Mem. S.A.It. Vol. 90, 327

**Memorie** della

# Machine learning techniques meet binaries

G. Traven[1], K. Čotar[2], T. Merle[3], M. Van der Swaelmen[4],
Y.-S. Ting[5,6,7]⋆, and the GALAH team

[1] Lund Observatory, Department of Astronomy and Theoretical Physics, Box 43, SE-221 00 Lund, Sweden, e-mail: gregor.traven@astro.lu.se
[2] Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia
[3] Institut d'Astronomie et d'Astrophysique, Université Libre de Bruxelles, CP. 226, Boulevard du Triomphe, 1050 Brussels, Belgium
[4] INAF - Osservatorio Astrofisico di Arcetri, Largo E. Fermi, 5, I50125 Firenze, Italy
[5] Institute for Advanced Study, Princeton, NJ 08540, USA
[6] Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA
[7] Observatories of the Carnegie Institution of Washington, 813 Santa Barbara Street, Pasadena, CA 91101, USA

**Abstract.** We briefly review the various machine learning methods and discuss how they can be used in efficient identification and analysis of spectroscopic binary stars. They can be treated as complementary to conventional methods, and we argue that some amount of human oversight is always needed and in fact highly beneficial when employing machine learning. We propose that a general dimensionality reduction technique can serve to diagnose and classify a given data set, and in case of GALAH spectra, our method quite effectively reveals a population of SB2 and SB3 systems. Once identified, the binary spectra can be analysed with the help of generative models, which can be constructed using machine learning techniques such as The Cannon and The Payne. Furthermore, in the case of spectroscopically unresolved multiple stars, we can recover the multiple contributions to an observed spectrum by reversing the process and proceeding from analysis to identification.

**Key words.** Stars: binaries: close – Stars: binaries: spectroscopic – methods: data analysis – methods: numerical – techniques: radial velocities – techniques: spectroscopic

## 1. Introduction

Binary stars are ubiquitous in stellar populations across the universe. They offer exceptional insight to star formation and evolution (Duchêne & Kraus 2013), while also establishing distances across our Galaxy and to its satellites (Pietrzyński et al. 2013), providing critical tests in asteroseismology (Huber 2015), and furnishing benchmark systems with accurate fundamental stellar properties (Popper 1980; Stassun et al. 2009). The list goes on, and although no one disputes the astrophysical importance of multiple systems, they are often overlooked simply because their nature can be quite effectively concealed in observational data.

⋆ Hubble Fellow

However, the advances in machine learning, and the unprecedented volume of stellar spectra from modern surveys (e.g. RAVE Steinmetz et al. 2006, APOGEE (Majewski et al. 2017), *Gaia*-ESO (Gilmore et al. 2012), GALAH (De Silva et al. 2015), LAMOST (Luo et al. 2015), WEAVE (Dalton et al. 2012), 4MOST de Jong et al. 2012), enable us to devise new approaches for discovery and characterisation of these fascinating associations of stars, building upon the effort of previous studies (Matijevič et al. 2010; Gao et al. 2014; Merle et al. 2017; Skinner et al. 2018; El-Badry et al. 2018; Kounkel et al. 2019).

Due to ever increasing observed samples of stars, it is now feasible to focus on characterising the large scale statistical properties of Galactic binary stars. This necessitates a variety of observational techniques to identify them across the full range of primary masses, mass ratios, orbital periods, separations, eccentricities, and ages/evolutionary states for different stellar populations and Galactic environments (Duchêne & Kraus 2013). Thus acquired knowledge of the fundamental statistics of stellar multiplicity beyond the solar neighbourhood can lead to a better understanding of the outcome of star-formation process as modelled by molecular cores and star-forming regions (Parker & Meyer 2014; Lomax et al. 2015; Moe et al. 2019; Bate 2019). It can additionally provide insight into the impact of binary populations on high-redshift galaxy radiation and reionization, compact objects evolution, interpretation of multi-messenger astronomy, near-Universe constraints on the Hubble constant, local measures of dark-matter substructure masses, demographics of planetary systems, and many other aspects of Astronomy (for further explanation see Breivik et al. 2019).

Besides scientific data, large sample sizes are nowadays present in many aspects of our lives, and they range from our medical and financial records to political beliefs and all kinds of personal information that we willingly share on a multitude of social platforms. Thus the term "Big Data" came into common parlance, and it represents information of greatly varying dimensionality, possibly collected automatically, and prepended with a timestamp. This has proven to be a goldmine for newly developed "smart" mathematical algorithms, that we sometimes mistakenly call A.I. (artificial intelligence), which are to certain extents able to tell apart pictures of dogs and cats, translate texts, track insects, predict aircraft failures, find new particles and new antibiotics, perform surveillance, target the right consumers and voters, or determine chemical abundances in the atmospheres of stars. These algorithms are what we refer to as "Machine Learning - ML", and a graphical representation of the tasks they encompass is in Figure 1. There are many kinds of machine learning algorithms, and their use for data analysis and knowledge discovery in astronomical research dates back to the previous century, with a continuously growing popularity (Figure 2).

In this work, we discuss how distinct machine learning approaches can be merged to efficiently discover and characterise double-lined binary stars (SB2s) in a volume of stellar spectra collected by the GALAH spectroscopic survey. GALAH is an ongoing project with the aim to unveil the Milky Way's history by studying the fossil record of ancient star formation and accretion events preserved in stellar light. It is close to achieving its goal of obtaining spectra of $\sim$ 1 million stars and measuring up to 32 elemental abundances, which enables the technique of chemical tagging (Freeman & Bland-Hawthorn 2002). The survey is targeting a randomly selected, magnitude limited ($12 \leq V \leq 14$) sample of stars, collecting spectra in three visual and one infra-red band, with a resolution of R $\approx$ 28 000 and a typical SNR $\sim$ 100 (Buder et al. 2018). GALAH therefore produces a highly suitable data set for extraction of a largely unbiased sample of spectroscopic binary stars which can be identified by a line-of-sight velocity offset between the individual components.

