Astronomical and Astrophysical Transactions (*AApTr*), 2024, Vol. 34, Issue 4, pp. 389–414,
ISSN 1055-6796, Photocopying permitted by license only, ©Cambridge Scientific Publishers

An adaptive algorithm for detecting binary stars in astrometric surveys

Mikhail V. Sazhin[1,†], Valerian Sementsov[1,*], Sergey Sorokin[2], Dan Lubarskiy[3], Alexander Raikov[4]

[1] *Sternberg Astronomical Institute, Lomonosov Moscow State University, Universitetsky pr., 13, Moscow 119234, Russia*
[2] *Tver State University, Zhelyabova 33, Tver, Russia*
[3] *Scienteco, Inc., Boston, MA, USA*
[4] *Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*

Received 8 November, 2024; accepted 28 November, 2024

The paper develops a method for detecting extremely close visual double stars based on the use of astrometric catalogues in combination with machine learning (ML) methods. The low probability of random occurrence of such close neighbours allows us to speak about the physical coupling of the detected stars. These stars can therefore be considered binary. A computational experiment was carried out on the example of the Hipparcos mission catalogue and the Pan-STARRS (PS1) catalogue by applying the suggested method. It has shown that the reliability of predicting a stellar multiplicity reaches 90–95%. We note the prospects and effectiveness of creating a proprietary research platform – Cognotron.

Keywords: Astrometry, catalogues, (stars:) binaries: general, methods: data analysis

1 Introduction

The purpose of this work is to increase the accuracy of estimating the ratio of the number of binaries over single stars by applying artificial intelligence (AI) methods to the classical astronomical techniques.

Binary stars form dynamic systems that rotate under the gravitational attraction around a common centre of masses. Binary stars are subdivided into visual binary stars, astrometric binaries, spectral binary stars, eclipsing variable stars etc. (Hilditch, 2001, pp. 17–21). Methods of their detection are, therefore, divided into astrometric, spectral, and photometric. The development of classical methods of astronomical measurements calls for increasingly complex models, which may include more than ten free parameters per object. As a result, the proportion of discovered

*Email: valera@sai.msu.ru

stars demonstrating non-linear motions is constantly increasing and, accordingly, the risk of an erroneous estimation of the ratio of binary and single stars is growing. In addition, the classical methods of processing astrometric measurements suffer from the "curse of dimensionality" when the computational costs to determine the non-linear and free parameters grows exponentially with respect to their number.

In this paper, it is proposed to supplement the classical methods of estimating the aforementioned ratio with machine learning (ML) methods. The latter allow us to work with a set of poorly defined parameters and reduce the exponential growth of the volume of calculations, arising from their number, to polynomial. However, modern ML methods, as it is well known, do not always enable us to get an explanation of the results obtained. For the purpose of this paper, this is not an obstacle, since the verification of the results derived from the data by ML, as in the classical case, is carried out based on direct observations.

Applications of ML have an appealing potential, which is constantly increasing, primarily due to the development of ML methods and analytical methods of processing big data, as well as the growth of computer power. New versions of ML are being introduced, and the computing capabilities for their implementation are increasing. The development of methods is heading in the direction of providing opportunities to self-study, adapt to the dynamics of the external environment, solve interdisciplinary tasks, plan, analyse, give explanations, consider the subjective, non-local and wave aspects of the behaviour of the research objects (Wang et al. (2018); Raikov (2021)).

As an example of a successful application of ML in astronomy, one can mention the work of Becker et al. (2020), presenting a classifier of astronomical events in real time for Automatic Learning System for the Rapid Classification of Events (ALeRCE). The article of Carrasco-Davis et al. (2021) discusses photometric and spectroscopic observations of rapidly variable sources formed after the explosion of an astronomical object. To classify signals from the nuclei of galaxies, supernovae, asteroids, etc., Convolutional Neural Network (CNN) is used. At the same time, metadata is added to astronomical images in the form of a priori known functions and indicators, which helps to achieve a high level of accuracy (≈ 94 %). There are two types of classification methods – based on a template and a light curve. The first is able to distinguish a richer taxonomy of events, the second uses only the first event warning. The classification of events by templates is based on the use of CNN (Carrasco-Davis et al. (2021)). CNN input requires images and metadata about the properties of objects from various catalogues. The standard shape of the event template is 63×63 pixels. The list of alert metadata includes about 15 indicators.

The authors of ALeRCE (Förster et al. (2021)) proposed a template-based classification method for distinguishing five different classes of events within the framework of ALeRCE, the detection and alert of supernova explosions (SNE), using prediction and distinguishing between SNE, as well as other complex classes of events. Template classification is necessary for morphological differentiation of galactic nuclei, SNEs, stars, asteroids and false warnings. It uses rotational invariance of images. CNN is trained using entropy data and additional information that experts could use to evaluate candidates. Modern astronomical instruments are able to assess the level of chaos caused by the explosion of objects (Reyes et al. (2018)), to estimate the size of a companion star (Jiang et al. (2017)), to recognize, annotate and classify big data obtained from survey telescopes.

The article is structured as follows. First, the Hipparcos catalogue is discussed, as well as an overview of the labour of the detection of binaries in this catalogue. At the same time, the general principles of allocation of unresolved binary stars in the astrometric survey are discussed. Following this, a review of a calculation experiment using machine learning and AI methods in astronomical research is described. At the same time, considering the features of ML, significant and control signs of duality are distinguished, these signs are selected classifying single and binary stars. The stability of the constructed classification to changes of the observational selection of the training sample is checked. In conclusion, an estimate of the proportion of binary stars to single stars in the catalogue is given.

2 Ground-based observations and cataloguing of binary stars

The idea of the existence of physical binary and multiple star systems in the Universe was first considered by Michell (1767). He applied then new statistical methods to the study of stars and demonstrated that many more stars occur in pairs or groups than a random distribution can explain. For the Pleiades cluster, Mitchell estimated that the probability of such a close group of stars is $5 \cdot 10^{-5}$. He concluded that stars in such binary or multiple star systems can attract each other, which is the first proof of the very fact of the existence of binary stars and star clusters. His work on binary stars may have influenced William Herschel's research on the same topic, which took shape in the first catalogue of binary stars (Herschel (1785)).

2.1 Modern astrometric, photometric, and spectral methods for determining the multiplicity of stars

Identifying objects with a small angular distance as binary stars was initially proposed (the so-called visual double stars, whose random proximity in the sky is realizable with very little probability). Later it was supplemented by other methods, primarily astrometric, analysing the proper motions of individual stars or pairs of stars (see e.g. Makarov and Kaplan (2005)). Long-term observations of optical and astrometric binaries make it possible to determine the orbits of the components in some cases. We are currently talking about catalogues with hundreds of such objects (Söderhjelm (1999); Hartkopf et al. (2001)), with the exception of Gaia DR3, which contains approximately 135,000 systems (van Leeuwen et al. (2021)). The percentage of detected stars (roughly one thousand double stars out of four million stars in the Astrographic Catalogue compared to 150 thousand double stars out of two billion stars in Gaia) shows that the problem is only getting worse. This is the only direct method of determining the physical mass of stars.

Astrometric methods are good for investigating sufficiently wide star pairs. For closer components, which are unresolved, astrophysical methods (photometric or spectroscopic) are more effective.

In most cases, the duality of an object is revealed using criterion similar to Rayleigh's criterion (Arenou and Söderhjelm (2005); Griffin and Griffin (1986)). Optical binary stars were discovered by applying Rayleigh criterion: if the distance between the components exceeds the half-width of the so-called point spread function

(PSF), according to earlier methods, one considers the star to be optical binary. It is essential that for more than a hundred years the accuracy of determining the astrometric parameters of stars has been much better than PSF. This is related to the development of astrometric methods for detecting binary stars based on the features of their proper motions. Our work is also aimed at clarifying the criteria for detecting the duality of astrometric binary stars, which would be orders of magnitude more sensitive than the Rayleigh criterion.

