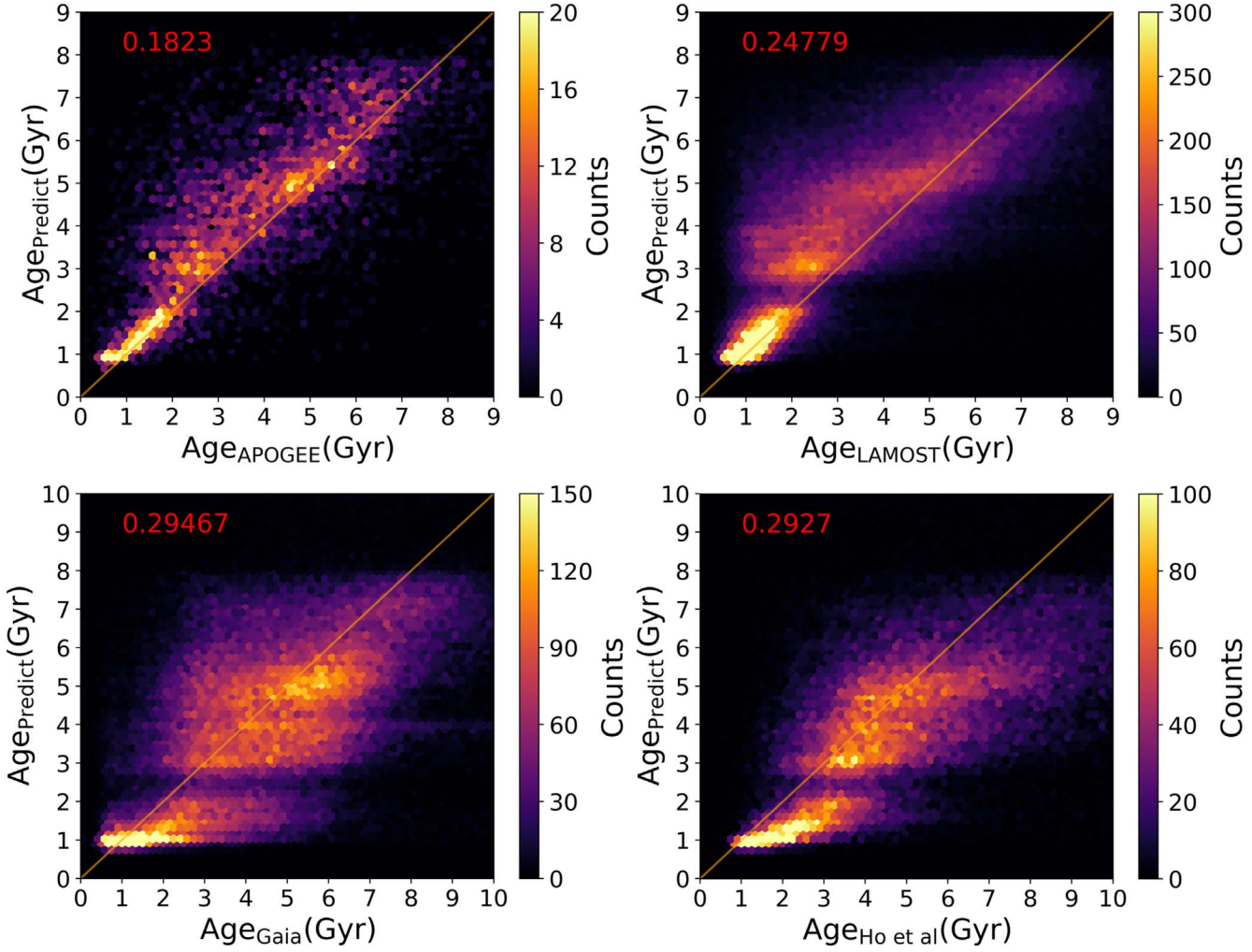

Figure 11. Comparing our predicted age with other works using LAMOST, APOGEE, and Gaia data. On the top left is the age of APOGEE data using a different method (Ting & Rix 2019), the top-right panel is the age of LAMOST data (Ting et al. 2018), the bottom-left panel is the age of Gaia data (Sanders & Das 2018), and the bottom-right panel is Ho et al. (2017). The median value of relative error is shown on the top left and the consistency is acceptable. We have fewer stars around 2 Gyr in the training data set so there are apparently disconnected features.

Figure 16 shows the age distribution on each panel of different stellar parameters. From top left to bottom left these are: $\Delta\nu$ versus [Ti/Fe], [C/Fe] versus [N/Fe], [Mn/Fe] versus $T_{\rm eff}$, [Ba/Fe] versus [Mg/Fe], [Na/Fe] versus log g, [Ni/Fe] versus [Co/Fe], [α/Fe] versus ΔP, [Ca/Fe] versus [Si/Fe], [O/Fe] versus [Cr/Fe], and [Fe/H] versus [α/Fe]. These panels show that all of them have a correlation, more or less, with age, either positive or negative. In particular for the last one, [α/Fe] and [Fe/H], we can see a thick-disk population with the red patch and a thin-disk population with the blue patch.

As we mentioned, almost all of the parameters are correlated with age, but why do we only choose the first six to nine parameters for our method and why do the other parameters shown in Figure 2 not have a high importance? The reason is that we find they are related to the properties of the RF method. This means that when there are correlations for multiple features, the RF will extract the one with the greatest contribution, and then the importance of other features might become not very relevant artificially (e.g., [Fe/H]).

As a test, we attempt to use the first six stellar parameters in importance to independently predict other stellar parameters in the RC-age sample and check the predicted results. As shown in Figure 17, we find that other stellar parameters can be predicted by using the first six stellar parameters. Because the first six features are more or less related to other features, the importance of the other features is not as significant when we make the related analysis.

Conversely, we also randomly choose several other relevant stellar parameters to empirically predict age in order to compare with our previous results, as shown in Figure 18. Obviously, we find that even though we use other parameters to predict the age, a similar precision could be reached. All of these results show that our method for age and mass estimation is reasonable and we could make full use of many parameters to estimate age and mass for other catalogs, even though we are lacking some chemical stellar parameters.