We first give a short general introduction to machine learning techniques in Section 2, and then present the procedure for identification of double-lined stellar spectra in Section 3. The prospect of analysing detailed parameters of both stars in a binary system is shortly

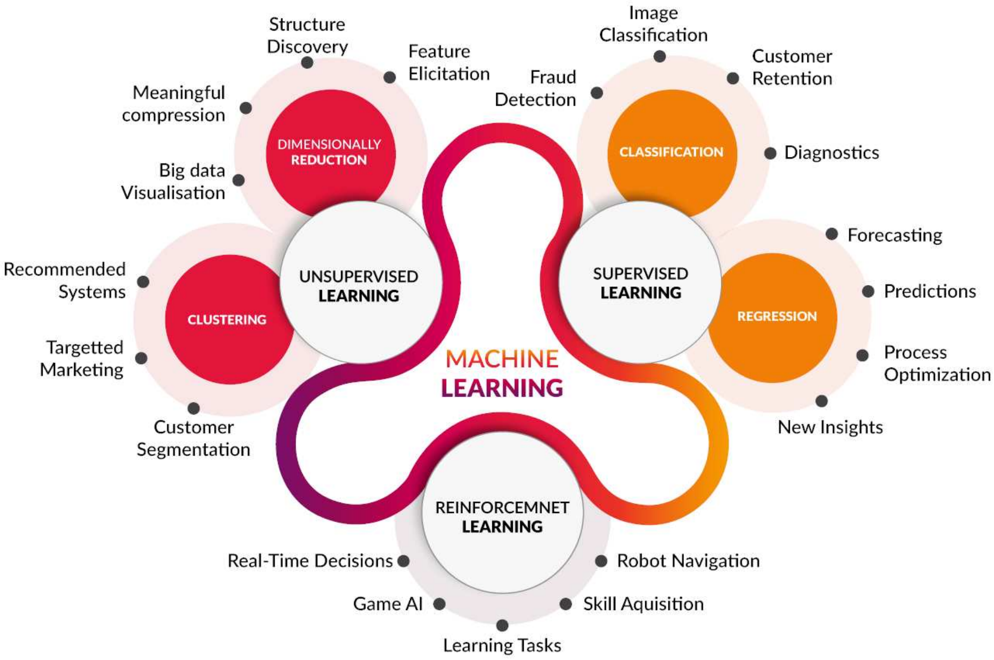


**Fig. 1.** General scheme of machine learning utility (source: www.cookieegroup.com)

discussed in Section 4, and we conclude with Section 5.

## 2. Machine learning techniques

We can try and describe machine learning in one sentence: it is a process whereby we employ an algorithm, which, when presented with real world data, is able to figure out its properties, and then we use the resulting tool (e.g. decision maker, classifier, generative model of data) to extract knowledge from a given data set. Machine learning techniques can be grouped in different ways, for example by the amount of effort (apart from technical implementation) that we have to put in to make the algorithm work. Hence we can first distinguish between **unsupervised** and **supervised** methods.

### *2.1. Supervised methods*

As the name implies, the supervised methods require some kind of prior training on (preferably) real data. Training data thus becomes an essential factor in determining the final utility of the method, and defining the right training sample can become quite an obstacle since it is a task specific to a problem that we are trying to solve. However, the advantage of supervised methods is that we can obtain **quantitative** results for a given data set. This can be in the form of values for the labels (parameters), the probability that data belongs to a certain category, or the actual reconstruction of the original data. This brings us to a distinction between obtaining some information (labels) from the data or the reverse, creating data from the labels, as explained below.

#### 2.1.1. Discriminative algorithms

The discriminative type of algorithms are conventionally used for the purpose of classification or prediction of the label values for some data. They are relatively fast as they directly map some test data into label space. There is a plethora of different methods that fall under this category, e.g. logistic regression, linear discriminant analysis, support vector machines, conditional random fields, random forests.

#### 2.1.2. Generative algorithms

Quite the opposite of discriminative algorithms, generative algorithms **produce the data** based on some input labels. They can also be used for classification or prediction purposes, however in such a case we have to introduce additional computational tasks for comparison of the data to the model. Generative algorithms are gaining in importance or at least popularity, with an obvious desire to produce models which can replicate reality (e.g. generative adversarial networks, variational autoencoders).

### *2.2. Unsupervised methods*

The unsupervised machine learning algorithms[1] are not pre-trained on any data, and therefore have no prior knowledge of the specific context, e.g. physical laws. These methods can be broadly described by the similar tasks of **clustering** and **dimensionality reduction**, and again there is a wealth of linear and non-linear techniques available to perform them, e.g. singular value decomposition, principal component analysis, self organising maps, locally linear embedding, DBSCAN, autoencoders, UMAP, t-SNE. The immense potential of unsupervised methods lies in the ability to detect the unknown, may it be unwanted outliers, noisy data, or new and unexpected discoveries.

We will show how we can make a bridge between the unsupervised methods and the "supervised" end-product such as classification, with a help of some human intervention.