2.2 Binary star catalogues, multiplicity in stellar surveys

Catalogues of double stars have been published since the end of the XVIII century (Herschel (1785)). The development of the situation is shown in Table 1 for the following publications: Burnham (1906); Aitken and Doolittle (1932); Jeffers et al. (1963); Lipaeva et al. (2014); Worley C.E. (1984); Worley and Douglass (1997); Mason et al. (2001, 2021); Dommanget and Nys (1994); Fabricius et al. (2002). It is easy to see that the growth of number of discovered binary systems is quite moderate: at the beginning of the XX century, astronomers used visual catalogues containing a total of about a million stars (Schönfeld (1886); Gill and Kapteyn (1896, 1897, 1900); Argelander (1903)), in the second half of the century, a photographic Carte du Ciel with about 4.5 million stars (Eichhorn (1980); Sémirot (1972); Stock and Cova S. (1983)) became available electronically, the development of space astronomy required catalogues of tens of millions of objects (Lasker et al. (1990); Russell et al. (1990); Jenkner et al. (1990); Morrison et al. (2001)), and in the XXI century, electronic versions of photographic catalogues with a volume of about a billion stars were introduced to the scientific community (Lasker et al. (2008); Monet et al. (2003)). During this period, the volume of double star catalogues has grown only a few times. We intentionally do not include the Gaia results in this review because it is meaningless to expect that a machine learning method that used for its tuning a sample of 100,000 stars observed 35 years ago on a 30-cm telescope would outperform the shaft of a modern survey of 2 billion stars on a much better instrument. It was mentioned above that the basic problem of increasing the percentage of detectable binaries cannot be solved

Table 1 Catalogues of double stars.

ID	Year	Total number of multiple systems	Total number of individual components	References
Hershel	1785	434		Hershel (1785)
BDS	1906	13665		Burnham (1906)
ADS	1932	17180		Aitken and Doolittle (1932)
IDS	1961	56572 (29965)	69819 (36861)	Jeffers et al. (1963), Lipaeva et al. (2014)
WDS	1994	73610	154333	Worley (1984), Worley and Douglas (1997), Mason et al. (2001, 2021)
CCDM	1994,2002	34031	74861	Dommanget and Nys (1994)
Tycho-3	2001	32631	103259	Fabricius et al. (2002)

by increasing the aperture.

The above situation is typical of survey catalogues limited to the weakest observed stellar magnitude. Most of the stars in such a catalogue as Kharchenko (2001) are slightly brighter than the detection limit and identifying duality in this case is problematic.

A survey of the star catalogues, compiled according to a somewhat different principle: identification of all objects within a given volume of space (Gliese and Jahreiß (1991)) significantly changes the statistics (Bastian et al. (2010); Duquennoy et al. (1991); Duquennoy and Mayor (1991)). Dimmer objects of late spectral classes are beginning to prevail in the samples of the nearest to the Sun stars, and the proportion of stars with signs of duality is approaching 50%.

2.3 Theoretical models of the emergence of multiple stars, the percentage of binary stars in the vicinity of the Solar System

There is no common opinion in the research of the mods of the collapse of protostellar clouds yet. For the appearance of single stars, theoretical models are well-regarded, and a sufficiently convincing initial mass function is obtained, which then gives reasonable results in further population calculations. So, according to Kroupa (2002), there are the following equations for different masses of stars:

$$\xi(M) = \begin{cases} k_0 \left(\frac{M}{m_0}\right)^{-\alpha_0} & , \quad m_0 < M \leqslant m_1 \\ k_1 \left(\frac{M}{m_1}\right)^{-\alpha_1} & , \quad m_1 < M \leqslant m_2 \\ k_2 \left(\frac{M}{m_2}\right)^{-\alpha_2} & , \quad m_2 < M \end{cases}$$

where $m_0 = 0.01 m_\odot$, $m_1 = 0.08 m_\odot$, $m_2 = 0.5 m_\odot$, $\alpha_0 = 0.3$, $\alpha_1 = 1.3$, $\alpha_2 = 2.3$. The last equation is the initial mass function of Salpeter (1955).

There is no generally accepted model of binary star formation yet. There are several theoretical models of the collapse of protostellar clouds with an initial rotation. In such models gas compression must lead to the formation of a toroidal structure. This structure then decays into separate protostars, which form a multiple star system (in the simplest case, a binary one). Taking into account various physical mechanisms (for example, the degree of influence of the magnetic field) leads to significantly different results, none of which is fully confirmed by observations.

Our study of binary stars statistics is carried out under conditions (Gliese and Jahreiß (1991)) of extremely high probability of duality of objects (about 50%) and the absence of a generally accepted astrophysical theory.

3 Hipparcos main catalogue

The Hipparcos spacecraft became the first specialized astrometric satellite. High-precision optical measurements require a long-focus instrument, which therefore has a small field of view.

3.1 Observations technique and the structure of the catalogue. Sources of information and data on multiplicity

The task of carrying out high-precision measurements throughout the entire celestial sphere determined the instrument design with two fields of view with a diameter of about 0.9° each, spaced from each other by ≈ 58°. The device rotated with a period of 120 minutes around an axis perpendicular to the plane in which the entrance pupils lie (Perryman et al. (1989a)), and the axis itself (according to the observation plan) slowly precessed along the cone of 43° around the direction to the Sun. All these movements combined led to a uniform coverage of the observations of the celestial sphere (according to Perryman et al. (1989a)) and, on the other hand, did not allow the satellite's solar panels to deviate significantly from the direction of the Sun and lose power. On the third side, due to the obvious anisotropy of the directions of scanning the celestial sphere, this will limit the detection of very close pairs of stars, but this is already a specific feature of the observational material.

To improve the geometric stability of the measuring instrument and statistical performance, it wasn't just the passage of a star in the field of view of an electric vacuum device that was being recorded, but its passage through a lattice of 2688 slits. The recorded coordinates thus ended up being one-dimensional, they were tied to a great circle, that remained the same for several revolutions of the satellite. During the processing stage, the parameters of the circles were linked to each other for the entire celestial sphere, then the spherical coordinates of individual stars, their parallaxes and proper motions were calculated (Perryman et al. (1989b)).

The results of the satellite observation (ESA (1997)) have shown the accuracy of coordinates better than of ground observations (Perryman et al. (1989c)) by about 100 times. During the experiment, this made it possible to determine the coordinates alongside parallaxes and proper motions with high accuracy of about 1 msec of arc. That is, an experiment of 3.5 years yielded a result comparable to the astrometric activity of an entire century and in some cases (in terms of high-precision parallaxes), even surpassing the results. We have to note that, strictly speaking, this is correct only of the objects of the Hipparcos program, with respect to a posteriori confirmation of the accepted source model.

The duality of objects in the Hipparcos main HIP catalogue was initially established by their belonging to the Catalogue of Components of Double and Multiple Stars (CCDM) (Dommanget and Nys (1994)). At the same time, new optical binaries reliably detected during the observations, separated by the Rayleigh-type criterion, were added to this catalogue (about 5% of its volume).

3.2 Further work on the detection of double stars in HIP

After the publication of the main Hipparcos catalogue and the additional Tycho catalogue to the original program, their in-depth research began. The latter catalogue was obtained based on observations not of the main Hipparcos spacecraft photodetector, but of the signals from service stellar sensors, initially intended to monitor the appearance of an object that is a part of the observation program in the field of view immediately before the start of the main measurements. This additional material allows us to get the coordinates of about a million stars with an accuracy of an order of one to one and a half magnitude worse than in the main HIP catalogue.

Considerable efforts of researchers were aimed at improving the accuracy of the Tycho catalogue' proper motions and increasing its size based on satellite observations records and the use of Astrographic catalogue, so Tycho-2 (Høg et al. (2000)) catalogue appeared. Later on, this information formed an additional catalogue of binary stars Tycho (Fabricius et al. (2002)) and the discovery of new binary stars in the main HIP catalogue (Makarov and Kaplan (2005)). We will use the latter list below to test the operation of machine learning algorithms.

3.3 General principles of detection of unresolved binaries in an astrometric survey

It is apparent that many different criteria for detecting the duality of an object by measuring any one parameter, such as the ellipticity of the visible image, the deviation of proper motion from a straight line, anomalous photometry, bifurcation of spectral lines have already been used, generally speaking, by the authors of observations or their closest followers and cannot yield anything new. The general principle of duality detection turns out to be similar: the Rayleigh criterion for spectral lines in the second case or the image ellipticity significant compared to the point spread function (PSF) in the first case are similar and do not allow duality detection if the separation (or anomalous proper motion) of the components is smaller than the errors of the spectral line width or the same PSF.

