Observatório
Nacional

TESE DE DOUTORADO

APRENDIZADO DE MÁQUINA APLICADO A FOTOMETRIA MULTIBANDA

EDUARDO MACHADO PEREIRA

RIO DE JANEIRO
2025

Ministério da Ciência, Tecnologia, Inovações e Comunicações
Observatório Nacional
Programa de Pós-Graduação

Tese de Doutorado

APRENDIZADO DE MÁQUINA APLICADO A FOTOMETRIA MULTIBANDA

por

Eduardo Machado Pereira

Tese submetida ao Corpo Docente do Programa de Pós-graduação em Astronomia do Observatório Nacional, como parte dos requisitos necessários para a obtenção do Grau de Doutor em Astronomia.

Orientador: Dr. Simone Daflon

Co-orientador: Dr. Vinicius Placco

Rio de Janeiro, RJ – Brasil
Setembro de 2025

M149 Machado Pereira, Eduardo
Aprendizado de máquina aplicado a fotometria multibanda
[Rio de Janeiro] 2025.
xx, 194 p. 29,7 cm: graf. il. tab.

Tese (doutorado) - Observatório Nacional - Rio de Janeiro,
2025.

1. Astrofísica estelar. 2. Aprendizado de máquina. 3.
Fotometria multibanda. 4. Aglomerados abertos. I.
Observatório Nacional. II. Título.

CDU 000.000.000

“APRENDIZADO DE MÁQUINA APLICADO A FOTOMETRIA MULTIBANDA”

EDUARDO MACHADO PEREIRA

TESE SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM ASTRONOMIA DO OBSERVATÓRIO NACIONAL COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM ASTRONOMIA.

Aprovada por:

Dr. Simone Daflon – Observatório Nacional
(Orientador)

Dr. Vinicius Placco – NSF NOIRLab
(Co-orientador)

Dr. Hélio Perottoni – Observatório Nacional

Dr. Leandro Beraldo e Silva – Observatório Nacional

Profa. Paula Coelho – IAG/USP

Prof. Charles Bonatto – IF/UFRGS

RIO DE JANEIRO, RJ – BRASIL
26 DE SETEMBRO DE 2025

Dedico este trabalho à memória
de minha avó, cujo maior anseio
era presenciar, agora sim, este
momento.

Agradecimentos

Este relato é mais uma autorreflexão guiada pela tentativa de reconhecer contribuições que vão muito além do que se é possível trazer em palavras. Por isso, a quem se interessar, desculpo-me antecipadamente pela extensão.

Em geral, os agradecimentos em teses pendem para um lado acadêmico de uma forma que eu não tenho como seguir aqui. O desfecho do trabalho realizado ao longo dos últimos quatro anos e meio, para culminar neste documento, passa por pilares que, apesar de emaranhados, vão muito além desses aspectos puramente. Em compensação a essa limitação, e na realidade fugindo à minha abordagem usual, inicio essa lista destacando o trabalho extraordinário dos dois orientadores que viabilizaram tornar realidade um dos maiores anseios, se não o maior, de minha saudosa avó.

Sou eterna e profundamente grato a Simone e Vinícius. Além de profissionais altamente competentes e grandes referências na área, enxergam com sensibilidade ímpar as relações em um ambiente cheio de tantas fontes desumanizadoras. Dessa forma, minha gratidão abarca não só o tempo e a dedicação investidos em vários âmbitos deste projeto, mas também a paciência, a serenidade, e o cuidado constantes. Sem essa estrutura, sem dúvida a tarefa teria sido imensamente mais espinhosa, se mesmo possível.

Simone me levou a Vinícius, que me levou a Christina e Paul (e Smokey). Minha estada em Tucson foi uma vivência sem igual, e inclusive de uma tranquilidade incrível, graças a essas almas de coração gigante. Estendo minha gratidão a todos os demais envolvidos em um círculo social muito especial composto por múltiplas nacionalidades. Um pouco mais distante, mas ainda ali perto, impossível agradecer o suficiente por todo o suporte e carinho dos tão queridos Thiago, Gabi e Peter.

Ainda longe, graças a tia Jaisa, tia Wilma, tio Cláudio, tia Soraia, tia Dinda, dentre vários outros e de várias formas, desfrutei da melhor experiência que já passei. Foi incrível vivenciar uma semana de estrada conhecendo lugares incríveis e colecionando memórias que vou levar para o resto da vida.

Ponto alto dessas memórias foi poder olhar para o lado e não perder de vista meu farol. Agradeço a Giovanna todos os dias pelo privilégio que é juntar minha vida com a dela e dar cada passo em conjunto, entre conquistas e frustrações, atalhos e barreiras, viagens e entraves, mas com muito mais brindes do que aflições. Obrigado pela segurança desse colo e esse abraço que chamo de lar. Também faço questão de exaltar o papel fundamental de sua família, que me acolheu com um carinho extraordinário.

Não existem palavras para descrever meu núcleo familiar e os valores que até hoje seguem me ensinando. Posso dizer que minha mãe e meu pai representam minha essência, sendo portanto responsáveis diretos pelas relações que pude construir e que definitivamente me conduziram a essa ocasião tão comemorada.

A propósito, é certo que a estrela em que minha avó foi morar está em festa. Suspeito que vai ser assim que eu vou descobrir onde é. Até lá, certo da incapacidade de concluir a tarefa, vou tentando reconhecer toda sua dimensão.

A impossibilidade de elencar todos que deveriam ser citados aqui é refletida pelo fato de ter sido esta a última parte deste documento a ser finalizada. Eu sou francamente grato por cada conversa, conselho, cerveja, e tantas outras interações que me ensinam tanto e dão sentido a essa vida muitas vezes tão angustiante. Um brinde a "*viver as relações humanas, as únicas que transcendem: o amor, a amizade, aventura, solidariedade, família*".

Por fim, que a educação siga, em suas diferentes formas, a ascender desafortunados, a destruir mitos, e, sobretudo, a libertar.

“Whatever you do in this life,
it’s not legendary unless your friends are there to see it”

Eduardo Machado Pereira

APRENDIZADO DE MÁQUINA APLICADO A FOTOMETRIA MULTIBANDA

RESUMO

A determinação precisa dos parâmetros atmosféricos estelares é uma tarefa central em astrofísica, já que a temperatura efetiva ($T_{\rm eff}$), a gravidade superficial ($\log g$) e a metalicidade ([Fe/H]) constituem os principais observáveis que descrevem a estrutura e a evolução estelar. Tradicionalmente, essas quantidades são derivadas da espectroscopia, mas o crescimento no volume de dados de grandes levantamentos tem motivado a exploração de alternativas que aproveitem sua ampla cobertura. Técnicas de inteligência artificial (IA) e aprendizado de máquina (ML) são particularmente úteis nesse contexto, pois podem mapear padrões complexos entre fotometria e parâmetros físicos, permitindo inferências precisas para grandes amostras.

Este trabalho desenvolve um framework de ML para estimar $T_{\rm eff}$, $\log g$ e [Fe/H] de estrelas do levantamento fotométrico multibanda J-PLUS, integrando também dados dos catálogos Gaia e CatWISE, bem como espectroscopia do levantamento LAMOST. Os modelos utilizam o algoritmo de gradient boosting LightGBM e são treinados com cores corrigidas de extinção e magnitudes absolutas. Uma estratégia sistemática de seleção de features foi implementada para identificar features mais relevantes, totalizando 67, 27 e 35 para $T_{\rm eff}$, $\log g$ e [Fe/H], respectivamente. Reduções adicionais não justificam a perda de informação, de forma que esses foram os conjuntos de features adotados.

Os valores previstos foram comparados com os do LAMOST, resultando em erros absolutos médios competitivos de 42 K em $T_{\rm eff}$, 0,06 dex em $\log g$ e 0,06 dex em [Fe/H]. A aplicação a uma amostra selecionada de 154 membros de aglomerados abertos e grupos co-móveis demonstra que as metalicidades dos aglomerados em geral concordam com os reportados na literatura, embora as medidas sejam em média ligeiramente subestimadas ($-0,12$ dex). Além da análise individual das estrelas, a abordagem fornece assim um método uniforme com base em IA, fotometria e incertezas para se estimarem metalicidades de aglomerados quando a espectroscopia não está disponível, possibilitando follow ups fundamentados e estudos comparativos entre aglomerados.

Para o futuro, perspectivas incluem explorar combinações adicionais de índices de cores e previsões de outros parâmetros relevantes, como [α/Fe], fundamental no estudo da evolução química da Galáxia. Como a metodologia se baseia em princípios generalizáveis de ML e na combinação de fotometria multibanda com astrometria, pode ser aplicada a levantamentos semelhantes, como o S-PLUS e o J-PAS, ampliando seu impacto futuro e fortalecendo o papel de métodos orientados por IA na astrofísica estelar moderna.

Eduardo Machado Pereira

MACHINE LEARNING APPLIED TO MULTIBAND PHOTOMETRY

ABSTRACT

The accurate determination of stellar atmospheric parameters is a central task in astrophysics, as effective temperature ($T_{\rm eff}$), surface gravity (log g), and metallicity ([Fe/H]) are the primary observables that describe stellar structure and evolution. Traditionally, these quantities are derived from spectroscopy, but the growth in the volume of large survey data has motivated the exploration of alternatives that take advantage of their broad coverage. Artificial intelligence (AI) and machine learning (ML) techniques are particularly useful in this context, as they can map complex patterns between photometry and physical parameters, enabling precise inferences for large samples.

This work develops an ML framework to estimate $T_{\rm eff}$, log g, and [Fe/H] for stars from the J-PLUS multiband photometric survey, also integrating data from the Gaia and CatWISE catalogs as well as spectroscopy from the LAMOST survey. The models employ the LightGBM gradient boosting algorithm and are trained with extinction-corrected colors and absolute magnitudes. A systematic feature selection strategy was implemented to identify the most relevant features, totaling 67, 27, and 35 for $T_{\rm eff}$, log g, and [Fe/H], respectively. Further reductions are not justified by the resulting information loss, and these were therefore adopted as the final feature sets.

The predicted values were compared with those from LAMOST, resulting in competitive mean absolute errors of 42 K in $T_{\rm eff}$, 0.06 dex in log g, and 0.06 dex in [Fe/H]. Application to a selected sample of 154 members of open clusters and co-moving groups shows that the metallicities of the clusters are generally consistent with those reported in the literature, although the estimates are on average slightly underestimated (-0.12 dex). Beyond the individual analysis of stars, the approach thus provides a uniform and uncertainty-aware AI-based photometric method to estimate cluster metallicities when spectroscopy is not available, enabling guided follow-ups and comparative studies across clusters.

Future perspectives include exploring additional combinations of color indices and extending predictions to other relevant parameters, such as [α/Fe], which is fundamental to the study of Galactic chemical evolution. Since the methodology is based on generalizable ML principles and on the combination of multiband photometry with astrometry, it can also be applied to similar surveys such as S-PLUS and J-PAS, broadening its future impact and strengthening the role of AI-driven methods in modern stellar astrophysics.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The undeniable impact of artificial intelligence (AI) in pivotal aspects of society is revolutionary. It has been deeply transforming human life on multiple fronts, from large language models used in daily tasks to autonomous vehicles and satellite imaging. Inevitably, AI likewise stands at the vanguard of many scientific fields.

Advances in science and technology have marked the emergence of the big data era, characterized by the generation, efficient storage, and broad availability of unprecedented volumes of information. This transformation has reshaped many research fields, giving rise to new methodologies capable of handling the scale and complexity of modern datasets.

Astronomy has been at the forefront of this shift, particularly with the expansion of large observational facilities and systematic surveys producing an ever-growing wealth of data (e.g. DJORGOVSKI *et al.*, 2013). Estimates place its yearly storage requirements on the order of exabytes, comparable to the scales of genomics and surpassing those of social media platforms (STEPHENS *et al.*, 2015), already reaching the petabyte regime and moving toward the exabyte era.

Over the past decades, repeated and extensive coverage of large sky regions has enabled numerous large-scale astronomical surveys. The systematic collection and processing of vast data volumes have paved the way for increasingly ambitious spectroscopic and photometric efforts. Precise determinations of stellar atmospheric parameters, typically provided by large spectroscopic surveys, are critical for addressing a wide range of questions in stellar astrophysics. This work builds on exploring photometric efforts to derive such parameters, in particular employing machine learning techniques to map correlations between input features and desired quantities.

In this chapter, Section 1.1 presents a brief overview of the astrophysical foundations most relevant to the development of this thesis. More specifically, particular emphasis is given to stellar atmospheres and its related observables, the so-called atmospheric parameters. Next, goals and motivations of this work are presented in Section 1.2.

The rest of the thesis is organized as follows. Chapter 2 introduces the concept of large surveys, both photometric and spectroscopic, providing brief historical overviews and highlighting modern examples of these catalogs. In Chapter 3, the methodological framework is presented, covering machine learning, ensemble methods, and gradient boosting, followed by a detailed discussion of LightGBM, the tool employed in this work. Chapter 4 describes the final training sample and then outlines the construction of the final models, including their optimal features and hyper-parameters, along with performance results. Chapter 5 develops the science case of open cluster metallicities, contextualizing their astrophysical importance and describing the cluster sample used here. Finally, Chapter 6 summarizes the main results, presents the concluding remarks of this thesis, and discusses perspectives for further developments of the method.

1.1 Astrophysical foundations

Describing a star is a comprehensive task that can assume many complex forms. However, a particular quantity dictates its life even before its early stages. While the mass of a star constitutes the fundamental parameter that ultimately dictates its structure, evolution, and fate, even the onset of formation is itself conditioned by mass, through the requirement that a fragment in a molecular cloud exceeds the so-called Jeans mass (JEANS, 1902, 1928, c.f. BONNOR 1957). This critical limit establishes the minimum threshold for the collapse of a molecular cloud, allowing a protostar to emerge.

Once formed, the initial stellar mass (usually larger than the Jeans mass of the fragment that seeded it because the protostar can continue accreting material until feedback halts the process) determines the trajectory followed during nuclear burning phases and subsequent evolution. On a population scale, the distribution of initial stellar masses is described by the initial mass function (IMF), which characterizes the relative frequency of stars of different types that are born. Although processes such as stellar winds and advanced evolutionary phases can substantially reduce the mass of a star, its initial value remains the dominant factor shaping its destiny.

It is then well established that mass is identified as the primary parameter driving the internal structure, energy generation, and evolutionary timescales of stars. In addition, the environment where they are born directly influences their formation, determining their chemical composition, which in turn fixes key properties such as opacity, nuclear reaction rates, and population characteristics. This reasoning is condensed in the so-called Vogt-Russell theorem, which states that the structure of an isolated spherical, non-rotating star, without magnetic fields, can be determined from its total mass and chemical composition (MACIEL, 2016).

Although based on such simplistic assumptions, this heuristic approach proved incredibly powerful, as all higher-order stellar properties follow without additional free parameters, including radius, luminosity, internal temperature and density profiles, nuclear burning rates, and evolutionary timescales. Moreover, MACIEL (2016) suggests a modified version that includes the age, considering the modifications introduced in the stellar structure by the evolution process itself. In this way, the structure of a star is determined primarily by its mass, chemical composition, and age.

Importantly, stellar observables are commonly characterized by a standard set of quantities known as atmospheric parameters. This terminology underscores the limitation that observations probe only the outermost stellar layers, namely the atmospheres. Yet, despite this restriction, the escaping photons retain a substantial wealth of information, sufficient to precisely determine these parameters, such as effective temperature (T_{eff}), surface gravity ($\log g$), and metallicity ([Fe/H]). These are detailed in the next section.

1.1.1 Stellar atmospheric parameters

As stellar light can be separated into its constituent wavelengths, their spectra constitute one of the most valuable tools in astronomy. In fact, it is commonly stated that the field of astrophysics essentially originated with spectroscopy in the mid-19th century (e.g. GARS, 2015), since astronomical spectroscopy was the first method to reveal the chemical composition, temperature, and motion of distant celestial objects. This transformative shift, specifically linked to the study of these astronomical spectra, greatly advanced the study of the universe from mere position measurements to understanding the physical nature of galaxies and stars, for which the following three parameters are paramount. Note that this discussion is predominantly based on the content in GRAY (2005).

Effective temperature (T_{eff})

The wealth of information imprinted in distinct features of stellar light turns out to be of utmost value for atmospheric parameters, which means that spectroscopy is a go-to task employed for their determination. Because stars are good approximations for blackbodies, the shape of their spectral continuum is a first-order indicator to describe stellar spectra. It is intricately associated with the effective temperature (T_{eff}) of stars, which is defined as the temperature of a blackbody that emits the same total energy flux per unit surface area as the star.

Blackbodies are described by the spectral radiance B_ν in Planck's law:

$$B_\nu(T) = \frac{2h\nu^3}{c^2}\left[\exp\left(\frac{h\nu}{k_{\text{B}}T}\right) - 1\right]^{-1}, \tag{1.1}$$

or, equivalently in wavelength,

$$B_\lambda(T) = \frac{2hc^2}{\lambda^5}\left[\exp\left(\frac{hc}{\lambda k_{\text{B}}T}\right) - 1\right]^{-1}, \tag{1.2}$$

where h is the Planck constant, c the speed of light, k_{B} the Boltzmann constant, and T

the temperature.

The temperature dependence of B_λ (or B_ν) governs the slope and peak of the continuum, which is why it alone can already provide a strong constraint on stellar temperature. Its dominant role in shaping the continuum follows directly from the location of the maximum of the Planck function. Maximizing $B_\lambda(T)$ with respect to λ yields Wien's displacement law:

$$\lambda_{\max} T = b \approx 2.898 \times 10^{-3}\ \mathrm{m\,K}, \tag{1.3}$$

which depends only on T. As T increases, the peak shifts to shorter wavelengths and the continuum becomes bluer, so the global morphology of the continuum and broad band photometric colors are set primarily by temperature, with other atmospheric properties entering mainly through spectral lines. Figure 1.1 illustrates this behavior more explicitly for various temperatures.

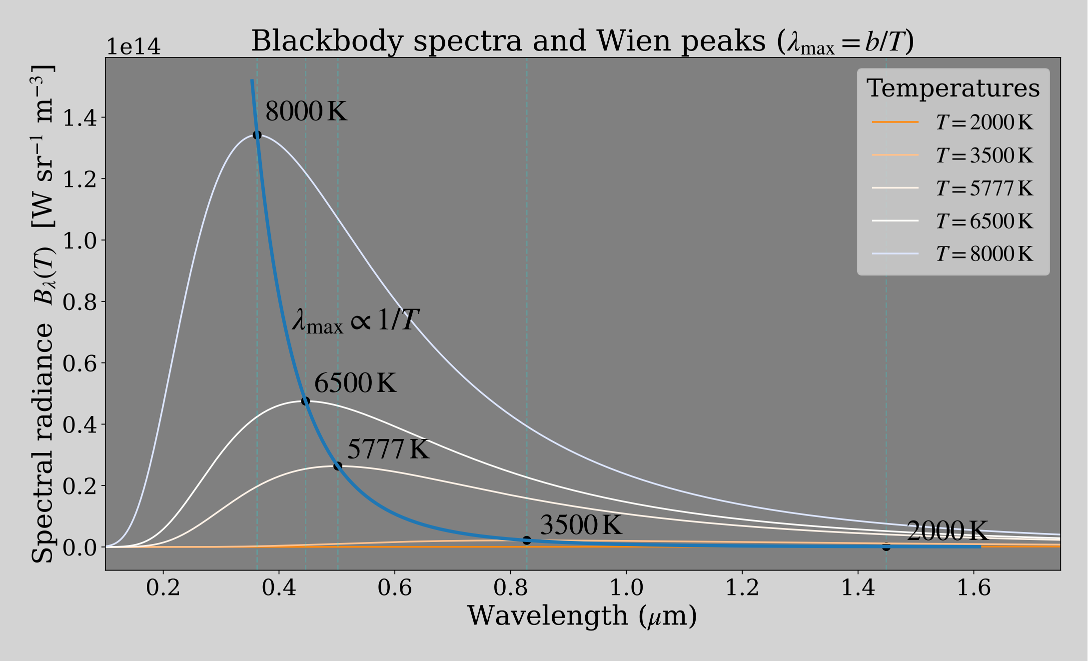

Figure 1.1: Blackbody spectra for $T = \{2000, 3500, 5777, 6500, 8000\}$ K. Vertical guide lines mark $\lambda_{\max}$ for each temperature according to Wien's displacement law. Markers indicate the spectral peaks, and the smooth curve traces the locus, $\lambda_{\max}(T)$, $B_\lambda(\lambda_{\max}, T)$, illustrating the inverse relation between the peak wavelength and temperature.

One can appropriately anticipate the spectral radiance to be linked to stellar flux, F, which in fact is defined as the energy radiated E per unit area A per unit time t:

$$F = \frac{dE}{dA\,dt}. \tag{1.4}$$

Simplifying terminology and skipping the steps of working its derivation, the emergent flux of a star F is defined as the radiated power (intrinsic luminosity L) per surface area

A:

$$F = \frac{L}{A}. \tag{1.5}$$

A straightforward dimensional analysis of Planck's law shows that integrating it over all frequencies and in every direction above the stellar surface also yields flux. Omitting the heavier math, the Stefan-Boltzmann law emerges as:

$$F = \pi \int_0^\infty B_\nu(T)\,\mathrm{d}\nu = \sigma T^4, \tag{1.6}$$

where σ is the Stefan-Boltzmann constant.

This way, by equating both flux expressions, it is shown that the stellar luminosity depends on radius and effective temperature via $L \propto R^2 T_{\text{eff}}^4$. Taking the empirical mass-luminosity relation $L \propto M^\alpha$, where α ranges from ≈ 2.3 for low-mass stars to ≈ 6.0 for the most massive stars (e.g. CUNTZ & WANG, 2018; WANG & ZHONG, 2018), recall the Vogt–Russell theorem for contextual anchoring to compare both luminosities and solve for mass:

$$M \propto (R^2 T_{\text{eff}}^4)^{1/\alpha}. \tag{1.7}$$

All that reasoning is to show that temperature can be a key observable quantity to constrain stellar mass, bridging T_{eff} directly with stellar structure.

A quick disclaimer: the mathematical formulation (Eqs. 1.1–1.7) was presented to introduce blackbody radiation through Planck's law, which provides a straightforward analytical description of the continuum. In contrast, observational signatures arising from spectral lines require a substantially more complex treatment, invoking from the outset the Saha and Boltzmann equations for detailed level populations and ionization balance, together with considerations of line formation, line broadening, and possible non-local thermodynamic equilibrium (NLTE) effects. For this reason, spectral lines are treated here by describing the qualitative roles of the parameters in shaping line profiles, rather than by presenting the underlying physics.

Besides its dominant role in shaping the continuum, temperature also affects the strengths of numerous spectral features, so both continuum morphology (accessible to photometry) and line diagnostics (accessible to spectroscopy) provide complementary constraints. Classic continuum indicators include the Balmer jump and the Paschen slope, while line-based diagnostics rely on line-depth ratios, such as V I λ6251.83 to Fe I λ6252.57 and Si I λ6125.03 to Ti I λ6126.22 (e.g. GRAY & JOHANSON, 1991), or the wings of Balmer lines (e.g. BARKLEM, 2002; CAYREL *et al.*, 2011).

In this context, local thermodynamic equilibrium (LTE) is usually assumed, so that excitation and ionization equilibria are governed by the Boltzmann and Saha equations. These expressions set the populations of atomic states and ionization stages, which in turn control the depth and dominance of spectral lines. Because these populations depend sensitively on temperature (in addition to electron pressure), the observed spectrum couples $T_{\rm eff}$ with composition: abundances directly set line strengths and indirectly affect continuum opacity through the electron supply. As line strengths respond sharply to these equilibria, even small errors in $T_{\rm eff}$ can propagate into systematic shifts in derived abundances.

To summarize, some observational techniques to estimate $T_{\rm eff}$ in practice include photometric colors or full spectral energy distributions (e.g. DAVIES *et al.*, 2013), which exploit the continuum slope; the infrared flux method (IRFM) (e.g. RAMÍREZ & MELÉNDEZ, 2005), which relates bolometric to infrared flux; and continuum discontinuities such as the Balmer jump (e.g. SOKOLOV, 1998; WOLFF *et al.*, 1968) or the Paschen slope (e.g. SCHILD *et al.*, 1971; WOLFF *et al.*, 1968). Spectroscopic approaches make use of temperature-sensitive line–depth ratios (e.g. GRAY & JOHANSON, 1991), particularly from pairs of lines with different excitation potentials, as well as the wings of Balmer lines (e.g. CAYREL *et al.*, 2011; FUHRMANN *et al.*, 1993, and references therein), which respond strongly to temperature in FGK stars. For cooler stars, the strength of molecular bands such as TiO or VO can also serve as effective thermometers (e.g. OESTREICHER & SCHMIDT-KALER, 1998).

In short, $T_{\rm eff}$ is a fundamental atmospheric parameter that dictates the continuum shape and can be derived from both photometry and spectroscopy. In either case, $T_{\rm eff}$ remains central by the Vogt–Russell theorem, both constraining stellar mass and encoding information about composition. The former, in turn, relates more closely to the next atmospheric parameter, surface gravity.

Surface gravity (log g)

Stellar mass is inherently connected to surface gravity, another one of the key atmospheric parameters usually pursued. Having $g \propto M/R^2$ from Newton's gravitation law, gravity is naturally tied to mass, hence directly encoding information on mass in the context of describing stellar structure. It is mostly seen in its logarithm form, log g, as g spans several orders of magnitude across different types of stars. log g condenses a huge range[1] into a more manageable scale: roughly 0 to 5 in most cases, i.e. (super)giants and main sequence stars, reaching larger values for white dwarfs and other compact objects. This makes it easier to compare stars on the same plot or in tables without awkward scientific notation.

Surface gravity may leave weak direct signatures in the spectrum, yet it governs the pressure structure of the stellar atmosphere. In hydrostatic equilibrium the weight of the overlying layers is balanced by the gas pressure gradient, so stars with higher log g have more compressed atmospheres and higher pressures.

---

[1] In cgs (centimeters, grams, seconds), from 10 cm s^{-2} in luminous supergiants up to 10^9 cm s^{-2} in dense white dwarfs (MACIEL, 2016).

While stronger gravity raises the pressure throughout the atmosphere, the actually observed information comes mostly from the continuum-forming layers. These are the regions where most continuum photons escape, typically defined by an optical depth $\tau = 2/3$, meaning absorption and emission are in balance[2]. The optical depth is simply a measure of how much radiation is absorbed or scattered as it travels through the atmosphere. At deeper layers ($\tau \gg 2/3$) photons are efficiently absorbed and re-emitted, while at shallower ones ($\tau \ll 2/3$) the gas is too tenuous to interact significantly with radiation. Moreover, note that $\tau \approx 2/3$ layers may occur at different physical depths for different frequencies, since opacity inside stars varies with wavelength.

Now back to the continuum-forming layers, the goal is to enlighten that this emergent radiation carries information on the pressure conditions in the $\tau = 2/3$ layers, where the photons escape. In this context, it is convenient to consider the partial pressure exerted by free electrons, P_e, which contributes to the total gas pressure (P_g): isolating it is useful for modeling ionization and opacity, even though it is physically embedded within P_g[3].

Working with P_e allows concluding that pressure, and hence gravity, is directly linked to the observable flux distribution. Reaching this conclusion is not trivial, but it can be clarified by considering two fronts.

First, opacity affects the emergent continuum of stars, since it measures how effectively that gas blocks the passage of radiation through absorption and scattering. Second, the Saha equation predicts the ratio of ionized to neutral atoms (ionization equilibrium) as a function of temperature and electron pressure.

These statements point in different directions: opacity connects the physical structure of the atmosphere to the escaping radiation, while the Saha equation links local thermodynamic conditions to the availability of free electrons. To establish their relation, one must follow how the main sources of opacity depend on electron pressure.

Stellar atmospheres are composed of mostly neutral hydrogen in photospheric layers, but a modest pool of free electrons (supplied mainly by easily ionized metals) can attach to neutral hydrogen to form H^- ions. Their large absorption cross-section (in the optical and near-infrared), together with the abundant supply, in solar-type photospheres, of neutral hydrogen ready to capture free electrons, make these ions a dominant source of continuum opacity in these stars. This way, calling on the first front, H^- is an efficient absorber of photons in the optical and near-infrared, hence affecting the emergent stellar continuum.

---

[2] The value $\tau = 2/3$ arises from the Eddington approximation to the radiative transfer equation in a gray atmosphere, whose solution yields the effective photosphere at this depth (e.g. CARROLL & OSTLIE, 2017; MIHALAS, 1978).

[3] Radiation pressure, P_{rad}, also contributes to the total pressure balance of a stellar atmosphere alongside P_g. While P_g arises from particle collisions (including electrons, ions, atoms, and molecules), P_{rad} comes from the momentum transfer of photons. It is usually negligible in stars like the Sun, but in hot, massive stars it can become comparable to or even exceed P_g, significantly influencing atmospheric structure and stability.