[1] Sometimes it is hard to decide whether an algorithm is supervised or unsupervised (e.g. Generative adversarial network - GAN)

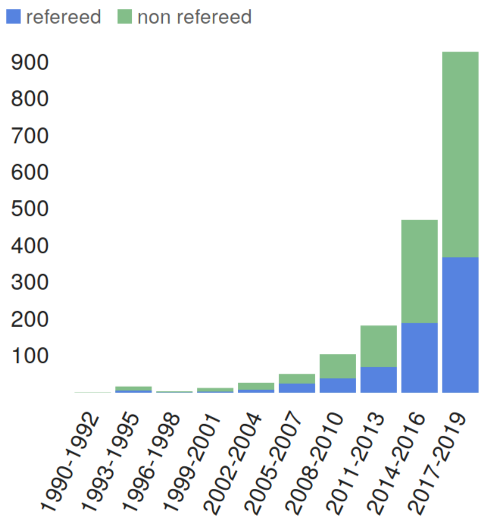


**Fig. 2.** Astronomical publications with abstracts containing "machine learning" (source: ADS).

This can give us the best of the two worlds: the speed and efficiency of the machine learning algorithm and the profound understanding of a human user.

## 3. Detecting binary stars

Binary stars come in different flavours, and we can distinguish them both by their physical and observational properties. From the times of Benedetto Castelli and Galileo Galieli, through the discoveries of William Herschel, and looking onwards to the *Gaia* harvest of binary stars, we have coined a range of observational expressions for them: *visual*, *resolved*, *common proper motion*, *astrometric*, *spectroscopic*, *photometric*, *eclipsing*. Although these names refer to observational techniques used to detect binary stars, they are also intimately linked to a fundamental property of binary systems - their orbital period, itself tightly linked to the semi-major axis of the binary orbit.

The widely accepted log-normal distribution of binary star periods is shown in Figure 3, and it has been confirmed by a range of authors (Duquennoy & Mayor 1991; Duchêne & Kraus 2013 and references therein). However, we have yet to agree on the precise parame-

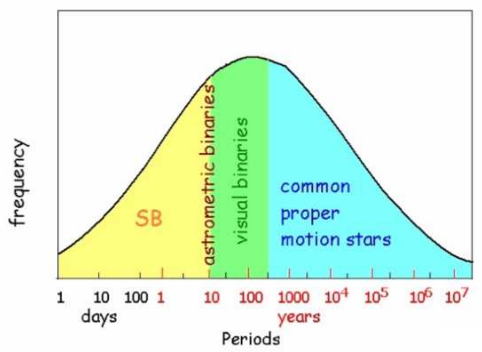


**Fig. 3.** Log-normal distribution (frequency) of binary systems given their orbital period (source: Boffin 2017).

ters of this distribution and in fact, on its validity when applied to diverse environments and evolutionary histories. Nevertheless, this period distribution predicts the number of binary systems that we can expect to detect with different observational techniques as indicated in Table 1.

Besides focusing on only one type of observational data, we are nowadays offered a variety of large catalogues, which can be mined to more efficiently extract as many binary stars as possible, by e.g. looking for connections between astrometric and radial velocity variability, or common proper motion and chemical signatures of observed stars. However, as long as we define exactly what such a mining algorithm should extract from available catalogues, in terms of e.g. parameter ranges or their combinations, we can not call that machine learning, as the algorithm is not actually trying to figure out anything on its own.

A simple way of putting machine learning to use, would be to merge information about objects on the sky, as measured from diverse all-sky surveys, and feed that to a clustering algorithm, which is able to reveal the underlying structure of the data. Thus, the algorithm can potentially identify many clumps of objects which share similar properties. Such clustering can be an essential advantage, since it is inconceivable for a human to explore the multi-dimensional space of original data (each property of an object representing one dimension). If the algorithm is instead performing dimensionality reduction, or in other words mapping the data from the original N-dimensional space to a 2-dimensional map, the groups of similar objects can become apparent also to the human eye. This process can then lead to an efficient classification of the data, and a powerful way to diagnose and discover all kinds of features or abnormalities.

The distinction between conventional and machine learning approaches for discovery and characterisation of **spectroscopic** binary stars is further illustrated in the following chapters. However, addressing the rest of the binary population, indicated by the orbital period distribution in Figure 3, is beyond the scope of this work.

### *3.1. Conventional methods for detection of spectroscopic binaries*

Spectroscopic binary (multiple) stars are detected by the Doppler shift of absorption/emission lines in a spectrum. A notation (SB$n$; $n > 0$) is agreed upon, which indicates the number of distinct sets of lines. An SB1 spectrum means we only see spectral lines of the brighter component in a binary system, an SB2 means we can resolve both components, and higher numbers denote the triple, quadruple, etc. sets of detected lines, which correspond to the number of stars in a multiple stellar configuration.

SB1 binary systems can therefore be identified only with repeated exposures of the same object through radial velocity variability, whereas higher SB$n$s can be readily identified in only one acquired spectrum per object. There is no SB0 notation, thus if we know the object is a multiple system but cannot be identified as such through the Doppler shift in spectral lines, we call it a spectroscopically unresolved binary system. Likewise, in the case of SB$n$ systems, it is possible there are additional components contributing to the combined light in the spectrum but are not resolved.