On the other hand, it is obvious that the optical separation of the components of the assumed binary at distances greater than the errors of the coordinate measurements should somehow manifest itself in the results of the observations. Our work consists of verifying the fact that for unresolved, as well as for resolved, binary stars, anomalies in the errors of the measured values will appear.

The proposed method relies on using the widest possible set of data for each object. The next section shows how, based on a training sample, various artificial intelligence algorithms are trained, significant signs of duality and signs of negligible significance are highlighted. An additional check is also performed on the sampling effect, the influence of the chosen machine learning method, the influence of the observational selection in the training sample and the dependence on the imperfection of the reduction procedure of the main catalogue. In addition, some of those parameters of stars that should not depend on duality are to be analysed for the purpose to check the algorithms used.

4 Application of ML methods for detection of binary star systems

In this work we present the results of applying modern ML methods and AI systems to the data of HIP catalogue to increase the quality of the data received via the astrometric satellite. More specifically, we used neural networks and decision tree-based models.

4.1 Feature extraction

The HIP catalogue contains many characteristics of stars from the main list (there are 77 data fields in the catalogue (ESA (1997))). From them, we excluded fields like references to other catalogues or data sources. We also excluded astronomical coordinates, because on the one hand, the catalogue covers only a small volume of the Galaxy in a close neighbourhood of the Sun, where the structure of the Galaxy is not prominent. On the other hand, exact values of star coordinates are enough to uniquely identify a star, so the ML models will just remember which stars are binary and which are not, without trying to extract dependencies from other fields.

The spectral class field of a star is presented in the HIP Catalogue as a text field, which prevents its direct use as an input to ML models. To use information from this field, the text format was converted to a synthetic (manually created) feature vector which describes the spectral data in a format, suitable for ML models. A complete set of features of this vector is shown in Appendix A. Thus, the catalogue data was mapped in a space of 73 features, including 34 numeric values taken directly from the HIP catalogue and 39 features from the spectral classification fields. The table showing all catalogue fields used in the analysis is shown in Appendix B.

The duality flag obtained by the authors of the catalog (H61 in ESA (1997)) at this stage is excluded from consideration because it is a value derived in an undescribed way and depends on other parameters of the star that we consider.

Non-empty value of CCDM field of HIP catalogue was used as a target feature.

4.2 Data analysis models

In the AI research platform we have experimented with two kinds of models to process data: ensemble of neural networks and ensemble of XGBoost (eXtreme Gradient Boosting, Chen and Guestrin (2016)) decision-tree trained models.

Neural network ensembles were built from feed-forward networks composing of two fully-connected hidden layers of 200 and 100 neurons with ReLU (Rectified Linear Unit) activation function, followed by an output layer of a single neuron with a sigmoid activation function. Before feeding data to the neural network, it was passed through a normalization layer. Networks were trained using adaptive moment estimation (Adam) algorithm (Kingma and Ba (2015)), optimizing for a binary cross-entropy metric. Neural networks were implemented and trained using Keras (Chollet et al. (2015)) library, included in Tensorflow deep learning framework (Abadi et al. (2015)).

To create an ensemble of neural networks, we split the original HIP catalogue dataset into 25 random subsets, while preserving the proportion of target multiplicity feature in each subset. This was done using StratifiedShuffleSplit function from sklearn library (Pedregosa et al. (2011)). Then, each single subset was used as a training datum for a neural network, while the remaining 24 subsets where jointly used as a test set. Those 25 networks were grouped into an ensemble, which was used to calculate the final output value for all stars in the HIP Catalogue.

An ensemble of decision-tree models was trained using the same technique. Base models were created using XGBClassifier implementation of the extreme gradient boosting algorithm from XGBoost library (Chen and Guestrin (2016)). The following parameters were used for each of 25 classifiers: `learning_rate=0.01`; `n_estimators=`

1811; max_depth=6;min_child_weight=4; gamma=0.4; subsample=0.9; colsample_bytree=0.8; objective= 'binary:logistic'; nthread=4; scale_pos_weight=1; seed=27. Parameters essential for learning were found by a cross-validation method.

Since there is reason to believe that not all multiple systems are marked in the Hipparcos catalogue, we propose to use an approach that is often used in data analysis to identify labelling errors, to identify candidates for binary systems: train a classifier on an initial (incomplete in terms of labelling) data set and then consider objects with the maximum error as candidates for incorrect labelling (Brodley and Friedl (1999); Zhu et al. (2003); Angelova et al. (2005); Huang et al. (2019)).

In the context of our task this means that we are looking for stars, which are not marked as binary in Hipparcos catalogue, but, at the same time, show a high probability of being binary based on the results of the ML models, trained on the datasets from the catalogue. For such an object, it holds that, despite the fact that it was used in the training process as "not a binary star", i.e. the ML model was instructed that the probability of its multiplicity equals 0, the ML algorithm, nevertheless, insists on its duality, based on the patterns it derives from the data.

4.3 Importance of features required for classifying a star as binary

The XGBoost library provides a means to access various model statistics, including the significance of features for the trained classifier. Figure 1 shows that various statistical parameters from the catalogue are significant, while, for example, the spectral parameters are not. It means that ML algorithms confirmed the well-known from the catalogues of double and multiple stars result, that duality of a star is only loosely correlated with the spectral characteristics of the pair. Thus, we can conclude that training dataset used is consistent in this regard.

Importance score shown on y-axis of Fig. 1 represents the number of times the feature was used as a decision variable in trees, averaged over ensemble members. Our calculation experiments show that this feature is consistent across ensemble members

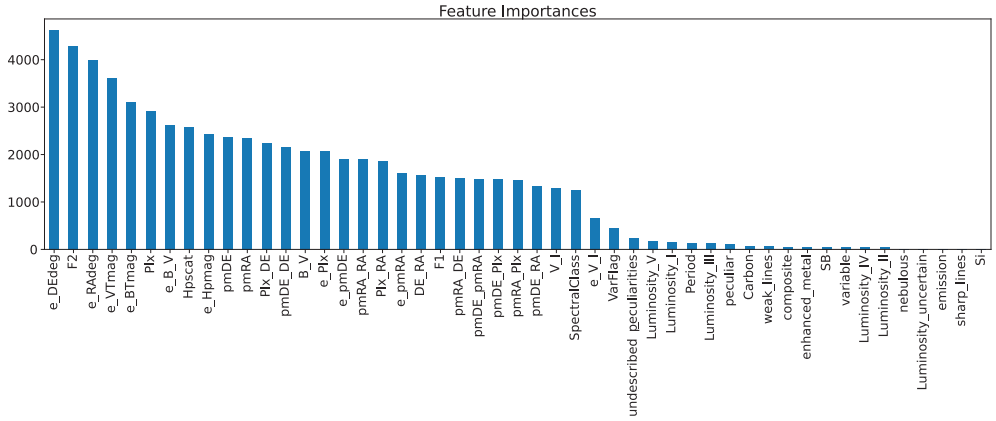


Figure 1 Features from the fields of Hipparcos catalogue ranked by importance to produce the probability of duality of a star.

and resilient to changes at random stages of algorithm training, like a split of the ensemble members into individual training groups, seed parameter of XGBClassifier.

4.4 Robustness of the classification algorithm

The training set for ML algorithms consists a subcategory of catalogue' stars in ESA (1997) assigned a non-empty value of CCDM (Dommanget and Nys (1994)) field, i.e. the star is recognized as a double. The very procedure for establishing multiplicity based on the anomalous proximity of neighbouring stars (see above) is subject to a very strong observational selection: multiplicity according to catalogues (Dommanget and Nys (1994); Mason et al. (2021)) for bright stars (of which there are few) turns out to be much more probable than for more numerous faint stars. In order to test the influence of this effect on the work of ML algorithms, an independent training was carried out, during which photometric values from model input were excluded. The specific parameters that were used at this stage are given in Appendix B. The star colour indices and photometric errors were retained as input parameters: the former, to control the adequacy of the classification algorithm; the latter, as one of the indicators of the possible duality of the object.