Now it remains to tie the occurrence of H^- to the Saha equation. Invoking this second front is straightforward: the Saha relation governs the balance between ionized and neutral atoms. Here lies the turning point, since the electron pressure P_e reappears as the bridge linking these quantities:

$$\text{surface gravity} \rightarrow P_g \rightarrow P_e \text{ (Saha)} \rightarrow \text{H}^- \rightarrow \text{opacity} \rightarrow \text{emergent continuum.}$$

Explaining the passage $P_g \rightarrow P_e$, which had not yet been stated explicitly: because the ionization fraction depends on local gas pressure (via Saha), higher gas pressure implies a higher electron pressure. Finally, the goal of this reasoning was to establish the chain by which pressure, and hence gravity, is connected to the observable flux distribution

In short, in continuum-forming layers the gas pressure controls the electron pressure, and with it the abundance of H^- ions, which dominate the continuum opacity in the optical and near-infrared. Stronger gravity therefore enhances H^- absorption, and by this mechanism the Balmer jump and the Paschen continuum slope acquire their sensitivity to pressure. In practice, flux ratios measured across well-chosen continuum and line regions provide a robust continuum-based estimate of $\log g$(e.g. GRAY & JOHANSON, 1991). This diagnostic becomes even more reliable when complemented by line-profile fitting of the wings of strong metal lines, which are highly sensitive to surface gravity through pressure broadening (e.g. AYRES & JOHNSON, 1977; EDVARDSSON, 1988).

Outside solar-type stars, the continuum responds differently to gravity. In hotter atmospheres the main opacity sources are hydrogen bound–free (a photon ejects an electron from a bound state) and hydrogen free–free (a free electron absorbs a photon while scattering off a proton without being captured), together with electron scattering. The role of H^- is therefore diminished and continuum features carry little sensitivity to pressure (e.g. KRTÍCKA *et al.*, 2024; LIND *et al.*, 2012). In cooler dwarfs the picture is further complicated by molecular absorption (for example TiO and VO) and collision-induced absorption of H_2, which dominate the opacity and can mask gravity trends. In these regimes continuum slopes or discontinuities only weakly influence surface gravity (e.g. GORLOVA *et al.*, 2003).

In spite of all these mechanisms, unlike temperature, gravity does not imprint too strongly on the continuum. Instead, more decisive diagnostics arise from spectral lines, whose shapes and broadening respond more directly to atmospheric pressure effects.

Higher gravity compresses the atmosphere, increasing particle densities and enhancing collisional and electric field interactions that broaden lines (e.g. FUHRMANN *et al.*, 1997; LIND *et al.*, 2012). Stark broadening shapes the wings of the hydrogen Balmer series(e.g. CATANZARO *et al.*, 2004), while pressure broadening is important in strong metallic features such as the Mg I b triplet around 5170 Å(e.g. FUHRMANN, 1998; FUHRMANN *et al.*, 1997). By contrast, Doppler broadening reflects thermal motion and turbulence rather than pressure, and therefore carries no direct information on gravity. Those line wings, when modeled with care, can therefore serve as valuable indicators for yielding reliable estimates of log g often outperforming continuum-based methods in precision (e.g. FUHRMANN, 1998; FUHRMANN *et al.*, 1997).

Additional gravity sensitivity arises from the balance between neutral and ionized species, with the Fe I/Fe II pair being the classical example (e.g. FUHRMANN, 1998). Because ionization fractions depend on both temperature and electron pressure, mismatches between predicted and observed line strengths reveal inconsistencies in assumed surface gravity; adjusting log g until both stages give consistent abundances is the usual correction. Velocity fields also complicate the picture. Microturbulence introduces small-scale motions that broaden saturated lines by redistributing absorption from the core into the wings, increasing equivalent width. This is corrected by adjusting the microturbulent velocity in the analysis (e.g. LIND *et al.*, 2012). Macroturbulence broadens entire line profiles and is modeled as a large-scale velocity field during fitting (e.g. FUHRMANN, 1998). Finally, non-LTE (NLTE) effects, which alter level populations in low-gravity atmospheres, are mitigated by applying corrections from detailed radiative transfer calculations (e.g. LIND *et al.*, 2012). Accounting for these effects is essential to obtain reliable values of log g.

The same pressure physics applies outside the solar-type regime, but the most useful diagnostics change with temperature and ionization. In hotter stars, Balmer and helium lines are broadened predominantly by the Stark effect and serve as primary indicators of gravity (e.g. CATANZARO *et al.*, 2004; KRTÍCKA *et al.*, 2024), while in cool dwarfs collisional broadening of the strong Na I and K I resonance lines provides the clearest signatures (e.g. GORLOVA *et al.*, 2003).

In summary, evaluating log g in practice is a complex task, either through spectroscopy or photometry. Spectroscopic strategies make use of line diagnostics, most notably the wings of strong features such as the Balmer lines or the Mg I b triplet, which broaden under higher pressures and provide reliable indicators of surface gravity (e.g. CATANZARO *et al.*, 2004; FUHRMANN, 1998). Ionization balance, classically between Fe I and Fe II, also constrains log g by requiring consistent abundances across neutral and ionized species (e.g. BERGEMANN *et al.*, 2012). Photometric strategies focus instead on continuum flux ratios, especially across features such as the Balmer jump or the Paschen slope (e.g. BESSELL, 2008). In solar-type stars, their shapes depend on electron pressure because free electrons attach to hydrogen to form H^- ions. Stronger gravity increases the supply of free electrons, enhancing H^- absorption and thus deepening these continuum signatures. While such diagnostics can reveal gravity trends, their precision is modest, since they are degenerate with temperature and metallicity and are further affected by log g (SMALLEY, 2005).

A distinct approach comes from time-series photometry, where high-precision light curves reveal stellar oscillations. The frequencies of these oscillations depend on the star's internal density and structure, providing surface gravity estimates of remarkable precision. In practice, asteroseismology stands as the most reliable method for measuring log g (CHRISTENSEN-DALSGAARD, 2021; KURTZ, 2022), though it is limited to stars with high-quality time-series data from space missions such as *Kepler* (BORUCKI *et al.*, 2010) or *TESS* (RICKER *et al.*, 2015).

From the above exposition, notice that log g determinations are inherently entangled with composition, circling back to the pivoted concept behind the Vogt-Russell theorem. While mass is inherently embedded within surface gravity, pressure sensitive spectral lines and populations are simultaneously affected by elemental abundances, thus bringing composition up. Besides anchoring this atmospheric parameter to describing stellar structure, this coupling naturally motivates the discussion of metallicity, the third of the fundamental atmospheric parameters.

Metallicity ([Fe/H])

Stellar metallicity refers to a first-order description of a star's chemical composition. In astrophysics, the term usually refers to all elements heavier than hydrogen and helium, which together constitute most of the detectable (non-dark) baryonic matter in the Universe. Within the framework of the Vogt–Russell theorem, metallicity practically stands all alone for composition itself, yet its effects are deeply entangled with those of stellar mass. At the same time, the other two parameters are also affected by composition, so this coupling complicates their interpretation, but concurrently ensures that the combined influence is reflected in the atmospheric parameters that can be measured.

In the present work, metallicity, [Fe/H], is expressed as the logarithmic abundance ratio of iron to hydrogen relative to the Sun, as is customary in stellar studies:

$$[\mathrm{Fe/H}] = \log_{10}\left(\frac{N_{\mathrm{Fe}}}{N_{\mathrm{H}}}\right)_{*} - \log_{10}\left(\frac{N_{\mathrm{Fe}}}{N_{\mathrm{H}}}\right)_{\odot}, \tag{1.8}$$

where N_{Fe} and N_{H} denote the number densities of iron and hydrogen atoms, the asterisk ($*$) indicates the star under study, and the solar symbol ($\odot$) refers to the corresponding solar ratio. Other metallicity measures include the ratio [M/H], where M represents all metals instead of only Fe, and the mass fraction $Z = 1 - X - Y$, where X and Y are the mass fractions of hydrogen and of helium, respectively, while Z constitutes all the remaining chemical elements. Mapping between [Fe/H], [M/H], and Z must assume a reference abundance pattern, meaning the relative proportions of all metal species must be fixed to a common scale. Moreover, throughout this thesis, referring to metallicity means [Fe/H] unless stated otherwise.

Iron is adopted as the tracer of metallicity because it combines several advantages: it is cosmically abundant and stably preserved once produced, it is synthesized and dispersed in large quantities by supernovae, and it reflects the integrated nucleosynthetic history of stellar populations. Observationally, it is also spectroscopically rich, with thousands of strong and measurable absorption lines across most spectral types, which makes it a practical and robust proxy for the overall metal content.

Metallicity is measured relative to hydrogen because it is by far the most abundant element and provides a stable reference baseline. Adopting the solar composition as a reference enables straightforward comparisons across different stars, while the logarithmic scale efficiently accommodates the wide dynamic range of stellar abundances. Consequently, it is a dimensionless quantity, expressed in decades units, or dex. This way, [Fe/H] $= -1$ dex corresponds to one tenth of the solar iron-to-hydrogen ratio, [Fe/H] $= 0$ dex to same amount, and [Fe/H] $= 1$ dex to one ten times larger.

Observationally, [Fe/H] is derived from photospheric iron lines. For the solar type stars emphasized here, iron is not significantly altered by internal nucleosynthesis or mixing over the main evolutionary phases, so the measured [Fe/H] reflects the heavy-element content of the gas from which the star formed. Surface convection circulates the outer envelope and keeps the photospheric composition representative of the bulk surface layers[4]. Thus [Fe/H] encodes a star's chemical heritage.

Because each star largely preserves the iron content of its birth gas, [Fe/H] can be used to trace the chemical history of the Milky Way via spatial patterns like metallicity gradients and distributions, along with age-metallicity relations (e.g. CARRARO *et al.*, 1998; CHIAPPINI *et al.*, 1997; EDVARDSSON *et al.*, 1993; TWAROG, 1980a,b). Adding key elements that are sensitive to enrichment timescales, namely α elements[5], [Fe/H] and [α/Fe] can distinguish various galactic populations, anchoring the chronology of galactic evolution (TINSLEY, 1979; WALLERSTEIN, 1962).

Arguably the most evident manifestation of stellar metallicity is seen in spectra. In most cases, absorption lines are clearly visible, preventing stellar spectral energy distributions (SEDs) from matching pure blackbodies. In general, from the Saha-Boltzmann expressions, a larger value of [Fe/H] means more metal lines (often from iron) permeate a stellar spectrum. Additionally, in some spectral regions, the continuum may be depressed by the combined absorption of many metal lines, an effect referred to as line blanketing (addressed ahead). Conversely, metal-poor stars are characterized by a shortage in metallic lines, making the spectrum appear cleaner, but dominated by the hydrogen series and a few very strong features.

In spectroscopic analyses, [Fe/H] is typically determined from Fe I and Fe II lines in a high-resolution, high signal-to-noise (S/N) spectrum. Equivalent widths of carefully selected lines with minimal blends are compared with synthetic spectra computed from

---

[4] In other regimes (such as O and B stars, A-type peculiar stars, M dwarfs, and systems altered by accretion or mass transfer) surface abundances can be altered by transport or atmospheric processes, so [Fe/H] may not trace the composition of the gas present at formation with the same fidelity.

[5] Even-Z (atomic number Z; not to be confused with mass fraction Z) nuclei produced in the He-burning α-process, including O, Mg, Si, S, Ca and Ti

one-dimensional (1D) LTE model atmospheres, with $T_{\rm eff}$, $\log g$, and microturbulence adjusted to satisfy excitation equilibrium of Fe I and ionization balance between Fe I and Fe II (e.g. GRAY, 2005; PAGEL, 2009). As an alternative, full spectral synthesis over broad wavelength ranges can simultaneously constrain [Fe/H] together with other atmospheric parameters (e.g. PAGEL, 2009; SNEDEN, 1973; SNEDEN *et al.*, 2012). In either case, diagnostics weaken at very low [Fe/H] (metal lines too few or too weak to detect) or saturate at high [Fe/H] (strong lines become optically thick and insensitive to further abundance increases) (e.g. PAGEL, 2009, and references therein).

For high-quality, high-resolution spectra, typical internal precisions can be as good as about 0.05 to 0.10 dex, while systematic differences between line lists, model atmospheres, or pipelines can be larger (e.g. PAGEL, 2009). For stars that are alike enough, a differential technique can be explored to reach unparalleled accuracies on the order of 0.01 dex (MELÉNDEZ *et al.*, 2009). Subtleties associated with 3D NLTE models can introduce additional corrections that depend on spectral type and the lines adopted (LIND & AMARSI, 2024); these are only noted here and not discussed further.

Metallicity sensitivity is strongest in the blue and near-ultraviolet, where line blanketing is most pronounced (e.g. KARATAŞ & SCHUSTER, 2006, and references therein). Line blanketing is the cumulative blocking of a star's continuum flux by a dense forest of overlapping (mostly metallic) absorption lines, which raises the effective opacity, suppresses short-wavelength flux, and redistributes energy toward longer wavelengths. This way, although much more weakly, composition also imprints on photometric indices that respond to metal line blanketing (ROMAN, 1954). With careful dereddening, these indices can be calibrated onto previously known spectroscopic scales to recover broad metallicity trends (WALLERSTEIN & CARLSON, 1960), though precision and accuracy are limited by coupling with $T_{\rm eff}$ and $\log g$, and by residual reddening (PAGEL, 2009). Achieving better sensitivity is possible via dedicated systems such as Strömgren-Crawford (CRAWFORD, 1958; STRÖMGREN, 1956, 1966), which rely on intermediate-band filters focused on line blanketing features (e.g. PAUNZEN, 2015, and references therein). More recently, improvements on photometric [Fe/H] estimates have been enabled by multiband efforts particularly exploring narrow-band filters (YUAN *et al.*, 2023), often coupled to the power of machine learning tools (e.g. GALARZA *et al.*, 2022; LOPES *et al.*, 2025; WHITTEN *et al.*, 2019).

Isochronal curves are also commonly employed in photometric metallicity estimates, since the underlying stellar models are parameterized usually by the metal mass fraction Z (BRESSAN *et al.*, 2012; CHOI *et al.*, 2016; HIDALGO *et al.*, 2018; PIETRINFERNI *et al.*, 2004; YI *et al.*, 2001). Fitting dereddened color–magnitude diagrams of star clusters, or other resolved populations, offers an alternative approach, but it often suffers from an age–metallicity degeneracy (LASTENNET & VALLS-GABAUD, 2002; WORTHEY, 1999) and requires a reference abundance scale to convert Z into [Fe/H].

Either in photometry or spectroscopy, the unifying driver for metallicity diagnostics is opacity because it governs the radiative transfer in stellar atmospheres, linking abundance patterns to both the H^- continuum and metal-line absorption. By measuring how effectively matter absorbs or scatters radiation at a given frequency, it is a fundamental physical quantity connecting the chemical composition of the gas to the emergent spectrum. Within this context, distinct types of atomic interactions that occur within the atmospheres of stars set different sources of opacity that can affect spectral lines or the continuum in different proportions.

Metals donate free electrons, strengthening the H^- continuum that shapes the optical and near infrared flux in solar-like stars. Their numerous bound-bound (photoexcitation) and bound-free (photoionization) transitions also enhance line opacity, producing the effect known as line blanketing, which suppresses short wavelength flux and redistributes energy toward longer wavelengths. By modulating opacity, metallicity also adjusts the atmospheric temperature structure and, at fixed mass, modestly shifts the dominant mode of energy transport in the envelope, as well as the evolutionary timescales from the main sequence through later phases.

Hence, ultimately, [Fe/H] is not only a chemical label but a fundamental control parameter. It directly relates to atmospheric opacity, which in turn shapes the temperature structure and the observed light, making metallicity reachable through both detailed spectroscopy and integrated broad-band photometry. In short, the composition and the emergent spectrum are tied by opacity through both lines and continuum.

However, these properties introduce observational challenges because metallicity and effective temperature intertwine. Higher metal content strengthens both the H^- continuum and line blanketing, reddening the emergent SED, while a lower T_{eff} likewise shifts flux toward longer wavelengths. As a consequence, an increase in [Fe/H] can mimic a decrease in T_{eff} in photometric colors or low-resolution spectra, resulting in a [Fe/H]-T_{eff} degeneracy (JR. *et al.*, 2014; KUDRITZKI *et al.*, 2008; MATSUNO *et al.*, 2024).

This degeneracy hinders [Fe/H] estimates more than T_{eff}, because T_{eff} is anchored by the broad continuum morphology that can be recovered at modest resolution and signal to noise and is relatively insensitive to fine line modeling. Metallicities, in turn, are read from small differential line depths and ratios that demand accurate continuum placement, careful dereddening, and clean line deblending, which makes them more vulnerable to calibration and noise, resulting in lower precision for metallicity measurements. This limitation is especially pronounced in photometric estimates, which are central to the present work.

All things considered, stellar metallicity serves as a powerful diagnostic tool, providing a first-order approximation of a star's overall chemical composition. This single parameter offers profound insights into the internal physics of stars, influencing structure, opacity, and energy transfer, which in turn shapes their temperature, pressure, and evolutionary pathways. Furthermore, the metallicity of stellar populations underpins galactic archaeology, as it encodes a fossil record of a galaxy's chemical heritage and star formation history, allowing us to reconstruct the evolutionary timeline of the Milky Way. More importantly, [Fe/H] completes the set of fundamental stellar atmospheric parameters alongside T_{eff} and $\log g$. Together, these primary observational constraints fulfill the promise of the Vogt-Russell theorem, enabling astronomers to determine a star's mass, evolutionary state, and composition, thereby providing a solid description of its structure and evolution.

Short summary

In summary, superimposed on the thermal continuum, absorption features respond to temperature, surface gravity, and chemical composition, thereby refining the atmospheric parameters. Taken individually, each atmospheric parameter diagnoses a distinct aspect of a star.

Effective temperature sets the star's energy output and spectral energy distribution, fixing the continuum slope and broadband colors while controlling excitation and ionization equilibria that determine which lines appear. Surface gravity measures the weight of the atmosphere per area and thus the pressure structure; higher $\log g$ compresses the photosphere, increases particle and electron pressures, strengthens H^- continuum opacity, and most clearly reveals itself through pressure and broadened line wings, which help distinguish evolutionary stages and constrain radius when combined with temperature. Metallicity, usually traced by [Fe/H], quantifies the heavy element content that supplies free electrons and line opacities, shaping both continuum and line strengths. In parallel, it encodes the star's birth environment and chemical history, significantly contributing to unveil the Milky Way's evolution.

The sensitivity to physical diagnostics (continuum or spectral lines) and main observational domain (photometry or spectroscopy) are very roughly outlined in Table 1.1. This is meant to serve as a heuristic quick-look consultation rather than a rigorously accurate description.

Both broadband photometry and spectroscopy provide excellent and direct measurements of T_{eff}, often used jointly, with photometry constraining the overall spectral energy distribution and spectroscopy refining the value through temperature-sensitive line diagnostics. For $\log g$, standard broadband photometry offers only an indirect handle, whereas asteroseismology from time-series photometry represents the gold standard for direct and precise measurements. Metallicity, on the other hand, cannot be reliably obtained from photometry alone: line blanketing produces only a rough estimate of [Fe/H], and accurate determinations require spectroscopy.

Furthermore, in light of the Vogt Russell theorem, stellar mass is most closely tied to T_{eff} and $\log g$, whereas chemical composition is characterized by [Fe/H]. As presented above, these three observables and the two fundamental properties act on one another

Table 1.1: Relative sensitivity of atmospheric parameters to continuum and line diagnostics, and how each domain responds to its constrains.

Parameter	Sensitivity		Domain performance	
	Continuum	Spectral lines	Photometry	Spectroscopy
$T_{\rm eff}$	Very strong	Strong	Excellent	Excellent
$\log g$	Moderate	Strong	Moderate[a]	Excellent
[Fe/H]	Mild	Very strong	Poor	Excellent

[a] Standard broadband photometry provides only an indirect constraint. In contrast, high-quality time-series photometry underlie asteroseismology, which is the gold-standard method for direct and precise $\log g$ measurement.

in an ever-turning wheel of interaction, together shaping the structure and evolution of stars. This coupled behavior results in a sophisticated system that maintains countless self gravitating bodies, all around the sky, in relative equilibrium for billions of years. This thesis develops a novel method to infer these three atmospheric parameters, which form the foundational set of observables measured for millions of stars by modern large scale surveys.

1.2 Motivation and objectives

Over the past few decades, the exponential growth in computational power has unlocked a new era of data-intensive astronomy. Alongside this, the increasing availability of high-quality multiband photometric data from large surveys such as S-PLUS, J-PLUS, and (mini)J-PAS has encouraged a wave of exploratory works employing machine learning and artificial intelligence to extract physical insights from vast datasets. An overview of a few of these surveys is presented in Chapter 2.

As already presented, the present work proposes a new method for estimating atmospheric parameters, but it also addresses an application for a real science case: open clusters. Despite the momentum on recent efforts just mentioned, studies focusing on open clusters within the Javalambre survey family remain scarce. Recent advances in astrometric precision, particularly from the Gaia mission, have led to the discovery of thousands of new open cluster candidates (e.g. CANTAT-GAUDIN & ANDERS, 2020; CASTRO-GINARD *et al.*, 2018, 2019), increasing the confirmed sample to about 5600 and the candidate pool to approximately 14 000 (PERREN *et al.*, 2023). Nevertheless, the current census is still incomplete (BONATTO *et al.*, 2006), and further efforts are required to better understand their stellar content and evolutionary properties.

In this context, multiband photometry, especially including narrow-band filters, offers a valuable tool for characterizing open clusters and their member stars. These filters provide sensitivity to specific spectral features, allowing for more accurate estimates of stellar properties and extinction effects. Previous studies combining JPLUS-like systems with Gaia data have demonstrated the advantages of such approaches, particularly in SED fitting and cluster analysis.

The main objective of this work is to develop a tool capable of estimating stellar atmospheric parameters from photometric data using machine learning and artificial intelligence methods. This is pursued through the construction of supervised models trained on matched photometric and spectroscopic data. The parameters of interest, namely effective temperature (T_{eff}), surface gravity ($\log g$), and metallicity ([Fe/H]), are inferred for individual stars in open clusters selected from the literature, located within the J-PLUS DR3 footprint. The models are based on 17 photometric bands: twelve from J-PLUS, two from CatWISE ($W1$, $W2$), and three from Gaia (G, B_{P}, and R_{P}). The spectroscopic labels are drawn from LAMOST DR8 (low resolution).

In addition to providing parameter estimates, this work aims to (i) identify the most informative photometric features for the task, (ii) deliver reliable uncertainty estimates for each prediction, and (iii) explore the potential of deriving cluster level metallicities based on member stars. The broader goal is to bridge the current gap in open cluster studies using J-PLUS photometry while showcasing the viability of data driven approaches for stellar parameter estimation.

1.2.1 Scope and limitations of the method

The photometric machine learning models presented in this work were developed to estimate a complete set of stellar parameters (T_{eff}, $\log g$, and [Fe/H]) from multiband photometry. Their intended use is not to replace spectroscopic determinations, but to provide a homogeneous and uncertainty-aware alternative in regimes where spectroscopy is unavailable, incomplete, or heterogeneous. In particular, when regarding clusters, they are useful to derive provisional cluster means when only a few spectra exist, to establish a uniform baseline for comparative studies across many clusters, and to act as a consistency check on membership and photometric quality.

Nevertheless, a few limitations must be acknowledged. First, the models are constrained by the parameter space of the training set. Predictions for stars outside or near the edges of this range should be interpreted with caution, as uncertainties increase and systematic deviations are likely. Second, the models inherit the global biases of the training distribution which must be accounted for in applications. Third, spectroscopic abundances remain superior for precise chemical studies, and the present approach is best suited to exploratory work, survey planning, and population-level comparisons.

In summary, the method contributes a practical pathway to approximate stellar parameters from photometry when spectroscopy is lacking, providing internally consistent values with propagated uncertainties. Its use should be restricted to the parameter ranges where the training set offers robust coverage, and its outputs are most valuable when employed as complementary context rather than as a substitute for high-quality spectra.

Chapter 2

Large surveys

Sky surveys are of paramount importance in astronomy. Since antiquity, civilizations have systematically catalogued celestial objects and their motions to understand and predict astronomical phenomena. In this chapter, I present modern photometric (Section 2.1) and spectroscopic (Section 2.2) surveys, providing the necessary context and laying the groundwork for the motivation behind this work.

2.1 Photometric surveys

Photometry, the measurement of stellar brightness, has long played a central role in astronomy. This section outlines the development of photometric surveys, beginning with a brief historical overview and progressing toward the era of modern, large-scale sky mapping. Afterwards, in Sections 2.1.1 and 2.1.2, particular emphasis is given to the surveys most relevant to this work: J-PLUS (alongside J-PAS and S-PLUS, although these do not play a central role here), Gaia, and CatWISE.

Historical background

Systematic observation of the sky is one of humanity's oldest scientific endeavors. Long before the advent of modern instruments, ancient civilizations developed sophisticated methods to track celestial events. Babylonian astronomers, active since at least the second millennium before the Common Era (BCE), compiled remarkably accurate records of planetary motions, eclipses, and transient phenomena, establishing some of the earliest known longitudinal datasets in astronomy (STEPHENSON, 1997). Importantly as well, Chinese (LI, 2000), Egyptian (DEPUYDT, 1997), and Mesoamerican (AVENI, 2001) cultures constructed detailed calendrical systems anchored in astronomical cycles.

The Hellenistic period introduced more systematic approaches to sky mapping. In the 2nd century BCE, Hipparchus compiled one of the first known star catalogs and proposed a six-point scale for stellar brightness, ranking stars from first to sixth magnitude according to their apparent intensity. This system, based on visual perception, was later formalized in Ptolemy's influential *Almagest* (2nd century CE), which preserved and extended Hipparchus's work and became a cornerstone of astronomical tradition (GRASSHOFF, 1990).

Between these eras, during the Islamic Golden Age, astronomers al-Ṣūfī (964 CE) and Ulugh Bēg (1437 CE) preserved and refined the star catalogs of antiquity. Their contributions are best known through al-Ṣūfī's *Book of Fixed Stars*, which updated Ptolemaic positions and magnitudes based on new observations, and Ulugh Bēg's *Zīj-i Sultānī*, a catalog of nearly 1000 stars compiled at the Samarqand Observatory with exceptional pre-telescopic precision. Their works improved both positional and brightness estimates by employing more advanced instruments and more systematic observational methods (HAFEZ, 2010). These efforts helped maintain continuity between the classical tradition and the scientific transformations that unfolded in early modern Europe.

In 1856, Norman Pogson provided a quantitative foundation for the stellar magnitude system by defining a five-magnitude difference as equivalent to a brightness ratio of exactly 100 to 1, a formalization of the six-point scale first introduced by Hipparchus (POGSON, 1856). While the original system was rooted in visual perception, Pogson's logarithmic formulation closely mirrored the nonlinear response of the human eye.

By the late 19th century, the advent of astrophotography enabled more objective and archival imaging of the sky. For the first time, large areas could be imaged with uniform sensitivity and preserved as permanent records. This advance facilitated precise and repeatable brightness measurements, as demonstrated by early large-scale photographic surveys such as the Harvard College Observatory plate archive and the international Carte du Ciel project (see LAYCOCK *et al.* 2010 and VICENTE *et al.* 2007, for each, respectively). These early photographic surveys marked a turning point in astronomical data collection, offering a level of uniformity, depth, and archival value that visual methods could not achieve.

Modern photometric surveys

The legacy of photographic surveys laid the foundation for a new technological era, when photometric data could be acquired with unprecedented precision, efficiency, and scalability. This shift enabled the emergence of truly systematic, large-scale surveys that would redefine the scope of astrophysical studies.

A decisive breakthrough came with the introduction of charge-coupled devices (CCDs) in the 1970s and their widespread adoption in the following decade. Replacing photographic plates, CCDs offered dramatic improvements in sensitivity, linearity, and dynamic range. These detectors allowed for more accurate and reproducible measurements of stellar fluxes across different bands, enabling uniform calibration, error control, and automated data acquisition. With these advancements, modern photometric surveys could efficiently cover vast areas of the sky while producing homogeneous datasets suitable for statistical analysis. This technological shift set the course for a new generation of sky mapping efforts, in which stellar magnitudes and colors could be systematically measured and analyzed at scale, powering discoveries across nearly every field of astronomy.

Crucially, many of these surveys were built upon and carefully aligned with the standardized framework of the Johnson–Morgan *UBV* system (JOHNSON & MORGAN, 1953) and its subsequent *RI* (JOHNSON *et al.*, 1966; KRON *et al.*, 1953) and *JKLMN* expansions (JOHNSON, 1964). Additionally, complementary systems such as the four-color Strömgren–Crawford $uvby\beta$ filters (CRAWFORD, 1958; STRÖMGREN, 1956, 1966) offered finer diagnostic power for stellar classification. For several decades, the Johnson and Johnson–Cousins systems served as the dominant framework for optical photometry, guiding the design of small- and medium-scale observational programs worldwide. This dominance persisted until the advent of large-scale digital surveys, which introduced new filter systems tailored to CCD detectors and automated data pipelines.