The standard techniques for detection of spectroscopic multiple stars have mostly relied on examining the scatter of the radial velocity values for the same object (SB1s; Matijevič

**Table 1.** Observational techniques for detection of binary stars.

| Binary star type | Observational technique for detection |
|---|---|
| visual, resolved | imaging |
| common proper motion | proper motions, radial velocity |
| astrometric | epoch astrometry (positions) |
| spectroscopic | doppler shift of spectral lines |
| photometric, eclipsing | variability in the light curve, eclipses |

et al. 2011; Merle et al. 2020; Birko et al. 2019), or the multiple peaks of the cross-correlation function (SB*n*s; Matijevič et al. 2010; Merle et al. 2017; Kounkel et al. 2019). Detection of SB1s can be hampered by many instrumental, data reduction, and other factors that originate from the physics of observed stars (Merle et al. 2020). Analysis of the cross-correlation function (CCF) can suffer from similar effects, however, advanced schemes, like the one presented by Merle et al. (2017) and improved in Van der Swaelmen et al. (in prep.; see also Van der Swaelmen et al. 2018), can improve the situation significantly. For example, Figure 4 demonstrates that exclusion of strong hydrogen lines significantly enhances the visibility of a double peaked profile, the signature of SB2s.

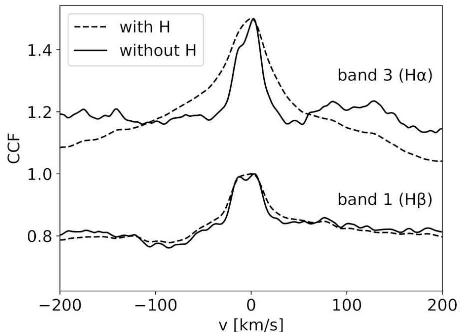


**Fig. 4.** Effect of the exclusion of H$\alpha$ and H$\beta$ in the solar template on the width of the CCF for an example GALAH SB2 candidate spectrum. The decrease of the CCF peak width reveals the presence of a second close component difficult to detect when H$\alpha$ and H$\beta$ are included (demonstrated here on the first and third GALAH spectral bands which include H$\beta$ and H$\alpha$, respectively).

The work by El-Badry et al. (2018) has recently introduced a method for efficient detection of **unresolved** binary components in stellar spectra. In short, this is achieved by comparing whether a double component model is a better fit to data than a single component model. Although defining the criteria for this distinction is a manual effort, machine learning was employed in this work for construction of a data-driven generative model of single star spectra (see Section 4). Their method makes it possible to identify many long-period binaries like the one in Figure 5, in which the velocity offset between the two stars is negligible.

### *3.2. Machine learning techniques for detection of spectroscopic multiple stars*

All spectroscopic binaries can not be attacked by the same identification approach, if not for other reasons, the SB1s for example necessitate an extra dimension in observations - time (temporal sampling). We hereby focus only on those objects, where the light in a single spectrum can reveal the multitude of gravitationally bound emitting bodies. Detection of spectroscopically unresolved multiple stars will be briefly addressed in Section 4, whereas the rest of this section is dedicated to detection of SB*n*s for $n \geq 2$.

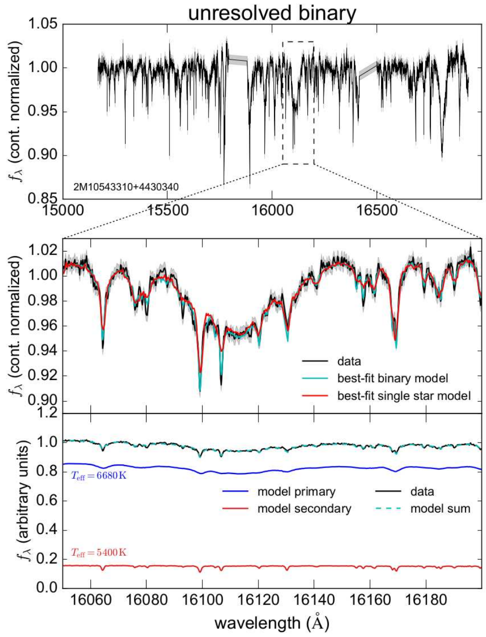


**Fig. 5.** Spectrum of an unresolved main-sequence binary with mass ratio $q = m_2/m_1 \approx 0.7$ as observed by APOGEE. Top panel shows the full normalised spectrum. Middle panel shows the spectrum and best-fit binary and single-star models. The binary model fits the data significantly better than the single-star model. Bottom panel shows the two components of the best-fit binary model (adapted from El-Badry et al. 2018).

In general, we would like to have an algorithm, that given some input data, decides whether that data is an SB$n$ or not. We can accomplish this with a classifier, which can sort all input data into a set of categories. However, in order to train a machine learning algorithm to reliably sort our data, we need to manually define all the classification categories and feed the method a large enough sample of representative objects for each category - the training set. Besides the significant amount of manual work needed for that, the downside of this **supervised** approach is the inability of the classifier to properly recognise any kind of data that doesn't fit into any of the defined categories.

We propose a more unsupervised approach, which can handle both known and unexpected types of data, and at the same time "catches" SB2 or higher order multiples. It consists of:

1. a smart visualisation of the whole data set,
2. manual inspection and definition of classification categories,
3. automatic selection of groups of similar objects.

This is not a direct approach of mapping data to labels as mentioned in Section 2.1.1, but rather a scheme, where the user is actively engaged, which we argue is still much more reliable than letting the machines figure out everything.

### 3.2.1. Visualisation of data by dimensionality reduction

When presenting data points of dimensionality (number of features) higher than e.g. 3, we usually face a challenge of condensing all that information in the same plot. However, dimensionality reduction can efficiently map high dimensional data on a regular 2-dimensional surface - a projection map. In the case of stellar spectra, dimensionality of original flux space can be very high (e.g. 10 000 dimensions – pixels), however a dimensionality reduction algorithm can still preserve the important features of the spectra by mapping them in the "right" place on the 2-dimensional map. In such a process, some information is inevitably lost, however if the algorithm is "smart" enough, the important information is retained, revealing the global as well as the local structure of our data.