The results of the work of the Hipparcos consortium published in ESA (1997) included the solution of a system of non-linear reduction equations for determining the kinematic parameters of stars. The system of equations was solved by the iteration method. One iteration in 1995–1996 took about half a year of calculations, and it was this iteration that was subsequently published. After 10 years, one of the members of the consortium repeated the processing of the observational material that had been preserved, somewhat improving the reduction scheme and by increasing the number of iterations (at that time, one iteration took about a week of computing time). The resulting new Hipparcos reduction was published (van Leeuwen (2007b,a)) and

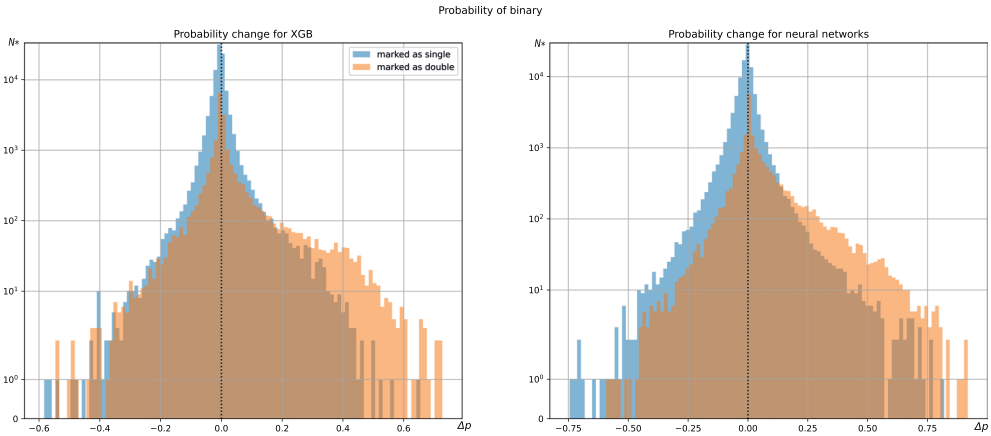


Figure 2 Two histograms of changes of probabilities of classifying a star as a single or double by different ML algorithms are plotted. The x-axis shows the differences in the probability of a certain classification of a star in ML for samples of different sizes (24/25 and 1/25 of the entire training sample, respectively).

showed slightly better accuracy, especially for bright stars. It has not been accepted as a coordinate standard, but is actively being used in scientific research.

To test the approach proposed in this paper, this new reduction is important, since it allows one to check the stability of ML algorithms to the non-Gaussian nature of the errors of the parameters being analysed. Since the convergence of the solution of the system of non-linear equations was not achieved in the HIP catalogue (it was only possible to check the absence of noticeable divergence), the errors of the quantities being determined will inevitably have a non-zero mean, which indicates that they will be non-Gaussian.

The training of ML models was carried out using data from the original HIP catalogue and data from its new reduction together. A specific list of fields taken from the new reduction (van Leeuwen (2007a)) is given in Appendix C. Also, two Boolean fields were added, indicating the presence of a 7- and 9-component solution (the presence of a star in the **hip7p.dat** and **hip9p.dat** tables from van Leeuwen (2007a)). The specific values of these solutions were not used, since they are available only for a small part of the catalogue stars, and it is unlikely that they will be effectively used by the ML models in this regard.

The experiment with machine learning based on the material in van Leeuwen (2007a) showed stability of the values of the main statistical characteristics in the catalogue for the purpose of classifying objects as binary stars (Fig. 2). The resilience of the results to the sampling effect was also tested.

4.5 Prediction of duality probability of HIP catalogue stars by ML models

Since the output of the proposed models is the probability of duality for each of the stars of the Hipparcos catalogue, it is natural to consider stars, for which probability exceeds a certain threshold as candidates for binary stars. Figure 3 shows how many new candidates for binary systems can be identified using the proposed models for different threshold values. The solid lines correspond to binary system candidates identified in comparison with the labelling of the original HIP catalogue, and the dotted lines correspond to those found in work published by Makarov and Kaplan (2005). The graph at the bottom of the figure shows what the percentage of multiple systems in the catalogue will be at different values of the probability threshold.

As outputs of two models – ensemble of neural networks and ensemble of decision trees differ from each other, it is a natural next step to group these models together into one joint ensemble including models of both types. In Fig. 3, along with the results of the ensemble of neural networks and the ensemble of decision trees, the results of the combined model are also presented.

The data in Fig. 3 correspond to the results obtained when training models using the data from both the original HIP catalogue (Perryman et al. (1997)) and the new reduction (van Leeuwen (2007a)), using a set of variables that limits the effect of observational selection. The methodology used in this work does not imply the identification of all binary star candidates in one run. Solving the problem of data analysis under conditions of partially incorrect labelling (Brodley and Friedl (1999); Zhu et al. (2003); Angelova et al. (2005); Huang et al. (2019)) implies an iterative process, during which after identifying the most probable labelling errors (in our case, previously unidentified double stars), it is necessary to check the corresponding

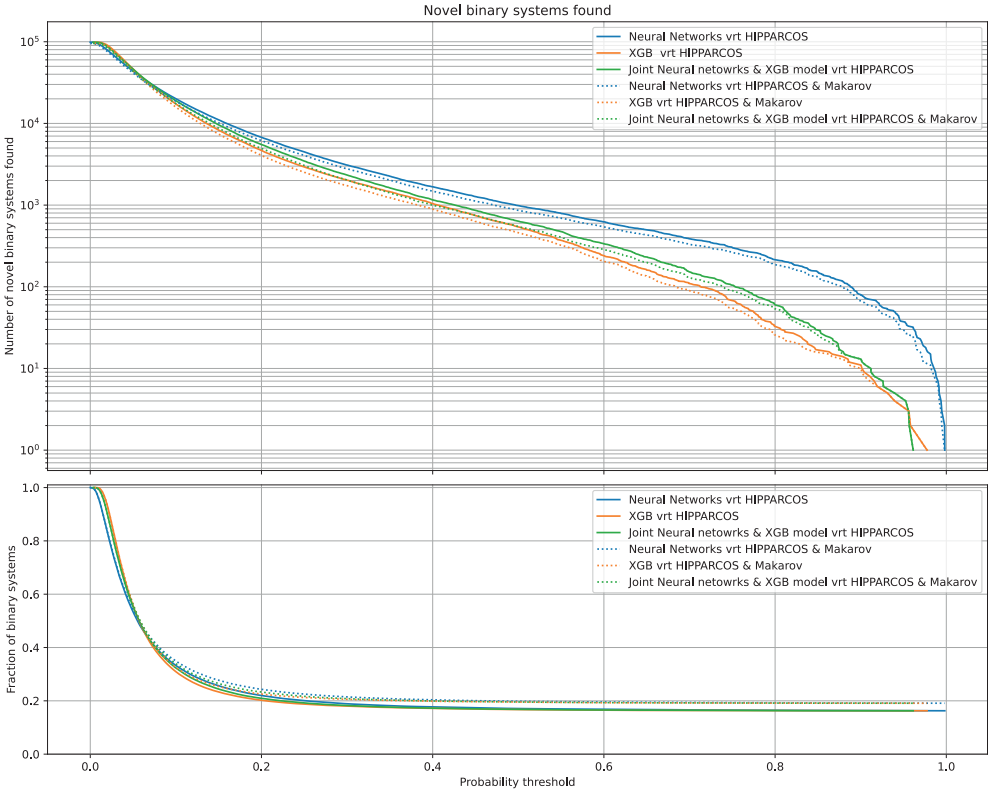


Figure 3 Number of candidates to double systems as a function of threshold on model probability.

objects, correct the labelling and repeat training. This cycle can be repeated several times until a satisfactory result is achieved. This article presents the first iteration of this process.

4.6 Comparison of ML methods' results with other publications

The comparison was carried out on the most extensive work on the identification of additional double stars in the HIP catalogue – the paper by Makarov and Kaplan (2005), where the astrometric method was used, with additional information from the Tycho-2 catalogue (Høg et al. (2000)).

Figure 4A (left panel) shows the probabilities computed by our ML models for stars identified as double in Makarov and Kaplan (2005). The axes of the graph correspond to probabilities of duality calculated by the ensemble of neural networks and the ensemble of tree classifiers.

In Fig. 4B (right panel) there are also stars that are not designated as double either in Makarov and Kaplan (2005) or in the original Hipparcos catalogue along with the stars identified in Makarov and Kaplan (2005), with a multiplicity probability estimates of $> 60\%$ from the ML models.