The Sloan Digital Sky Survey (SDSS; YORK *et al.*, 2000) introduced a novel *ugriz* filter set, but provided well-characterized transformations to traditional UBVRI bands, ensuring cross-survey consistency. In the near-infrared, the Two Micron All Sky Survey (*2MASS*; SKRUTSKIE *et al.*, 2006) offered all-sky coverage in the JHK_s bands, becoming a foundational dataset for various kinds of studies. The *SkyMapper* Southern Survey (KELLER *et al.*, 2007) adopted a custom set of six filters (*uvgriz*), calibrated to connect with both SDSS and Johnson systems while enhancing sensitivity to metallicity and temperature. Similarly, the *Pristine* survey (STARKENBURG *et al.*, 2017) employs a narrow-band Ca H & K filter in combination with SDSS-like *gri* photometry to efficiently identify extremely metal-poor stars across the Galactic halo.

In particular, recent efforts dedicated to multi-band photometry are central to the present work. The Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS; BENITEZ *et al.*, 2014), the Javalambre Photometric Local Universe Survey (J-PLUS; CENARRO *et al.*, 2019), and the Southern Photometric Local Universe Survey (S-PLUS; MENDES DE OLIVEIRA *et al.*, 2019) refine the visible part of the spectrum with multiple filters, including strategically placed narrow ones, as detailed in the following section.

2.1.1 J-PAS, J-PLUS, S-PLUS

J-PAS, J-PLUS, and S-PLUS are closely related narrow- and intermediate-band photometric surveys that share a common instrumental design. While J-PAS and J-PLUS are conducted at the Observatorio Astrofísico de Javalambre in Spain, S-PLUS operates from the southern hemisphere, using the NSF Cerro Tololo Inter-American Observatory (CTIO), the identical telescope, camera, and filter set as J-PLUS. Although originally conceived for cosmological and local universe studies, these surveys' dense spectral sampling across the optical range also enables a broad range of stellar astrophysics applications.

J-PAS

The Javalambre–Physics of the Accelerating Universe Astrophysical Survey (J-PAS; BENITEZ *et al.*, 2014) is an ambitious imaging program designed to map approximately 8500 deg^2 of the northern sky using a dedicated 2.5-meter telescope (JST/T250) at the Observatorio Astrofísico de Javalambre (OAJ; CENARRO *et al.*, 2010, 2012, 2014). It employs a unique set of 54 overlapping narrow-band filters ($\approx$ 150 Å), complemented by two mid-band filters ($\approx$ 600 Å) covering the blue and near-infrared ends of the optical window. With one additional broad-band Sloan-like filter (iSDSS), this configuration results in 57 filters in the Early Data Release[1] (EDR), offering nearly continuous sampling from 3300 to 9600 Å. The planned footprint and observed regions in EDR are depicted in Figure 2.1.

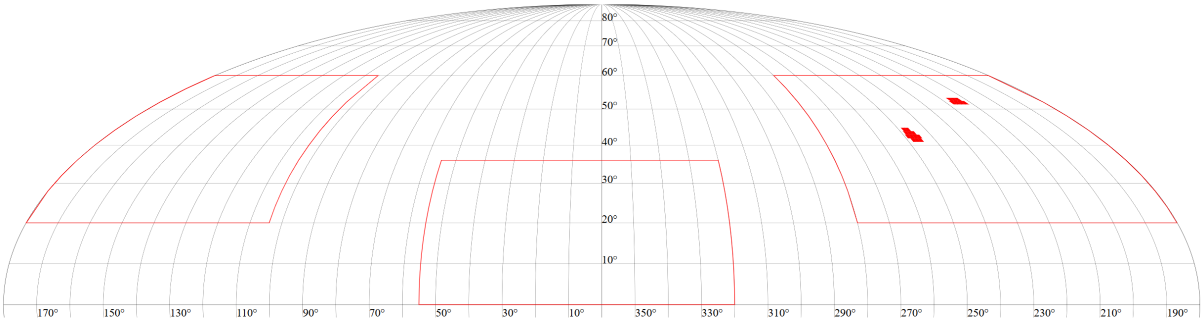

Figure 2.1: Planned sky coverage of the J-PAS survey in Aitoff projection (thin red outlines), showing the survey's full intended footprint. The filled red regions indicate the area observed and released in the Early Data Release (EDR). This scheme is available under the option `Coverage map` in EDR Data Access Services at J-PAS web site.

J-PAS filters are named according to the approximate central wavelength of their transmission curve in nanometers following the letter `J` (e.g., J1007 peaks near 10070 Å), while legacy or calibration filters retain conventional labels such as uJAVA (a custom ultraviolet filter tailored to the Javalambre system, distinct from uSDSS)[2]. The two mid-band filters, uJAVA and J1007, are designed to improve sensitivity to the Balmer and 4000 Å breaks, and to anchor the spectral energy distribution (SED) in regions where stellar continua show strong gradients, respectively. The broad-band filter, iSDSS, provides enhanced signal-to-noise ratio (SNR) in the red and facilitates cross-survey compatibility, particularly for weak lensing and photometric calibration. The full set of filters has a typical full width at half maximum (FWHM) of 145 Å, where the FWHM denotes the width of a filter's transmission curve measured at 50% of its peak throughput. This specification

---

[1] `https://www.j-pas.org/datareleases/jpas_early_data_release`

[2] Filters named with a letter followed by *SDSS* indicate Sloan-like filters.

provides a practical estimate of the spectral resolution and is constrained by the minimum width achievable by the manufacturer while preserving transmission homogeneity across the field of view (BENITEZ *et al.*, 2014).

Although primarily motivated by cosmological goals, such as precise photometric redshifts for baryon acoustic oscillation studies, J-PAS data also serve broader astrophysical purposes, including stellar populations and Galactic structure. Initial science verification and technical validation of the full J-PAS observing strategy were carried out using the Pathfinder camera on JST/T250, yielding the miniJPAS dataset (BONOLI *et al.*, 2021). This pilot survey focused on the well-studied Extended Groth Strip (EGS) field (e.g. DAVIS *et al.*, 2007), covering an area of approximately 1 $\deg^2$, and employed an extended filter set comprising the 57 EDR bands plus three additional broad filters: uJ-PAS, gSDSS, and rSDSS. The miniJPAS configuration thus totals 60 filters, enhancing photometric depth and blue-end coverage for stellar and extragalactic studies.

In the context of stellar astrophysics, YUAN *et al.* (2023) demonstrated the capabilities of the miniJPAS dataset for deriving stellar atmospheric parameters using its full complement of 56 J-PAS filters plus four SDSS-like bands. Effective temperatures were obtained through SED fitting using the VOSA tool (BAYO *et al.*, 2008), achieving uncertainties below 150 K and showing good agreement with spectroscopic estimates from LAMOST and photometric estimates from J-PLUS. They also constructed metallicity-dependent stellar loci for FGK dwarfs, achieving a precision of $\sim$0.1 dex and identifying specific filters (e.g., J0378, J0390, J0510) as particularly sensitive to chemical abundances. Their work further showed the feasibility of using J-PAS photometry to distinguish dwarfs from giants through machine learning classification.

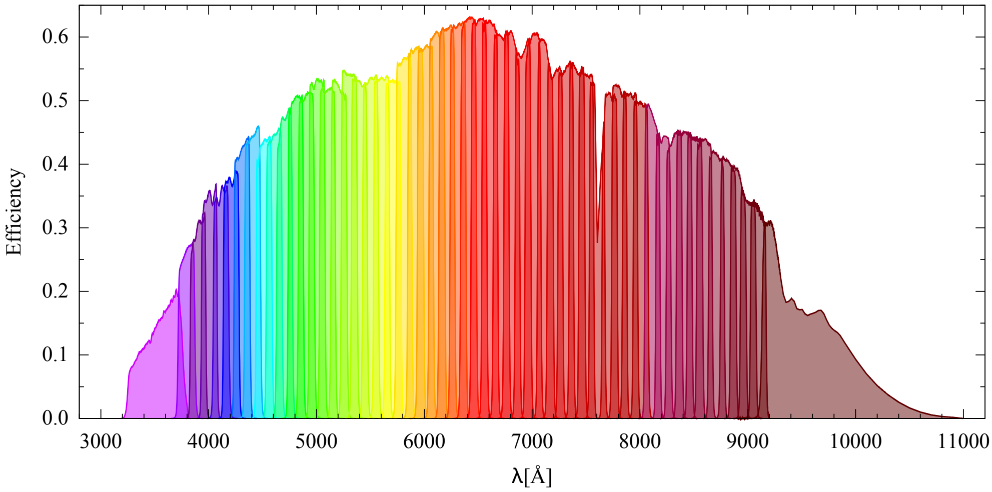

Figure 2.2: Measured transmission curves of the J-PAS filters as implemented in miniJPAS. Effects of the CCD quantum efficiency, the entire optical system of the JST/T250 telescope, and sky absorption are included. Figure 2 from BONOLI *et al.* (2021).

Figure 2.2 shows the measured transmission curves of the complete filter set used in miniJPAS. It includes the combined effects of the CCD quantum efficiency, the full optical system of the JST/T250 telescope, and atmospheric transmission. A complete listing of all filters used in miniJPAS, including their central wavelengths and widths, is provided in Appendix A.

Originally conceived to support the photometric calibration of J-PAS, J-PLUS was developed as a complementary project sharing the same observatory infrastructure, although distinct instrumental design. A detailed description follows in the next section.

J-PLUS

While J-PAS represents the most ambitious effort within the Javalambre survey family, the Javalambre Photometric Local Universe Survey (J-PLUS; CENARRO *et al.*, 2019) serves as a foundational and complementary project, offering essential calibration data and independent scientific value. J-PLUS is conducted with a dedicated 0.83-meter telescope (JST/T80) at the OAJ and is designed to observe over 8500 deg^2 of the northern sky. The planned footprint and observed regions up to DR3 are depicted in Figure 2.3.

Its filter system consists of 12 optical bands, including five broad-band filters akin to those used in SDSS (*ugriz*), two medium-width filters ($\approx 200 - 400$ Å), and five narrow-band filters (≈ 100 Å) strategically centered on key astrophysical features. This configuration enables both precise photometric calibration and a wide range of science applications, from stellar population studies to emission-line object searches, all within a lower-resolution spectral mapping framework.

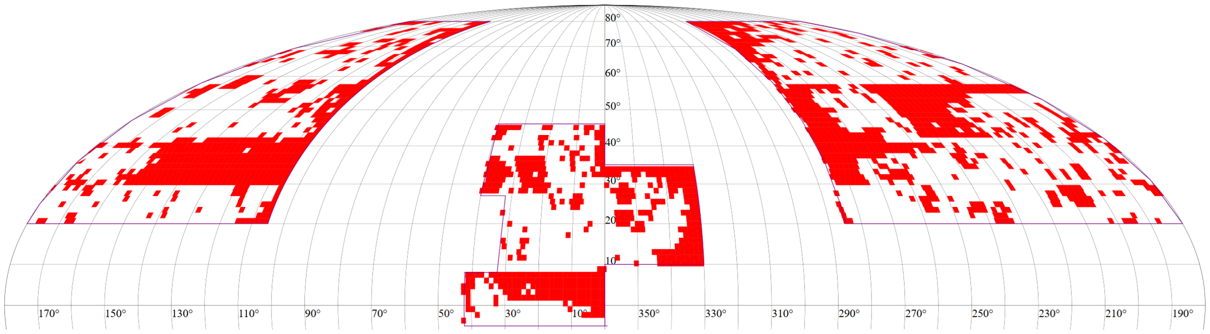

Figure 2.3: Planned sky coverage of the J-PLUS survey (thin purple outlines) in Aitoff projection, showing the survey's full intended footprint. The filled red regions indicate the area observed and released up to the Third Data Release (DR3). This scheme is available under the option `Coverage map` in DR3 Data Access Services at J-PLUS web site.

Table 2.1: Filter characteristics of J-PLUS. Table adapted from CENARRO *et al.* (2019).

Filter	λ_{central} (Å)	FWHM (Å)	Comments
uJAVA	3485	508	(a); (b)

Filter	λ_{central} (Å)	FWHM (Å)	Comments
J0378	3785	168	[O II]; (a)
J0395	3950	100	Ca H + K
J0410	4100	200	Hδ; (a)
J0430	4300	200	G band; (a)
gSDSS	4803	1409	SDSS
J0515	5150	200	Mg b triplet
rSDSS	6254	1388	SDSS
J0660	6600	138	Hα; (a)
iSDSS	7668	1535	SDSS
J0861	8610	400	Ca triplet
zSDSS	9114	1409	SDSS

Notes. (a) In common with J-PAS; (b) Sloan-like Javalambre blue mid-band filter. Rest-frame key spectral features matching the location of narrow- and intermediate-band filters are also indicated.

As in the J-PAS system, J-PLUS filters are named based on either their central wavelength or their calibration role. Broad-band filters adopt conventional labels such as u, g, r, i, and z, consistent with SDSS nomenclature, while the medium- and narrow-band filters are labeled according to their central wavelengths in nanometers, as in J-PAS (e.g., J0378, J0515). The medium- and narrow-band filters are specifically centered on prominent stellar absorption or nebular emission features. J0378 targets the [O II] λ3727 line and contributes to Balmer jump sensitivity, while J0395 covers the Ca II H + K lines. J0410 and J0430 are centered on the Hδ absorption and the G band, respectively. J0515 samples the Mg b triplet, and J0660 isolates the Hα line. The intermediate-width J0861 probes the near-infrared Ca II triplet, supporting metallicity and surface gravity diagnostics. This strategic placement of filters allows J-PLUS to recover low-resolution spectral information critical for stellar parameter estimation, star formation studies, and emission-line object detection. The filter system achieves a typical FWHM of approximately 100 Å for narrow bands and 200–400 Å for medium bands. This setup balances spectral resolution and photometric depth, optimizing the survey for stellar, nebular, and extragalactic science (CENARRO *et al.*, 2019). The J-PLUS filter system is described in Table 2.1 and illustrated in Figure 2.4

In the scope of stellar astrophysics, several studies using J-PLUS data have been conducted. In particular, different machine learning approaches were employed for various kinds of investigations across the three data releases available in J-PLUS, including measurements of stellar parameters (GALARZA *et al.*, 2022; HUANG *et al.*, 2024; WANG *et al.*, 2022; WHITTEN *et al.*, 2019; YANG *et al.*, 2022) and abundances (HUANG *et al.*, 2024; YANG *et al.*, 2022), identification of metal-poor stars (GALARZA *et al.*, 2022;

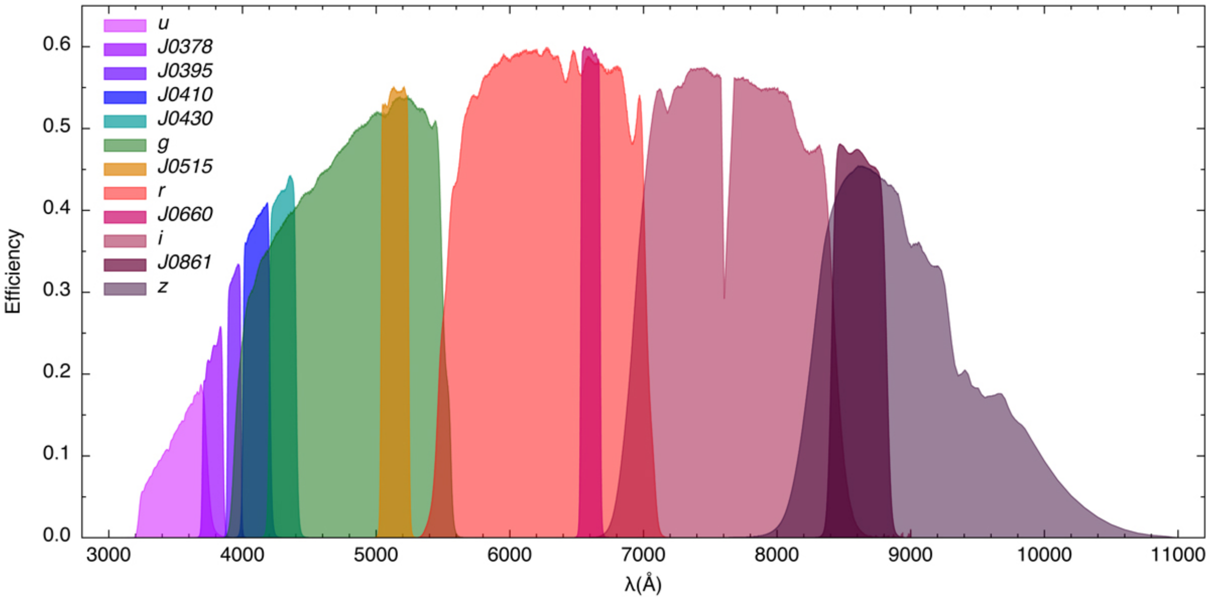

Figure 2.4: Overall transmission curves measured for the set of 12 J-PLUS filters, including systematic effects of sky, mirrors, lenses, and CCD. Figure 3 from CENARRO *et al.* (2019).

WHITTEN *et al.*, 2019) and ultra-cool dwarfs (MAS-BUITRAGO *et al.*, 2022; SOLANO *et al.*, 2019), high velocity stars characterization (QUISPE-HUAYNASI *et al.*, 2023), multiple populations in globular clusters (BONATTO *et al.*, 2019), and calibration studies (LÓPEZ-SANJUAN *et al.*, 2021; XIAO *et al.*, 2023). In Chapter 3, a few works which employed machine learning techniques for predicting atmospheric parameters will be further detailed, as well as the data used in this work.

Conducted from the southern hemisphere, the Southern Photometric Local Universe Survey (S-PLUS; MENDES DE OLIVEIRA *et al.*, 2019) extends J-PLUS coverage by implementing similar instrumental configuration and observational strategy at a different site. S-PLUS is detailed in the next section.

S-PLUS

Building on the instrumental design and many of the scientific goals of J-PLUS, the Southern Photometric Local Universe Survey (S-PLUS; MENDES DE OLIVEIRA *et al.*, 2019) broadens sky coverage and enables complementary science from the southern hemisphere, with a plan of covering 9300 deg^2. S-PLUS is conducted with the T80-South telescope (T80S), a dedicated 0.83-meter instrument installed at the NSF CTIO in Chile, employing an identical instrumental configuration to its northern counterpart. It uses the same camera design and the same set of 12 optical filters—comprising five broad bands, two medium-width bands, and five narrow bands—strategically selected to capture key spectral features.

This setup enables cohesive multi-band photometry across the entire sky and supports low-resolution spectral mapping over a wide area of the southern sky. Figure 2.5

demonstrates the versatility of the 12-filter system by showing example spectra of diverse astrophysical sources (a quasar, a galaxy, an A0-type star, a planetary nebula, and a symbiotic system) convolved with the S-PLUS filters. The resulting photometric points illustrate how spectral features can be captured with significantly greater detail than in broad-band photometry.

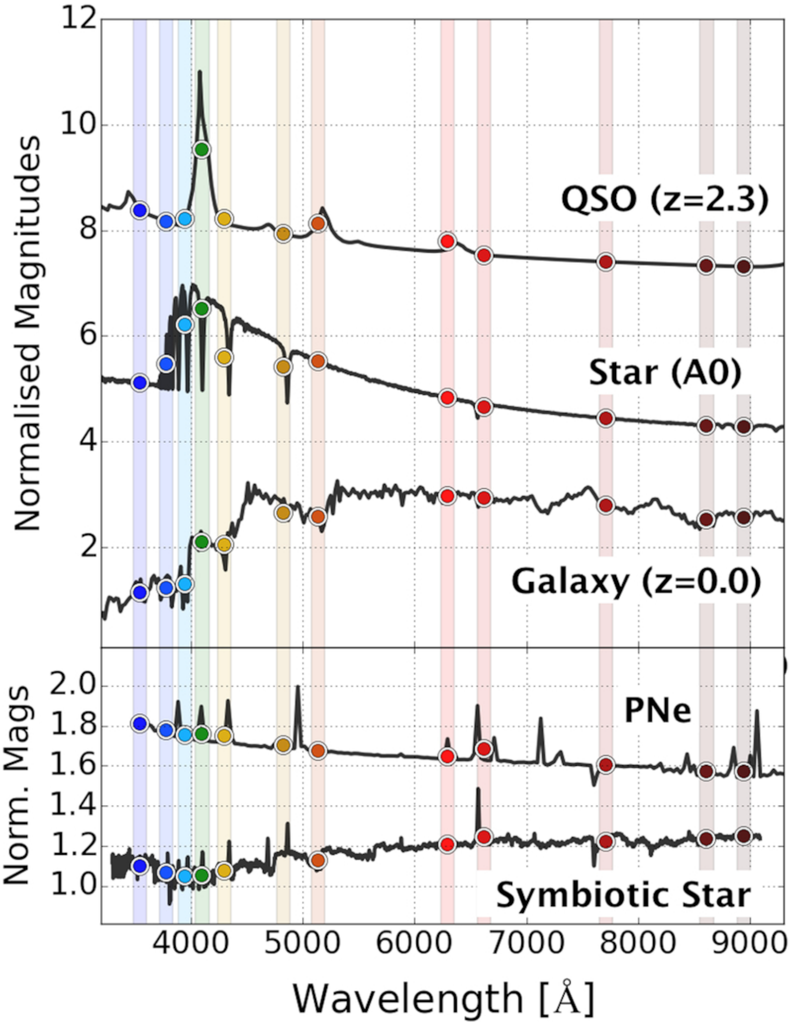

Figure 2.5: Examples of spectral energy distributions (SEDs; black lines) convolved with the S-PLUS 12-filter photometric system (colored dots). From top to bottom: a quasar at $z = 2.3$, a main-sequence star of spectral type A0, a local galaxy ($z = 0.0$), a planetary nebula (PNe), and a symbiotic star. Vertical shaded regions represent the effective wavelength ranges of the S-PLUS filters, and the colored dots mark the corresponding synthetic magnitudes obtained by convolving each spectrum with the respective filter transmission curve. Figure 7 from MENDES DE OLIVEIRA *et al.* (2019).

Table 2.2: Summary of S-PLUS filters as in Table 2 from MENDES DE OLIVEIRA *et al.* (2019).

Filter name	λ_{eff} (Å)	$\Delta\lambda$ (Å)	Comment
uJAVA	3563	352	Javalambre u
J0378	3770	151	[OII]
J0395	3940	103	Ca H + K
J0410	4094	201	Hδ
J0430	4292	201	G band
gSDSS	4751	1545	SDSS-like g
J0515	5133	207	Mgb Triplet
rSDSS	6258	1465	SDSS-like r
J0660	6614	147	Hα
iSDSS	7690	1506	SDSS-like i
J0861	8611	408	Ca Triplet
zSDSS	8831	1182	SDSS-like z

As in J-PLUS, the filter system in S-PLUS achieves a typical FWHM of approximately 100 Å in narrow-band filters and 200–400 Åin medium-width filters. The survey footprint and current coverage are shown in Figure 2.6. In Table 2.2, Table 2 is reproduced from MENDES DE OLIVEIRA *et al.* (2019) to illustrate the similarities with the J-PLUS filter system.

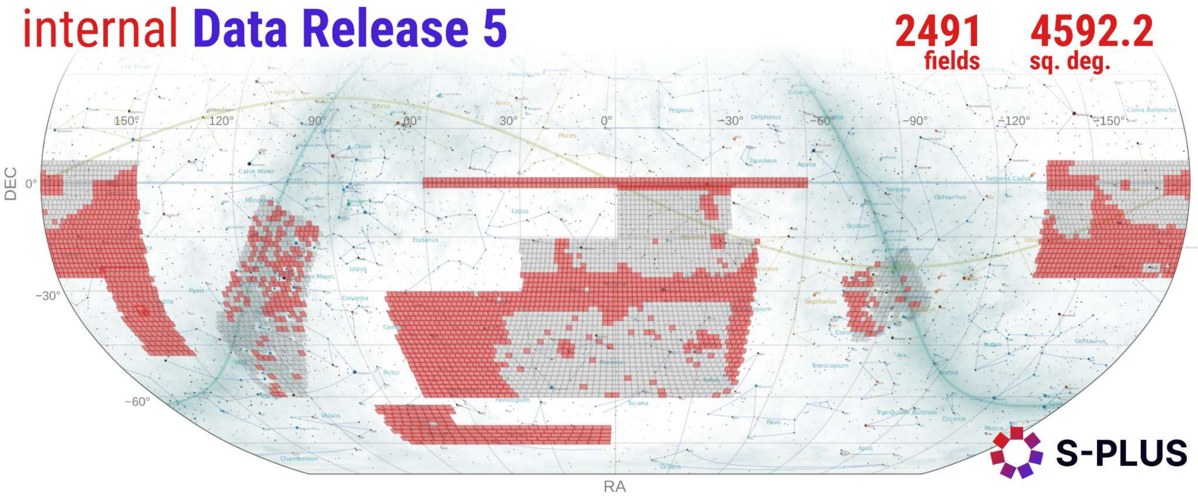

Figure 2.6: Planned sky coverage of the S-PLUS survey (gray squares) in Aitoff projection, showing the survey's full intended footprint. Red squares indicate the area observed and released up to the internal Fifth Data Release (iDR5). Figure available in the iDR5 Documentation page in the S-PLUS Data Access Interface (S-PLUS Cloud).

To best support the various scientific goals pursued within the S-PLUS collaboration, the overall survey is divided into five sub-surveys: the Main Survey, the Ultra-Short Survey, the Variability Fields, the Galactic Survey, and the Marble Field Survey. While sharing J-PLUS's science drivers such as stellar population studies, emission-line object searches, and photometric redshift estimation, S-PLUS additionally targets unique southern structures, including the galactic bulge—within the Galactic Survey—and the Magellanic Clouds—within the Main and Ultra-Short Surveys. In particular, the work of PEROTTONI *et al.* (2024) describes the first release (DR1) of the Ultra-Short Survey (USS), a photometric survey with short exposure times[3] that will cover approximately 9300 deg^2 of the Southern sky. The USS DR1 contains data for 163 observed fields totaling $\sim$ 324 deg^2 along the Celestial Equator.

Furthermore, as of the writing of this thesis, S-PLUS has surveyed approximately half of its planned footprint. The internal Data Release (iDR5) observed $\approx$ 4500 deg^2, as indicated in Figure 2.6, and is available only for members of the collaboration. Meanwhile, publicly available DR4 covered $\approx$ 3000 deg^2 and is accessible through S-PLUS Data Access Interface[4] (OLIVEIRA SCHWARZ, 2022) under the documentation section.

J-PAS, J-PLUS, and S-PLUS provide complementary optical photometric coverage of the northern and southern skies, composing a seminal dataset for a wide range of topics. However, the scientific goals of this work also rely on astrometric and infrared information not available in these surveys. The following section presents two supporting datasets that provide essential positional and photometric information used throughout this thesis: *Gaia* and CatWISE.

2.1.2 (Cat)WISE, Gaia

Two additional surveys relevant to this work are briefly discussed in this section: CatWISE and *Gaia*. The former was selected to incorporate infrared photometric information into our models, while the latter provides astrometric measurements of unmatched precision. Each catalog is introduced separately below.

(Cat)WISE

The CatWISE2020 catalog (MAROCCO *et al.*, 2021), one of the key datasets used in this work, represents the culmination of several efforts to extract proper motions and mid-infrared photometry from the Wide-field Infrared Survey Explorer mission. Originally launched as *WISE* (WRIGHT *et al.*, 2010), this space-based observatory mapped the entire sky in four mid-infrared bands, providing crucial data for a wide range of astrophysical investigations.