A plethora of options are available for the task of dimensionality reduction, from linear to non-linear ones, and from those that are useful for the process of classification described in this work to those that fail. We will first comment on a technique that is in principle promising, but does not produce the desired result in our use case. We consider a type of a neural network called an autoencoder, whose graphical representation is in Figure 6. This technique of reducing dimensions is instructive, as it is fairly intuitive in what it does, and at the same time introduces the basic functionality of neural networks. Explained in the most simplistic terms, the neurons (circles) in the network in Figure 6 are fully connected to each

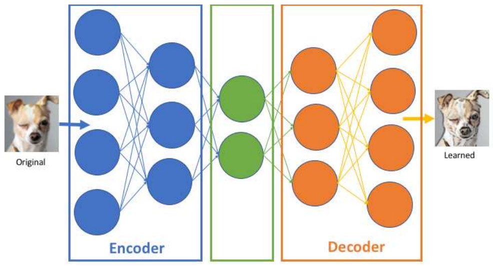


**Fig. 6.** Structure of an autoencoder, one specific type of an artificial neural network that can be used for the task of dimensionality reduction.

other, and these connections are weighted. The weights are optimised so that the information from the input (left side) is transferred through the network to the output in such a way, that the output reproduces the input as faithfully as possible. Figure 6 shows that some information is, as predicted, inevitably lost. We must now imagine that there are hundreds of images on the input. If the structure of the autoencoder is optimised, the encoder embeds the most important features of each image in the two neurons (2 dimensions) in the middle layer in such a way, that the decoder can reproduce the original image from those two values.

There is a lot that can be tuned in a general neural network (e.g. the number of layers or neurons, the type of activation functions between neurons). However, our experiments with autoencoders did not produce satisfactory results. The 2-dimensional map with the values from the middle two neurons of an autoencoder fed by GALAH spectra is shown in Figure 7. The position of points in the map is optimised such that the corresponding spectra are faithfully reproduced at output by the autoencoder. However, for a human, it is difficult to recognise groups of similar data points in this map without the help of colour-encoding. The situation is even worse for categories other than those marked in the two panels.

This is because the autoencoder is tasked with reproducing the original data at the output, in whatever configuration of the network that achieves that goal. It is not encouraged to produce a visually appealing and to the human user more useful mapping. Luckily, there are other methods which do just that, and we argue that t-SNE (t-distributed Stochastic Neighbour Embedding) is good at fulfilling this task. The mathematical details of the t-SNE method are explained in the original work by van der Maaten & Hinton (2008), and we only offer a brief summary here.

The objective for t-SNE is to minimise the divergence between pairwise similarities of data points in the original data space and the corresponding pairwise similarities in the projection space[2]. Therefore, two similar stellar spectra will be placed close together in the projection map, and the very dissimilar ones will be far apart. In addition, t-SNE solves the issue of crowding the data points in a small area of the map, such as seen in Figure 7, and which plagues many other dimensionality reduction techniques. The t-SNE projection of GALAH spectra is in Figure 8 (for more information see Traven et al. 2017 and Buder et al. 2018).

t-SNE performs extremely well in revealing the global as well as the local structure of the data[3]. Figure 8 for example shows a clear distinction between regions of dwarfs and giants, or regions of hotter and cooler stars, since these properties are one of the most important features in the GALAH spectra. More locally though, we can observe a progression of less prominent features, such as for example metallicity. Furthermore, the clear boundaries between different large islands of data points enable us to distinguish between significantly diverse spectra, and we describe this process below.

### 3.2.2. t-SNE Explorer - a user-friendly interface to data

Without any colour-encoding such as the one in Figure 8, it is difficult to predict what kind of stellar spectra are projected in which island of points. That is why we designed a powerful visual interface to a t-SNE projection map,

[2] Technically we are minimising the Kullback-Leibler divergence (information entropy loss)

[3] Even in the case of very few data points, one can synthesise mock ones to investigate where they are positioned in the t-SNE map.

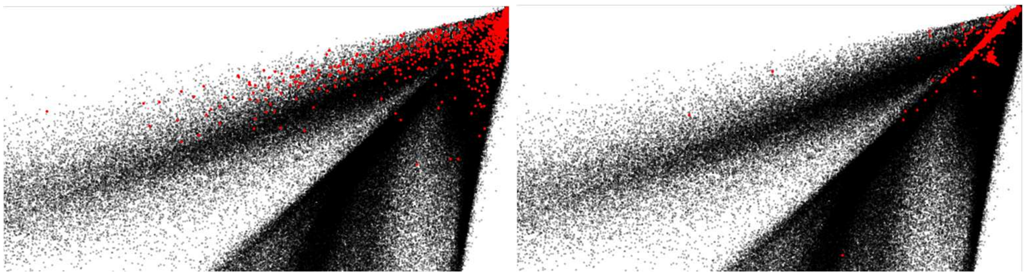

**Fig. 7.** Dimensionality reduction of GALAH spectra produced by an autoencoder (see Figure 6). Each point represents one spectrum. Previously classified spectra are in red (left: SB2s, right: cool giant stars).

that we nickname the t-SNE Explorer. It is a web–based tool, featuring the t-SNE map split into hexagons. Clicking on these small slices of the map, we immediately examine the average shape of the spectra contained inside, and can view them individually as well. Colour coding can be changed depending on which information is available about the investigated objects, and it can help guiding the user to specific areas of the map (e.g. hot stars).