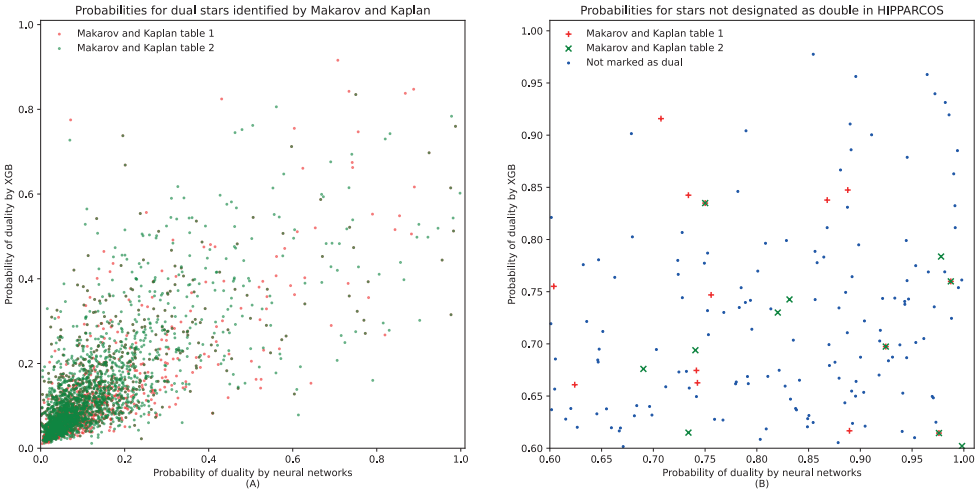


Figure 4 Comparison of ML algorithms' results with Makarov and Kaplan.

Analysing Fig. 4 (left panel), one can see that most of the multiple stars identified in Makarov and Kaplan (2005) received low probabilities of duality when assessed using the ML system. On the other hand, Fig. 4 (right panel) shows that most of the stars proposed as candidates for duality by the ML models were not identified in Makarov and Kaplan (2005). Thus, we can conclude that the patterns identified by machine learning models and the corresponding candidate stars differ from the results of Makarov and Kaplan (2005). This is not surprising, since in the mentioned work the basis of the method of detection of double stars is the use of additional information from the catalogue Tycho-2.

In this regard, it may be of interest to include the binary stars identified in Makarov and Kaplan (2005) in the training set, to train and study the resulting models.

4.7 Observation-based model verification

To verify the obtained results of applying ML to the objects of the Hipparcos catalogue, the lists of the most probable double stars were compared to the double stars from the Pan-STARRS catalogue of objects (Chambers et al. (2017)). The search for neighbours was carried out in the vicinity of the Hipparcos star with a radius of $5''$. This is the size of the working area of the photodetector that took the measurements.

Thus, the developed mechanism gives a 90–95% probability of correct prediction of duality (recall that a priori the probability of duality of a randomly selected star is about 50%).

Table 2 Identification of double star candidates in PS1.

Selection criteria	Number of objects	Multiplicity found (components found)	Components not found	Star not found*
NN, $p > 0.8$	214	142 (895 stars)	10**	72
XGB, $p > 0.7$	109	69 (430 stars)	5***	50

Notes:
*The catalogue was obtained from the observations of the telescope installed in the Hawaiian Islands, part of the sky is not available for observation
**Mostly spectroscopic binaries, 3 stars with large proper motion, 1 star against the background of the galaxy
***Spectroscopic binaries

Conclusions

The methodological approach developed by the authors was summarized by them in a research platform called "Cognotron" and presented in the paper, as well as the experiments performed, and shows that the use of machine learning methods on the Hipparcos catalog data allows us to extract additional information and identify double and multiple star systems that previously could not be detected by classical methods. This is the result of a discovery of complex relationships between astrometric and photometric characteristics and the errors of these characteristics by machine learning methods. Classical methods are based on an analysis of isolated characteristics, or small groups of characteristics, and are limited by the accuracy of their measurements. The combination of a larger number of characteristics and their errors in the analysis, which is achieved by using machine learning methods, is equivalent to using a larger number of accumulated light quanta during long-term observation, which makes it possible to increase the accuracy of detecting binary stars.

The disadvantage of the proposed method is that it does not allow us to introduce a strict criterion for the duality of stars. In classical methods, criteria of this kind are formulated on the basis of known physical laws prior to the analysis. However, it is not theoretically possible to formulate such a criterion that would describe the relationship between several dozen characteristics. Machine learning methods in this case rely on an automated extraction of dependencies from labelled data. But, in the case in picture, there is no correctly labelled training sample, and any patterns are extracted from data in which a significant part of binary stars are not labelled as such.

An immediate consequence of this situation is that the output values of the data analysis models proposed in this paper cannot be considered as probabilities of stellar duality. Since the proportion of binary stars in the training sample is underestimated, it should be expected that the output values will also be underestimated compared to the true probabilities.

The procedure for detecting binary stars is also becoming more complicated. To fully use the potential of the method proposed in this paper, it is necessary to implement an iterative procedure, during which the binary star candidates proposed by the ML models will need to be independently verified; the labelling of the training data will change based on this verification, new models will be trained that will offer the next set of candidates and so on.

This article, in fact, presents the results of one such iteration. It is important that this work demonstrates that during such iteration, it is indeed possible to select objects with a high probability of duality (90% of the objects were confirmed according to the PS1 catalogue). It means that:

- in the Hipparcos data, there are indeed significant dependencies indicating the duality of stellar systems;
- existing machine learning methods allow such dependencies to be detected and used to identify dual stars that were not detected by classical methods.

Significant features identified by ML algorithms for the dual star classification are of high interest as well. In the process of training and verifying ML models, it turned out that machine learning algorithms quite reliably identify a group of significant features associated with the statistical characteristics of the observed values. It is also shown that the identification of duality only loosely depends on the spectral characteristics of the pair. The methods turned out to be resilient to observational selection in the training set itself. In addition, the parameters of objects in the Hipparcos catalogue, which, according to the studies of other authors, are not related to the multiplicity/duality of stars (for example, various spectral characteristics), showed low significance, which was additional evidence of the effectiveness of this method.

The application of the digital ML methods-based approach proposed in this paper to data and catalogues of other missions (for example, Gaia, van Leeuwen et al. (2021)) is also possible and promising, but so far seems premature. Combining data across multiple missions is of theoretical interest, but may be accompanied by difficulties due to the difference in the values obtained by different instruments and due to different observation schemes. In addition, their statistical characteristics will also differ, which may introduce difficulties for ML algorithms. An additional complication arises from the difference in the operating ranges of stellar magnitudes, although an intersection does take place. Such a combination is also complicated by the fact that, according to pre-flight plans, the publication of data on relatively complex objects were scheduled to publish at the final stages of the experiment.

Despite the difficulties and obstacles that arise, the ML approach proposed in this paper is more accurate in its measurements and can help extract new knowledge and stimulate the generation of new ideas, compared to the classical approach. A specific AI platform named "Cognotron" was used in our study.

Data availability

The catalogues used in the article can be found on specialized websites:

- Hipparcos – https://cdsarc.cds.unistra.fr/ftp/I/239/;
- Hipparcos, the new reduction – https://cdsarc.cds.unistra.fr/ftp/I/311/;
- PanSTARRS PS1 partially available on request – http://vizier.u-strasbg.fr/viz-bin/VizieR?-source=II/349

All machine learning algorithms used in the article are listed in the references section with current Internet addresses.

The "Cognotron" AI platform data underlying this article will be provided upon reasonable request to the corresponding author.

Acknowledgements

The authors cordially thank the anonymous reviewers of the Astronomical Journal and MNRAS, whose comments helped to greatly improve the initial text (Sazhin et al. (2022)).

The research was carried out within the framework of the state assignment of Lomonosov Moscow State University, the theme "Ground and space astrometry, astrometric support of research".

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Aitken, R. G. and Doolittle, E. (1932). *New General Catalogue of Double Stars within* 120° *of the North Pole*. Washington, D.C.: Carnegie institution of Washington.

ALeRCE (2022). Automatic Learning for the Rapid Classification of Events.

Angelova, A., Abu-Mostafa, Y. S., and Perona, P. (2005). Pruning training sets for learning of object categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 494–501. IEEE Computer Society.