---

[3] Approximately 1/36th of the nominal values used in the S-PLUS survey,

[4] `splus.cloud`

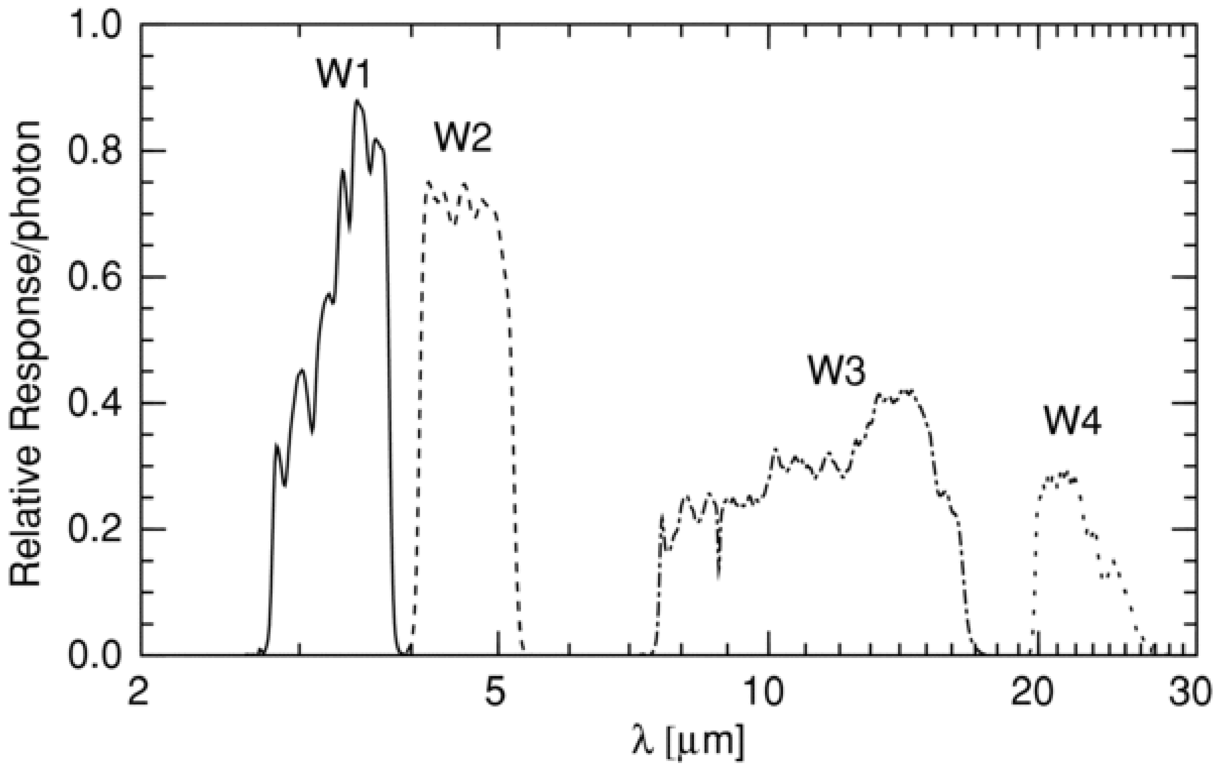

Figure 2.7: Relative response curves for the four *WISE* filters: W1 (3.4 μm), W2 (4.6 μm), W3 (12 μm), and W4 (22 μm). Effects of the telescope optics, detectors, and filter transmission are included. Figure 6 from (WRIGHT *et al.*, 2010).

WISE was launched in December 2009 with a 40 cm cryogenically cooled telescope and surveyed the whole sky in four mid-infrared bands centered at 3.4, 4.6, 12, and 22 μm—designated W1 through W4—with angular resolutions of 6.1″, 6.4″, 6.5″, and 12.0″, respectively (WRIGHT *et al.*, 2010). Each sky region was observed multiple times, increasing the depth and reliability of the final photometric catalog. The *WISE* All-Sky Data Release (CUTRI *et al.*, 2012) includes over 563 million objects, enabling fundamental research in a startling variety of fields. Figure 2.7 displays the system transmission curves of the four *WISE* mid-infrared photometric bands.

Following the depletion of its cryogen, the mission was reactivated as *NEOWISE* Post-Cryo (MAINZER *et al.*, 2011), continuing observations in the W1 and W2 bands and enabling time-domain analyses of motion and variability. The intermediate AllWISE data release (CUTRI *et al.*, 2014) combined exposures from *WISE* and early *NEOWISE* epochs, improving source detection and photometric depth. AllWISE provides proper motion estimates, variability metrics, and a multiepoch photometry database, as well as substantial improvements in sensitivity that led to enhanced photometric and astrometric accuracy compared to earlier releases. The AllWISE Source Catalog (WRIGHT *et al.*, 2019) offers accurate positions, apparent motion measurements, four-band fluxes, and flux variability statistics for over 747 million objects detected on the deep, coadded Atlas Images.

Despite its advances, the AllWISE catalog remained limited in depth and performance in crowded regions. To address these issues, the unWISE Catalog (SCHLAFLY *et al.*, 2019) was developed by coadding all available W1 and W2 imaging, including data from the *NEOWISE-Reactivation* mission. This comprehensive coaddition enables the detection of sources approximately 0.7 magnitudes fainter than AllWISE and improves source modeling in crowded regions. While unWISE does not include magnitude estimates directly, it provides flux measurements which can be converted to magnitudes using standard relations. The catalog encompasses positions and fluxes for over 2.2 billion objects across the entire sky, substantially enhancing the census of faint and distant infrared sources.

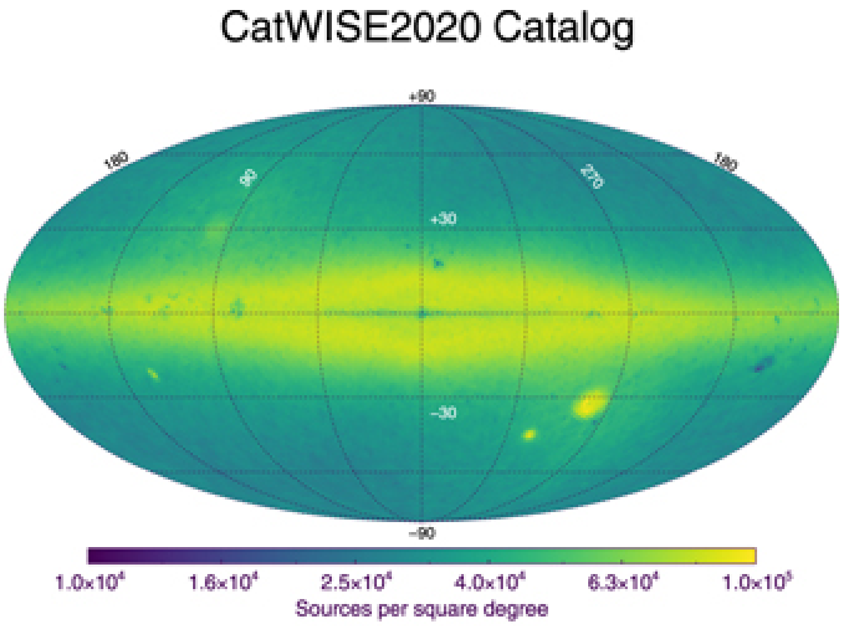

Figure 2.8: All-sky distribution of the number of sources per deg^2 in the CatWISE2020 Catalog, shown in Galactic coordinates. Adapted from Figure 3 of MAROCCO *et al.* (2021).

Whereas the unWISE Catalog significantly expanded source detection across the sky, it did not include motion measurements. The CatWISE project was designed to extract this temporal information by tracking positional changes of sources over time using W1 and W2 imaging. The CatWISE Preliminary Catalog (EISENHARDT *et al.*, 2020) provided proper motion estimates for nearly 900 million objects, achieving astrometric precision about ten times better than AllWISE. The final CatWISE2020 Catalog (MAROCCO *et al.*, 2021) incorporated additional epochs, improving completeness, particularly in dense fields, and reaching fainter limits. As shown in Figure 2.8, it achieves high source densities even in crowded regions like the Galactic plane, which reflects the improvements in completeness over earlier releases. The CatWISE2020 Catalog reports motions, positions, and fluxes for over 1.8 billion sources, establishing itself as a key resource for identifying faint and high-proper-motion objects in the infrared sky.

Gaia

It goes without saying that the *Gaia* mission (GAIA COLLABORATION *et al.*, 2016b) revolutionized modern astronomy. Launched by the European Space Agency (ESA) on December 19, 2013, and operational since 2014, the *Gaia* space telescope is widely recognized for its astrometric precision, having provided unparalleled measurements of stellar positions, parallaxes, and proper motions. Yet beyond its astrometry, *Gaia* has also delivered broadband photometric data of remarkable quality down to $G \approx 21$ mag, which has become increasingly valuable for stellar characterization. *Gaia*'s wealth of data has been explored to unveil the Galaxy's structure in unprecedented detail and shedding light on its complex formation history. The number of sources is expected to reach approximately 2.7 billion in the fourth data release[5].

The first *Gaia* data release (DR1; GAIA COLLABORATION *et al.*, 2016a), based on 14 months of observations, provided astrometry and G-band photometry for over 1 billion sources, including positions, parallaxes, and proper motions for 2 million bright stars through the Tycho-Gaia Astrometric Solution (TGAS). The second data release (DR2; GAIA COLLABORATION *et al.*, 2018a), based on 22 months of data, represented a significant leap in both precision and scope, delivering five-parameter astrometry for over 1.3 billion sources solely from *Gaia* measurements. Gaia DR2 introduced $G_{\rm BP}$ and $G_{\rm RP}$ photometry and was the first release to include radial velocities from the Radial

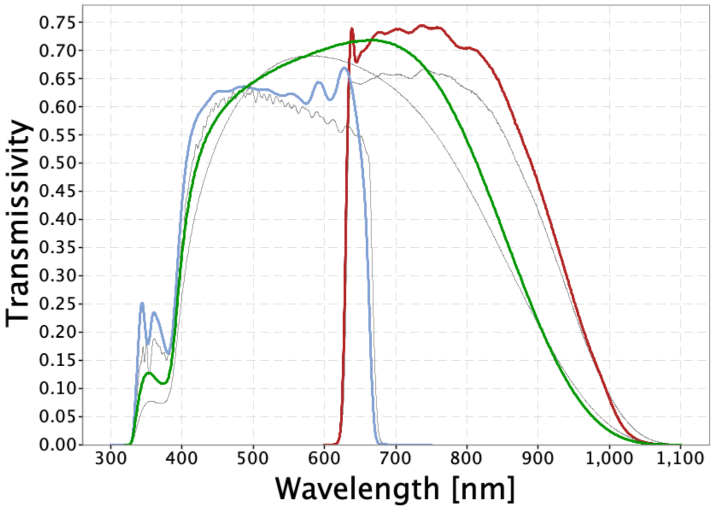

Figure 2.9: Gaia (E)DR3 passbands as produced by Coordination Unit 5 (CU5) of the Gaia Data Processing and Analysis Consortium (DPAC). The colored lines in the figure show the Gaia EDR3 photometric system defined by the G (green), $G_{\rm BP}$ (blue), and $G_{\rm RP}$ (red) passbands. The thin, grey lines show the nominal, pre-launch passbands published in JORDI *et al.* (2010), used for *Gaia* DR1. Figure available in the Gaia DR3 Photometric System page. Figure credit: ESA/Gaia/DPAC, P. Montegriffo, F. De Angeli, M. Bellazzini, E. Pancino, C. Cacciari, D. W. Evans, and CU5/PhotPipe team.

---

[5] `https://www.cosmos.esa.int/web/gaia/dr4`

Velocity Spectrometer (RVS; CROPPER *et al.*, 2018), providing median radial velocities for 7 million stars and estimates of astrophysical parameters for up to 161 million sources. Additionally, it explored a large number of extragalactic objects to present the celestial reference frame of Gaia DR2 (Gaia-CRF2), the first optical realization of a reference frame at sub-milliarcsecond (mas) precision (GAIA COLLABORATION *et al.*, 2018b). The transmission curves for the Gaia early third data release (EDR3, introduced in the next paragraph) are exhibited in Figure 2.9.

Gaia's Early Data Release 3 (EDR3; GAIA COLLABORATION *et al.*, 2021), published in 2020, provided significantly improved astrometry and broad-band photometry for over 1.8 billion sources, based on 34 months of observations. EDR3 offered enhanced precision and accuracy in positions, parallaxes, proper motions, and colors ($G_{\rm BP}$, $G_{\rm RP}$) for approximately 1.5 billion sources, and retained cleaned radial velocity data from DR2. The sky distribution of these sources is illustrated in Figure 2.10, which highlights the high density and near-all-sky coverage of Gaia observations. Its source list was reused in the full *Gaia* Data Release 3 (DR3; GAIA COLLABORATION *et al.*, 2023), published in 2022, which built upon EDR3's astrometry and photometry. DR3 delivered radial velocities for nearly 34 million stars, low-resolution BP/RP and RVS spectra for subsets of sources, and astrophysical parameters for nearly half a billion objects. Moreover, owing to substantial differences in the source list evolution, data processing, and calibration between releases, it is advised that data releases should be treated as independent, particularly regarding source identifiers (GAIA COLLABORATION *et al.*, 2018a, 2021, 2023).

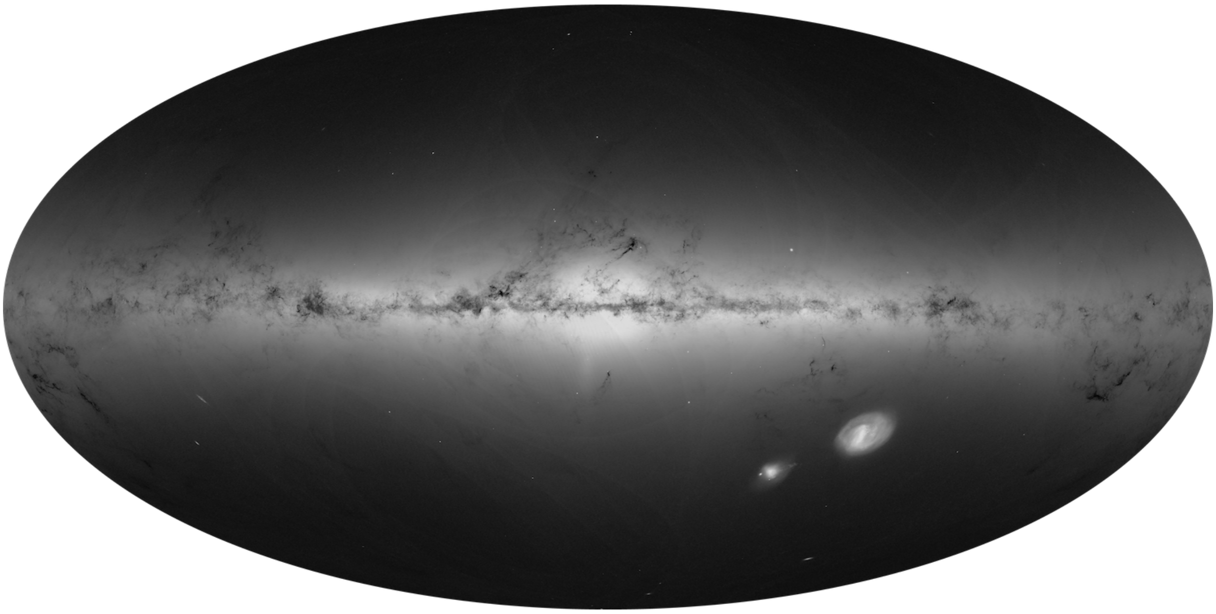

Figure 2.10: Sky map of source density of Gaia's Early Data Release 3, equally applicable to Gaia DR3. Figure available in the "Where are the stars?" option, under the Gaia DR3 stories menu within the Gaia DR3 overview web page. Figure credit: ESA/Gaia/DPAC, CC BY-SA 3.0 IGO.

As addressed later on, this work makes use of the third data release, with main interest in information on distances. In particular, we employ distances derived by the work of BAILER-JONES *et al.* (2021), who used a probabilistic approach to infer geometric and photogeometric distances for over 1.4 billion sources in *Gaia* EDR3. These distances offer improved reliability, especially for faint or distant stars with low parallax SNR. It is important to note that this catalog is not an official Gaia Collaboration product, but it has been widely adopted by the community due to its methodological rigor and extensive validation (e.g. EL-BADRY *et al.*, 2023; KORDOPATIS *et al.*, 2023; MYEONG *et al.*, 2022).

2.2 Spectroscopic surveys

As previously discussed in Section 1.1, spectroscopy marked a major paradigm shift in astronomy, incidentally giving rise to the field of astrophysics. After a quick historical background, this Section describes some modern spectroscopic surveys, particularly further highlighting LAMOST in Section 2.2.1, the most relevant one to this work.

Historical background

The systematic cataloging of stellar spectral information represents one of the foundational transitions in modern astronomy. Before the late 19th century, astronomical classification was largely restricted to positions and visual brightness, with little insight into the physical nature of stars. This began to change through the pioneering work of Henry Draper, a physician and amateur astronomer who, in the 1870s, was among the first to successfully obtain photographic spectra of stars. In 1872, he captured the first photographic spectrum of a star (Vega) using a prism attached to a telescope and a photographic plate, revealing clear absorption lines and demonstrating the potential of spectroscopy to reveal stellar compositions and properties (WEBB, 1999).

Following Draper's death in 1882, his widow, Mary Anna Draper, funded a major initiative at the Harvard College Observatory to advance his work. Under the direction of Edward C. Pickering, this effort led to the development of the *Henry Draper Catalogue* in the 1880s, which systematically classified the spectra of over 225 000 stars (CANNON & PICKERING, 1918; WEBB, 1999). The resulting classification scheme laid a foundational framework for modern stellar astrophysics.

Central to this achievement was a group of women known as the *Harvard Computers*, including Williamina Fleming, who developed an early classification scheme (FLEMING, 1890); Antonia Maury, who introduced a more physically informed system based on line widths (MAURY, 1897); and Annie Jump Cannon, whose refinements led to the OBAFGKM sequence and were applied to the full Draper Catalogue (CANNON, 1901).

Despite their pivotal contributions, they received little recognition during their lifetimes, with much of their work published under Pickering's name. Their roles remained largely overlooked for decades, reflecting the broader marginalization of women's intellectual labor in science (OGILVIE, 2000). Only in recent decades has the scientific community begun to acknowledge the foundational nature of their work, thanks in part to initiatives like Project PHaEDRA[6] and renewed historical interest (SOBEL, 2016; SPARKE & GALLAGHER, 2007), helping to establish their rightful place in the history of astronomy.

In parallel, European efforts such as those at the Astrophysical Observatory of Potsdam (VOGEL, 1900) contributed to the early photographic classification of stellar spectra, emphasizing the growing international momentum behind spectroscopic methods.

Collectively, these early surveys, though limited in spectral resolution, laid the observational and methodological groundwork that would sustain future stellar spectroscopy. More closely related to the context of this work, those initiatives provided the first large-scale empirical basis for studying stellar temperatures, compositions, and evolutionary sequences.

Modern spectroscopic surveys

For much of the 20th century, spectroscopic efforts remained largely confined to small-scale programs using photographic plates, whose low sensitivity and manual processing limited both scope and efficiency. Although stellar classification systems and radial velocity techniques did advance significantly during this period, the overall pace of progress was incremental, building upon frameworks established by the early pioneers.

As discussed earlier in the context of photometry, the introduction of CCDs marked a pivotal shift in astronomical data collection, and their adoption had a profound effect on spectroscopy as well. Replacing low-sensitivity photographic methods, CCDs brought dramatic improvements in quantum efficiency, linearity, and digital readout, allowing for precise and efficient acquisition of spectral data. These advances turned spectroscopy from a slow, manual process into a scalable and automatable technique, paving the way for survey-scale operations in which hundreds or thousands of spectra could be captured, calibrated, and analyzed systematically. This transition opened new possibilities across a wide range of research areas, from the detailed study of stellar atmospheres and chemical abundances to broader spectroscopic investigations of the Milky Way's dynamics, extragalactic systems, and the large-scale structure of the universe.

The digital revolution in spectroscopy paved the way for a new generation of large-scale surveys. High-resolution efforts expanded the scope of stellar parameter determination and chemical tagging. In parallel, other complementary approaches provided crucial low- to medium-resolution spectroscopic coverage across vast regions of the sky. Building

---

[6] https://platestacks.cfa.harvard.edu/project-phaedra

on these foundations, the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) emerged as a key player, combining wide field-of-view and high multiplexing capabilities to deliver what is now the largest publicly available stellar spectroscopic dataset.

2.2.1 LAMOST

The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST), also known as the Guoshoujing Telescope, is a 4-meter reflecting Schmidt telescope located at Xinglong Station, China. It features a wide field of view of 20 deg^2 and is equipped with 4000 fibers distributed across 16 spectrographs (250 fibers each), enabling simultaneous spectroscopic observations of up to 4 000 targets per exposure. The LAMOST survey is estimated to cover up to approximately 15 000 deg^2 of the sky.

LAMOST operates in two main spectral resolution modes. The low-resolution mode (LRS) covers the wavelength range from 3700 Å to 9000 Å with a resolving power of $R \approx 1800$ at 5500 Å, and a limiting magnitude of $r \approx 19$ mag (CUI *et al.*, 2012). The medium-resolution mode (MRS, $R \approx 7500$; LIU *et al.*, 2020) targets narrower wavelength intervals (4950–5350 Å and 6300–6800 Å) with a limiting magnitude of $G \approx 15$ mag.

The LAMOST Stellar Parameter Pipeline (LASP) estimates fundamental stellar parameters in LRS, including $T_{\rm eff}$, log g, and [Fe/H], by employing template matching and machine learning techniques (WU *et al.*, 2014). For MRS data, these parameters are provided both by LASP and by a convolutional neural network (CNN)-based pipeline trained to extract $T_{\rm eff}$, log g, and [Fe/H] directly from normalized spectra.

At the beggining of this study, Data Release 8 (DR8) was the latest available data release, providing the most comprehensive and up-to-date dataset for our analysis. DR8 offers over 10 million spectra, including $\sim$ 224 000 galaxy spectra and $\sim$ 72 000 quasar spectra. The stellar parameter catalog includes $\sim 10^6$ AFGK-type stars (from LASP catalogs) and over 600 000 M-type stars (DING *et al.*, 2022). The sky coverage for DR8 is shown in Figure 2.11, where colored circles represent the annual progression of observed fields throughout the survey, and the green dotted lines indicate galactic coordinates. The image is from the LAMOST DR8 LRS observing plan[7].

While DR8 remains a robust resource within the astronomical community, subsequent data releases have expanded upon this dataset. Public releases DR9 and DR10 have increased the number of stellar spectra and AFGK-type parameter determinations, while internal versions DR11, DR12, and DR13 reflect the ongoing growth and refinement of the survey. We highlight a few key releases in Table 2.3, where general information and summarizing numbers are presented.

In this work, we focus on the low-resolution data, rather than mid-resolution, due to

---

[7] https://www.lamost.org/dr8/v2.0/plan

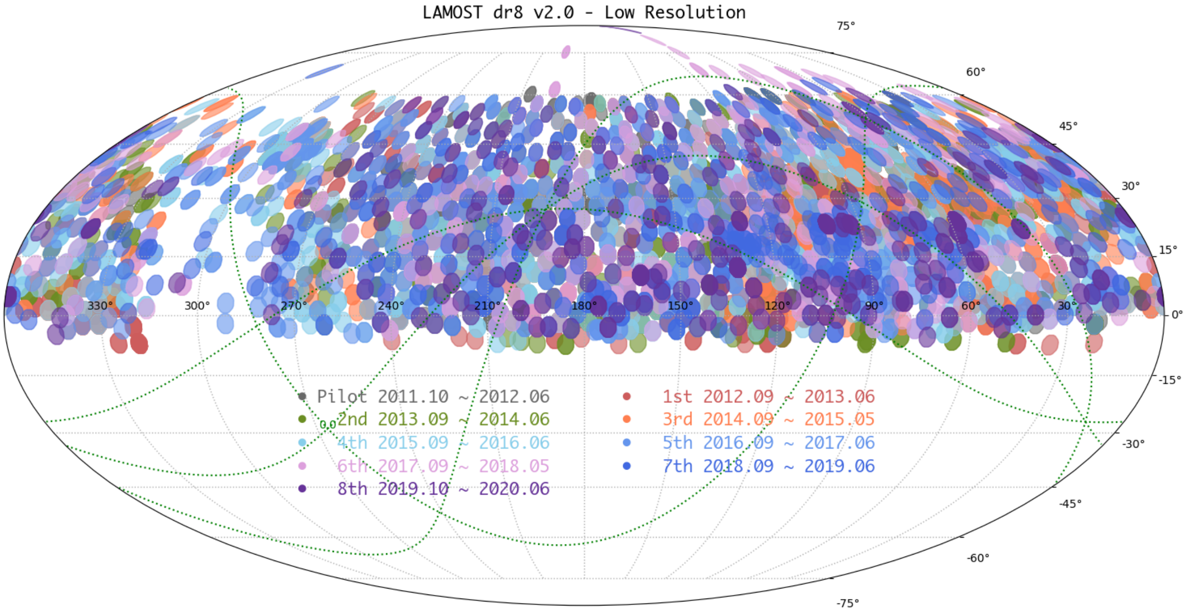

Figure 2.11: Sky distribution of LAMOST LRS in equatorial coordinates highlighting the sky coverage and the seasonal progression of the survey. Green dotted lines indicate galactic coordinates. Image from LAMOST DR8 LRS observing plan.

i) its significantly larger sample size ($\sim$ 6.6 *vs.* 1.2 million stars with available stellar parameters), which is better suited for training data-driven models (e.g. SETO *et al.*, 2022), and **ii**) its broader spectral coverage, which retains more astrophysical features (3700–9000 Å *vs.* 4950–5350 Å and 6300–6800 Å).

Table 2.3: Summary of recent LAMOST data releases. DR8 was the latest available public version at the start of this work.

Release	Version	Comment[a]	Access	Last Update	Total Spectra	Stars	AFGK Stellar Parameters
DR8	v2.0	Used in this work	Public	2022-09-30	10 633 515	10 336 752	6 684 413
DR10	v2.0	Most recent publicly available	Public	2024-09-29	11 441 011	11 100 139	7 450 303
DR12	v1.0	Most recent update	Internal	2025-03-26	12 602 390	12 231 890	8 370 041
DR13	v0	Last data release	Internal	2025-01-25	336 482	329 421	247 081

Note. (a) This column explains why each release is included for contextual purposes; while not directly used in this work, they help outline the current state of LAMOST. DR9, DR11, and releases prior to DR8 were excluded as they do not add relevant features beyond those already represented.

The parameter space covered by LAMOST is extensive, with $T_{\rm eff}$ ranging from approximately 2900 to 35 000 K, log g spanning from -0.24 to 5.9 dex, and [Fe/H] extending from -2.5 to $+1.0$ dex. However, it is important to note that the precision of these parameters varies, with typical uncertainties of 45 K for $T_{\rm eff}$, 0.25 dex for log g, and 0.22

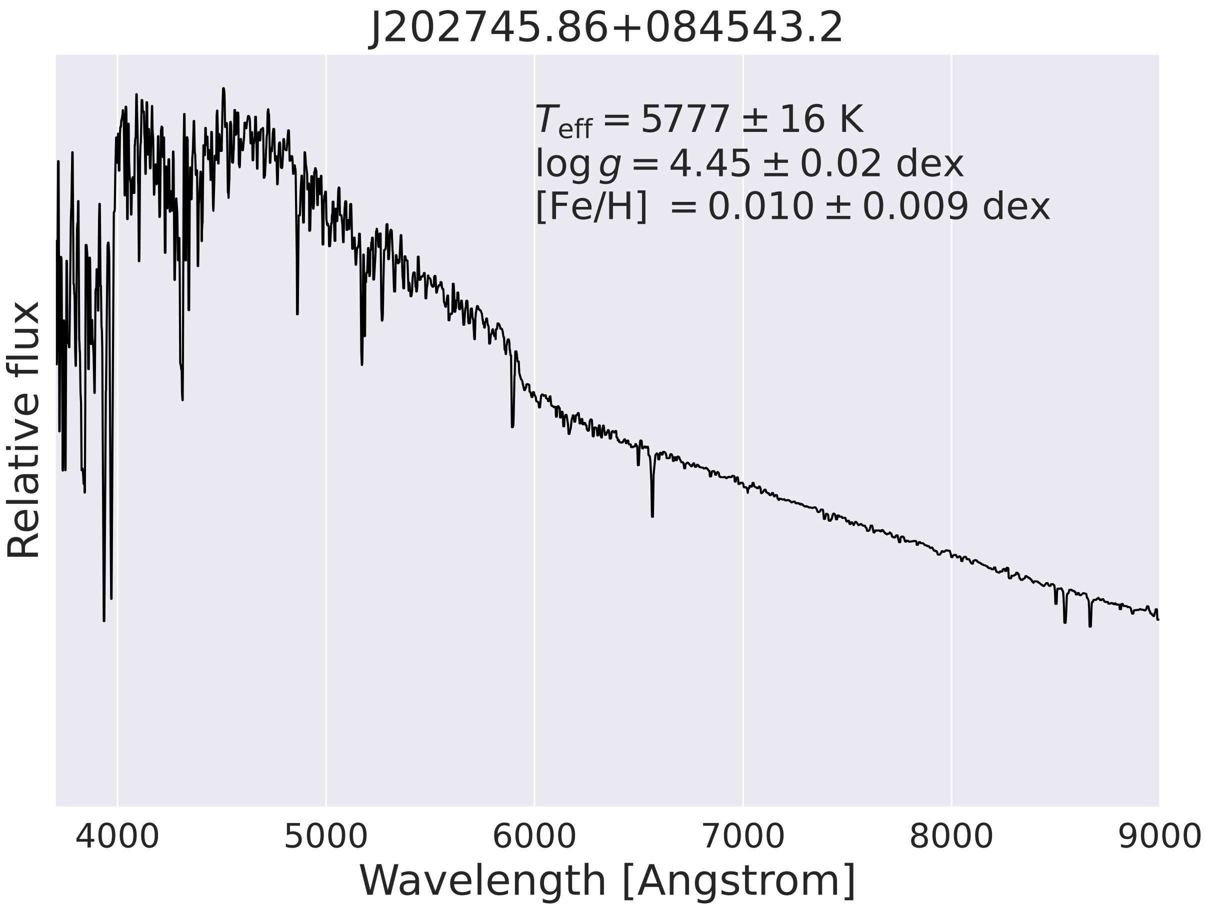

Figure 2.12: Example of LAMOST spectrum of a solar-like star with indicated atmospheric parameters and the star designation in the title.

dex for [Fe/H], particularly for M-type stars (LI *et al.*, 2022). Additionally, the accuracy of metallicity measurements decreases at the lower end of the [Fe/H] range, and users should interpret results for extremely metal-poor stars with caution.