5. Discussion

5.1. Comparisons for Age Prediction Using Different Catalogs

In this paper, we choose the RC-age of APOKASC-2 as the training data set because it is a high-resolution asteroseismology sample. In order to compare the age based on APOKASC-

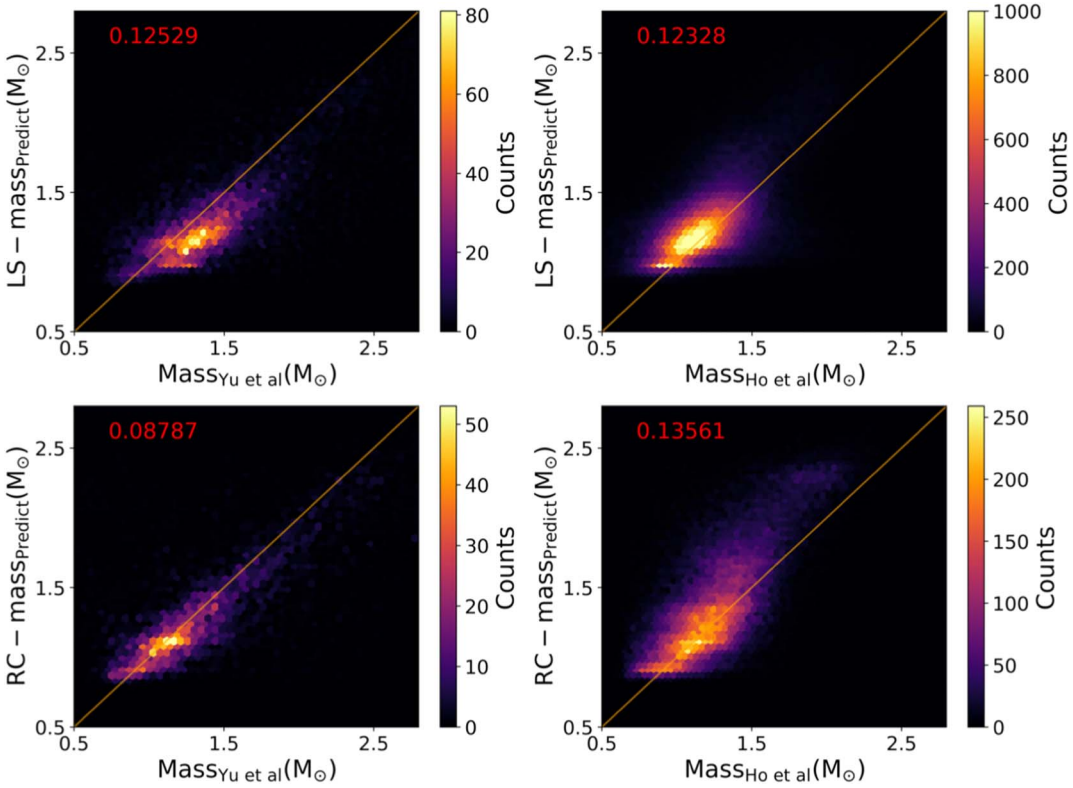

Figure 12. The figure shows the mass comparisons between this work, Yu et al. (2018), and Ho et al. (2017). The median value of relative error is shown on the top left of each panel and the consistency is acceptable.

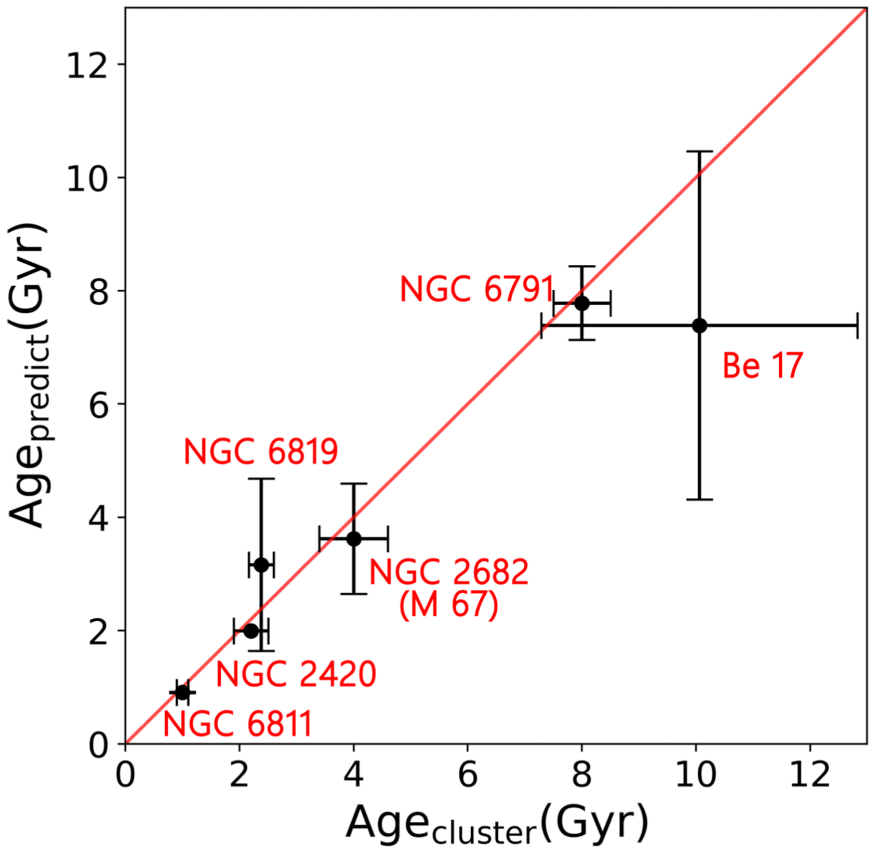

Figure 13. Open-cluster comparisons for our determinations and literature values based on Bragaglia et al. (2006), Geller et al. (2008), Grundahl et al. (2008), Jacobson et al. (2011), Janes et al. (2013), Brewer et al. (2016), Stello et al. (2016), and references therein. The comparison is quite good except for the older last cluster, and the error bars are calculated by the Gaussian dispersion or literature values.

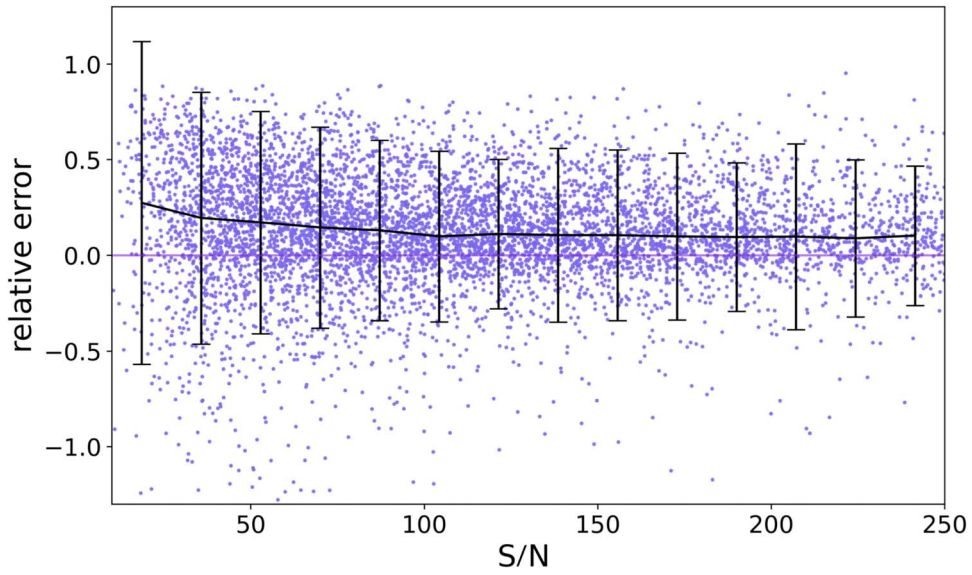

Figure 14. The relative error of RC-age vs. S/N in this work; the error bars represent the standard deviation in each bin.