Arenou, F. and Söderhjelm, S. (2005). Gaia Data Reduction Tasks for Double and Multiple Stars. In Turon, C., O'Flaherty, K. S., and Perryman, M. A. C., editors, *The Three-Dimensional Universe with Gaia*, volume 576 of *ESA Special Publication*, page 557.

Argelander, F. W. A. (1903). Bonner Durchmusterung des nordlichen Himmels. *Eds Marcus and Weber's Verlag*, page 0.

Bastian, N., Covey, K. R., and Meyer, M. R. (2010). A Universal Stellar Initial Mass Function? A Critical Look at Variations. ARA&A, 48:339–389.

Becker, I., Pichara, K., Catelan, M., Protopapas, P., Aguirre, C., and Nikzat, F. (2020). Scalable end-to-end recurrent neural network for variable star classification. MNRAS, 493(2):2981–2995.

Brodley, C. E. and Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research (JAIR)*, 11(1):131–167.

Burnham, S. W. (1906). *A General Catalogue of Double Stars within 121° of the North Pole*. Carnegie institution of Washington; [Chicago, University of Chicago press].

Carrasco-Davis, R., Reyes, E., Valenzuela, C., Förster, F., Estévez, P. A., Pignata, G., Bauer, F. E., Reyes, I., Sánchez-Sáez, P., Cabrera-Vives, G., Eyheramendy, S., Catelan, M., Arredondo, J., Castillo-Navarrete, E., Rodríguez-Mancini, D., Ruz-Mieres, D., Moya, A., Sabatini-Gacitúa, L., Sepúlveda-Cobo, C., Mahabal, A. A., Silva-Farfán, J., Camacho-Iñiguez, E., and Galbany, L. (2021). Alert Classification for the ALeRCE Broker System: The Real-time Stamp Classifier. AJ, 162(6):231.

Chambers, K. C. and et al. (2017). VizieR Online Data Catalog: The Pan-STARRS release 1 (PS1) Survey - DR1 (Chambers+, 2016). *VizieR Online Data Catalog*, page II/349.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Chollet, F. et al. (2015). Keras. `https://keras.io`.

Dommanget, J. and Nys, O. (1994). Catalogue des composantes d'etoiles doubles et multiples (CCDM) premiere edition - Catalogue of the components of double and multiple stars (CCDM) first edition. *Communications de l'Observatoire Royal de Belgique*, 115:1.

Duquennoy, A. and Mayor, M. (1991). Multiplicity among solar-type stars in the solar neighbourhood. II - Distribution of the orbital elements in an unbiased sample. A&A, 500:337–376.

Duquennoy, A., Mayor, M., and Halbwachs, J. L. (1991). Multiplicity among solar type stars in the solar neighbourhood. I. CORAVEL radial velocity observations of 291 stars. A&AS, 88:281–324.

Eichhorn, H. (1980). On the Construction of a Comprehensive General Catalogue of Star Positions. *Celestial Mechanics*, 22(2):127–134.

ESA (1997). *The HIPPARCOS and TYCHO catalogues. Astrometric and photometric star catalogues derived from the ESA HIPPARCOS Space Astrometry Mission*, volume 1200 of *ESA Special Publication*. ESA.

Fabricius, C., Høg, E., Makarov, V. V., Mason, B. D., Wycoff, G. L., and Urban, S. E. (2002). The Tycho double star catalogue. A&A, 384:180–189.

Förster, F., Cabrera-Vives, G., Castillo-Navarrete, E., Estévez, P. A., Sánchez-Sáez, P., Arredondo, J., Bauer, F. E., Carrasco-Davis, R., Catelan, M., Elorrieta, F., Eyheramendy, S., Huijse, P., Pignata, G., Reyes, E., Reyes, I., Rodríguez-Mancini, D., Ruz-Mieres, D., Valenzuela, C., Álvarez-Maldonado, I., Astorga, N., Borissova, J., Clocchiatti, A., De Cicco, D., Donoso-Oliva, C., Hernández-García, L., Graham, M. J., Jordán, A., Kurtev, R., Mahabal, A., Maureira, J. C., Muñoz-Arancibia, A., Molina-Ferreiro, R., Moya, A., Palma, W., Pérez-Carrasco, M., Protopapas, P., Romero, M., Sabatini-Gacitua, L., Sánchez, A., San Martín, J., Sepúlveda-Cobo, C., Vera, E., and Vergara, J. R. (2021). The Automatic Learning for the Rapid Classification of Events (ALeRCE) Alert Broker. AJ, 161(5):242.

Gill, D. and Kapteyn, J. C. (1896). The Cape Photographic Durchmusterung: for the Equinox 1875. Part 1: Zones -18 deg. to -37 deg. *Annals of the Cape Observatory*, 3:1.

Gill, D. and Kapteyn, J. C. (1897). The Cape Photographic Durchmusterung: for the Equinox 1875. Part 2: Zones -38 deg. to -52 deg. *Annals of the Cape Observatory*, 4:1.

Gill, D. and Kapteyn, J. C. (1900). The Cape Photographic Durchmusterung: for the Equinox 1875. Part 3: Zones -53 deg. to -89 deg. *Annals of the Cape Observatory*, 5:1.

Gliese, W. and Jahreiß, H. (1991). Preliminary Version of the Third Catalogue of Nearby Stars. On: The Astronomical Data Center CD-ROM: Selected Astronomical Catalogs.

Griffin, R. and Griffin, R. (1986). Composite spectra. Paper 1 : HR 6902. *Journal of Astrophysics and Astronomy*, 7:195–223.

Hartkopf, W. I., Mason, B. D., and Worley, C. E. (2001). The 2001 US Naval Observatory Double Star CD-ROM. II. The Fifth Catalog of Orbits of Visual Binary Stars. AJ, 122(6):3472–3479.

Herschel, W. (1785). Catalogue of double stars. by william herschel, esq. f. r. s. *Philosophical Transactions of the Royal Society of London*, 75:40–126.

Hilditch, R. W. (2001). *An Introduction to Close Binary Stars*. Cambridge University Press.

Høg, E., Fabricius, C., Makarov, V. V., Urban, S., Corbin, T., Wycoff, G., Bastian, U., Schwekendiek, P., and Wicenec, A. (2000). The Tycho-2 catalogue of the 2.5 million brightest stars. A&A, 355:L27–L30.

Huang, J., Qu, L., Jia, R., and Zhao, B. (2019). O2u-net: A simple noisy label detection approach for deep neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Jeffers, H. M., van den Bos, W. H., and Greeby, F. M. (1963). *Index Catalogue of Visual Double Stars, 1961.0*. Publications of the Lick Observatory, Mount Hamilton: University of California, Lick Observatory.

Jenkner, H., Lasker, B. M., Sturch, C. R., McLean, B. J., Shara, M. M., and Russel, J. L. (1990). The Guide Star Catalog. III. Production, database organisation, and population statistics. AJ, 99:2082–2154.

Jiang, J., Doi, M., ichi Maeda, K., Shigeyama, T., Nomoto, K., Yasuda, N., Jha, S. W., Tanaka, M., Morokuma, T., Tominaga, N., Željko Ivezić, Ruiz-Lapuente, P., Stritzinger, M. D., Mazzali, P. A., Ashall, C., Mould, J., Baade, D., Suzuki, N., Connolly, A. J.,

Patat, F., Wang, L., Yoachim, P., Jones, D. O., Furusawa, H., and Miyazaki, S. (2017). A hybrid type ia supernova with an early flash triggered by helium-shell detonation. *Nature*, 550:80–83.

Kharchenko, N. V. (2001). All-sky compiled catalogue of 2.5 million stars. *Kinematika i Fizika Nebesnykh Tel*, 17(5):409–423.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kroupa, P. (2002). The Initial Mass Function of Stars: Evidence for Uniformity in Variable Systems. *Science*, 295(5552):82–91.

Lasker, B. M., Lattanzi, M. G., McLean, B. J., Bucciarelli, B., Drimmel, R., Garcia, J., Greene, G., Guglielmetti, F., Hanley, C., Hawkins, G., Laidler, V. G., Loomis, C., Meakes, M., Mignani, R., Morbidelli, R., Morrison, J., Pannunzio, R., Rosenberg, A., Sarasso, M., Smart, R. L., Spagna, A., Sturch, C. R., Volpicelli, A., White, R. L., Wolfe, D., and Zacchei, A. (2008). The Second-Generation Guide Star Catalog: Description and Properties. AJ, 136(2):735–766.