Despite its extensive contributions to stellar spectroscopy, LAMOST data are subject to certain limitations. For instance, the SNR varies across observations, and lower SNR can impact the reliability of derived parameters. Moreover, while the LASP pipeline provides automated parameter estimations, manual verification and cross-matching with other surveys are recommended to ensure accuracy, especially for peculiar or rare stellar types. For illustration, an example of LAMOST spectrum for a solar-like star is shown in Figure 2.12.

In summary, LAMOST has significantly advanced the field of stellar spectroscopy by providing an unparalleled volume of spectroscopic data. Its ongoing data releases continue to support a wide array of astrophysical research endeavors.

2.2.2 Other large spectroscopic surveys

While LAMOST stands out for its immense volume and broad sky coverage, other large-scale spectroscopic surveys provide significant contributions through complementary approaches. Some have prioritized higher spectral resolution to enable more detailed stellar analyses, while others operate at resolutions comparable to LAMOST but with distinct observational strategies, as further discussed ahead. Together, these efforts have diversified the landscape of stellar spectroscopy and enriched our understanding of the Galaxy from multiple perspectives.

APOGEE

Among the major spectroscopic surveys, the Apache Point Observatory Galactic Evolution Experiment (APOGEE; MAJEWSKI *et al.*, 2017), part of the third phase of SDSS (SDSS-III), distinguishes itself by operating in the near-infrared H-band, enabling observations through dust-obscured regions of the Milky Way. Using high-resolution ($R \sim 22\,500$) cryogenic spectrographs, it collects data from both hemispheres via facilities at Apache Point Observatory (USA) and Las Campanas Observatory (Chile). Its latest data release, DR17, contains over 657 000 unique stellar spectra with typical SNR around 100, covering 2261 fields and targeting the bulge, disk, halo, star clusters, dwarf galaxies, and time-domain mission footprints. The spectral coverage spans three near-infrared windows—1514–1581 nm, 1585–1644 nm, and 1647–1696 nm—each recorded on separate detectors. The spatial distribution of APOGEE targets is shown in Figure 2.13, overplotted over a 2MASS (SKRUTSKIE *et al.*, 2006) all-sky image.

Stellar atmospheric parameters and detailed chemical abundances in APOGEE are derived using the APOGEE Stellar Parameters and Chemical Abundances Pipeline (ASPCAP; HOLTZMAN *et al.*, 2018; JÖNSSON *et al.*, 2020; PÉREZ *et al.*, 2016), which performs full-spectrum fitting based on a library of synthetic models generated with the `FERRE` code (PRIETO *et al.*, 2006). In addition to global parameter estimation, the pipeline also performs targeted fitting of individual lines to determine abundances for up to 18 elements. The number of elements measured depends on spectral quality, but the survey's consistent analysis and broad coverage make it one of the most chemically informative resources available for Galactic archaeology.

GALAH

In addition to APOGEE's high-resolution infrared spectroscopy, the GALactic Archaeology with HERMES (GALAH; BUDER *et al.*, 2021) survey provides a major optical counterpart aimed at unraveling the Milky Way's formation history through chemical tagging. GALAH focuses on obtaining high-resolution ($R \sim 28\,000$) optical spectra using the HERMES spectrograph on the Anglo-Australian Telescope, thereby covering the southern

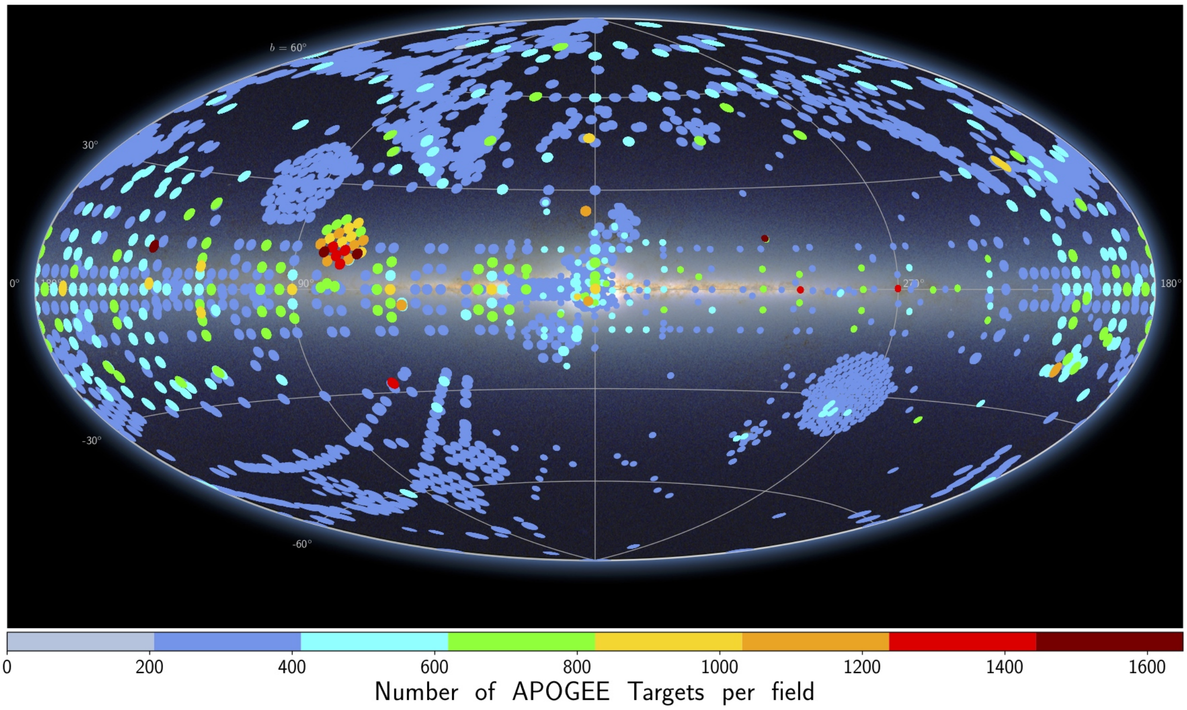

Figure 2.13: Sky distribution of APOGEE DR17 fields in Galactic coordinates, shown over a 2MASS all-sky image. Colors indicate the number of observed targets per field. Figure from APOGEE DR17 synopsis. Image credit: C. Hayes.

sky, and targeting primarily nearby stars across a broad range of Galactic latitudes. The survey was conducted in two main phases: Phase 1 prioritized high-latitude fields to avoid the Galactic plane, whereas Phase 2 extended the coverage to include lower latitudes. By analyzing the chemical abundances of up to 30 elements and combining them with precise kinematic data from *Gaia*, GALAH aims to chemically break down the Galaxy into its constituent stellar populations.

The current data release, GALAH DR4 (BUDER *et al.*, 2025), offers over 600 000 high-quality spectra with robust stellar parameters and elemental abundances. This release includes an updated analysis pipeline and improved validation procedures, notably enhancing the reliability of its stellar labels in comparison with earlier releases. The survey's extensive coverage is shown by the spatial distribution of GALAH targets in Figure 2.14, further highlighting synergies with catalogs K2 (HOWELL *et al.*, 2014) and TESS (RICKER *et al.*, 2015).

Gaia(-ESO)

Beyond ground-based efforts like GALAH and APOGEE, the Gaia mission represents a paradigm shift in Galactic studies through its space-based, all-sky survey combining astrometry, photometry, and spectroscopy. While primarily known for its unparalleled astrometric precision, Gaia also delivers spectrophotometric data via its low-resolution

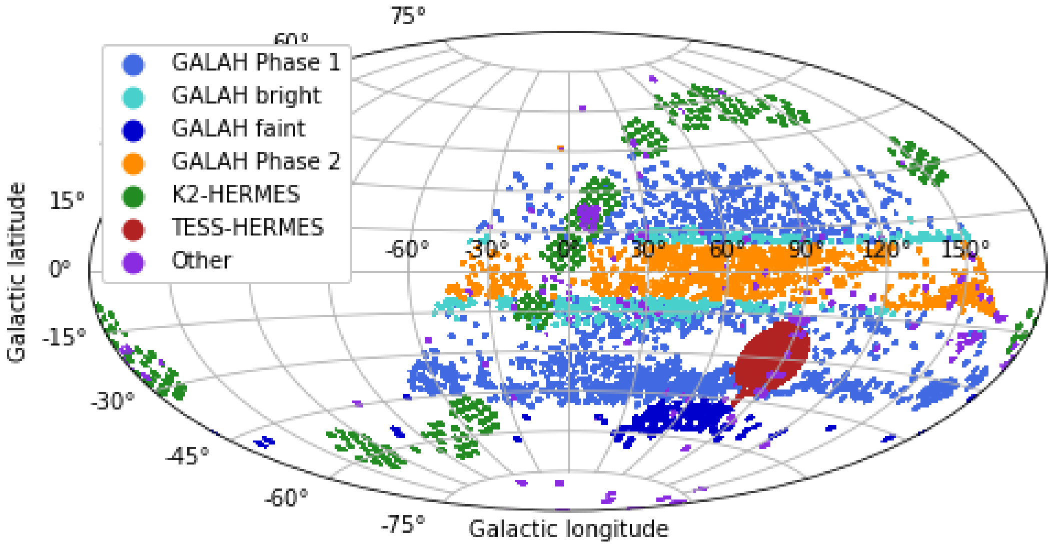

Figure 2.14: Sky distribution of GALAH DR4 targets in Galactic coordinates. Different colors indicate subsamples from GALAH Phases 1 and 2, as well as follow-up programs including K2-HERMES, TESS-HERMES, and other GALAH-related projects. Figure from GALAH DR4 overview.

Blue and Red Photometers (BP and RP), collectively referred to as XP spectra, and radial velocity measurements through its onboard Radial Velocity Spectrometer (RVS). These capabilities provide an essential link between stellar kinematics, chemical composition, and evolutionary status across the full sky. The BP and RP instruments produce low-resolution prism spectra that encode broad spectral energy distributions for over 200 million stars. Despite their modest resolution, XP spectra have proven valuable for estimating astrophysical parameters and for training data-driven models that connect photometric and spectroscopic domains (ANDRAE *et al.*, 2023; MONTEGRIFFO *et al.*, 2023; ZHANG *et al.*, 2023).

In parallel with the data collected by Gaia itself, the Gaia-ESO Survey (GILMORE *et al.*, 2012; RANDICH *et al.*, 2013) complements the mission by providing high-resolution ground-based spectroscopy for a representative sample of Milky Way stellar populations. Conducted with the FLAMES multi-object spectrograph at the Very Large Telescope (VLT), Chile, Gaia-ESO targeted over 100 000 stars across all major Galactic components, including open clusters, the bulge, halo, and field stars in the thin and thick disks (GILMORE *et al.*, 2022; RANDICH *et al.*, 2022). The survey utilized both the GIRAFFE and UVES spectrographs, achieving resolving powers of approximately 20 000 and 47 000, respectively. By delivering precise radial velocities and detailed chemical abundances, the survey was designed to calibrate and augment Gaia's astrometric and photometric data. Its homogeneous analysis and dedicated calibration strategy also make it a cornerstone reference for validating stellar parameters across other surveys. The final data release,

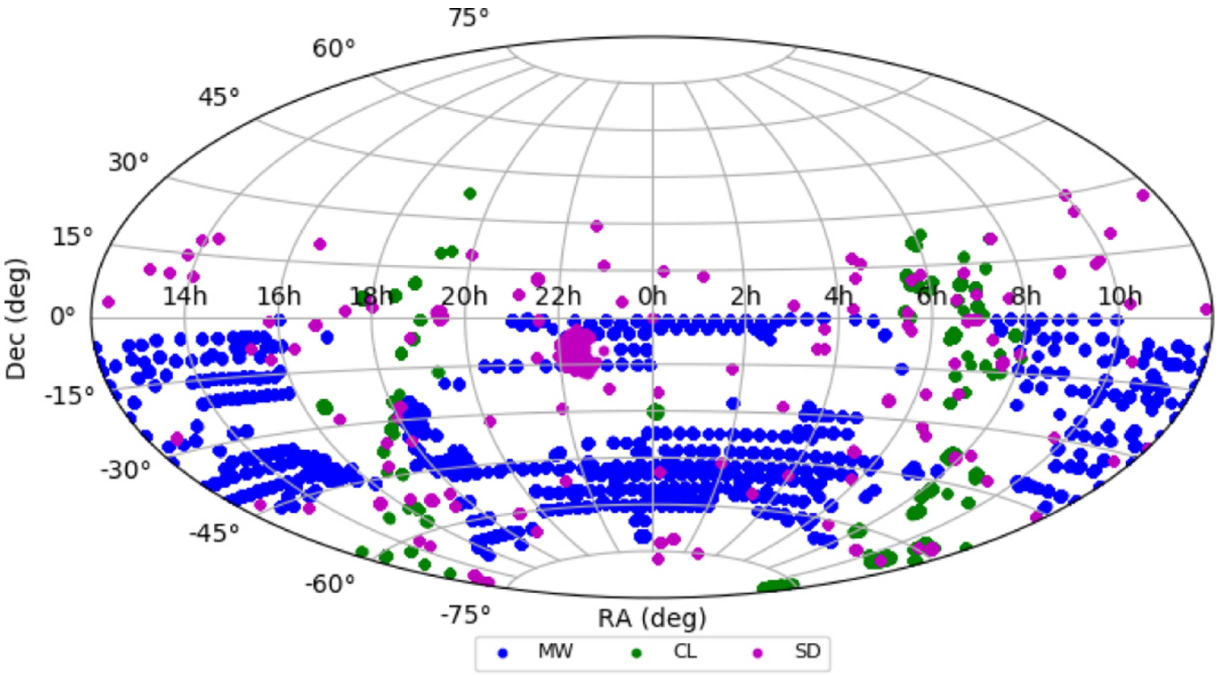

Figure 2.15: Sky distribution of Gaia-ESO Survey observed fields broken down by field-type: Milky Way fields (MW), clusters (CL), standards and calibration fields (SD). Figure 6 from RANDICH *et al.* (2022).

DR5.1, encompasses data for over 10^5 stars, providing an unprecedented, homogeneous overview of kinematic distributions and elemental abundances, significantly enhancing our understanding of Galactic and stellar evolution (HOURIHANE *et al.*, 2023). The spatial distribution of observed targets in Gaia-ESO Survey is illustrated in Figure 2.15.

SDSS/SEGUE

In the low- to mid-resolution regime, the Sloan Extension for Galactic Understanding and Exploration (SEGUE; NEWBERG *et al.*, 2002; YANNY *et al.*, 2009) represented a major advance in the scope and accessibility of stellar spectroscopic data. SEGUE-1, the initial phase of the program, collected spectra for nearly 230 000 unique stars spanning a wide range of spectral types to probe the structure of the Milky Way. Building on this foundation, SEGUE-2 (ROCKOSI *et al.*, 2022) extended the survey by targeting an additional $\sim 119\,000$ stars, with a specific focus on the in situ stellar halo at Galactocentric distances between 10 and 60 kpc. Both surveys delivered spectra with a resolution of $R \sim 2000$ over the 3800–9200 Å range, providing atmospheric parameter estimation via the SEGUE Stellar Parameter Pipeline (SSPP; LEE *et al.*, 2008a,b).

Combined, SEGUE-1 and SEGUE-2 provided a spectroscopic census of approximately 347 000 unique stars, publicly released in SDSS-III DR 8 (AIHARA *et al.*, 2011). DR9 (AHN *et al.*, 2012) introduced an updated version of the SEGUE Stellar Parameter Pipeline, along with new value-added catalogs. Although no further updates to SEGUE

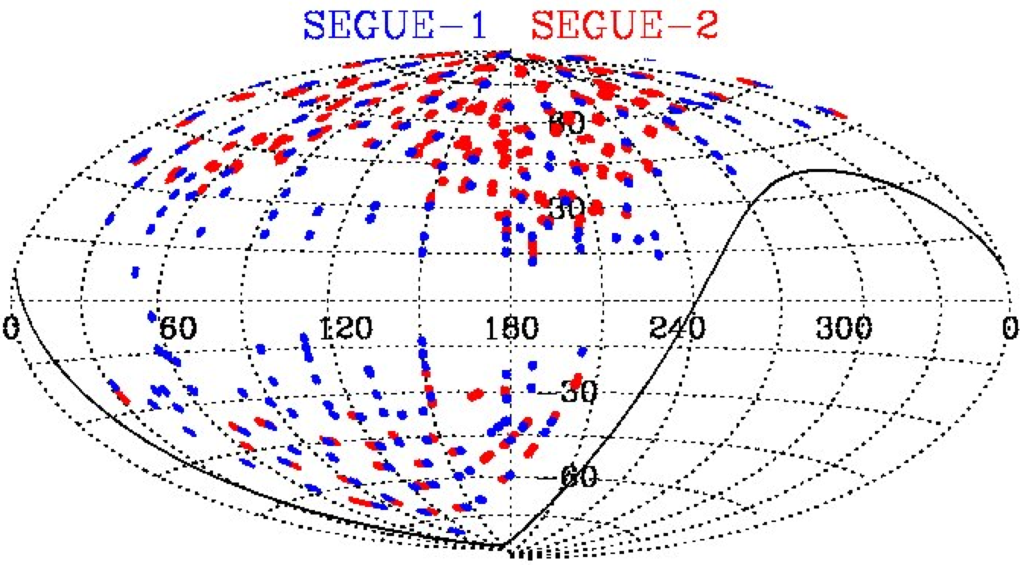

Figure 2.16: SEGUE-1 and SEGUE-2 fields displayed in blue and red, respectively. The map is in Galactic coordinates. Figure from SDSS III/SEGUE-2 overview. Credit: M. Strauss.

data or processing were made in subsequent releases, the cumulative nature of SDSS data releases ensures that the refined SEGUE dataset from DR9 remains available in later releases. The most recent one, DR18 (ALMEIDA *et al.*, 2023), contains targeting catalogs for SDSS-V, supporting the Black Hole Mapper, Milky Way Mapper, and open fiber science programs. The sky coverage of both SEGUE-1 and SEGUE-2 is illustrated in Figure 2.16, with fields shown in blue and red, respectively.

Chapter 3

Methodology

This chapter details the machine learning approach used in this work. First, essential background on ensemble methods is provided in Section 3.1, focusing on gradient boosting and the LightGBM framework. Then, the astrophysical context and motivation for applying machine learning to photometric data are established in Section 3.2, supported by key literature examples.

3.1 Machine learning

Machine learning (ML) is a central pillar of modern artificial intelligence (AI), which is an umbrella term for a broad scientific effort of developing computational systems capable of solving problems by emulating aspects of human cognition, such as learning, perception, and reasoning. These systems may operate autonomously or semi-autonomously, adapting their behavior based on the information they receive.

While AI comprises diverse approaches, ML has become its most prominent and dynamic subfield due to its success in handling large-scale, high-dimensional datasets. It relies on data-driven statistical models that improve performance through experience without being explicitly programmed for each specific task, encompassing a range of algorithms designed to recognize patterns, learn from data, and make predictions or decisions with minimal human intervention.

The conceptual foundations of ML, although often regarded as a cornerstone of modern data science, trace back much further than one might expect. The very core of ML is intricately connected to statistics. As early as the 18th century, the formulation of Bayes' theorem[1] (BAYES, 1763; LAPLACE, 1812) introduced principles of probabilistic reasoning that lie at the heart of many of today's most sophisticated learning algorithms. Alongside the Bayesian tradition, frequentist methods such as maximum likelihood estimation (FISHER, 1922) also played a central role in shaping early techniques. The

---

[1] Although commonly attributed to Thomas Bayes, the work was posthumously edited and expanded by Richard Price, who provided substantial contributions (MCGRAYNE, 2011).

evolution from classical statistical theory to modern ML reflects not only a broadening of scope, from foundational inference principles to flexible frameworks for analyzing complex, high-dimensional data, but also the convergence of advances in optimization, computational power, and data availability.

While these statistical principles formed the conceptual basis, practical ML development remained limited until the mid-20th century, when emerging computing technology and theoretical breakthroughs sparked renewed interest. This period gave rise to foundational advances, including the invention of the first neural model (MCCULLOCH & PITTS, 1943), Turing's proposal of a "learning machine" capable of artificial intelligence (TURING, 1950), the construction of the first neural network machine (SNARC; MINSKY, 1954; MINSKY & EDMONDS, 1951, see RUSSELL & NORVIG 2016), Rosenblatt's development of the Perceptron, the first algorithmically trainable neural network (ROSENBLATT, 1957), and the creation of the first machine learning programs, along with the introduction of the term "machine learning" by SAMUEL (1959).

Although these breakthroughs demonstrated promise, progress stalled during the so-called "AI winter", a period of reduced interest and funding in AI research during the 1970s and early 1980s (CREVIER, 1993, cf HAIGH 2023; TOOSI *et al.* 2021). In essence, extending networks capabilities from one simple layer to many (multilayer perceptrons, or MLP) was interpreted as a sterile idea to be put in practice since a while before (MINSKY & PAPERT, 1969). The backpropagation algorithm (RUMELHART *et al.*, 1986a,b) provided an efficient method for training multi-layer networks, overturning this earlier skepticism and eventually revitalizing neural network research. This process iteratively refines predictions by applying the chain rule to compute error gradients, adjusting the initially random weights connecting the input and hidden layers. Repetition gradually improves accuracy. In later editions, MINSKY & PAPERT (1988) acknowledged this breakthrough as key to training MLPs effectively.

The success of backpropagation in revitalizing neural networks helped establish methodological grounding and a sense of stability in the field, contributing to a more balanced and sustained development of AI. As the discipline matured, these early concepts evolved into a wide range of learning strategies. In contemporary terms, ML methods can be broadly classified into three major paradigms as addressed below, depending on the nature and availability of the information provided during training.

- **Supervised learning:** algorithms are trained on labeled data, learning to map input features to known target outputs; this is the most widely adopted paradigm in astronomical applications, including the estimation of stellar parameters from photometry or spectra.

- **Unsupervised learning:** no labeled outputs are provided; the algorithm seeks to uncover hidden patterns or structures within the data itself, often through clustering or dimensionality reduction.

- **Reinforcement learning:** a learning agent interacts with an environment, mapping its problem space, and learns a strategy through trial and error guided by a reward signal.

Within these learning paradigms, methodologies range from traditional ML to modern deep learning approaches.

Traditional machine learning typically operate on structured feature representations, allowing direct integration of domain knowledge through careful feature selection and engineering. While less suited to raw unstructured data, they provide native interpretability (e.g., feature importance scores, decision rules), computational efficiency (CPU-based training), and strong performance on tabular data. Well-established methods include linear regression, logistic regression, support vector machines (SVMs), k-nearest neighbors (kNNs), decision trees, and principal component analysis (PCA), as well as ensemble approaches such as random forests and gradient boosted trees.

Deep learning (DL), in contrast to traditional machine learning, is particularly well suited to problems involving high-dimensional, unstructured data. It relies on artificial neural networks (ANNs) composed (after the backpropagation breakthrough) of multiple processing layers that learn hierarchical feature representations directly from raw inputs, bypassing the need for human-curated data treatment. This capability enables DL models to uncover complex and abstract patterns that are difficult to capture with traditional approaches. However, their ability to escape manual feature engineering comes at the cost of interpretability and substantial computational demands.

Model explainability is particularly valuable in astrophysics, where linking predictions to physical principles is as critical as achieving high accuracy. For structured datasets (e.g., spectral catalogs, stellar parameters tables), traditional ML methods can match or exceed DL performance while offering greater transparency and computational efficiency (BORISOV *et al.*, 2024). In particular, the combination of predictive power, interpretability, and lower resource demands makes ensemble methods notably interesting in the context of the present work.

3.1.1 Ensemble methods

Among the benefits and costs of traditional ML and DL approaches, tree-based models and their ensemble extensions offer a balance between predictive power and relative transparency. Ensemble methods combine the predictions of multiple individual models to produce a more robust and accurate outcome than any single model alone. By aggregating over diverse learners through averaging, voting, or more complex strategies, they reduce variance, mitigate overfitting, and often improve generalization. Many ensemble techniques rely on decision trees as their base learners, making trees a natural starting point for understanding the structure and behavior of more advanced models.

Decision trees stand out for their intuitive structure and interpretability (BREIMAN *et al.*, 1984; MORGAN & SONQUIST, 1963). These models operate by recursively splitting the data based on feature values, forming a hierarchy of decisions that leads to a final prediction. Their logic mimics human reasoning in many ways, making them especially appealing for both exploratory analysis and model transparency (for a review, see LOH, 2014). A decision tree is exemplified by Figure 3.1, showing a simple scheme for classifying stellar temperatures into hot, warm, or cool categories through sequential threshold decisions on the B−V color index.

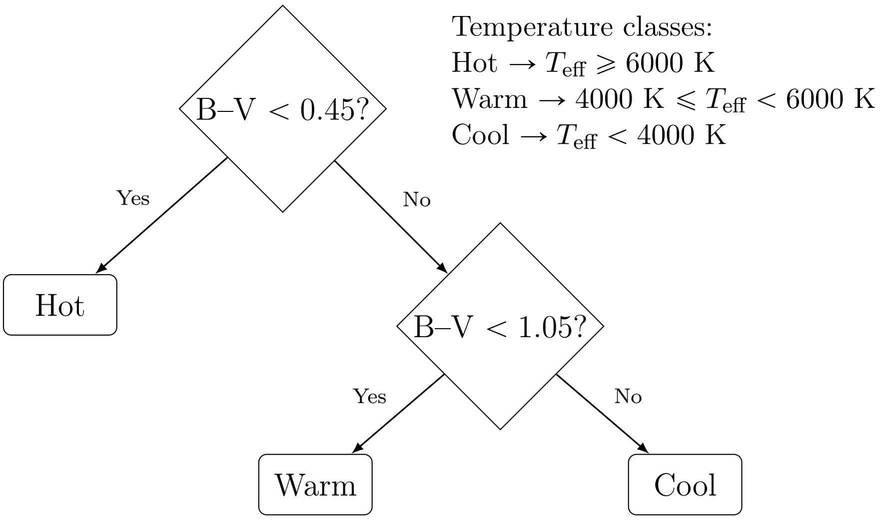

Figure 3.1: Simple visualization of a decision tree for classifying stellar temperatures based on the $(B - V)$ color index.

Despite these advantages, individual decision trees typically exhibit high variance and limited generalization capability. To address these limitations, ensemble methods (LEAMER, 1978) have emerged as a powerful extension that combines the predictions of multiple models (typically composed of many decision trees) to enhance robustness and predictive accuracy. The rationale is rooted in the principle that collective judgments tend to outperform individual ones (CONDORCET, 2014; DIETTERICH, 2000), an idea widely known as the "wisdom of the crowd" (e.g. LANDEMORE & ELSTER, 2012).

Ensemble implementation strategies are broadly categorized into stacking, bagging, and boosting (e.g. ZHOU, 2012). Stacking is an approach that combines predictions from diverse base models using a separate meta-model trained to optimize their integration (WOLPERT, 1992). It explores the complementary strengths of different algorithms to improve performance. By blending the outputs of varied models (e.g. decision trees, neural networks, support vector machines) through a meta-learner (often linear regression or logistic regression), stacking can capture complex patterns that might be missed by individual models. Bagging (bootstrap aggregating) improves model stability by training multiple versions of the same algorithm on different bootstrapped subsets of the data (BREIMAN, 1996). Their outputs are then aggregated, typically via voting (classification) or averaging (regression). A key implementation is the Random Forest (BREIMAN, 2001; HO, 1995), which expands upon the bootstrap aspects of bagging by generating a set of decision trees from these bootstrap samples, further adding random feature selection to reduce correlation between trees and improve generalization.

Boosting is a technique that constructs models sequentially. Each new learner in the sequence focuses specifically on correcting errors made by previous models through adaptive reweighting of training instances. This iterative error correction process progressively reduces bias, effectively transforming weak learners into a strong predictor (KEARNS, 1988; KEARNS & VALIANT, 1989). While decision trees are standard due to their adaptability, other weak learners such as splines or linear models (BÜHLMANN & HOTHORN, 2007) can also be implemented, albeit with trade-offs in flexibility and scalability. Unlike bagging methods that train models independently, boosting's sequential approach provides particular advantages. Algorithms like AdaBoost (FREUND & SCHAPIRE, 1997), XGBoost (CHEN & GUESTRIN, 2016), and LightGBM (KE *et al.*, 2017) implement this strategy, specializing in handling difficult-to-predict examples and achieving state-of-the-art performance on structured data for both regression and classification tasks. While boosting can theoretically be susceptible to overfitting with noisy data, such modern implementations have demonstrated remarkable robustness through techniques like early stopping, regularization, and sub-sampling. This resilience holds even in high-dimensional settings (GHOJOGH & CROWLEY, 2019), making boosting one of the most powerful tools for contemporary machine learning challenges.