### 3.2.3. Automatic selection of similar objects

Once we decide on the classification category to assign to a certain island of points in a t-SNE map, we can automatically select them on the basis of their density structure. A general clustering algorithm DBSCAN (Ester et al. 1996) proved to be the most efficient in selecting these islands of points in the projection map without a priori knowledge of the number of points or the overall structure of the island.

The functionality of t-SNE, t-SNE Explorer, and DBSCAN allows the user to efficiently determine the classification categories for the investigated data set, and with the help of colour coding by previous classification efforts, to update the classification as the data set grows. The final result of the above described process for classification of GALAH spectra contained in GALAH DR2 is in Figure 8.

Two types of SB*n* systems are identified in the t-SNE map in Figure 8, SB2s being counted in thousands and SB3s in dozens. The t-SNE detection of SB2 objects can be complemented with the CCF detection as described in Section 3.1. We show the detected SB2 systems in an HR diagram in Figure 9. We see differences between both approaches to detection, on the one hand, in the number of detected SB2s, and on the other hand, in the efficiency of detection in different parts of the parameter space of GALAH data. This implies that both approaches of detection not only produce false positives, but also miss some obvious candidates (false negatives).

Our philosophy for detection of SB*n* systems is that different methods can be merged, in order to have as little missed detections as possible, however we can live with many false positives as they can be rejected further on in the analysis process.

## 4. Analysing binary stars

We can benefit significantly from machine learning techniques not only in the identification process, but also during analysis of our binary candidates. For the latter, our philosophy is to use all available observational data and a good mixture of model and data driven methods to extract as much information as possible from the detected SB2 candidates (Traven et al.(submitted)Traven, Feltzing, Čotar, team, et al. 2019).

In the case of Bayesian analysis of SB2 objects as presented in Traven et al.(submitted)Traven, Feltzing, Čotar, team, et al. (2019), we use spectroscopic, photometric, and astrometric information to derive a comprehensive set of parameters