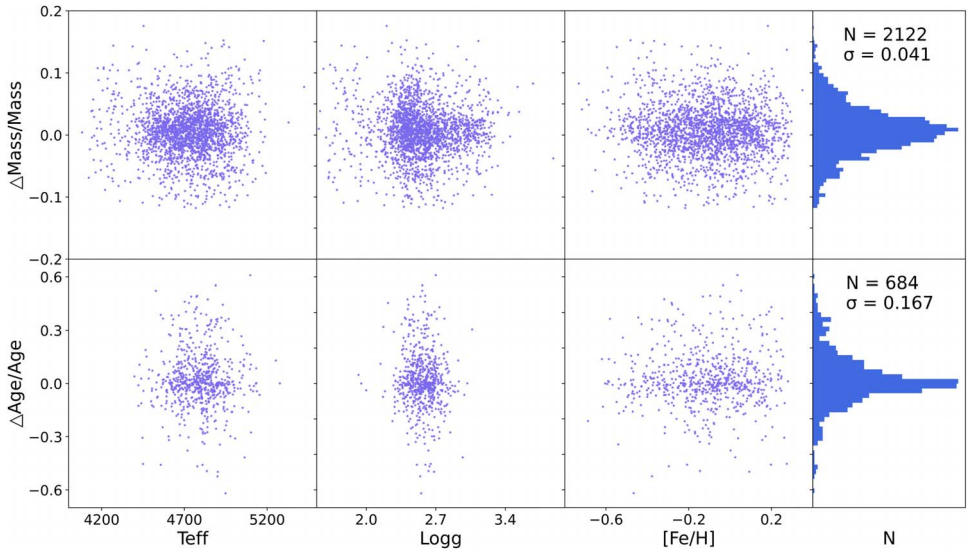

Figure 15. The distribution of relative errors of LS-mass and RC-age for our test data set we predict along with $T_{\rm eff}$ (kelvin), log g (dex), and [Fe/H] (dex).

2 and APOGEE (Ting et al. 2019), we use these two different catalogs to predict age, as shown in Figure 19. In the figure, the x-axis is age trained by APOGEE and the y-axis is trained by APOKASC-2; they have a different stellar number. We find that for older stars, the age predicted by APOKASC-2 is systematically higher than for the stars predicted by APOGEE, which is possibly caused by the different precision of the data sets. And as can be seen from Figure 20, showing the relative error analysis for these two catalogs, the age based on APOGEE is systematically smaller than that based on APOKASC-2. With the increase of age, the difference becomes more and more obvious, however for the overall trend almost of all difference is within 10%, which is acceptable and implies that the precision of the prediction is dependent on the quality of the data set.

5.2. Comparison of Common Stars between Two Different Mass Predictions in This Work

We have predicted the mass of two groups of samples, the LS-mass with the convex-hull algorithm and the RC-mass without the convex-hull algorithm. After crossmatching we find 155,532 common stars and then we compare the two slightly different mass prediction methods.

As can be seen from Figure 21, the value of relative errors is 8%, which shows that the mass difference predicted by the two methods is small and self-consistent.

5.3. Comparison of Different Machine-learning Methods

Different machine-learning methods used in this work have their own characteristics but there should be no absolute

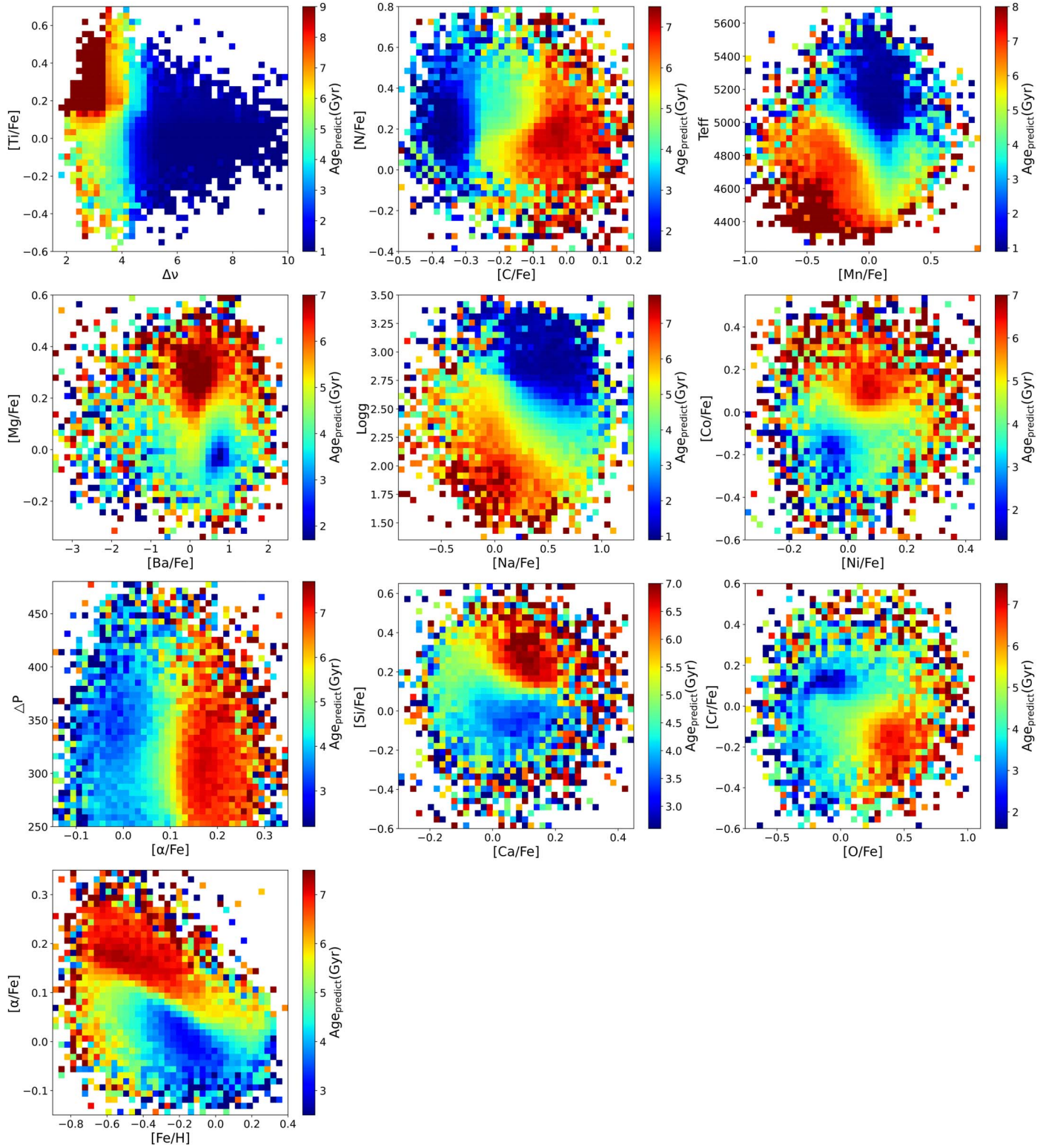

Figure 16. The distribution of our predicted ages over every two stellar parameters. From top left to bottom left these are: $\Delta\nu$ vs. [Ti/Fe], [C/Fe] vs. [N/Fe], [Mn/Fe] vs. $T_{\rm eff}$, [Ba/Fe] vs. [Mg/Fe], [Na/Fe] vs. log g, [Ni/Fe] vs. [Co/Fe], [α/Fe] vs. ΔP, [Ca/Fe] vs. [Si/Fe], [O/Fe] vs. [Cr/Fe], and [Fe/H] vs. [α/Fe].

difference for the advantages and disadvantages, which are dependent on the specific purposes. The reason why we choose RF is that after many attempts, we find that it is better in line with our expectations. The quantitative comparison of the six machine-learning methods including Bayesian linear regression (BYS), gradient-boosting decision tree (GBDT), multilayer perceptron (MLP), multiple linear regression (MLR), RF, and support vector regression (SVR) is shown in this subsection.