A distinct method, which plays a central role in the present work, gradient boosting (FRIEDMAN, 2001) is further addressed ahead. Moreover, ensemble methods are extensively addressed in detail in several works (e.g. MIENYE & SUN, 2022; SAGI & ROKACH, 2018; ZHANG & MA, 2012; ZHOU, 2025).

3.1.2 Gradient boosting

Gradient boosting is a powerful ensemble technique that builds predictive models step by step. It belongs to the family of boosting methods, as it iteratively improves predictions by correcting the errors of previous estimations. At the same time, it extends the concept of gradient descent (CAUCHY, 1847) to function spaces (MASON *et al.*, 1999), treating the model itself as a function to be refined.

In this context, a function refers to a rule applied to an input feature (such as a star's color) to return a target output (such as an estimated temperature). While such a rule is often expressed as a mathematical formula, here it is implemented using a decision tree. Despite lacking an explicit equation, a decision tree defines a function by partitioning the input space into regions and assigning a constant prediction to each. For example, a tree might say: “if feature A > 1.2 and feature B < 0.7, predict $T_{\text{eff}} = 5800$ K”. This way, it maps each input to an output value, much like a standard equation-based function.

The concept of gradient descent, in turn, guides how the model is updated. It is an optimization algorithm used to minimize an analytical objective function, typically a loss function such as the mean absolute error (MAE), mean squared error (MSE), or root mean squared error (RMSE). The core idea is to reduce the loss by iteratively adjusting the model’s parameters. At each step, the algorithm computes the gradient of the loss with respect to the parameters, which indicates the direction of steepest increase. To minimize the loss, the parameters are updated in the opposite direction. These updates are applied in small increments, allowing the algorithm to gradually approach a local minimum of the loss function.

Conversely, in the functional framework of gradient boosting, models are represented by decision trees. Each new tree acts as a corrective function, pushing the overall model in the direction that most reduces the remaining error. By combining these weak learners additively, the ensemble gradually converges toward a more accurate approximation of the target function.

In other words, since the model itself is treated as a function $F(x)$, represented by a non-analytical decision tree, each step adds a new function $h_t(x)$ designed to reduce the remaining errors, or “pseudo-residuals.” That is, $h_t(x)$ is trained to approximate[2] the negative gradient of the loss function with respect to the current model $F_{t-1}(x)$. This leads to an updated model:

$$F_t(x) = F_{t-1}(x) + \eta\, h_t(x), \tag{3.1}$$

where η is a learning rate that controls the step size, and $h_t(x)$ represents the decision tree added at iteration t. Each $h_t(x)$ acts as a small adjustment that incrementally improves the model’s predictions.

In essence, at iteration t, for a set of features $\mathbf{x}$, with given a loss function $\mathcal{L}(y_i, F(\mathbf{x}_i))$, the algorithm takes the following steps.

1. Computes pseudo-residuals $r_t^i = -\left[\frac{\partial \mathcal{L}(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F=F_{t-1}}$;
2. fits a weak learner $h_t(\mathbf{x})$ to $\{r_t^i, \mathbf{x}_i\}_{i=1}^T$;
3. updates the model: $F_t(\mathbf{x}) = F_{t-1}(\mathbf{x}) + \eta \cdot h_t(\mathbf{x})$.

After T iterations, the final model is obtained by expressing the iterative updates as

---

[2] Here, “approximate” means that the decision tree attempts to mimic the direction and structure of the analytical gradient using discrete, piecewise-constant predictions over the input space.

a cumulative sum of these individual corrections:

$$F_T(x) = F_0(x) + \sum_{t=1}^{T} \eta \, h_t(x), \tag{3.2}$$

where $F_0(x)$ is the initial guess (often a constant such as the mean of the target values), and x denotes the input features (e.g. photometric colors).

While gradient boosting is a highly effective and flexible method, it can also suffer from poor generalization if model complexity is not properly controlled. In this context, underfitting occurs when the model is too simple to capture the underlying relationships in the data, leading to poor performance on both the training and validation sets. Conversely, overfitting arises when the model becomes too complex and adapts too closely to the training data, including its noise and fluctuations. This typically results in excellent performance on the training set but degraded accuracy on unseen data, as the model fails to generalize. In gradient boosting, such behavior is often caused by the use of overly deep trees, too many boosting iterations, or large update steps.

To mitigate these issues, several strategies are used to control model complexity and improve generalization. These include: limiting the impact of individual trees by scaling their contribution to the overall model (learning rate shrinkage), which reduces the risk of overfitting by encouraging gradual learning and avoiding large, overconfident updates; restricting how much detail each tree can capture by limiting its depth (maximum tree depth), which helps prevent overfitting by avoiding the memorization of overly specific patterns in the training data; discouraging overly complex or unbalanced tree structures by adding penalties to the size of the tree and the magnitude of its predictions (regularization), which further reduces the risk of overfitting by favoring simpler models; and training each tree on a randomly selected subset of the data (stochastic sub-sampling), a variance-reducing technique related to bagging that improves generalization by injecting randomness into the learning process.

In practice, these strategies are often tuned jointly: for example, using shallow trees and a small learning rate together typically requires more boosting rounds but yields more robust results. Typical configurations use tree depths between 3 and 6 and learning rates in the range of 0.01 to 0.1 (HASTIE *et al.*, 2009). Although these strategies are primarily intended to prevent overfitting, excessively conservative values may lead to underfitting, which can be addressed by increasing model capacity, loosening regularization, or raising the learning rate. When appropriately combined, these techniques enable gradient boosting models to achieve high accuracy while maintaining good generalization to unseen data.

Ultimately (and at the risk of what is admittedly some repetition, but this is intentional and meant to reinforce clarity), gradient boosting's effectiveness relies on three foundational principles, which are further detailed below.

- **Additive expansion, whereby each learner incrementally corrects the ensemble's residuals.**
  The ensemble $F_m(\mathbf{x})$ at iteration m is constructed by sequentially adding new weak learners $h_m(\mathbf{x})$ with learning rate η:

  $$F_t(\mathbf{x}) = F_{t-1}(\mathbf{x}) + \eta \cdot h_t(\mathbf{x})$$

  where $F_{t=0}(\mathbf{x})$ is typically initialized as a constant value minimizing the loss.

- **Gradient-based optimized updates, guiding learning toward loss minimization.**
  At each iteration, the weak learner h_t is fit to the negative gradient (pseudo-residuals) of the loss function $\mathcal{L}$ with respect to the current ensemble:

  $$r_{im} = -\left.\frac{\partial \mathcal{L}(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right|_{F=F_{t-1}}$$

- **Generalization control strategies, which temper overfitting through diverse techniques.**
  Implemented through:

  a) scaling the contribution of each tree to the overall model (learning rate shrinkage);
  b) restricting model complexity through structural constraints (e.g., maximum tree depth, regularization);
  c) and training each tree on a randomly selected subset of the data (stochastic sub-sampling).

The most successful implementations of gradient boosting algorithms include AdaBoost (FREUND & SCHAPIRE, 1997), XGboost (CHEN & GUESTRIN, 2016), and LightGBM (KE *et al.*, 2017), the latter being outlined ahead.

3.1.3 LightGBM

LightGBM (Light Gradient Boosting Machine) is an open-source framework (KE *et al.*, 2017) for the efficient implementation of gradient boosting without sacrificing predictive power. Its strong performance with large datasets and high-dimensional feature spaces has contributed to its widespread adoption in regression, classification, and ranking tasks across both scientific research and industry. While it is beyond the scope of this work to provide a detailed technical description, a brief conceptual overview is offered here, following the official LightGBM documentation[3].

---

[3] `https://lightgbm.readthedocs.io/en/stable/index.html`

LightGBM builds on the foundational principles of gradient boosting, including additive modeling, decision tree learners, and iterative optimization in function space. It also introduces algorithmic innovations that make it particularly effective for structured, high-dimensional data. In the context of this work, where thousands of stars are described by dozens of photometric attributes, LightGBM was selected for its combination of computational efficiency, interpretability, and scalability.

A key feature that distinguishes LightGBM from other boosting frameworks, such as XGBoost and AdaBoost, is its use of histogram-based training. Instead of evaluating each possible split on continuous input features, LightGBM pre-processes the data by grouping feature values into discrete bins. This approach allows the training process to operate on histograms rather than raw values, significantly reducing both memory usage and the computational cost of split finding. Once the histograms are built, the split search becomes more efficient because it operates only over the number of bins, which is typically much smaller than the number of training samples (i.e., stars in the dataset).

Another central innovation is its leaf-wise tree growth strategy. While traditional decision tree algorithms expand trees level by level, LightGBM identifies the leaf with the highest contribution to loss reduction and grows the tree from that point. Instead of expanding all branches of the tree uniformly, LightGBM focuses on the leaf that currently offers the greatest potential to reduce the training loss. This results in trees that grow deeper only where needed, often leading to faster convergence. This strategy is known as best-first growth and tends to produce trees that reduce training loss more effectively with fewer iterations. The resulting trees are typically asymmetric and deeper where necessary, as illustrated in Figure 3.2. Although this strategy increases model flexibility, it can lead to overfitting if not properly regularized. Parameters such as the maximum depth of the tree or the minimum number of samples in a leaf are used to mitigate this effect.

In terms of scalability, LightGBM supports a range of parallel and distributed learning strategies. On a single machine, it leverages multi-threading to speed up training by distributing work across all available processing cores. For multi-machine training, it provides both feature-parallel and data-parallel modes, which are designed to minimize communication overhead and distribute computational effort (the former distributes the features while keeping the full dataset on each machine, whereas the latter splits the dataset by rows across machines).

Despite its focus on performance, LightGBM incorporates a wide range of regularization options to control model complexity. These include parameters to limit the number of leaves in a tree, enforce a minimum number of data points in each leaf, and define the minimum gain (that is, the minimum improvement in prediction quality) required to perform a split. LightGBM also supports row and column sub-sampling, which introduces randomness into the training process in a way conceptually similar to bagging. Additionally, it allows for early stopping based on validation performance and includes explicit

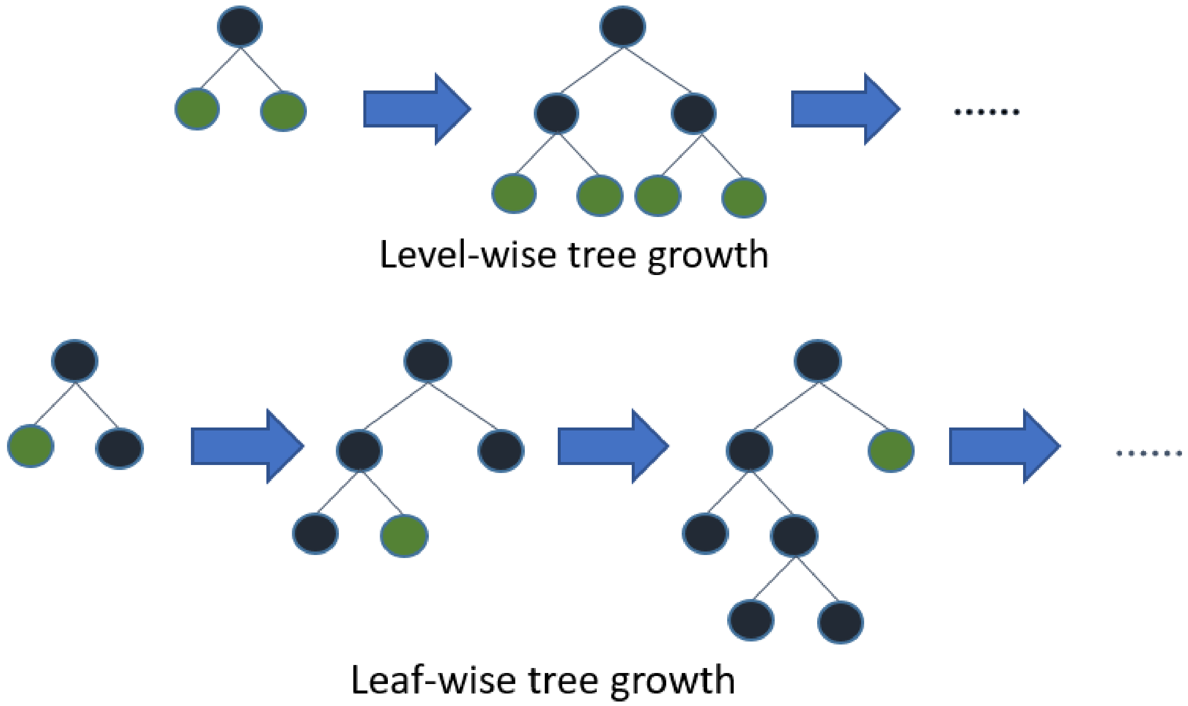

Figure 3.2: Illustration of tree growth strategies. Top: level-wise (depth-wise) expansion. Bottom: LightGBM's leaf-wise (best-first) strategy. Figure adapted from LightGBM documentation (last access on June 27, 2025).

penalties for model complexity, all of which help ensure that models generalize well to new data.

Interpretability remains a key advantage of LightGBM. Since the model is composed of decision trees, it is possible to analyze how different features contribute to predictions. Feature importances can be extracted based on gain or frequency of use, while other approaches can be optionally implemented by the user depending on the preferred mode of interpretation. In astrophysical contexts, such transparency is especially important for understanding how photometric features influence the derived stellar parameters. These importance metrics are used later in this work to inform feature selection and guide model interpretation.

In addition to its architectural design, LightGBM is highly configurable for different learning tasks. It supports various loss functions appropriate for regression, such as L2 (mean squared error), L1 (mean absolute error), quantile, Huber, and Poisson. This flexibility allows the user to adapt the training objective to the specific nature of the data, whether the data is sensitive to outliers, exhibit heteroscedasticity, or deviate from Gaussian assumptions.

LightGBM also includes functionalities that enhance its general applicability, such as support for sparse and categorical data, distributed training across multiple machines, and

GPU acceleration. These were not required in the present analysis, since missing values were explicitly handled during pre-processing (as addressed in Section 4.1), all input variables are continuous, and training was performed without GPU or multi-machine parallelism. Nonetheless, LightGBM is capable of constructing histograms using only non-zero values, partitioning categorical variables without one-hot encoding (a common strategy in which each category is converted into a separate binary variable), and scaling efficiently to large computing clusters when needed.

In summary, LightGBM extends the framework of gradient boosting through practical innovations that make it particularly well suited to scientific modeling. Its speed, memory efficiency, configurability, and support for distributed and GPU training align with the demands of large-scale photometric regression tasks. These properties justify its selection as the learning algorithm used throughout this thesis to predict stellar atmospheric parameters.

3.2 Machine learning in astronomy

Having touched on the main algorithmic tool employed in this thesis, this section turns to the broader scenario of ML in astronomical research. A brief overview is provided before stepping into the work done here (Chapter 4), with emphasis on the estimation of atmospheric parameters (Section 3.2.1), particularly via ensemble methods.

The early use of ML in astronomy dates back to the 1980s[4], when RUMELHART *et al.* (1986a)'s seminal work on backpropagation crucially revived interest in AI. In particular, two influential studies published in 1988 independently highlighted how ANNs could be applied to astronomical problems (SMITH & GEACH, 2023). One explored automated galaxy classification (RAPPAPORT & ANDERSON, 1988), while the other focused on the identification of point sources in infrared data (ADORF & JOHNSTON, 1988). These early demonstrations helped initiate a rise of interest in the so-called connectionist[5] models within astronomy, leading to a broadening of use cases throughout the early 1990s (MILLER, 1993, see SMITH & GEACH 2023). A few works followed, both pioneering decision trees (e.g. OWENS *et al.*, 1996; WEIR *et al.*, 1995; WEIR *et al.*, 1995) and advancing ANNs (FOLKES *et al.*, 1996; LAHAV *et al.*, 1996; ODEWAHN *et al.*, 1992; STORRIE-LOMBARDI *et al.*, 1992) in astronomy, while distinct prominence was earned by the distinguished software Source Extractor (SExtractor; BERTIN & ARNOUTS, 1996), which employs a multilayer feed-forward NN to estimate *a posteriori* probabilities

---

[4] While principal component analysis (PCA) has been widely applied in astronomy (see e.g. BARON, 2019; FLUKE & JACOBS, 2020) since DEEMING (1964), it is not included in this short overview, as this pre-ML statistical method lacks iterative learning or generalization capabilities, despite its occasional later adoption in ML workflows.

[5] Connectionism is a broad term in AI referring to the use of neuron-like connectors for learning modeling, such as in ANNs. It contrasts with symbolic AI, which relies on explicit rules and logic-based representations, such as in natural language processing.

for star/galaxy classification.

Highlighted applications of machine learning in astronomy began to emerge more frequently in the subsequent decade. The 2000s saw a surge in ML adoption, particularly driven, for instance, by the Sloan Digital Sky Survey (SDSS; YORK *et al.*, 2000), with ANNs and support vector machines (SVMs) dominating the landscape. Examples of key advancements included photometric redshift estimation (COLLISTER & LAHAV, 2004; WADADEKAR, 2005), automated classification of variable stars (WOŹNIAK *et al.*, 2004), solar flare prediction (QAHWAJI & COLAK, 2007), and galaxy morphology analysis (BALL *et al.*, 2004; BANERJI *et al.*, 2010). Ensemble methods also gained traction for tasks like morphological galaxy classification (BAZELL & AHA, 2001; DE LA CALLEJA & FUENTES, 2004) as well as galaxy spectral modeling (LU *et al.*, 2006). This moment set the stage for modern data-driven astronomy (for a review, see BALL & BRUNNER, 2010), bridging traditional techniques with the computational demands of data-intensive surveys.

The exponential growth of AI/ML use in astrophysics became increasingly evident in the following decade (RODRÍGUEZ *et al.*, 2022), as illustrated in Figure 3.3. Driven by advances in deep learning, large surveys, high performance computing, among others, this rapid progress made it impractical to comprehensively track the expanding literature. In this context, the studies cited here do not aim to exhaustively cover the field, but rather to provide representative examples, much like in previous paragraphs, though now with an even greater volume of work inevitably (but unwillingly) left out.

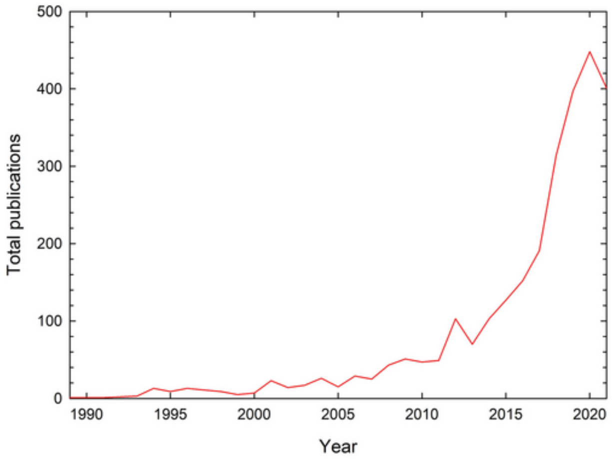

Figure 3.3: Chronological evolution of the number of publications related to the application of AI/ML in astronomy/astrophysics. Figure 1 from RODRÍGUEZ *et al.* (2022).

The 2010s witnessed the consolidation of ML as an essential tool across multiple astrophysical domains, transitioning from initially exploratory use cases to widespread adoption in scientific pipelines. Notable advances were achieved in time-domain event detection (e.g. NARAYAN *et al.*, 2018; RICHARDS *et al.*, 2011), photometric redshift estimation (e.g. D'ISANTO & POLSTERER, 2018; SADEH *et al.*, 2016), galaxy classification (e.g.

DIELEMAN *et al.*, 2015; DOMÍNGUEZ SÁNCHEZ *et al.*, 2018), gravitational lens modeling (e.g. HEZAVEH *et al.*, 2017; LANUSSE *et al.*, 2018), stellar parameter inference (e.g. NESS *et al.*, 2015; TING *et al.*, 2019), and exoplanet detection and characterization (e.g. SHALLUE & VANDERBURG, 2018; ZUCKER & GIRYES, 2018). In addition, the systematic implementation of ML-based strategies in key frameworks, such as large-scale surveys (e.g. BELLM *et al.*, 2019; KHAN *et al.*, 2019) and adaptive optics systems (e.g. SUÁREZ GÓMEZ *et al.*, 2019), became standard practice in many data-intensive efforts.

Astronomical literature since then saw expanded interest also in a broader set of algorithms often integrated into ML pipelines, including dimensionality reduction and similarity-based techniques. Particularly for tasks like feature compression and structure discovery in high-dimensional data, clustering methods such as (k)PCA (e.g. DAVIS *et al.*, 2017; ISHIDA & DE SOUZA, 2013; SUN *et al.*, 2021), t-SNE (e.g. ANDERS *et al.*, 2018; DA SILVA & SMILJANIC, 2025), and UMAP (e.g. GUTIÉRREZ-SOTO *et al.*, 2025; SUN *et al.*, 2021) began to be increasingly adopted. While these approaches are frequently associated with ML workflows (e.g. FOTOPOULOU, 2024), they differ in important ways from unsupervised, clustering learning algorithms, which also found growing use in the field, such as support vector classifiers (e.g. WHITTEN *et al.*, 2019), k-nearest neighbors (e.g. ZOU *et al.*, 2019), k-means (e.g. HOGG *et al.*, 2016), and density-based clustering methods like HDBSCAN (e.g. CASTRO-GINARD *et al.*, 2018; GUTIÉRREZ-SOTO *et al.*, 2025; HUNT & REFFERT, 2021). As the former do not involve learning a model that generalizes to new data in the usual sense, and given the limited scope of this overview, those were not included among the core methods discussed here.

In particular, several works using data from J-PLUS employed ML techniques for stellar astrophysics. WHITTEN *et al.* (2019) applied their artifical neural network, SPHINX, to J-PLUS photometry to estimate stellar effective temperatures and metallicities, achieving typical uncertainties of ~ 91 K in $T_{\rm eff}$ and ~ 0.25 dex in [Fe/H]. YANG *et al.* (2022) applied a cost-sensitive neural network (CSNet) to J-PLUS DR1 photometry to estimate stellar parameters and elemental abundances for two million stars. Their method achieved precisions of ~ 55 K in $T_{\rm eff}$, ~ 0.15 dex in $\log g$, and ~ 0.07 dex in [Fe/H], as well as uncertainties of $0.04 - 0.08$ dex for [C/Fe], [N/Fe], [Mg/Fe] and [α/Fe]. In a complementary approach, WANG *et al.* (2022) used support vector regression (SVR) on J-PLUS DR1 to estimate stellar parameters for 2.5 million stars, obtaining root mean square errors of ~ 160 K in $T_{\rm eff}$, ~ 0.35 dex in $\log g$, and ~ 0.25 dex in [Fe/H].

GALARZA *et al.* (2022) developed the SPEEM pipeline, using Random Forest and XGBoost to estimate $T_{\rm eff}$, $\log g$, and [Fe/H] from J-PLUS photometry. Applied to J-PLUS DR2, it achieved typical errors of 41 K for $T_{\rm eff}$, 0.11 dex for $\log g$, and 0.09 dex for [Fe/H] successfully identifying a number of very metal-poor (VMP) and extremely metal-poor (EMP) stars with a 64% success rate in spectroscopic validation. QUISPE-HUAYNASI *et al.* (2023) identified 28 stars with Galactocentric velocities exceeding 400 km s^{-1} and inferred stellar parameters using a convolutional neural network (CNN) trained on LAMOST data, achieving typical uncertainties of ~ 100 K in $T_{\rm eff}$, ~ 0.12 dex in [Fe/H], and ~ 0.18 dex in $\log g$. More recently, HUANG *et al.* (2024) extended these efforts to J-PLUS DR3, deriving [Fe/H], [C/Fe], [Mg/Fe], and [α/Fe] for five million stars using kernel principal component analysis, with results achieving typical precisions of $0.10-0.20$ dex for [Fe/H] and [C/Fe], and ~ 0.05 dex for [Mg/Fe] and [α/Fe]. Furthermore, given the resemblance in filter systems, S-PLUS has also been similarly used for stellar parameter estimation in the southern sky (e.g. LOPES *et al.*, 2025; QUISPE-HUAYNASI *et al.*, 2024; WHITTEN *et al.*, 2021)

Moreover, recent approaches have continued to build on this rationale, while extending the applicability of DL/ML across a broader range of astrophysical problems. Applications include exoplanets' identification (e.g. WANG, 2025) and habitability classification (e.g. KONG *et al.*, 2022), hybrid ML methods to infer stellar parameters of stellar spectra (e.g. SHANKAR *et al.*, 2025), multi-epoch simulations (e.g. MCGIBBON & KHOCHFAR, 2023) and image classification (e.g. TARSITANO *et al.*, 2022) of galaxies, gravitational wave signal modeling (e.g. ZHAO *et al.*, 2023), and simulation-based cosmology (e.g. LEMOS *et al.*, 2023). Looking ahead, ongoing integration of ML with cloud-based processing (e.g. THE MULTIMODAL UNIVERSE COLLABORATION *et al.*, 2024) and forward modeling frameworks (e.g. ANDRAE *et al.*, 2023) promises to further enhance scalability and physical interpretability in next-generation surveys, such as WEAVE (DALTON & THE WEAVE CONSORTIUM, 2012) and 4MOST (DE JONG & 4MOST CONSORTIUM, 2019), as well as those provided by the DESI (AGHAMOUSA & THE DESI COLLABORATION, 2016), Rubin Observatory (IVEZIĆ & THE LSST COLLABORATION, 2019), and Roman Space Telescope (MOSBY & THE ROMAN SPACE TELESCOPE DETECTOR TEAM, 2020).

This brief overview of early AI efforts aimed to provide context before transitioning to more focused topics. The next section focuses on the use of ML in stellar astrophysics, with particular emphasis on the estimation of atmospheric parameters. For recent comprehensive reviews on AI and ML applications in astronomy and astrophysics, see FLUKE & JACOBS (2020); KREMER *et al.* (2017); LONGO *et al.* (2019); SEN *et al.* (2022); WANG *et al.* (2018).

3.2.1 Atmospheric parameters from photometry and ensemble methods

This work employs the gradient-boosting framework LightGBM (Section 3.1.3) to estimate atmospheric parameters using photometric inputs. The following discussion briefly outlines the rationale for this approach, along with selected literature on related methods.

Ensemble methods improve robustness and generalization by combining multiple weak learners, offering a valuable advantage in the presence of noise and systematic uncertainties common in large photometric surveys. Simpler regression models often fall short in capturing the complex relationships between photometric inputs and atmospheric parameters, while deep learning techniques may require extensive training data or introduce unnecessary complexity. Ensemble approaches offer a balanced alternative, combining interpretability, computational efficiency, and competitive performance. These qualities make them particularly well suited to the task of atmospheric parameter estimation, where avoiding overfitting and managing high-dimensional inputs such as multi-band photometry is essential.

Nonetheless, while ensemble methods are capable of capturing nonlinear mappings, their effectiveness can be limited by physical degeneracies. These occur when different combinations of stellar parameters lead to similar observational signatures. In cases where metallicity effects are entangled with temperature, for example, photometry alone may not be sufficient for reliable parameter recovery. In this scenario, feature construction becomes essential, whether through human-curated means or automated techniques. This step is particularly rich and complex in astrophysics, given the reasonable expectation that features reflecting observable signatures of the underlying stellar physics should yield better performance. In light of this, input ML features may originate from a wide range of intrinsic stellar properties, including kinematics, spectral characteristics, pulsation modes, variability, and, notably, traditional photometric measurements.

In fact, photometry constitutes a benchmark for the application of ML in astrophysics. Photometric features typically consist of magnitudes and (pseudo-[6])colors, each potentially sensitive to different atmospheric properties. In this sense, and particularly for the estimation of the stellar parameters considered here, feature engineering directly guides the model's ability to extract meaningful patterns from data, given the physical interpretability of the input observational quantities.

Color indices can enhance the performance of ensemble models by capturing relative flux variations while mitigating distance effects. By emphasizing spectral energy distribution shapes rather than absolute flux measurements, colors allow models to outperform those trained on magnitudes alone in ML tasks (e.g. LIANG *et al.*, 2022). Variations in the overall continuum shape reflect changes in effective temperature, while pressure-sensitive features such as the Balmer discontinuity and molecular bands modulate colors in ways that correlate with surface gravity. Similarly, line blanketing effects in the blue and near-UV regions reduce the flux of metal-rich stars, altering colors in a metallicity-dependent manner. Tree-based models can learn these subtle yet consistent patterns across multi-

---

[6] Combinations of colors, often referred to as pseudo-colors, can also be explored. However, increasing complexity, such as varied operations across multiple magnitudes, may come at the expense of interpretability.

ple colors and capture complex interactions by identifying meaningful regions in feature space, thus effectively constraining T_{eff}, log g, and [Fe/H] using only photometric data.