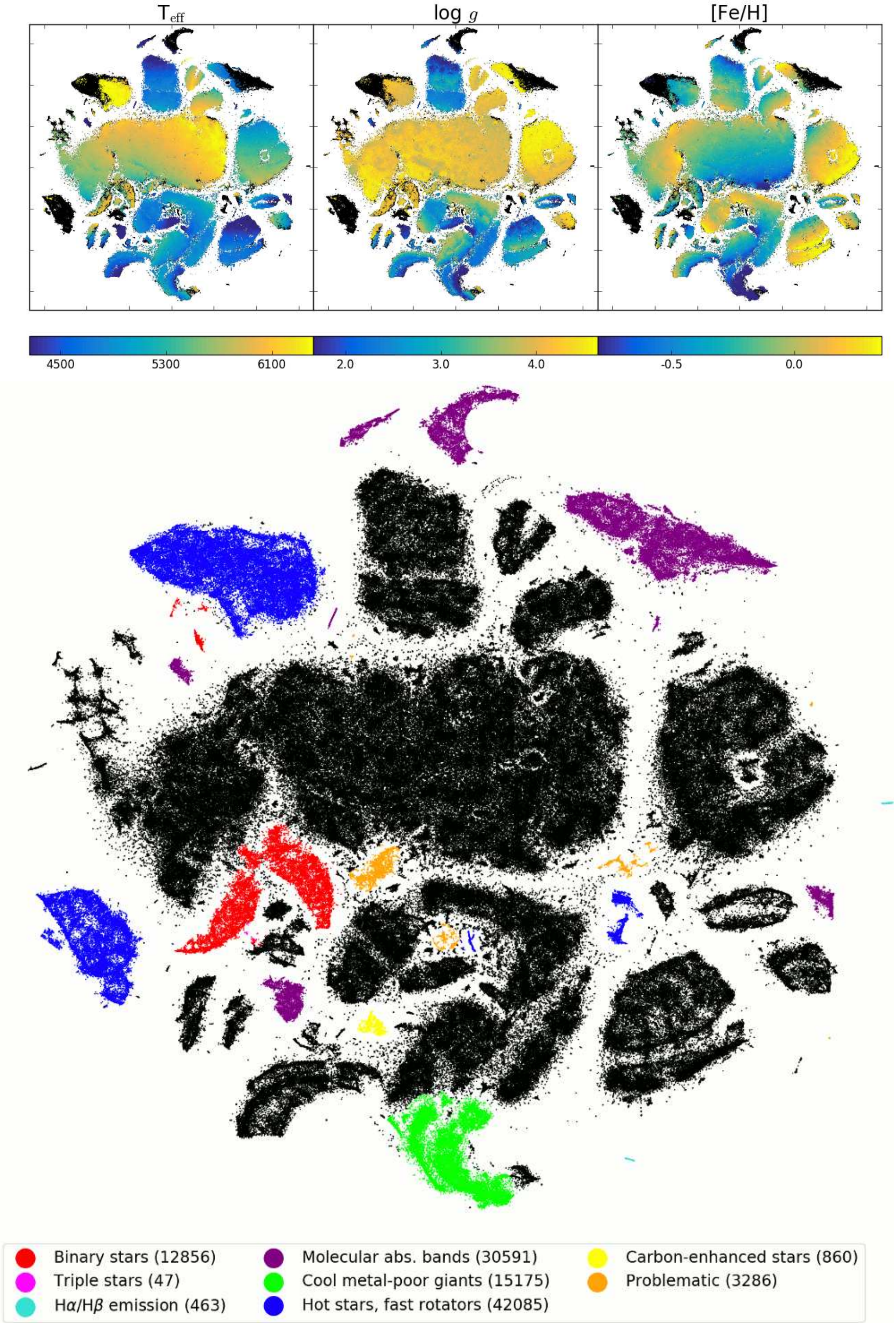


**Fig. 8.** t-SNE projection map of 587 153 spectra. Top: Colour coding of 413 920 spectra by stellar parameters as given in Buder et al. (2018), others are plotted as black points. Bottom: Colour coding by classification category. The majority of stars do not show peculiarities and are shown as black dots. The flagged triple stars (SB3) are few and hardly seen next to the lower left group of binary stars (SB2), whereas the H$\alpha$/H$\beta$ emission stars are on the far right and bottom right in the map. The count of spectra for each category is given in the legend.

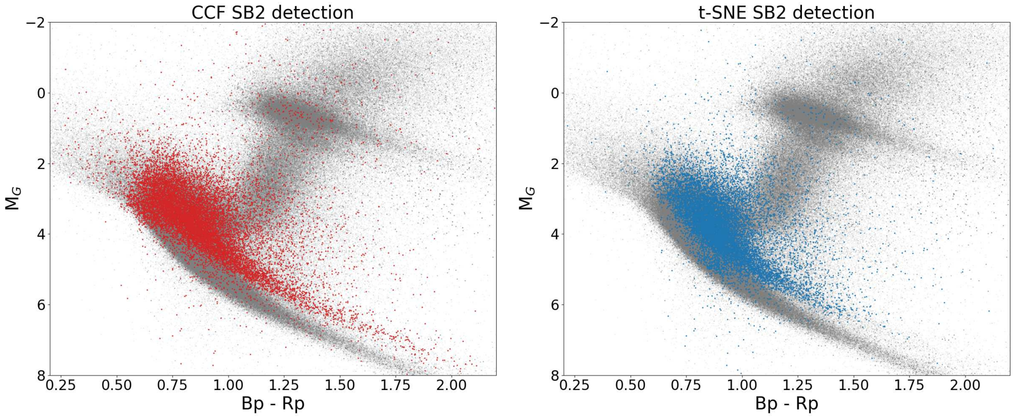


**Fig. 9.** HR diagram using only *Gaia* magnitudes and parallaxes for GALAH single stars (grey) with marked SB2 detections by CCF (left panel) and the t-SNE process (right panel) as explained in the text.

for each detected SB2 system (e.g. $T_{\mathrm{eff}[1,2]}$, $\log g_{[1,2]}$, [Fe/H], $V_{r[1,2]}$, $R_{[1,2]}$, $E(B-V)$). Among other ingredients, we also need a spectroscopic model to compare to the spectroscopic data. We therefore make use of a **generative** algorithm (**supervised** machine learning technique) called The Cannon (Ness et al. 2015) in order to construct a generative model of observed GALAH single-star spectra. Having defined the generative model, we can construct an SB2 template by summing together single-star spectra that the The Cannon model generates based on a set of labels for each of the binary components.

However, The Cannon is a polynomial interpolation method which tries to model the flux of observed spectra at each wavelength with a (usually quadratic) function of the labels. The Cannon model can be made more flexible by going to higher order functions of the labels. However, due to the complexity of the change in flux with varying labels, other methods have been proposed recently to accommodate the need for higher flexibility of the model. One of them is The Payne (Ting et al. 2018), approximating flux variation through neural networks, and a comparison of performance between The Cannon and The Payne is seen in Figure 10.

As demonstrated in El-Badry et al. (2018), a spectral model obtained by a machine learning method employing neural networks can be fitted to all observed stellar spectra, without a priori identification of multiplicity. In this way, the detection and analysis process is effectively reversed, but El-Badry et al. (2018) are thus able to identify spectroscopically unresolved binary stars, a huge advantage if we want to probe long-period binary systems (Figure 5).

## 5. Discussion and conclusions

We have shown how machine learning techniques can be applied in the discovery and analysis of multiple stars, where the signature of multiple components is embedded in their spectra. There are many ways to do that, and we discuss a few use cases where unresolved spectroscopic binary stars and those which exhibit multiple lines in the spectra are probed efficiently.

We demonstrate that classification of any observational data can be a powerful tool, helping us to (1) highlight all problematic data with unpredictable effects from either instrumentation or reduction stages, and (2) identify any peculiar spectra that are interesting per se and merit further investigation (e.g. binary stars). We present one way to efficiently perform classification by t–SNE (van der Maaten & Hinton 2008) reduction of spectral information, and we describe how modelling the spectral flux can be used to infer labels (parameters) of observed multiple stars.

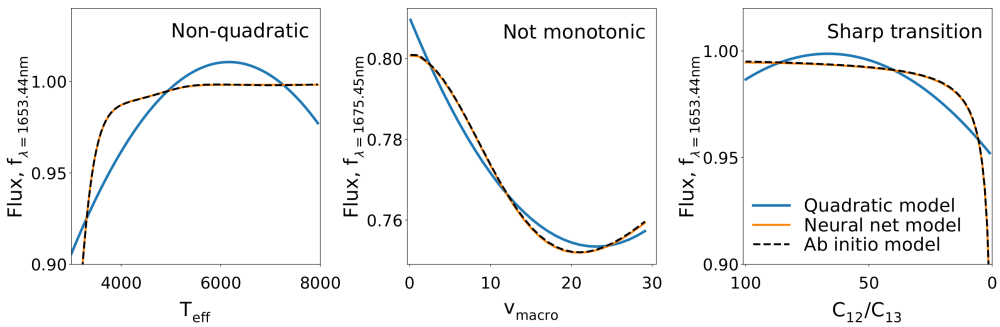


**Fig. 10.** High-fidelity spectral flux interpolation and prediction by The Payne (orange), compared to The Cannon (blue). The dashed line shows the expected flux variation at individual pixels (wavelengths) with different label variations. The three panels show three different scenarios where quadratic model does not approximate the flux well, whereas the neural network approach has no issues (adapted from Ting et al. 2018).