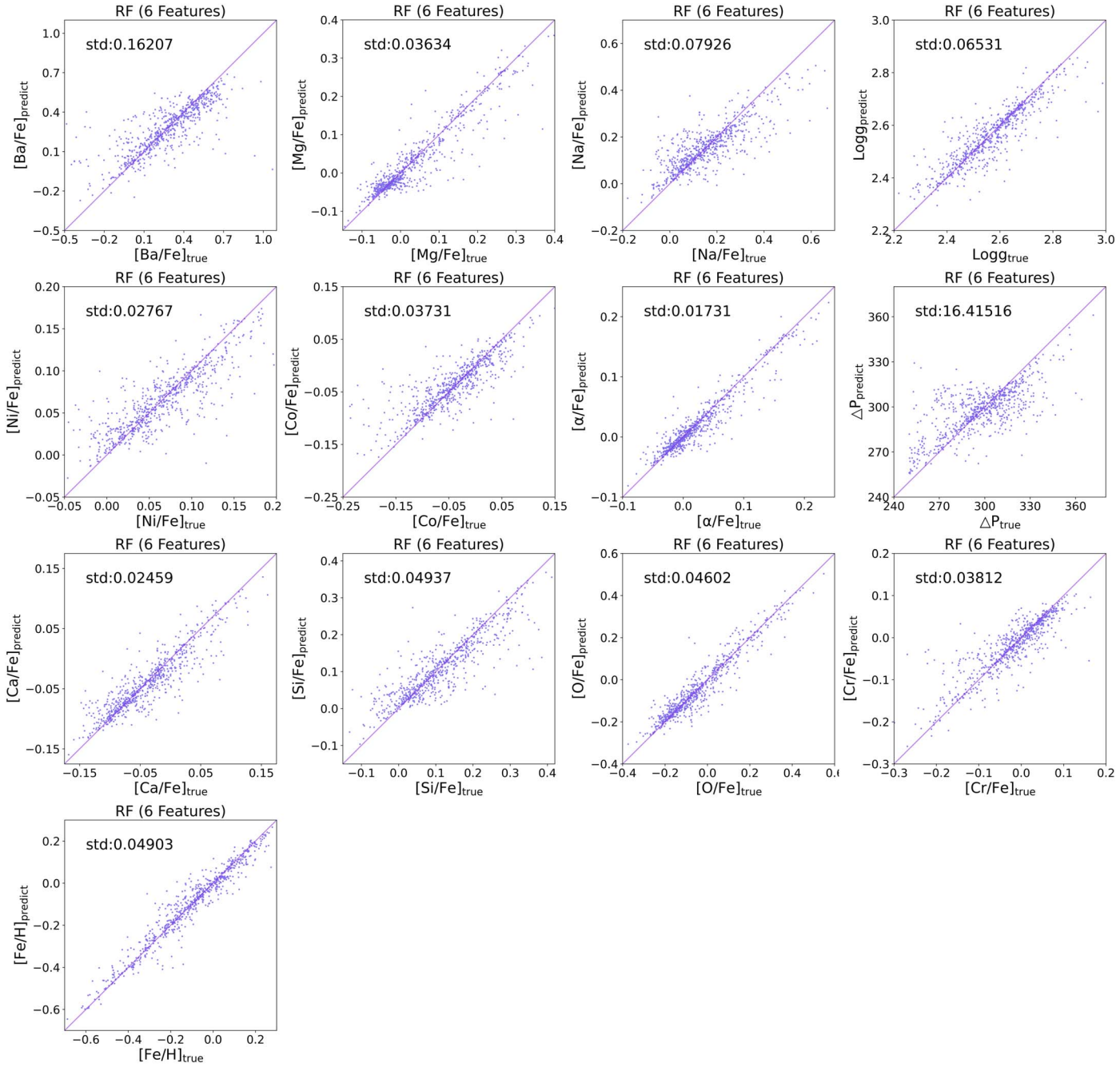

Figure 17. The predicted results for some chemical parameters using the first six stellar parameters shown in feature importance in Figure 2. The consistency is quite good and from top left to bottom left these are: [Ba/Fe], [Mg/Fe], [Na/Fe], log g, [Ni/Fe], [Co/Fe], [α/Fe], ΔP, [Ca/Fe], [Si/Fe], [O/Fe], [Cr/Fe], and [Fe/H].

Figure 22 shows the relation between the number of features used in the training model and the median relative error in different machine-learning methods. Meanwhile, Figure 23 shows the age prediction of different methods for the test data set. Based on the value labeled in the panels of these two figures, we can clearly see that BYS and MLR are relatively worse because both of them have a higher median relative error of ∼28% and larger dispersions of 0.97 Gyr, which might be caused by our prediction of RC stars being nonlinear, whereas BYS and MLR are linear models.

Among the other nonlinear methods, MLP is difficult to adjust during our experiments and the performance is hard to keep stable: the median relative error is 13% and the dispersion is 0.73 Gyr. The median relative error and dispersion of SVR are 14% and 0.74 Gyr, respectively. We can see from Figure 22 that the precision of GBDT is similar to RF with a median relative error of 10% and a dispersion of 0.68 Gyr, but more features of GBDT (10) are needed than RF (6) when the median relative error is becoming stable. In order to make our trained model applicable to more stars with fewer features, we decide to choose the RF for this work. More introductory text about the six machine-learning methods will be presented in the Appendix.