Lasker, B. M., Sturch, C. R., McLean, B. J., Russell, J. L., Jenkner, H., and Shara, M. M. (1990). The Guide Star Catalog. I. Astronomical Foundations and Image Processing. AJ, 99:2019.

Lipaeva, N. A., Sementsov, V. N., and Malkov, O. Y. (2014). Catalogue of bright IDS stars with extensive cross-identifications. *Baltic Astronomy*, 23:245–247.

Makarov, V. V. and Kaplan, G. H. (2005). Statistical Constraints for Astrometric Binaries with Nonlinear Motion. AJ, 129(5):2420–2427.

Mason, B. D., Wycoff, G. L., Hartkopf, W. I., Douglass, G. G., and Worley, C. E. (2001). The 2001 US Naval Observatory Double Star CD-ROM. I. The Washington Double Star Catalog. AJ, 122(6):3466–3471.

Mason, B. D., Wycoff, G. L., Hartkopf, W. I., Douglass, G. G., and Worley, C. E. (2021). The washington visual double star catalog : B/wds. `http://vizier.u-strasbg.fr/viz-bin/VizieR?-source=B/wds`.

Michell, J. (1767). An Inquiry into the Probable Parallax, and Magnitude of the Fixed Stars, from the Quantity of Light Which They Afford us, and the Particular Circumstances of Their Situation, by the Rev. John Michell, B. D. F. R. S. *Philosophical Transactions of the Royal Society of London Series I*, 57:234–264.

Monet, D. G., Levine, S. E., Canzian, B., Ables, H. D., Bird, A. R., Dahn, C. C., Guetter, H. H., Harris, H. C., Henden, A. A., Leggett, S. K., Levison, H. F., Luginbuhl, C. B., Martini, J., Monet, A. K. B., Munn, J. A., Pier, J. R., Rhodes, A. R., Riepe, B., Sell, S., Stone, R. C., Vrba, F. J., Walker, R. L., Westerhout, G., Brucato, R. J., Reid, I. N., Schoening, W., Hartley, M., Read, M. A., and Tritton, S. B. (2003). The USNO-B Catalog. AJ, 125(2):984–993.

Morrison, J. E., Röser, S., McLean, B., Bucciarelli, B., and Lasker, B. (2001). The Guide Star Catalog, Version 1.2: An Astrometric Recalibration and Other Refinements. AJ, 121(3):1752–1763.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Perryman, M. A. C., Hassan, H., Batut, T., and et al. (1989a). *The Hipparcos mission. Pre-launch status. Volume I: The Hipparcos satellite.*, volume 1. ESA Publication Division.

Perryman, M. A. C., Lindegren, L., Kovalevsky, J., Hog, E., Bastian, U., Bernacca, P. L., Creze, M., Donati, F., Grenon, M., Grewing, M., van Leeuwen, F., van der Marel, H.,

Mignard, F., Murray, C. A., Le Poole, R. S., Schrijver, H., Turon, C., Arenou, F., Froeschle, M., and Petersen, C. S. (1997). The Hipparcos Catalogue. A&A, 500:501–504.

Perryman, M. A. C., Lindegren, L., Murray, C. A., and et al. (1989b). *The Hipparcos mission. Pre-launch status. Volume III: The data reductions.*, volume 3. ESA Publication Division.

Perryman, M. A. C., Turon, C., Arenou, F., and et al. (1989c). *The Hipparcos mission. Pre-launch status. Volume II: The Input Catalogue.*, volume 2. ESA Publication Division.

Raikov, A. (2021). *Cognitive Semantics of Artificial Intelligence: A New Perspective*. SpringerBriefs in Applied Sciences and Technology. Springer Singapore.

Reyes, E., Estévez, P. A., Reyes, I., Cabrera-Vives, G., Huijse, P., Carrasco, R. A., and Förster, F. (2018). Enhanced rotational invariant convolutional neural network for supernovae detection. *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Russell, J. L., Lasker, B. M., McLean, B. J., Sturch, C. R., and Jenkner, H. (1990). The Guide Star Catalog. II. Photometric and Astrometric Models and Solutions. AJ, 99:2059.

Salpeter, E. E. (1955). The Luminosity Function and Stellar Evolution. ApJ, 121:161.

Sazhin, M. V., Sementsov, V., Sorokin, S., Lubarskiy, D., and Raikov, A. (2022). An adaptive algorithm for detecting double stars in astrometric surveys. *arXiv e-prints*, page arXiv:2208.03269.

Schönfeld, E. (1886). Bonner Durchmusterung des sudlichen Himmels. *Eds Marcus and Weber's Verlag*, page 0.

Sémirot, P. (1972). La carte du ciel. *Vistas in Astronomy*, 13(1):153–154.

Söderhjelm, S. (1999). Visual binary orbits and masses post HIPPARCOS. A&A, 341:121–140.

Stock, J. and Cova S., J. (1983). On the reduction of the Carte du Ciel catalogues. Rev. Mexicana Astron. Astrofis., 5:233–260.

van Leeuwen, F. (2007a). *Hipparcos, the New Reduction of the Raw Data*, volume 350 of *Astrophysics and Space Science Library*. Springer.

van Leeuwen, F. (2007b). Validation of the new Hipparcos reduction. A&A, 474(2):653–664.

van Leeuwen, F., de Bruijne, J., Babusiaux, C., Castañeda, J., Hobbs, D., Busso, G., Sartoretti, P., Utrilla, E., Luri, X., Marrese, P. M., Mora, A., Fabricius, C., González-Núñez, J., Hambly, N., Altavilla, G., Altmann, M., Antoja, T., Arenou, F., Bakker, J., Balbinot, E., Barache, C., Bastian, U., Bauchet, N., Bellazzini, M., Biermann, M., Blomme, R., Bombrun, A., Brown, A., Busonero, D., Butkevich, A., Cacciari, C., Carrasco, J. M., Cheek, N., Clotet, M., Creevey, O., Crowley, C., Cánovas, H., David, M., Davidson, M., De Angeli, F., Diakité, S., Drimmel, R., Duran, J., Evans, D. W., Fabrizio, M., Fernández-Hernández, J., Figueras, F., Findeisen, K., Garcia-Gutierrez, A., Gracia-Abril. G., Guerra, R., Gutiérrez-Sánchez, R., Helmi, A., Henar Sarmiento, M., Hernandez, J., Hutton, A., Jordi, C., Khanna, S., Klioner, S., Lammers, U., Leclerc, N., Lindegren, L., Löffler, W., Marinoni, S., Martín-Fleitas, J., Masana, E., Masip Vela, A., Masip, A., Messineo, R., Michalik, D., Mignard, F., Montegriffo, P., Muraveva, T., Nienartowicz, K., Pancino, E., Panem, C., Portell, J., Racero, E., Rainer, M., Ramos, P., Reylé, C., Ríos Diaz, C., Riva, A., Robin, A., Robin, A., Roegiers, T., Romero-Gómez, M., Rowell, N., Rybizki, J., Salgado, J., Sanna, N., Seabroke, G., Segovia, J. C., Siddiqui, H., Smart, R., Stephenson, C., Teyssier, D., Torra, F., Turon, C., Valero, J., Vallenari, A., van Leeuwen, M., and Weiler, M. (2021). GAIA EDR3 documentation. `https://ui.adsabs.harvard.edu/abs/2021gdr3.reptE....V`.

Wang, P., Li, X., and Hammer, P. (2018). Self in nars, an agi system. *Frontiers in Robotics and AI*, 5.

Worley, C. E. and Douglass, G. G. (1997). The Washington Double Star Catalog (WDS, 1996.0). A&AS, 125:523–523.

Worley C.E., D. G. (1984). The washington double star catalog. US Naval Obs., unpublished.

Zhu, X., Wu, X., and Chen, Q. (2003). Eliminating class noise in large datasets. In Fawcett, T. and Mishra, N., editors, *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 920–927. AAAI Press.