In this context, even though most AI applications in astronomy focus on neural networks, ensemble methods also figure among common choices for estimating stellar atmospheric parameters from photometric data (RODRÍGUEZ *et al.*, 2022). A number of studies have successfully applied these techniques to this task, most often employing random forests. MILLER *et al.* (2015) use random forest regression on photometric variability to recover effective temperature, surface gravity, and metallicity with rms errors of roughly 165 K, 0.39 dex, and 0.33 dex. SARRO *et al.* (2018) evaluate multiple machine learning regressors, including random forests, to show that models based on features selected based on genetic algorithms can identify subdwarfs and perform comparably to classical methods for T_{eff} and [M/H]. BAI *et al.* (2019) train a supervised regressor on over four million stars to predict Gaia DR2 T_{eff} with a precision of 191 K. CHEN *et al.* (2022) employ random forests to define line indices from LAMOST A-type spectra, enabling refined indices that improve SVR predictions under noise and extinction. Complementarily, BRETON *et al.* (2021) developed a random forest-based pipeline to extract stellar rotation periods from Kepler light curves, achieving over 95% agreement with reference values and demonstrating high reliability across large samples with minimal visual inspection.

In contrast to random forests, gradient boosting seems to remain underexplored even among ensemble applications (RODRÍGUEZ *et al.*, 2022), despite its well-known advantages in other domains. WANG *et al.* (2024a) apply gradient boosting decision trees to LAMOST RGB stars for age estimation, with median relative errors near 11.6% and total uncertainties between 15–30%. LIANG *et al.* (2022) combine PCA with LightGBM to estimate T_{eff}, log g, and [Fe/H], reporting rms errors of 90 K, 0.40 dex, and 0.20 dex, respectively. AVDEEVA *et al.* (2024) employ XGBoost, CatBoost, and LightGBM to assign quality flags to Gaia DR3's effective temperature estimates, reporting approximately 90% precision in selecting T_{eff} values accurate to around 250 K. Furthermore, EL-KHOLY & HAYMAN (2025) explore active learning to combine various methods, including gradient boosting, to optimize spectral sampling in the SDSS-IV MaStar library, demonstrating that this approach provides selecting samples with superior performance metrics in comparison to random sampling and even stratified samples, using fewer training instances.

Some works include both random forest and gradient boosting methods to illustrate widespread use rather than for direct comparison. In the context of the J-PLUS survey, the previously mentioned work from GALARZA *et al.* (2022) develop the SPEEM pipeline using both random forests and gradient boosting, reporting median absolute deviations of 41 K in T_{eff}, 0.11 dex in log g, and 0.09 dex in [Fe/H]. QU *et al.* (2024) analyze Gaia DR3 cool-dwarf atmospheric parameters with both methods, yielding error dispersions of 68 K, 0.05 dex, and 0.22 dex for T_{eff}, log g, and [Fe/H], respectively. Moreover, stellar parameters have also been estimated through other less common ensemble methods, such as stacking (KAMULALI *et al.*, 2025; MOYA & LÓPEZ-SASTRE, 2022) and extra trees (ANDRAE *et al.*, 2018; GAVEL *et al.*, 2021; KAMULALI *et al.*, 2025), while isolation forests have also been explored for outlier removal (BANSAL *et al.*, 2024; WANG *et al.*, 2024b).

Ultimately, although both ensemble methods, mostly random forests and gradient boosting, are widely used for stellar parameter estimation, their core differences shape how they are best applied. Random forests offer robustness, simplicity, and resistance to overfitting, making them well suited for exploratory analyses and noisy data. Gradient boosting, in contrast, refines predictions iteratively and can achieve higher accuracy in structured problems, though it requires more careful tuning. In practice, the choice between them often reflects priorities such as model stability, interpretability, or predictive precision rather than clear empirical superiority.

Given its flexibility and high precision, gradient boosting offers a compelling balance between performance and scalability. While it may require more tuning than random forests, it remains a strong choice when careful optimization of predictive accuracy is desired, especially in large, structured datasets such as multiband photometry. In this sense, LightGBM stands out by often outperforming in speed and memory efficiency while matching or exceeding accuracy compared to both random forests and other gradient boosting frameworks. It is therefore the primary method adopted in the present work.

Chapter 4

Data and models

This chapter presents the construction of the models developed to estimate stellar atmospheric parameters. Section 4.1 details the selection of the input sample, while Section 4.2 describes the final models. The procedures for selecting the most informative features, including those that allow physical interpretation, and for optimizing hyper-parameters are also outlined. Model performance is evaluated, and the methodology used to estimate the uncertainties of the predicted parameters is discussed. Importantly, a condensed version of the analysis presented here appears in MACHADO-PEREIRA *et al.* (2026), with slight differences in data processing and discussion.

4.1 Final training sample

The final training sample arose after considerable experimentation, mainly prioritizing simplicity and specifically avoiding significant quality loss. As in many similar studies, photometry here provides input features for the model, while gives provides the parameters to be reproduced, i.e., the *targets*. The following sections describe these selections in further detail: features in Section 4.1.1 and targets in Section 4.1.2.

In addition, although uncertainties in the input data are not explicitly included in model training, they are partially accounted for by aiming at the best-performing model, which means requiring the most accurate data available. Nevertheless, as mentioned earlier, individual uncertainty estimates for the predicted stellar parameters are provided (Section 5.2).

4.1.1 Features

The primary source of photometric data used in this work is J-PLUS DR3 catalog, with complementary photometry from CatWISE and Gaia also explored. The process of data retrieval and filtering is detailed ahead.

J-PLUS data

J-PLUS constitutes the primary data source explored in the present work. In particular, its third data release (DR3; LÓPEZ-SANJUAN *et al.*, 2024) includes 47.4 million sources across 3192 deg^2, delivering a unique multicolor dataset of the local universe with improved photometric calibration. Since July 2022, J-PLUS DR3 is fully available for the collaboration members. It comprises 1642 J-PLUS fields observed in twelve optical bands amounting to 3192 deg^2. J-PLUS DR3 is based on images collected from November 2015 to February 2022 by the JAST80 telescope and includes $\approx$ 30 million stars brighter than $r = 21$.

Data from the J-PLUS DR3 and previous data releases can be accessed from its website[1] in various ways, including Virtual Observatory (VO) services such as Table Access Protocol (TAP), e.g., through Astronomical Data Query Language (ADQL, OSUNA *et al.*, 2008) queries. However, the limit of rows that the TAP service returns is 1 million. To overcome this limitation and avoid the risk of not accessing all J-PLUS DR3 stars, the ADQL query showed in Appendix B was attached into a Python code to acquire the data in the form of blocks of reference tiles, so that the full query is split into several minor queries, as indicated by the user manual for the CEFCA's Catalogues Portal[2] (Section 4.4.2). The script applies quality cuts and retrieves useful information such as sky positions, magnitudes, and their respective uncertainties.

In particular, J-PLUS photometry offers various kinds of magnitudes, e.g. measured from aperture or point spread function techniques. The standard 6” aperture limits flux contamination from nearby sources and prevents dominance by background noise, but it is insufficient to capture the total flux of the stars. In this work, following LÓPEZ-SANJUAN *et al.* (2019), the magnitude used here is from a 6” diameter aperture after correction for aperture effects (column `APER_COR_6_0`).

For J-PLUS stars selection, the constraints applied to the DR3 database include the following (see Appendix B).

- Star/galaxy separation:
  `sglc_prob_star` > 0.7 to select objects with more than 70% probability of being a star.
- Magnitude error:
  `array_max_float(ps.mag_err_aper_cor_6_0)` < 0.5 to (leniently) exclude stars with obviously low S/N in the 6″ diameter in any of the twelve filters.
- Adequate exposure:
  `array_min_float(dualobj.norm_wmap_val)` > 0.8 to avoid sky regions not observed at full depth.

---

[1] https://www.j-plus.es/datareleases/data_release_dr3

[2] https://archive.cefca.es/doc/manuals/catalogues_portal_users_manual.pdf, version 1.29 was consulted.

- Quality flags:
  `array_max_int(dualobj.mask_flags) = 0` to select stars without any warning flags, meaning with good photometry.

This query returned 5 114 494 sources, which are used for cross-matching with complementary catalogs, as described below.

CatWISE and Gaia data

In addition to J-PLUS, two other surveys are explored for gathering features: CatWISE (MAROCCO *et al.*, 2021) and Gaia DR3 (GAIA COLLABORATION *et al.*, 2023).

Infrared photometry from CatWISE complements the optical coverage of J-PLUS by providing sensitivity to cooler stellar photospheres and enabling a better characterization of late-type stars, which are intrinsically redder. The longer wavelengths sampled by the CatWISE W1 and W2 bands trace higher atmospheric layers (i.e., the upper photosphere) compared to optical bands, and are also less affected by interstellar extinction. These properties offer additional discriminatory power for estimating stellar parameters. Including CatWISE data improves the feature set and can refine the model's ability to capture more complex relations in feature space, particularly helping to break T_{eff}–[Fe/H] degeneracies (e.g. ANDRAE *et al.*, 2023). Moreover, infrared data can be especially valuable when working with stars in dustier environments or at larger distances, where extinction becomes more severe and must be accounted for more carefully.

Data from Gaia DR3 were also incorporated in this work, primarily to obtain distances and derive absolute magnitudes to be used as features. From first principles, surface gravity is strongly tied to a star's intrinsic luminosity, so absolute magnitudes carry valuable information for predicting $\log g$. Consequently, their inclusion enhances the model's ability to separate stars of similar temperature but different evolutionary stages. This is not only expected based on physical reasoning but was also shown to hold empirically (CORDEIRO DA SILVA & BORGES, 2023). While simple parallax inversion is acceptable in very precise cases, distance estimates are taken from the photogeometric solutions of BAILER-JONES *et al.* (2021), which provide significantly more robust, precise, and accurate estimates. In this case, since Gaia data would already be part of the sample, its photometry was also included in the feature set.

The training sample was further constrained after all data sources (J-PLUS, CatWISE, Gaia; in addition to LAMOST, addressed ahead) were gathered, which resulted in a total of 17 magnitudes (twelve from J-PLUS, three from Gaia, and two from CatWISE). This process is described in Section 4.1.3, which also addresses the cross matches for building the final sample.

4.1.2 Targets

The selection of a spectroscopic sample for training was guided by some key criteria known to support robust performance in supervised learning. Such a sample is expected to be accurate, representative, clean, consistent, and sufficiently large. In other words, the ideal dataset should provide reliable stellar parameters with uncertainties (accuracy), span a adequate range range of atmospheric and photometric properties (representativeness), allow filtering of low-quality entries (cleanliness), be internally homogeneous (consistency), and include a large number of sources to provide statistical support, reduce variance, and improve generalization (sufficiency). These principles motivated the choice of LAMOST as the spectroscopic reference for this work, as it offers millions of uniformly processed stellar spectra, with quantified parameters and uncertainties homogeneously provided by LASP (WU *et al.*, 2014). These characteristics make it particularly well-suited for training data-driven models with general applicability.

Rather than following through with multiple spectroscopic surveys for target selection, only LAMOST was used mainly due to its substantially larger sample size. While including other spectroscopic surveys could, in principle, extend the parameter coverage or slightly increase the number of training examples, the gain would be marginal: LAMOST contains approximately 6 million stars, compared to fewer than $\lesssim 700\,000$ in any other single survey before any quality cuts. More importantly, combining different sources would introduce complexity in both homogenizing the data to a common parameter scale and interpreting model results, as any derived conclusions would no longer be directly associated with a specific survey. Although other surveys, such as Gaia-ESO or APOGEE, provide more precise stellar parameters for individual stars, their sample size is significantly smaller, increasing the risk of overfitting, less stable model performance, and poorer generalization.

Preliminary tests with other spectroscopic surveys were also performed during the early stages of this work. Although they used similar input selections and generally yielded comparable outcomes, the samples were not identical, and no direct comparison was carried out. These exercises nonetheless helped guide the decision to adopt LAMOST as the sole spectroscopic source, given its balance of scale, consistency, and practicality.

In addition, efforts based on homogenized datasets were also excluded, particularly the PASTEL catalog (SOUBIRAN *et al.*, 2016, the largest of its kind), primarily due to its limited size: it contains around 81 000 stars, of which only about 24 000 have both determinations and associated uncertainties for T_{eff}, log g, and [Fe/H]. **(PAULA) Similarly, stellar parameters stemmed from synthetic spectra were not employed in the present work, in favor of using real observed data.**

Acquiring LAMOST data

At the time this work was initiated (early 2023), the latest publicly available release of LAMOST was Data Release 8 (DR8), which was therefore adopted for the analysis. Although more recent catalogs have since been published, the incremental increase in the number of stars with available stellar parameters is relatively modest. For AFGK-type stars, the number of entries with parameters increased from 6 684 413 in DR8 version 2.0 (v2.0) to 6 921 466 in DR9 v2.0, and 7 450 303 in DR10 v2.0[3].

In LAMOST, stellar spectra are available in both low- and medium-resolution modes, each processed through distinct pipelines. While medium-resolution spectra offer higher precision, only 717 571 DR8 stars have published atmospheric parameters in this regime. In contrast, low-resolution spectra cover over 6 million stars, offering greater statistical support and coverage of the parameter space. For these reasons, and in line with the initial design and timeline of this work, the analysis was carried out using low-resolution data.

Data from DR8 v2.0 is publicly available from the LAMOST DR8 portal[4], and access is made through various means, including a simple search for objects in either low- or medium-resolution mode. For this work, a basic search was carried out using low-resolution data. The portal offers several selection constraints, such as position, internal designation, S/N and stellar parameters (AFGKM-type), among others. In this case, the only constraint applied was to exclude spurious values by requiring uncertainties on the parameters of interest (T_{eff}, log g, [Fe/H]) to be greater than or equal to zero. Although zero-error values are also uninformative and would ideally be excluded, the data access tool allows filtering only through minimum and maximum value fields. As such, setting a lower limit of zero was a practical choice that removed clearly invalid entries without requiring more elaborate post-processing.

This process resulted in almost the entirety of the available objects, with 6 684 026 entries. The main columns required were basic identifiers (including LAMOST internal designation and Gaia DR3 source identification), positions, S/N in SDSS *ugriz* bands and the derived stellar parameters (T_{eff}, log g, and [Fe/H], along with their respective uncertainties).

4.1.3 Cross matches and preprocessing

The sources presented in the previous sections were cross matched for assembling a preliminary sample, yet to be refined through preprocessing steps, as presented ahead.

---

[3] As of the time of writing (mid 2025), Data Releases 11 through 13 remain internal. Within each release, version 2.0 refers to the final reprocessed version.

[4] `https://www.lamost.org/dr8/v2.0/`

Cross matches

All cross-matching steps were performed within TOPCAT[5] (TAYLOR, 2005), which offers a built in functionality of sky cross-matching a local table with any table provided by the VizieR database. A maximum angular separation of 3" between matched sources was adopted.

The resulting list from J-PLUS query (Section 4.1.1) was cross-matched with CatWISE (The CatWISE2020 Catalog; MAROCCO *et al.*, 2021) to include *W1* and *W2* magnitudes in the model construction, as infrared photometry information from these bands helps reducing temperature–extinction degeneracy (ANDRAE *et al.*, 2023). The provided 5 027 068 stars were then cross-matched with Gaia data, more specifically with:

i) DR3 main source catalog for the usual Gaia magnitudes G, B_P and R_P (BABUSIAUX *et al.*, 2023; GAIA COLLABORATION, 2022; GAIA COLLABORATION *et al.*, 2023), returning 5 026 927 stars;

ii) with EDR3 distances catalog from BAILER-JONES *et al.* (2021), returning 5 004 860 stars;

iii) and with DR3 astrophysical parameters (CREEVEY *et al.*, 2023; GAIA COLLABORATION, 2022) for information on extinction corrections, returning 5 004 738 stars.

Recall that keeping only stars with valid uncertainties ($\sigma_{T_\mathrm{eff}}, \sigma_{\log\,g}$ and $\sigma_{[\mathrm{Fe/H}]} > 0$) in the LAMOST DR8 sample resulted in 6 684 026 stars. Cross-matching these with the list obtained from the previous cross-match step leaves the sample with 529 565 stars, which are further refined as detailed below. Table 4.1 summarizes the steps described above.

Table 4.1: Summary of steps in the construction of the training sample, with each row consecutively describing the cross matches performed (within a 3″ search radius from the J-PLUS RA and DEC).

Catalog	Sample size	Fraction	Notes
J-PLUS DR3	5 114 494	1.00000	Appendix B
CatWISE	5 027 068	0.98291	For $W1$ and $W2$
Gaia DR3	5 026 927	0.98288	For G, B_P, R_P
Gaia EDR3$^{(a)}$	5 004 860	0.97856	For distances$^{(a)}$
Gaia AP	5 004 738	0.97854	For extinctions
LAMOST DR8	529 565	0.10354	For spectroscopic targets

Notes. (a) Values from the work of BAILER-JONES *et al.* (2021).

---

[5] Tool for OPerations on Catalogues And Tables: `https://www.star.bris.ac.uk/~mbt/topcat/`; version 4.9 was used.

Preprocessing

Data preprocessing in ML involves cleaning and transforming raw data into a usable format for training models. Key steps typically include handling missing values, dealing with outliers, scaling numerical features, and splitting the data into training, validation, and testing sets; for classifying tasks, it also includes encoding categorical variables and stratified sampling to handle class imbalance.

First, the Gaia DR3 source identifiers were required to be identical across all Gaia catalogs and the LAMOST dataset. Next, stars with obviously low-quality spectra were initially filtered out by selecting S/N > 0 in all SDSS *ugriz* bands from LAMOST. This was still further refined after some experimentation, as will be addressed shortly. Finally, only the best parameters were kept by requiring uncertainties limited to 100 K in $T_{\rm eff}$ and 0.1 dex in both log g and [Fe/H]. These actions resulted in a list with 331 293 stars.

In order to ensure photometric reliability, only sources with magnitudes brighter than 21 were retained in all filters. This threshold is slightly conservative relative to the 5σ limiting magnitudes reported for a $3''$ aperture in J-PLUS DR3, which range from 20.4 to 21.8 depending on the filter. Although a $6''$ aperture was adopted in this work, which may alter the effective depth, the chosen limit acts as a practical proxy to avoid low S/N measurements, especially in the bluer filters where the limits are more restrictive. Reinforcing this rationale, photometric uncertainties were required to be below 0.1 mag in every band, including those from J-PLUS, CatWISE, and Gaia, further ensuring the quality of the input features. Of the 331 293 sources initially read, 316 474 met these conditions, indicating that the early selection was already effective in retaining the most representative sources.

Nevertheless, a few even more restrictive steps were taken, one particularly regarding S/N in LAMOST. To further ensure spectroscopic data quality, a set of S/N thresholds was applied across the available spectroscopic bands. A stricter cut of S/N > 50 was enforced for the g and r bands, for which the distributions in Figure 4.1 show a noticeable drop in parameter uncertainties above this level. Although u plays a central role in constraining metallicity, it typically suffers from lower S/N; applying the same threshold as in r would result in a substantial loss of sources. Since uncertainty distributions for u likely exceed those of r, the S/N$>$ 10 threshold adopted for u, as well as for i and z, reflects a compromise aimed at preserving sample size without severely compromising reliability. The same threshold (SNR > 10) was adopted for z and for the $W1$ and $W2$ bands from CatWISE, following typical minimum S/N practices to ensure photometric reliability, likewise despite the absence of a rigid criterion. This way, the limits are summarized as follows, preserving high-quality data whilst keeping the sample large enough for training, with 119 188 stars so far:

- $\mathrm{SNR}_g^{\min} = \mathrm{SNR}_r^{\min} > 50$;

- $\mathrm{SNR}^{\min}_{W1} = \mathrm{SNR}^{\min}_{W2} = \mathrm{SNR}^{\min}_{u} = \mathrm{SNR}^{\min}_{i} = \mathrm{SNR}^{\min}_{z} > 10$.

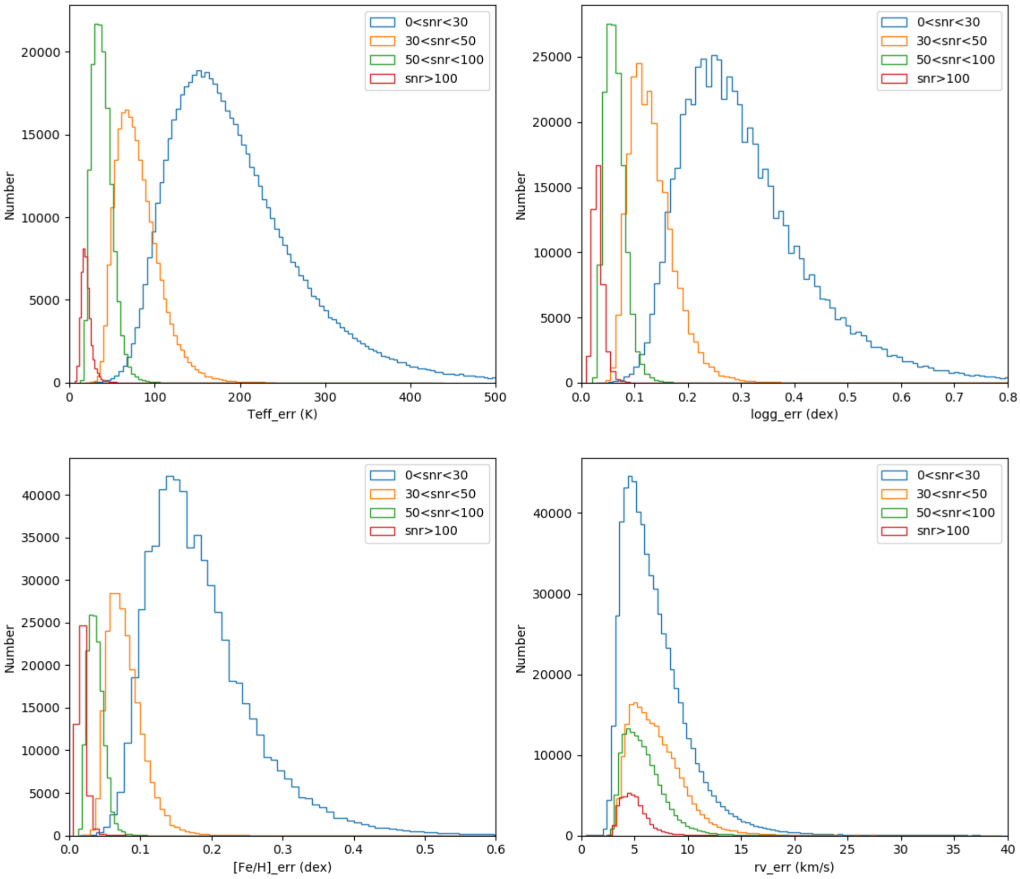

Figure 4.1: Distribution of LAMOST-derived parameter uncertainties as a function of signal-to-noise ratio (S/N) in each band, illustrating the improvement in parameter precision with increasing S/N. Radial velocities are also shown but are not used in this work. Figure from the LAMOST DR8 v2.0 Release Note.

Next, handling missing values, or not-a-number (NaN) entries, is a crucial step in data preprocessing, with common strategies falling into two main categories: deletion and imputation. Deletion removes rows or columns containing missing data, while imputation replaces missing values with estimated ones. In this work, rows with invalid values in key columns, such as distance, extinction, spectroscopic parameters, and their associated uncertainties, were excluded, since these columns had not been processed separately, thus staying at risk of containing NaNs. This filtering step reduced the sample by 12%, yielding a final set of 104 891 objects.

Additionally, constraints from Gaia DR3 on the parallax-over-error ratio (ϖ/σ_{ϖ}) and the re-normalized unit weight error (RUWE) were respectively set to $\varpi/\sigma_{\varpi} > 10$ and

RUWE < 1.4, avoiding likely poor quality astrometric solutions. The ϖ/σ_ϖ threshold ensures that only stars with well-constrained distances are retained, which is particularly important for deriving reliable absolute magnitudes and reducing propagation of uncertainties into the models. RUWE, on the other hand, is a diagnostic of the goodness-of-fit of Gaia's single-star astrometric model (see e.g. CASTRO-GINARD *et al.*, 2024). Values significantly above unity often indicate unresolved binaries, extended sources, or other deviations from the point-source assumption. The commonly adopted limit of 1.4 serves as a practical threshold to exclude most problematic cases while preserving the bulk of high-quality astrometric data. After these cuts, 88 421 stars compose a final sample to be used in training.

Although common in many ML workflows, feature scaling was not applied here. Tree-based models such as LightGBM are inherently insensitive to the scale of input features, since they rely on threshold-based splits rather than distance or gradient computations in the input space. In this case, gradient based models, such as neural networks and linear regression trained via gradient descent, compute gradients with respect to weights based on the relationship between inputs and targets, making them sensitive to the scale of both. In contrast, gradient boosting algorithms measure how the loss function changes with the model's predictions, not with respect to the input data. As a result, scaling in tree-based models only shifts the numeric value at which splits occur, without affecting the model's performance.

In this study, the full dataset is randomly partitioned into three subsets according to their role in the fitting process: 70% for training, 20% for validation, and 10% for testing. The training set enables the model to capture underlying patterns within the data; the validation set is specifically employed for hyper-parameter tuning, ensuring better generalization to new data by preventing over-fitting; the test set serves as the final evaluation benchmark, assessing the model's performance on data that is not part of the training or validation sets, providing a reliable internal measure of its present generalization capabilities to previously unseen portions of the dataset (IVEZIĆ *et al.*, 2020). Figure 4.2 presents the final distributions of the train (dotted lines), validation (dashed lines), and test (solid lines) samples within our full dataset. The mean and standard deviation values for the test sample are indicated in the legend and closely match those of the entire dataset, as shown in Table 4.2.

Table 4.2: Summary of final training sample with 88 421 objects

Parameter	Mean	SD	Minimum	Maximum
$T_{\rm eff}$ (K)	5857	465	4020	7499
log g (dex)	4.134	0.393	0.665	4.844
[Fe/H] (dex)	−0.216	0.306	−2.350	0.777

Interstellar extinction corrections for each of the 12 J-PLUS filters were performed

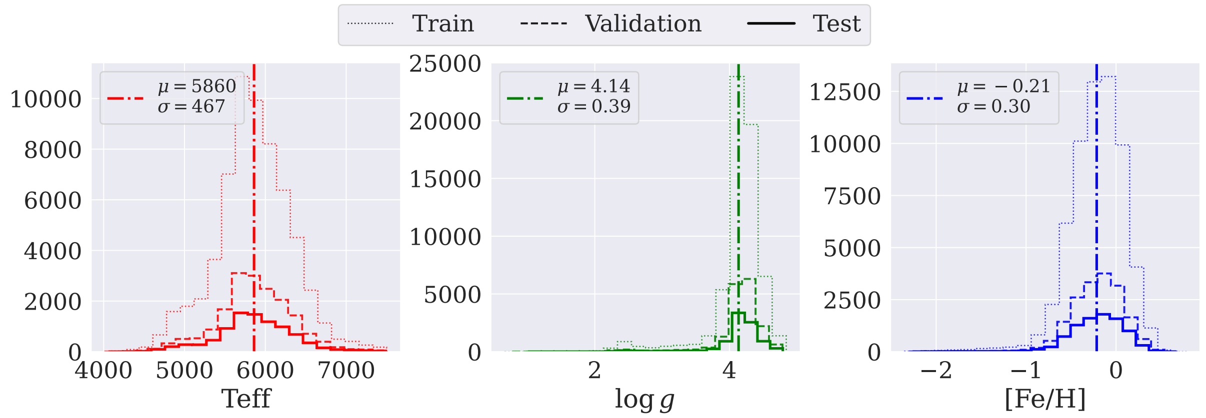

Figure 4.2: Distributions of the train (dotted lines), validation (dashed lines), and test (solid lines) sets within the full training sample. The mean and standard deviation of the test set are indicated by the legend and are as similar as expected to those of the train and validation sets, even though the latter are not explicitly shown.

using coefficients k_x provided in the J-PLUS catalog, where x represents each filter. For Gaia, extinction values were directly retrieved from DR3 catalog ($A_{\mathrm{G}}, A_{\mathrm{BP}}, A_{\mathrm{RP}}$; CREEVEY *et al.*, 2023; GAIA COLLABORATION, 2022), while coefficients from literature ($A_{\mathrm{W1}}, A_{\mathrm{W2}}$; YUAN *et al.*, 2013) were used to calculate CatWISE extinctions[6]. All cases assume an extinction factor of $R_{\mathrm{V}} = 3.1$.