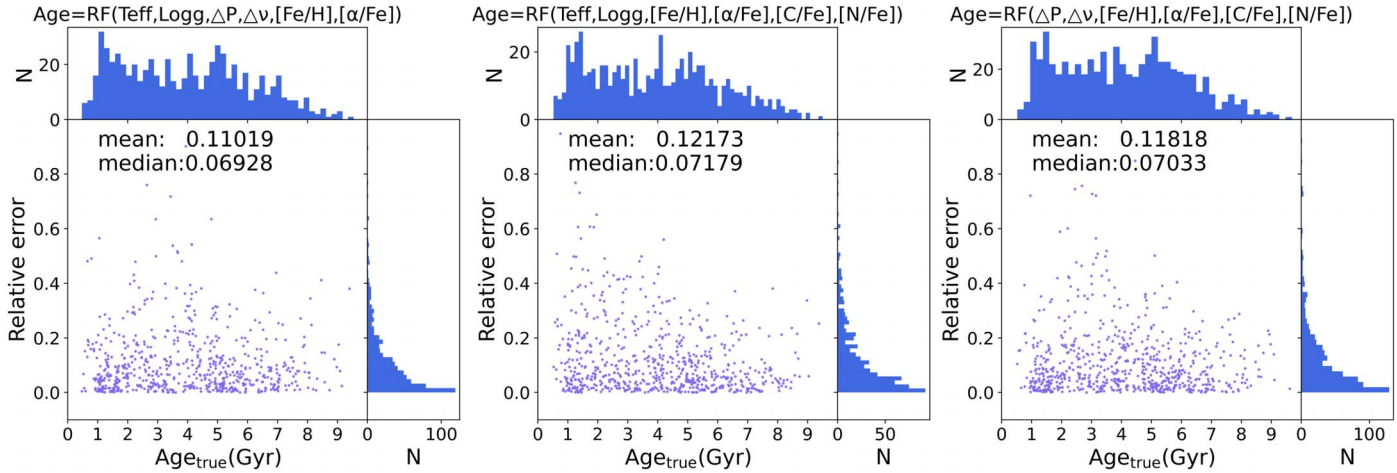

Figure 18. Age determination of RC stars for testing, with six empirical stellar parameters shown in the top of each panel. We can see the precision is almost the same.

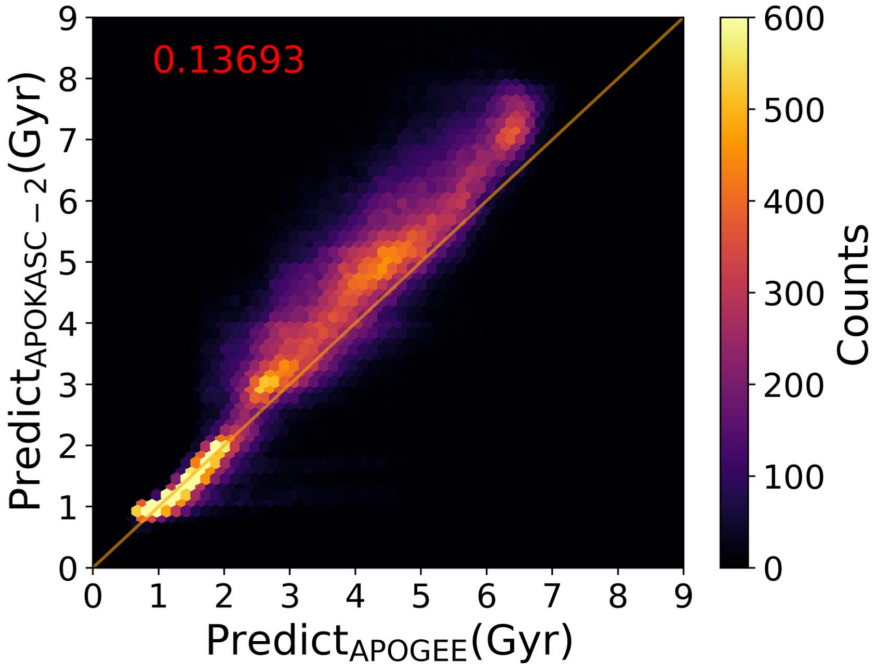

Figure 19. Comparison of age predictions using two different age catalogs. The *x*-axis is age trained by APOGEE and the *y*-axis is trained by APOKASC-2; the figure is colored by star counts.

6. Conclusions

In this paper, with the help of LAMOST, APOGEE, and asteroseismology data, we use RF to predict the mass of 948,216 large-sample stars, and the mass and age of 163,105 RC stars. We select stellar parameters with high correlation with mass and age to construct a training model, then we use these features, a convex-hull algorithm, and the RF method to determine the age and mass of the larger sample.

We find that the precision of the mass for large-sample stars could reach 3%, for RC stars it could reach 4%, and for RC-age precision it could be 7% for the test data set (shown in Figure 8). Compared with other high-quality samples, the precision for mass of large-sample stars could reach 13%, the mass precision of RC stars could reach 9%, and the age precision of RC stars could reach 18% for the median relative error. In general, our results could be compared well to recent works, in particular for open clusters, which could reach 9.5% for median relative error, so this strongly implies that we could make full use of this method in the future.

We also explore the performance of different machine-learning methods for the first time, in particular for age.

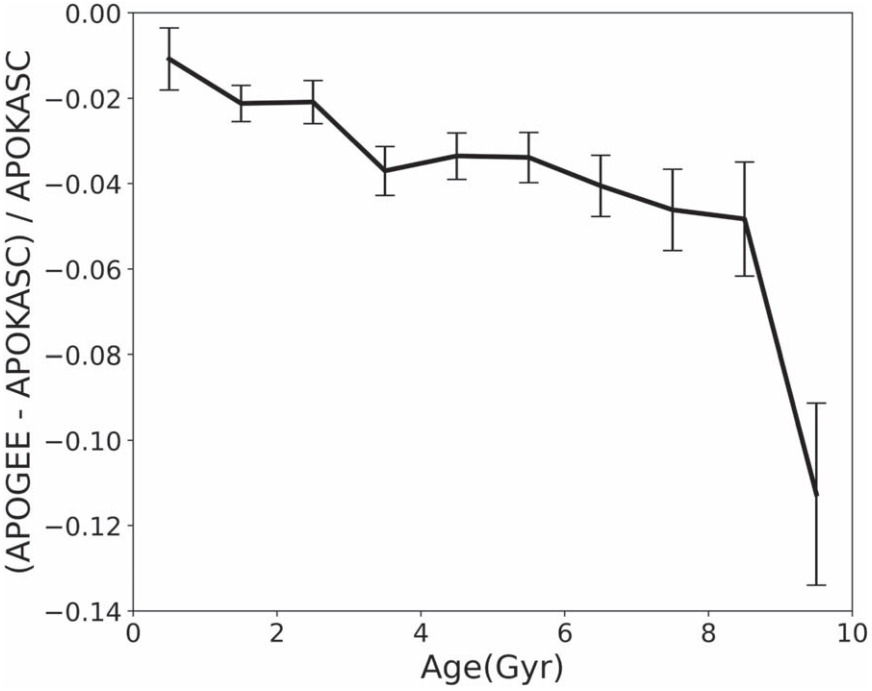

Figure 20. The relative age error of APOGEE and APOKASC-2 along with the age for common stars. We use two catalogs to make the prediction and find that the older the star, the more obvious the difference. The error bars are Poisson noise.