Whether we proceed from identification to analysis, or vice-versa, machine learning techniques (e.g. t-SNE, DBSCAN, The Cannon, The Payne) can aid in recognising the binarity and extracting binary parameters in a given data set. Nevertheless, human intervention in the whole process is still irreplaceable, as some methods require a certain amount of training or tuning, and even if left completely unsupervised, different effects can often mimic binarity (multiplicity) in observed data (e.g., Merle et al. 2017). We are still at the level of machine learning development where our methods can be easily fooled, however this field has an immense potential, and it can also be a lot of fun.

## References


Bate, M. R. 2019, MNRAS, 484, 2341

Birko, D., Zwitter, T., Grebel, E. K., et al. 2019, arXiv e-prints, `arXiv:1906.11486`

Breivik, K., Price-Whelan, A. M., D'Orazio, D. J., et al. 2019, arXiv e-prints, `arXiv:1903.05094`

Buder, S., Asplund, M., Duong, L., et al. 2018, MNRAS, 478, 4513

Dalton, G., Trager, S. C., Abrams, D. C., et al. 2012, in Ground-based and Airborne Instrumentation for Astronomy IV, ed. I.S. McLean, S.K. Ramsay, H. Takami (SPIE, Bellingham, WA), Proc. SPIE, 8446, 84460P

de Jong, R. S., Bellido-Tirado, O., Chiappini, C., et al. 2012, in Ground-based and Airborne Instrumentation for Astronomy IV, ed. I.S. McLean, S.K. Ramsay, H. Takami (SPIE, Bellingham, WA), Proc. SPIE, 8446, 84460T

De Silva, G. M., Freeman, K. C., Bland-Hawthorn, J., et al. 2015, MNRAS, 449, 2604

Duchêne, G. & Kraus, A. 2013, ARA&A, 51, 269

Duquennoy, A. & Mayor, M. 1991, A&A, 248, 485

El-Badry, K., Ting, Y.-S., Rix, H.-W., et al. 2018, MNRAS, 476, 528

Ester, M., Kriegel, H.-p., Jorg, S., & Xu, X. 1996, in Proceedings of the 2nd International Conference on KDD, ed. E. Simoudis, J. Han, U. Fayyad (AAAI Press), 226

Freeman, K. & Bland-Hawthorn, J. 2002, ARA&A, 40, 487

Gao, S., Liu, C., Zhang, X., et al. 2014, ApJ, 788, L37

Gilmore, G. et al. 2012, The Messenger, 147, 25

Huber, D. 2015, in Giants of Eclipse: The $\zeta$; Aurigae Stars and Other Binary Systems, ed. Ake T. B., Griffin E. (Springer, Cham), ASSL, 408, 169

Kounkel, M., Covey, K., Moe, M., et al. 2019, AJ, 157, 196

Lomax, O., Whitworth, A. P., Hubber, D. A., Stamatellos, D., & Walch, S. 2015, MNRAS, 447, 1550

Luo, A. L., Zhao, Y.-H., Zhao, G., et al. 2015, Research in Astronomy and Astrophysics, 15, 1095

Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94

Matijevič, G., Zwitter, T., Bienaymé, O., et al. 2011, AJ, 141, 200

Matijevič, G., Zwitter, T., Munari, U., et al. 2010, AJ, 140, 184

Merle, T., Van der Swaelmen, M., Van Eck, S., et al. 2020, A&A, 635, A155

Merle, T., Van Eck, S., Jorissen, A., et al. 2017, A&A, 608, A95

Moe, M., Kratter, K. M., & Badenes, C. 2019, ApJ, 875, 61

Ness, M., Hogg, D. W., Rix, H.-W., Ho, A. Y. Q., & Zasowski, G. 2015, ApJ, 808, 16

Parker, R. J. & Meyer, M. R. 2014, MNRAS, 442, 3722

Pietrzyński, G., Graczyk, D., Gieren, W., et al. 2013, Nature, 495, 76

Popper, D. M. 1980, ARA&A, 18, 115

Skinner, J., Covey, K., Bender, C., et al. 2018, American Astronomical Society Meeting Abstracts, #231, 231, 244.16

Stassun, K. G., Hebb, L., Lopez-Morales, M., & Prsa, A. 2009, arXiv e-prints, arXiv:0902.2548

Steinmetz, M., Zwitter, T., Siebert, A., et al. 2006, AJ, 132, 1645

Ting, Y.-S., Conroy, C., Rix, H.-W., & Cargile, P. 2018, arXiv e-prints, arXiv:1804.01530

Traven, G., Feltzing, S., Čotar, K., team, T. G., et al. submitted, A&A

Traven, G., Matijevič, G., Zwitter, T., et al. 2017, ApJS, 228, 24

van der Maaten, L. & Hinton, G. 2008, Journal of Machine Learning Research, 9, 2579

Van der Swaelmen, M., Merle, T., Van Eck, S., & Jorissen, A. 2018, in SF2A-2018: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, ed. P. Di Matteo, et al., 183