Appendix A: Spectrum description fields

Field	Range	Description	Notation in the field SpType
SpectralClass	0...8	Spectral class	O, B, A, F, G, K, M, L, T, C, S, SC, WN, WC, WO, WR, R, N, DA, DB, DC, DO, DZ, DQ, DG, DF, CN
Luminosity_I	Boolean	Luminosity class I	I, Ia, Iab, Ib
Luminosity_II	Boolean	Luminosity class II	II, IIa, IIb
Luminosity_III	Boolean	Luminosity class III	III, IIIa, IIIb
Luminosity_IV	Boolean	Luminosity class IV	IV, IVa
Luminosity_V	Boolean	Luminosity class V	V, Va, Vb
Luminosity_VI	Boolean	Luminosity class VI	VI
Luminosity_uncertain	Boolean	Luminosity class is not precisely defined	:
subdwarf	Boolean	Subdwarf	sd
WhiteDwarf	Boolean	White dwarf	DA, DB, DC, DO, DZ, DQ, DG, DF
W	Boolean	Wolf–Rayet star	WN
Carbon	Boolean	C-type star (carbon star)	C
S	Boolean	S-type star (zirconium star)	SC
R	Boolean	Spectral class R	R
N	Boolean	Spectral class N	N
MN	Boolean		MN
Nebulous	0, 1, 2	Wide spectrum lines	n, nn, n:
Enhanced_metal	Boolean	Strong metal lines	m, m:
Peculiar	Boolean	Spectrum peculiarities	p, p:, +pec
Shell	Boolean	Shell star	sh, +shell, shell
Emission	Boolean	Emission lines	e, e:, eq:, E
Weak_lines	Boolean	Weak lines	w, wk, wl
Sharp_lines	Boolean	Narrow lines	s, ss, s:
Variable	Boolean	Variable spectrum	v, va, var
Weak_helium	Boolean	Weak lines of helium	Hewk
NIIandHeIIEmission	Boolean	N III emission, absence or weak absorption of He II	(f)
HeIIabsorbtionNIIemission	Boolean	Strong He II absorption, weak N III emission	((f))
Composite	Boolean	Spectrally double star	comp
Undescribed_peculiarities	Boolean	Other features	., ..., .., +..., +.., +., +....
SB	Boolean	Spectroscopic binary	SB, SB1, SB:
Sr	Boolean	Spectral lines of strontium	Sr, Sr:
Cr	Boolean	Spectral lines of chromium	Cr
Eu	Boolean	Spectral lines of europium	Eu
Mn	Boolean	Spectral lines of manganese	Mn

continued on next page

Appendix A continued

Field	Range	Description	Notation in the field SpType
Hg	Boolean	Spectral lines of mercury	Hg, Hg:
Si	Boolean	Silicon spectral lines	Si
Li	Boolean	Lithium spectral lines	Li
Del	Boolean	Spectrum-like Delta Delphini	delDel, dDel, deltaDel
Nova	Boolean	Nova star	Nova

Appendix B: Used parameters from the Hipparcos catalog

Name	Description in the Hipparcos catalog	Used in the full set	Used in a set with a restriction of observational selection
Catalog	Catalogue (H=Hipparcos)		
HIP	Identifier (HIP number)		
Proxy	Proximity flag		
RAhms	Right ascension in h, m, s, ICRS (J1991.25)		
DEdms	Declination in °, ′, ″, ICRS (J1991.25)		
Vmag	Magnitude in Johnson V	+	
VarFlag	Coarse variability flag	+	+
r_Vmag	Source of magnitude		
RAdeg	alpha, degrees (ICRS, Epoch=J1991.25)		
DEdeg	delta, degrees (ICRS, Epoch=J1991.25)		
AstroRef	Reference flag for astrometry		
Plx	Trigonometric parallax	+	+
pmRA	Proper motion mu_alpha*cos(delta), ICRS	+	+
pmDE	Proper motion mu_delta, ICRS	+	+
e_RAdeg	Standard error in RA*cos(Dedeg)	+	
e_DEdeg	Standard error in DE	+	
e_Plx	Standard error in Plx	+	+
e_pmRA	Standard error in pmRA	+	+
e_pmDE	Standard error in pmDE	+	+
DE_RA	Correlation, DE/RA*cos(delta)	+	+
Plx_RA	Correlation, Plx/RA*cos(delta)	+	+
Plx_DE	Correlation, Plx/DE	+	+
pmRA_RA	Correlation, pmRA/RA*cos(delta)	+	+
pmRA_DE	Correlation, pmRA/DE	+	+
pmRA_Plx	Correlation, pmRA/Plx	+	+
pmDE_RA	Correlation, pmDE/RA*cos(delta)	+	+
pmDE_DE	Correlation, pmDE/DE	+	+
pmDE_Plx	Correlation, pmDE/Plx	+	+
pmDE_pmRA	Correlation, pmDE/pmRA	+	+
F1	Percentage of rejected data	+	+
F2	Goodness-of-fit parameter	+	+
H31	HIP number (repetition)		
BTmag	Mean BT magnitude	+	
e_BTmag	Standard error on BTmag	+	
VTmag	Mean VT magnitude	+	
e_VTmag	Standard error on VTmag	+	
m_BTmag	Reference flag for BT and VTmag		
B_V	Johnson B-V colour	+	+
e_B_V	Standard error on B-V	+	+
r_B_V	Source of B-V from Ground or Tycho		
V_I	Colour index in Cousins' system	+	+
e_V_I	Standard error on V-I	+	+

continued on next page

Appendix B continued

Name	Description in the Hipparcos catalog	Used in the full set	Used in a set with a restriction of observational selection
r_V_I	Source of V-I		
CombMag	Flag for combined Vmag, B-V, V-I		
Hpmag	Median magnitude in Hipparcos system	+	
e_Hpmag	Standard error on Hpmag	+	+
Hpscat	Scatter on Hpmag	+	+
o_Hpmag	Number of observations for Hpmag		
m_Hpmag	Reference flag for Hpmag		
Hpmax	Hpmag at maximum (5th percentile)		
HPmin	Hpmag at minimum (95th percentile)		
Period	Variability period (days)	+	+
HvarType	Variability type		
moreVar	Additional data about variability		
morePhoto	Light curve Annex		
CCDM	CCDM identifier		
n_CCDM	Historical status flag		
Nsys	Number of entries with same CCDM		
Ncomp	Number of components in this entry		
MultFlag	Double/Multiple Systems flag		
Source	Astrometric source flag		
Qual	Solution quality		
m_HIP	Component identifiers		
theta	Position angle between components		
rho	Angular separation between components		
e_rho	Standard error on rho		
dHp	Magnitude difference of components		
e_dHp	Standard error on dHp		
Survey	Flag indicating a Survey Star		
Chart	Identification Chart		
Notes	Existence of notes		
HD	HD number <III/135>		
BD	Bonner DM <I/119>, <I/122>		
CoD	Cordoba Durchmusterung (DM) <I/114>		
CPD	Cape Photographic DM <I/108>		
V_I_red	V-I used for reductions		
SpType	Spectral type*	+	+
r_SpType	Source of spectral type		

*The list of used fields in decoded form is given in Appendix A

Appendix C: Used parameters from the new reduction

Name	Description from the paper of Floor van Leeuwen	Used in the full set	Used in a set with a restriction of observational selection
HIP	Hipparcos identifier		
Sn	[0,159] Solution type new reduction		
So	[0,5] Solution type old reduction		
Nc	Number of components	+	
RArad	Right Ascension in ICRS, Ep=1991.25		
DErad	Declination in ICRS, Ep=1991.25		
Plx	Parallax	+	
pmRA	Proper motion in Right Ascension	+	
pmDE	Proper motion in Declination	+	
e_RArad	Formal error on RArad	+	+
e_DErad	Formal error on DErad	+	+
e_Plx	Formal error on Plx	+	+
e_pmRA	Formal error on pmRA	+	+
e_pmDE	Formal error on pmDE	+	+
Ntr	Number of field transits used	+	
F2	Goodness of fit	+	+
F1	Percentage rejected data	+	+
var	Cosmic dispersion added		
ic	Entry in one of the suppl.catalogues		
Hpmag	Hipparcos magnitude	+	
e_Hpmag	Error on mean Hpmag	+	+
sHp	Scatter of Hpmag	+	+
VA	[0,2] Reference to variability annex		
B-V	Colour index	+	
e_B-V	Formal error on colour index	+	+
V-I	V-I colour index	+	
UW	Upper-triangular weight matrix*	+	+

*Each element of the matrix is represented as a separate parameter