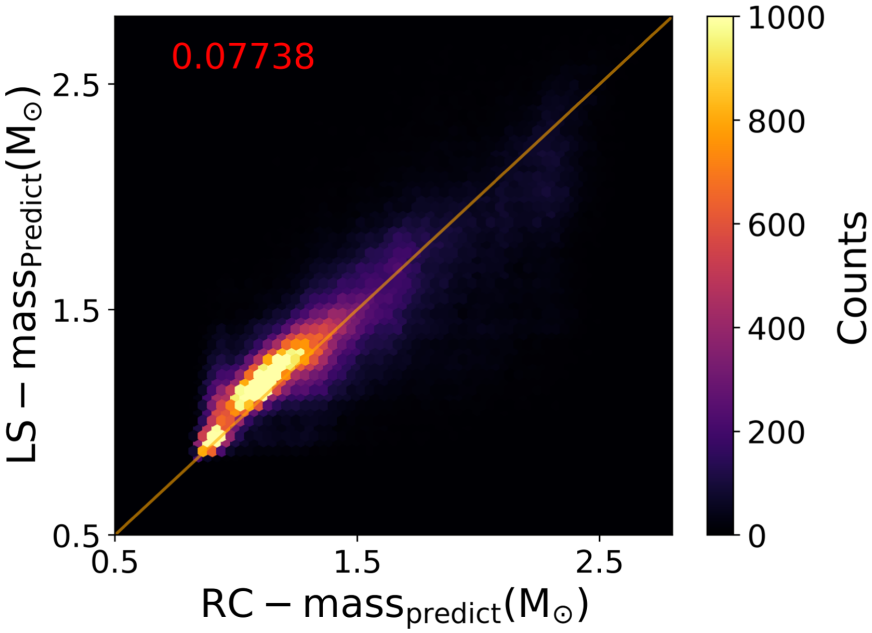

Figure 21. The comparison of common stars between the LS-mass and RC-mass.

There should be no absolute advantages and disadvantages between different machine-learning methods, and each method has its own applications dependent on purpose. After comparisons, we find that the nonlinear model is more in line with our expectations than the linear model, and the GBDT and RF are better. In order to make the model suitable for more stars, we choose the RF, which needs fewer feature numbers to achieve our scientific target in this work.

To some extent, this paper could be considered as the first paper of our series of works and the catalog is available at doi: 10.5281/zenodo.6949334. This method will be widely used for the other catalogs or surveys and we will also attempt to consider systematics and possible zero-points for age in the future.

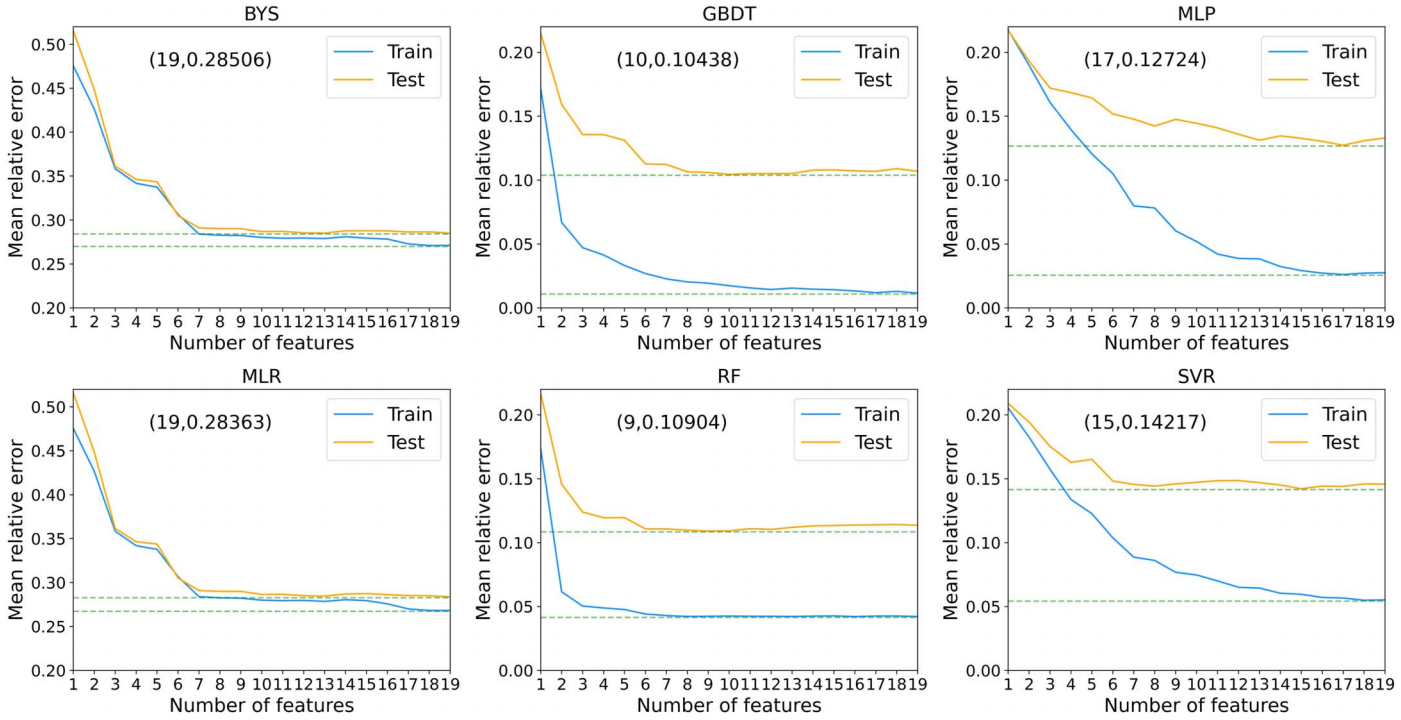

Figure 22. The relations between the number of training features and the mean relative error of the test data set for the six prediction models. Different panels are different machine-learning methods and different color lines are training and test data sets, respectively. The horizontal dashed lines are used to guide our eyes to the stable pattern. Notice that it is the minimum value that is labeled but it is not the final feature we adopt; we choose final features shown in Figure 23 empirically and reasonably.

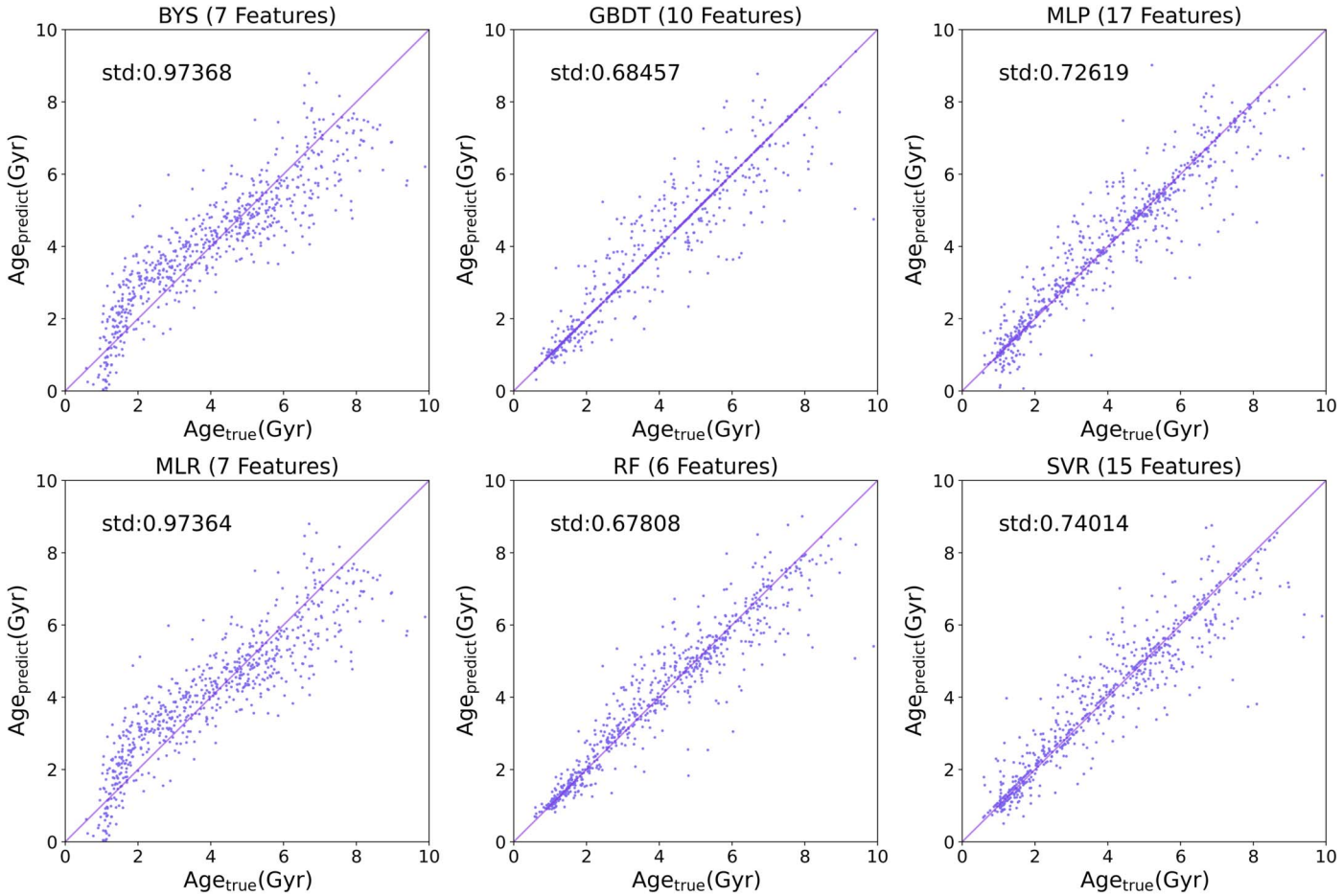

Figure 23. The comparison between our predicted ages and the true values we use in the catalog. Different panels are different machine-learning methods and the dispersions are labeled in each panel and diagonal lines are used for comparison. The final number of features (corresponding to Figure 2) we adopt to train each model is marked at the top of each panel.

We would like to thank the anonymous referee for the very helpful and insightful comments. Thanks also for the helpful comments from Martín López-Corredoira. H.F.W. is supported by the CNRS-K.C.Wong Fellow in France and we acknowledge the science research grants from the China Manned Space Project with Nos. CMS-CSST-2021-B03 and CMS-CSST-2021-A08. H.F.W. also acknowledges the support from the project "Complexity in self-gravitating systems" of the Enrico Fermi Research Center (Rome, Italy). L.Y.P is supported by the National Key Basic R&D Program of China via 2021YFA1600401, the National Natural Science Foundation of China (NSFC) under grant 12173028, the Chinese Space Station Telescope project: CMS-CSST-2021-A10, the Sichuan Science and Technology Program (grant No. 2020YFSY0034), the Sichuan Youth Science and Technology Innovation Research Team (grant No. 21CXTD0038), the Major Science and Technology Project of Qinghai Province (grant No. 2019-ZJ-A10), and the Innovation Team Funds of China West Normal (grant No. KCXTD2022-6).

H.F.W. is fighting for the plan "Mapping the Milky Way (Disk) Population Structures and Galactoseismology (MWDPSG) with large sky surveys" in order to establish a theoretical framework in the future to unify the global picture of the disk structures and origins with a possible comprehensive distribution function. We pay our respects to elders, colleagues, and others for comments and suggestions, thanks to all of them. The Guo Shou Jing Telescope (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope, LAMOST) is a National Major Scientific Project built by the Chinese Academy of Sciences. Funding for the project has been provided by the National Development and Reform Commission. LAMOST is operated and managed by National Astronomical Observatories, Chinese Academy of Sciences. This work has also made use of data from the European Space Agency (ESA) mission Gaia (https://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement.

Appendix
Machine-learning Methods Introduction

MLR is a linear model assuming that there is a simple weighted summation relation between the variable and the predicted parameter. It has good performance in some cases although the assumption is strong. BYS is applied Bayesian inference to the linear regression model. The parameters in the linear model are regarded as random variables and then we can calculate the posterior distribution; it has the basic properties of a Bayesian statistical model. In this experiment the data will be repeated and the overfitting will be prevented effectively, but it is computationally expensive. To be more specific for some details, in this work we do not change most of the default parameters in the sklearn software for MLR and the parameter $\mathrm{fit}_{\mathrm{intercept}}$ is set to be true, which means we could calculate the intercept value for this model. For BYS, we set the parameter $\mathrm{n}_{\mathrm{iter}} = 30$ and $\mathrm{tol} = 1 \times 10^{-3}$, the meaning of $\mathrm{n}_{\mathrm{iter}}$ is the maximum number of iterations and the tol setting could stop the algorithm if it has been converged.

Both RF and GBDT are based on decision trees, the difference is that the former uses bagging and the latter uses boosting. The final predicting result is dependent on the decision trees and is random due to the random sampling at the beginning. The correlation between different parameters can be easily identified with the help of information gain when we are training the model. Moreover, both of these methods have good robustness and overfitting could be avoided. Because of the good robustness, sometimes standardization or normalization might not be needed. Good robustness means that the machine-learning model could have good precision for parameters. For RF, we set the parameters $n_{\mathrm{estimators}} = 2000$, $n_{\mathrm{jobs}} = -1$, $\mathrm{max}_{\mathrm{features}} = \mathrm{auto}$, and $\mathrm{min}_{\mathrm{samples-leaf}} = 1$. Here $n_{\mathrm{estimators}}$ means the number of trees in the forest, n_{jobs} can change the number of jobs in order to run in parallel, $\mathrm{max}_{\mathrm{features}}$ defines the maximum number of features of each tree, and $\mathrm{min}_{\mathrm{samples-leaf}}$ is the minimum number of samples required at the node. For GBDT, we set the parameter $n_{\mathrm{estimators}} = 2000$, $\mathrm{learning}_{\mathrm{rate}} = 0.1$, $\mathrm{subsample} = 1.0$, $\mathrm{loss} = \mathrm{ls}$, $\mathrm{max}_{\mathrm{features}} = \mathrm{none}$, and $\mathrm{min}_{\mathrm{samples-leaf}} = 1$. The $n_{\mathrm{estimators}}$ means the number of boosting parameters, $\mathrm{learning}_{\mathrm{rate}}$ is the weight contribution of each tree, subsample defines the stellar fraction used for fitting the individual learners, the loss setting could optimize the loss function, and ls means least-squares regression in the method.

MLP is a neural-network model consisting of an input layer, hidden layer, and output layer. Each layer is closely connected to the neurons. It is sensitive to overfitting and it difficult to adjust its parameters with the computer time being proportional to the networks. SVR's regression is dependent on the hyperplane constructed from the data sets. Since a supervised-learning method is based on the symmetric loss function for training, one of the advantages is that the computational complexity does not depend on the dimension of the data, but when the dimensions are more than the number of data points the results might not be acceptable. For MLP, we set the parameter hidden-layer sizes = 147; this means there is only one hidden layer with 147 neurons because we found that networks that were too complex would not improve the prediction performance. For SVR, the Gaussian kernel has been used, and we set the parameter $C = 15$ (regularization parameter). Notice that if "C" is too large or too small, the prediction performance will be reduced. More details could be found in the Scikit-learn publicly available package (Anghel et al. 2019; Mediratta & Oswal 2019; Florescu & England 2020).

ORCID iDs

Hai-Feng Wang https://orcid.org/0000-0001-8459-1036
Yang-Ping Luo https://orcid.org/0000-0003-3736-6076
Qing Li https://orcid.org/0000-0001-5049-123X
Yuan-Sen Ting https://orcid.org/0000-0001-5082-9536

References

Anghel, A., Ioannou, N., Parnell, T., et al. 2019, arXiv:1910.06853
Brewer, L. N., Sandquist, E. L., Mathieu, R. D., et al. 2016, AJ, 151, 66
Bergemann, M., Ruchti, G. R., Serenelli, A., et al. 2014, A&A, 565, A89
Bland-Hawthorn, J., Krumholz, M. R., & Freeman, K. 2010, ApJ, 713, 166
Bland-Hawthorn, J., Sharma, S., Tepper-Garcia, T., et al. 2019, MNRAS, 486, 1167
Bragaglia, A., Tosi, M., Andreuzzi, G., et al. 2006, MNRAS, 368, 1971
Buder, S., Lind, K., Ness, M. K., et al. 2019, A&A, 624, A19
Ceillier, T., van Saders, J., García, R. A., et al. 2016, MNRAS, 456, 119
Chaplin, W. J., Basu, S., Huber, D., et al. 2014, ApJS, 210, 1
Edvardsson, B., Andersen, J., Gustafsson, B., et al. 1993, A&A, 500, 391
Florescu, D., & England, M. 2020, arXiv:2005.11251
Freeman, K., & Bland-Hawthorn, J. 2002, ARA&A, 40, 487
Gai, N., Basu, S., Chaplin, W. J., et al. 2011, ApJ, 730, 63

García, R. A., Ceillier, T., Salabert, D., et al. 2014, A&A, 572, A34
Geller, A. M., Mathieu, R. D., Harris, H. C., et al. 2008, AJ, 135, 2264
Grundahl, F., Clausen, J. V., Hardis, S., et al. 2008, A&A, 492, 171
Hawkins, K., Ting, Y.-S., & Walter-Rix, H. 2018, ApJ, 853, 20
Haywood, M., DiMatteo, P., Lehnert, M. D., et al. 2013, A&A, 560, A109
Ho, A. Y. Q., Rix, H.-W., Ness, M. K., et al. 2017, ApJ, 841, 40
Huang, Y., Schönrich, R., Zhang, H., et al. 2020, ApJS, 249, 29
Jacobson, H. R., Pilachowski, C. A., & Friel, E. D. 2011, AJ, 142, 59
Janes, K., Barnes, S. A., Meibom, S., et al. 2013, AJ, 145, 7
Mackereth, J. T., Miglio, A., Elsworth, Y., et al. 2021, MNRAS, 502, 1947
Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94
Martig, M., Fouesneau, M., Rix, H.-W., et al. 2016, MNRAS, 456, 3655
Mathur, S., Huber, D., Batalha, N. M., et al. 2017, ApJS, 229, 30
McQuillan, A., Mazeh, T., & Aigrain, S. 2014, ApJS, 211, 24
Mediratta, D., & Oswal, N. 2019, arXiv:1911.01217
Ness, M., Hogg, D. W., Rix, H.-W., et al. 2015, ApJ, 808, 16
Ness, M., Hogg, D. W., Rix, H.-W., et al. 2016, yCat, J/ApJ/823/114
Nordström, B., Mayor, M., Andersen, J., et al. 2004, A&A, 418, 989
Pinsonneault, M. H., Elsworth, Y. P., Tayar, J., et al. 2018, ApJS, 239, 32
Sanders, J. L., & Das, P. 2018, MNRAS, 481, 4093
Silva Aguirre, V., Stello, D., Stokholm, A., et al. 2020, ApJL, 889, L34
Soderblom, D. R. 2010, ARA&A, 48, 581
Stello, D., Saunders, N., Grunblatt, S., et al. 2022, MNRAS, 512, 1677
Stello, D., Vanderburg, A., Casagrande, L., et al. 2016, ApJ, 832, 133
Takeda, G., Ford, E. B., Sills, A., et al. 2007, ApJS, 168, 297
Ting, Y.-S., Conroy, C., Rix, H.-W., et al. 2019, ApJ, 879, 69
Ting, Y.-S., Hawkins, K., & Rix, H.-W. 2018, ApJL, 858, L7
Ting, Y.-S., & Rix, H.-W. 2019, ApJ, 878, 21
van Saders, J. L., Ceillier, T., Metcalfe, T. S., et al. 2016, Natur, 529, 181
Wang, H.-F., Carlin, J. L., Huang, Y., et al. 2019, ApJ, 884, 135
Wang, H. F., Hammer, F., Yang, Y. B., et al. 2022b, arXiv:2205.02306
Wang, H.-F., Huang, Y., Zhang, H.-W., et al. 2020c, ApJ, 902, 70
Wang, H.-F., Liu, C., Xu, Y., et al. 2018b, MNRAS, 478, 3367
Wang, H.-F., López-Corredoira, M., Carlin, J. L., et al. 2018a, MNRAS, 477, 2858
Wang, H.-F., López-Corredoira, M., Huang, Y., et al. 2020a, MNRAS, 491, 2104
Wang, H.-F., López-Corredoira, M., Huang, Y., et al. 2020b, ApJ, 897, 119
Wang, H.-F., Yang, Y.-B., Hammer, F., et al. 2022a, arXiv:2204.08542
Wu, Y., Xiang, M., Zhao, G., et al. 2019, MNRAS, 484, 5315
Xiang, M.-S., Liu, X.-W., Yuan, H.-B., et al. 2015, RAA, 15, 1209
Xiang, M., Liu, X., Shi, J., et al. 2017, ApJS, 232, 2
Xiang, M., Ting, Y.-S., Rix, H.-W., et al. 2019, ApJS, 245, 34
Yang, P., Wang, H.-F., Luo, Z.-Q., et al. 2022, arXiv:2205.09227
Yu, J., Huber, D., Bedding, T. R., et al. 2018, ApJS, 236, 42
Yu, Y., Wang, H.-F., Cui, W.-Y., et al. 2021, ApJ, 922, 80
Zasowski, G., Johnson, J. A., Frinchaboy, P. M., et al. 2013, AJ, 146, 81
Zhang, B., Liu, C., & Deng, L.-C. 2020, ApJS, 246, 9
Zhang, B., Li, J., Yang, F., et al. 2021, ApJS, 256, 14