Finally, all 136 possible color indices (subtracting bluer from redder magnitudes) were calculated from the 17 available magnitudes. Table 4.3 describes each filter, providing reference wavelength and width in Å. For J-PLUS filters, the reference wavelength is the central wavelength of the filter and the reference width is the filter amplitude, both as taken from the `jplus.Filter` table, available in J-PLUS Data Access Services. For Gaia and WISE filters, these quantities are retrieved from the Spanish Virtual Observatory (SVO) Filter Profile Service[7] (RODRIGO & SOLANO, 2020; RODRIGO *et al.*, 2012): the reference wavelength is taken to be the pivot wavelength (see Eq. A5 of RODRIGO *et al.* 2024), and the reference width is the effective width, defined as the width of a rectangle with a height equal to the maximum transmission and with the same area that the one covered by the filter transmission curve (see Eq. A4 of RODRIGO *et al.* 2024). This includes cross-survey combinations, such as J0430 − $W1$, $\mathrm{G}_{\mathrm{Gaia}} - J0660$ or $\mathrm{RP}_{\mathrm{Gaia}} - W2$.

Additionally, 17 absolute magnitudes were also calculated plugging photogeometric distances from the catalog of BAILER-JONES *et al.* (2021) in the following fundamental

---

[6] $A_x = k_x \times \mathrm{E}_{(\mathrm{B-V})}$, where x denotes the photometric band, k_x is the corresponding extinction coefficient for that band, and $\mathrm{E}_{(\mathrm{B-V})}$ is taken from J-PLUS DR3 for each object.

[7] https://svo2.cab.inta-csic.es/theory/fps/index.php?mode=browse

Table 4.3: List of the 17 photometric magnitudes used to build features in this work. Information on J-PLUS filters is from table `jplus.Filter` available in J-PLUS Data Access Services; information on Gaia and WISE filters is from the SVO Filter Profile Service.

Filter	λ_{ref} (Å)	W_{ref} (Å)	Note
uJAVA	3485.00	458.00	J-PLUS u
J0378	3785.00	138.00	J-PLUS [O II]
J0395	3950.00	70.00	J-PLUS Ca II H+K
J0410	4100.00	170.00	J-PLUS Hδ
J0430	4300.00	170.00	J-PLUS CH G-band
gSDSS	1250.00	1250.00	J-PLUS g SDSS-like
BP	5124.20	2333.06	Gaia G_{BP} filter$^{(a)}$
J0515	5150.00	170.00	J-PLUS Mg b Triplet
rSDSS	6250.00	1250.00	J-PLUS r SDSS-like
G	6251.50	4203.61	Gaia G filter$^{(a)}$
J0660	6600.00	100.00	J-PLUS Hα
iSDSS	7725.00	1250.00	J-PLUS i SDSS-like
RP	7829.65	2842.11	Gaia G_{RP} filter$^{(a)}$
J0861	8610.00	360.00	J-PLUS Ca II Triplet
zSDSS	9150.00	1300.00	J-PLUS z SDSS-like
W1	33526.00	6626.42	WISE $W1$ filter
W2	46028.00	10422.66	WISE $W2$ filter

Notes. (a) Values corresponding to pre-launch estimates, although Gaia DR3 data were used in the analysis. This only affected the sorting order and the sign convention in one color index (rSDSS−G, since λ_{ref} for G band in DR3 is 6217.59 Å), with no impact on the results.

distance modulus equation, solving for absolute magnitude M_0^{absolute}:

$$\begin{aligned} M_0^{\mathrm{absolute}} &= m_0^{\mathrm{apparent}} - 5\log_{10}\left(\frac{d/\mathrm{pc}}{10}\right) \\ &= m_0^{\mathrm{apparent}} - 5\log_{10}(d/\mathrm{pc}) + 5. \end{aligned} \tag{4.1}$$

In summary, there are 153 available features for the models (described in the next section) to map correlations and reproduce the spectroscopic targets, i.e. T_{eff}, log g and [Fe/H] of the 88 421 stars in the full training sample.

4.2 Final Models

This section describes the construction of the models developed to predict the target atmospheric parameters from photometric data. It includes details on the training configuration and model evaluation. The models adopted in this work were implemented using the gradient boosting framework LightGBM, as introduced in Section 3.1.3.

In machine learning, a loss function quantifies the difference between a model's predicted output and the actual (called ground truth) value. It's a crucial component for training models as it provides a metric to evaluate performance and guides the optimization process to improve accuracy. Among the commonly used regression losses, such as mean squared error (MSE), Huber loss, and mean absolute error (MAE), the present work adopted MAE (Equation 4.2, where y_i is the true value, $\hat{y}_i$ the predicted value, and n the number of samples) for its practical balance between simplicity and reliability.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.2}$$

Although no systematic comparison was carried out, this choice was motivated by two main factors. First, MAE is robust to outliers, which is desirable given that input data, despite undergoing quality control, may still contain non-representative or misleading cases. Second, it provides predictions directly expressed in the same units as the target variables, allowing for straightforward interpretation of model performance. While MAE yields constant gradients and lacks smooth differentiability, which can slow convergence in gradient boosting, it was still preferred for maintaining robustness to outliers and interpretability without introducing the added complexity of alternative loss functions.

For the training step, LightGBM (version 4.3.0) was used through the `scikit-learn` (version 1.4.1.post1) Application Programming Interface (API), via the `lightgbm.LGBMRegressor` function[8]. It offers a number of hyperparameters that can be tuned depending on the user's goals, some of which were optimized using `shap-hypetune`[9], a Python package developed for simultaneous hyperparameter tuning and feature selection, specifically tailored for gradient boosting models. Details of this process and the package are provided in the next section.

4.2.1 Optimal features and hyper-parameters

This section describes the full procedure adopted for selecting the most relevant input features and tuning the hyper-parameters of the predictive models. The approach integrates interpretability tools, optimization techniques, and cross-validation strategies to ensure robust and efficient model configuration. It begins with a general overview of the package used to orchestrate this process, `shap-hypetune`, followed by discussions on SHAP-based interpretability, Bayesian hyperparameter search, and implementation details. The section concludes with the application of RFECV and an analysis of the final feature importances. The focus is on conveying the practical and conceptual rationale of each step, while a formal mathematical treatment remains beyond the scope of this work.

---

[8] https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMRegressor.html

[9] https://github.com/cerlymarco/shap-hypetune

The `shap-hypetune` package

The package `shap-hypetune` provides a unified framework for simultaneous hyperparameter tuning and feature selection, addressing a process that is often conducted in separate and sequential steps. Originally designed for gradient boosting models such as LightGBM and XGBoost, it integrates smoothly with the `scikit-learn` library and supports both classification and regression tasks. The package offers flexible customization of the training process, including support for early stopping, and allows either task (tuning or feature selection) to be performed independently.

Search strategies for hyperparameter optimization include grid search, random search, and Bayesian optimization. Feature selection can be performed using recursive methods (elimination or addition) or the Boruta algorithm[10], and importance measures can be derived from either classical model-based scores (i.e., gain and split) or SHAP values, using the Python package SHAP[11].

SHAP-based interpretability in feature selection

SHapley Additive exPlanations (SHAP) values (LUNDBERG & LEE, 2017) are an explainability technique that breaks down a model's output into contributions from each feature based on Shapley values from cooperative game theory (SHAPLEY, 1951), where each feature is treated as a "player" in a game. In essence, SHAP values quantify how much each feature contributes to a specific prediction, enabling both global and local interpretability while preserving the original units of the input features. The SHAP value for each feature represents its average contribution to the prediction by considering all possible ways in which the feature can combine with other features. In other words, this method helps clarify how features influence specific predictions and their overall effect on the model's performance. Moreover, it has recently started being used in astronomical contexts as ML techniques have become more accessible (e.g. CHIERICHINI *et al.*, 2024; HE *et al.*, 2022; ROBINSON *et al.*, 2024).

The distinguishing attribute in `shap-hypetune` consists of offering the possibility to use SHAP-based importance for optimal feature selection while tuning parameters. The main advantage of SHAP over native tree-based importance measures (such as gain or split count) lies in its ability to provide a more stable and theoretically grounded assessment of feature contributions.

Particularly, *gain* measures the total improvement in the model's objective function brought by splits on a given feature, while *split* count simply tracks how often a feature is used to split nodes across all trees. Such traditional importance metrics are tied to the internal mechanics of greedy tree construction, which tends to favor continuous variables

---

[10] https://www.jstatsoft.org/article/view/v036i11

[11] https://shap.readthedocs.io/en/latest/

or categorical variables with many levels due to the larger number of possible split points they afford. This often leads to biased attributions of importance, particularly in datasets with mixed feature types, resulting in unreliable rankings and potential overfitting.

SHAP addresses this limitation by quantifying each feature's average effect on the model's predictions across many possible combinations of feature inclusion, regardless of whether or how often the feature was split on, potentially revealing relationships not captured by split frequency or gain alone. Although this study did not compare SHAP-based importances directly with classical metrics, such a comparison can be highly informative and will be considered in future work to further clarify how different attribution methods behave in this context.

It is, however, important to recognize the specific characteristics of the dataset considered here, which contains only continuous input features and continuous target variables. In such a setting, the classical bias associated with the greedy split selection process is significantly reduced, since all features are of similar type and scale, and thus equally subject to the mechanics of tree construction. Furthermore, potential overfitting due to misleading splits is actively mitigated through separate procedures such as regularization strategies, early stopping, and cross-validation (as addressed further ahead). As a result, the risk of distorted interpretability or overfitting due to biased feature selection is less pressing in this context.

What remains, however, is the possibility that the greedy nature of the split algorithm may underutilize certain informative features simply because they did not yield the highest immediate gain during training. While SHAP is unable to intervene in the training process or revise the model's structure, it provides a *post hoc* assessment of how much each feature truly influences model predictions (although still based on the fully trained model) rather than relying on how frequently or early it was used in the splitting process. It is in this sense that SHAP offers value: not by correcting bias at the source, but by offering a more consistent and model-agnostic criterion for identifying features that truly drive predictions.

Hyperparameter tuning from a Bayesian approach

Hyperparameters are configuration variables that are set before the training of an ML model and control the learning process. They are distinct from model parameters, which are learned during training. Common examples include the learning rate, number of layers in a neural network, and the number of trees in a random forest or gradient based decision trees (GBDT).

In order to allow for optimal performance, LightGBM (or more precisely the `lightgbm.LGBMRegressor` function) can take numerous (hyper)parameters, which in turn can be passed to `shap-hypetune` for optimization. This ease of use follows from the package being specifically designed for gradient boosting models and developed to integrate smoothly with the scikit-learn ecosystem. Some relevant and commonly used parameters include the following:

- `boosting_type`: specifies the boosting algorithm used;
- `num_leaves`: controls model complexity by limiting the number of leaves per tree;
- `max_depth`: limits the depth of each tree;
- `learning_rate`: determines the step size at each boosting iteration;
- `n_estimators`: sets number of boosting rounds, or trees to build;

among several others.

Finding the best hyperparameters for optimized performance is one of the most challenging aspects of machine learning. In the absence of empirical guidance or prior experimentation, tuning becomes largely uninformed and inefficient, particularly in high-dimensional spaces. In early machine learning systems, configurations related to model behavior and learning dynamics were often adjusted through manual trial and observation (e.g. SAMUEL, 1959). This empirical approach motivated the development of more systematic methods such as exhaustive search algorithms, Bayesian optimization, evolutionary strategies, and early-stopping techniques. While these modern techniques offer a more structured exploration of the search space, they still rely on user-defined ranges or scales, which mostly depend on intuition and/or previous experimentation. In that sense, empirical judgment continues to play a role, and a certain level of arbitrariness remains even in advanced optimization strategies. For the present work, one of the main advantages of `shap-hypetune` is its support for grid search, random search, and Bayesian search for hyperparameter tuning. While grid and random search explore predefined values or ranges in a fixed or stochastic manner, respectively, Bayesian search offers a more

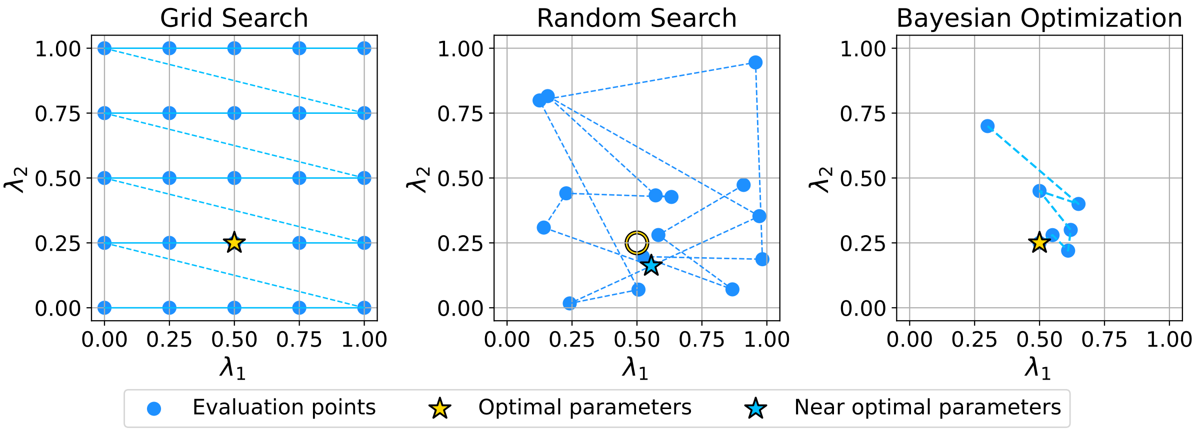

Figure 4.3: Illustrative comparison of search strategies for hyperparameter tuning. Grid search samples the parameter space exhaustively on a fixed grid, random search does so stochastically, and Bayesian optimization progressively concentrates evaluations in promising regions. This schematic visualization is not based on real model evaluations. Yellow markers (stars and circle) indicate optimal configurations, while the blue star represents a near-optimal solution.

guided strategy by relying on previous evaluations to inform subsequent steps. It builds a probabilistic model of the objective function to focus the search on promising regions of the hyperparameter space, typically requiring fewer evaluations to identify optimal configurations, as visually illustrated in Figure 4.3. This informed decision-making approach is generally less computationally expensive, and its exploration–exploitation[12] balance supports a more efficient and broader coverage of the search space. As a result, Bayesian search is also better suited for global optimization tasks aimed at identifying global rather than local optima, although success in doing so is not guaranteed.

Implementation with `shap-hypetune`

In Section 4.2, the discussion so far has focused on the principles behind the tool adopted in conjunction with LightGBM, namely `shap-hypetune`. This part now turns to the practical aspects of implementing these frameworks. All procedures described in this section were carried out using Python notebooks in a Jupyter-Lab interface (version 4.1.5) hosted on the NOIRLab's Astro Data Lab platform[13].

After all filtering and preprocessing steps described in Section 4.1, the final training sample comprises 88 421 carefully selected stars, described by 153 features in the form of color indices and absolute magnitudes. Recall that an important step in the process is splitting this set into training (70%), validation (20%), and test (10%) subsamples, which may lead to some ambiguity in naming. To avoid confusion, the complete set of 88 421 stars is referred to as the *full sample*, while the subsamples are referred to individually as simply the training, validation, and test samples.

The base model adopted for regression is a LightGBM estimator (`lightgbm.LGBMRegressor`, see footnote 8), configured to minimize the mean absolute error (`objective='mae'`) and executed with parallel processing (`n_jobs=3` as a compromise between computational efficiency and fair resource usage on the shared 64-core system). The boosting type was kept at its default setting, which is gradient boosting decision tree (`boosting_type='gbdt'`). Early stopping, a regularization technique that interrupts training if no improvements are observed on the validation set for a given number of rounds, was set to 5 (`early_stopping_round=5`). With all other parameters left in default values, this estimator was passed to the `BoostRFE` routine from the `shap-hypetune` package. This method integrates SHAP-based importance measures with hyperparameter tuning and applies recursive feature elimination (RFE), which recursively removes the least important features based on the estimator's internal ranking until a target number is reached.

---

[12] The exploration–exploitation dilemma (BERGER-TAL *et al.*, 2014) refers to the trade-off between choosing known options with predictable outcomes (exploitation) or trying new ones that may yield better results but carry risks (exploration).

[13] `https://datalab.noirlab.edu/`

For feature selection and hyperparameter optimization, `BoostRFE` operates iteratively by removing the least relevant features at each step based on SHAP values while evaluating candidate configurations through Bayesian search, using the package `hyperopt`[14]. A subset of LightGBM hyperparameters was selected for tuning based on initial exploratory tests, which informed both the choice of parameters and the definition of their search ranges as detailed in Table 4.4. These were chosen for their influence on model complexity and learning dynamics, while also balancing search space coverage and computational feasibility. Bagging-related parameters were fixed to empirically robust values (`bagging_fraction=0.8`, `bagging_freq=1`)[15], chosen through preliminary testing and aligned with common practice in gradient boosting. All other parameters were kept at their default values, as they showed limited impact on validation performance or introduced unnecessary tuning overhead.

Parameter	**Sampling distribution**	**Description**
`n_estimators`	$\mathcal{U}(30,\ 200)$	Number of boosting iterations
`learning_rate`	$\log\mathcal{U}(0.01,\ 0.2)$	Step size shrinkage at each iteration
`min_split_gain`	$\mathcal{U}(0,\ 15)$	Minimum loss reduction to perform a split
`min_child_samples`	$\mathcal{U}(5,\ 200)$	Minimum number of data points in a leaf

Table 4.4: Hyperparameter search space used in the `BoostRFE` routine.

As shown in Table 4.5, the optimal hyperparameter configurations and number of selected features vary across targets, reflecting model-specific adjustments made by the `BoostRFE` process.

`Parameter`	`n_estimators`	`learning_rate`	`min_split_gain`	`min_child_samples`
T_{eff}	181	0.1326	3.462	28
$\log g$	128	0.1659	1.325	83
[Fe/H]	173	0.1057	0.783	61

Table 4.5: Best hyperparameters found by the `BoostRFE` routine for each target.

RFECV and final features importances

Although `BoostRFE` provides both best hyperparameters and best features, the latter are further refined to identify the most informative ones, aiming to prevent overfitting, assuming irrelevant or redundant features can degrade performance, and potentially reduce computation time. To this end, a cross-validation strategy was employed alongside RFE to ensure robust feature selection, using the `RFECV`[16] algorithm from `scikit-learn`.

---

[14] `https://hyperopt.github.io/hyperopt/`

[15] `bagging_fraction` sets the proportion of data used in each boosting round, while `bagging_freq` controls how often bagging is performed.

[16] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html`

In this approach, the data is split into a user-defined number of subsets (typically up to $k = 10$), trained on $k-1$ folds, and validated on the remaining one. This process is repeated k times, with each fold used once as the validation set. RFECV integrates RFE with cross-validation by evaluating model performance at each elimination step, systematically determining the subset of features that yields the best average score. This results in the selection of the optimal number of features, which contributes to improved model accuracy and reduced overfitting, thereby enhancing robustness and ensuring that the selected features better generalize to unseen data. In the present work, RFECV was implemented through an iterative 5-fold cross-validation process.

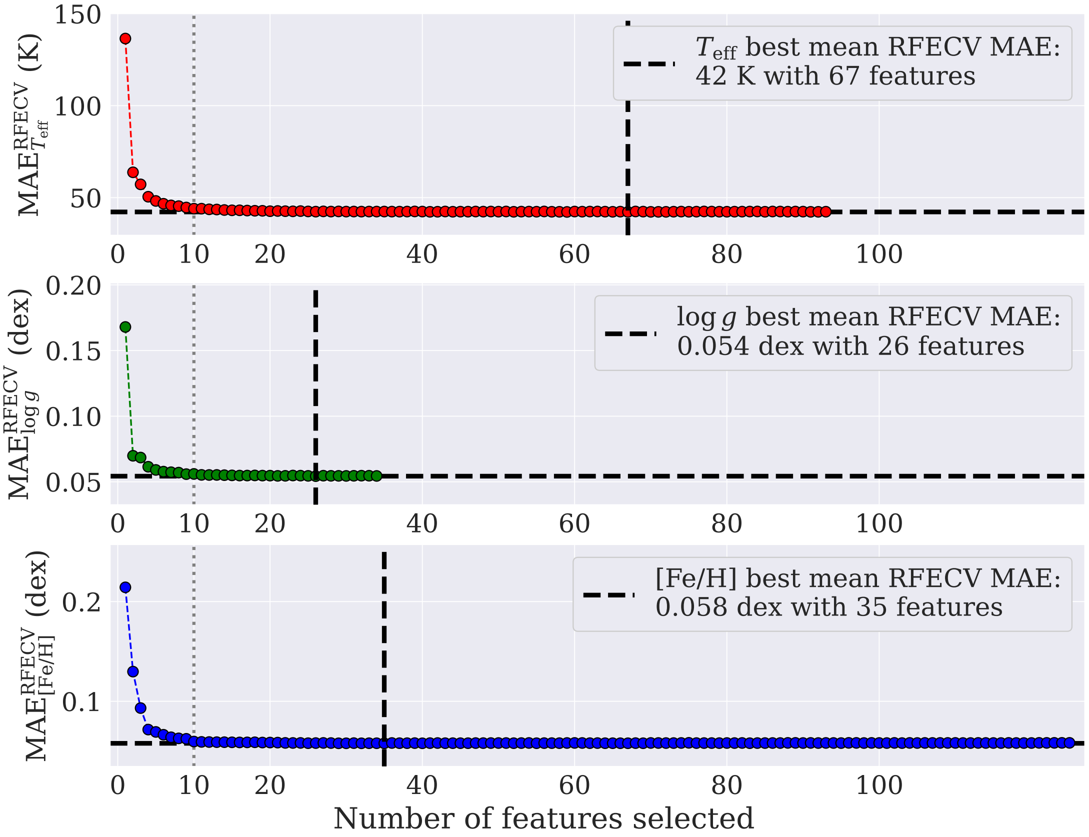

Figure 4.4: Mean RFECV MAE values as a function of the selected number of features. The number of points in each panel (93, 34, 125 features for T_{eff}, log g, and [Fe/H], respectively) are the results from the `BoostRFE` routine, while the black dashed lines in each panel mark the best mean RFECV MAE and the associated number of features (67, 26, 35 features for T_{eff}, log g, and [Fe/H], respectively).

The results of this step for T_{eff}, log g, [Fe/H], are shown in Figure 4.4, where the y-axis is the MAE and the number of features can be seen in the x-axis. The number of points in each panel (93, 34, 125 features for T_{eff}, log g, and [Fe/H], respectively) represents the results from the `BoostRFE` approach previously implemented by `shap-hypetune`, before the RFECV step. One can easily notice that, although the optimal number of features varies for each stellar parameter as indicated (black dashed lines mark 67, 26, 35 features and indicate MAE values of 42 K, 0.054 dex and 0.058 dex for T_{eff}, log g, and [Fe/H], respectively), in all cases a visual plateau is reached above around 10 features (dotted, vertical, gray lines). The horizontal black dashed line is plotted to mark the best mean RFECV MAE value in each panel (which is barely noticeable), but visually demonstrates this behavior.

Importantly, some standards regarding the names of the features shown in the following figures are: J-PLUS filters are the ones with J as first letter followed by four characters representing either the central wavelength of filter in nanometers (for the narrow bands, e.g. J0430) or an indication of which broad band (uJAVA or grizSDSS); CatWISE magnitudes are W1mproPM and W2mproPM, while Gaia magnitudes are BPmag, RPmag and Gmag; the majority of them are simple magnitudes subtractions; all magnitudes exhibit the subscript `_0` to indicate extinction correction; those with the prefix `M_` represent absolute magnitudes, which were calculated using photogeometric distances from BAILER-JONES *et al.* (2021), indicated in the feature name by `rpgeo`.

Figure 4.5 shows final feature importance values for the ten most important features for each parameter's model. The leading features exhibit physically meaningful quantities, lending confidence to the model's internal consistency. For $T_{\rm eff}$, the dominance of the (J0430 − W1) and (J0430 − J0861) colors, for instance, reflects the strong sensitivity of these filters to the shape of the stellar continuum for the spectral types of interest to this work, which is largely driven by temperature. In the case of $\log g$, the two most relevant features are the absolute magnitudes $\rm M_{W1}$ and $\rm M_{W2}$, as expected to correlate with surface gravity due to their dependence on stellar luminosity and radius. The third feature, (J0430 − J0515), involves the intermediate-band filter `J0515`, which partially covers the Mg triplet region around 5150 Å, exhibiting some sensitivity to surface gravity due to the pressure-broadened absorption lines in this spectral feature. This makes it particularly useful for distinguishing between dwarfs and giants in stellar populations. For [Fe/H], the dominance of features such as (uJAVA − J0430) and (J0378 − J0430) highlights the role of the blue part of the spectrum, where many metal absorption lines are concentrated. Additionally, colors combining J0395 (encompassing metallicity sensitive Ca II H and K lines), J0660 (centered on Hα) and J0861 (covering the Ca II triplet) often appear in the top 10 and further support the model's ability to capture metallicity effects, given the known sensitivity of these features to metal content.

Final feature importances are ranked in Figure 4.6, where relative (i.e. divided by their sum) feature importances from SHAP values are shown on the vertical axis and the ranking of feature positions on the horizontal axis. The first 15 most important features are indicated on each respective panel, with the first 10 highlighted in the color of the respective data points (red for $T_{\rm eff}$, green for $\log g$, and blue for [Fe/H]; same as those exhibited in Figure 4.5).

More importantly, Figure 4.6 shows the information on individual importances (translucent curves and crosses) and how their relative values accumulate (solid). The

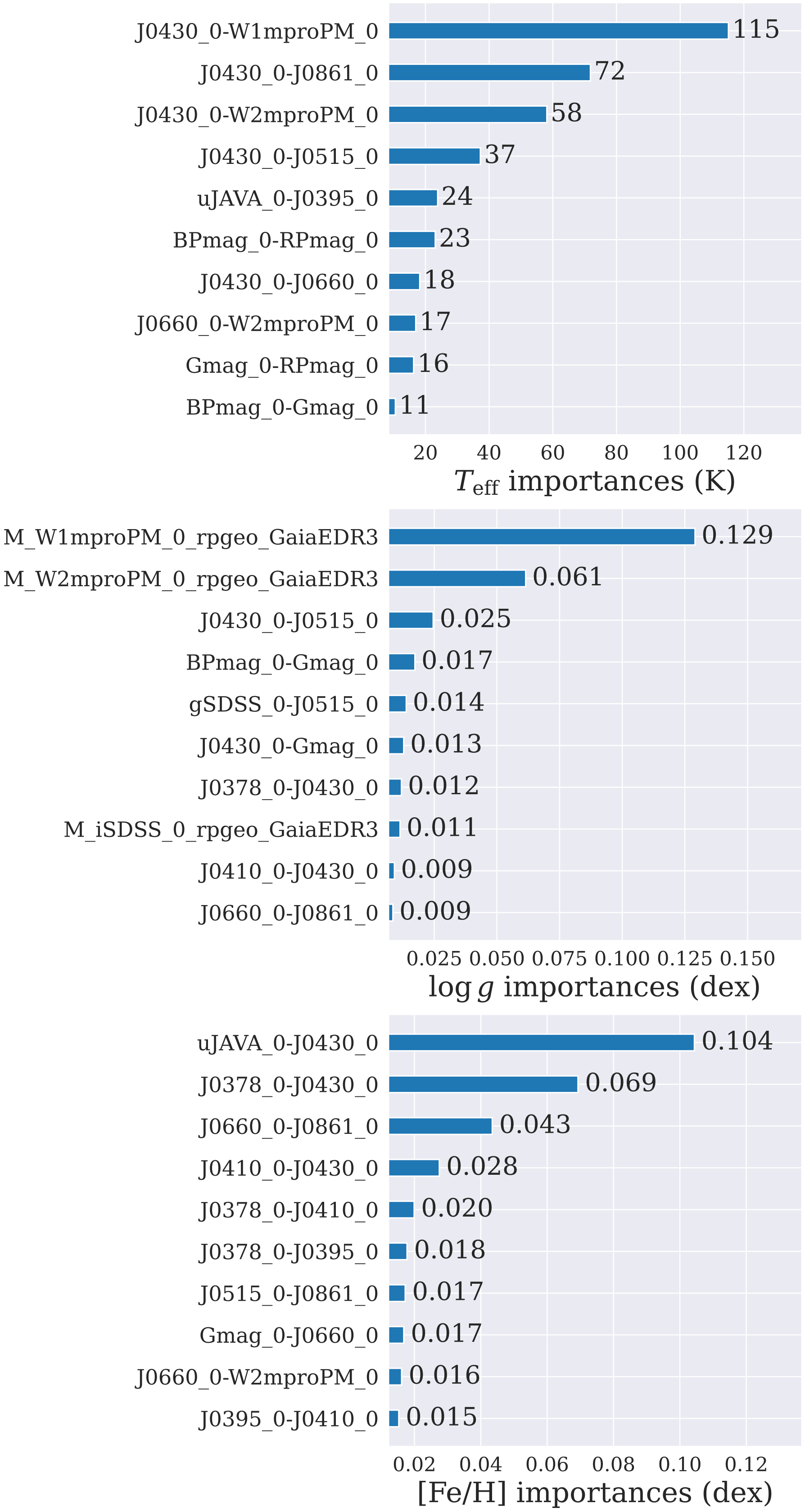

Figure 4.5: Bar plot of the top 10 most important features, with their corresponding importance values labeled.

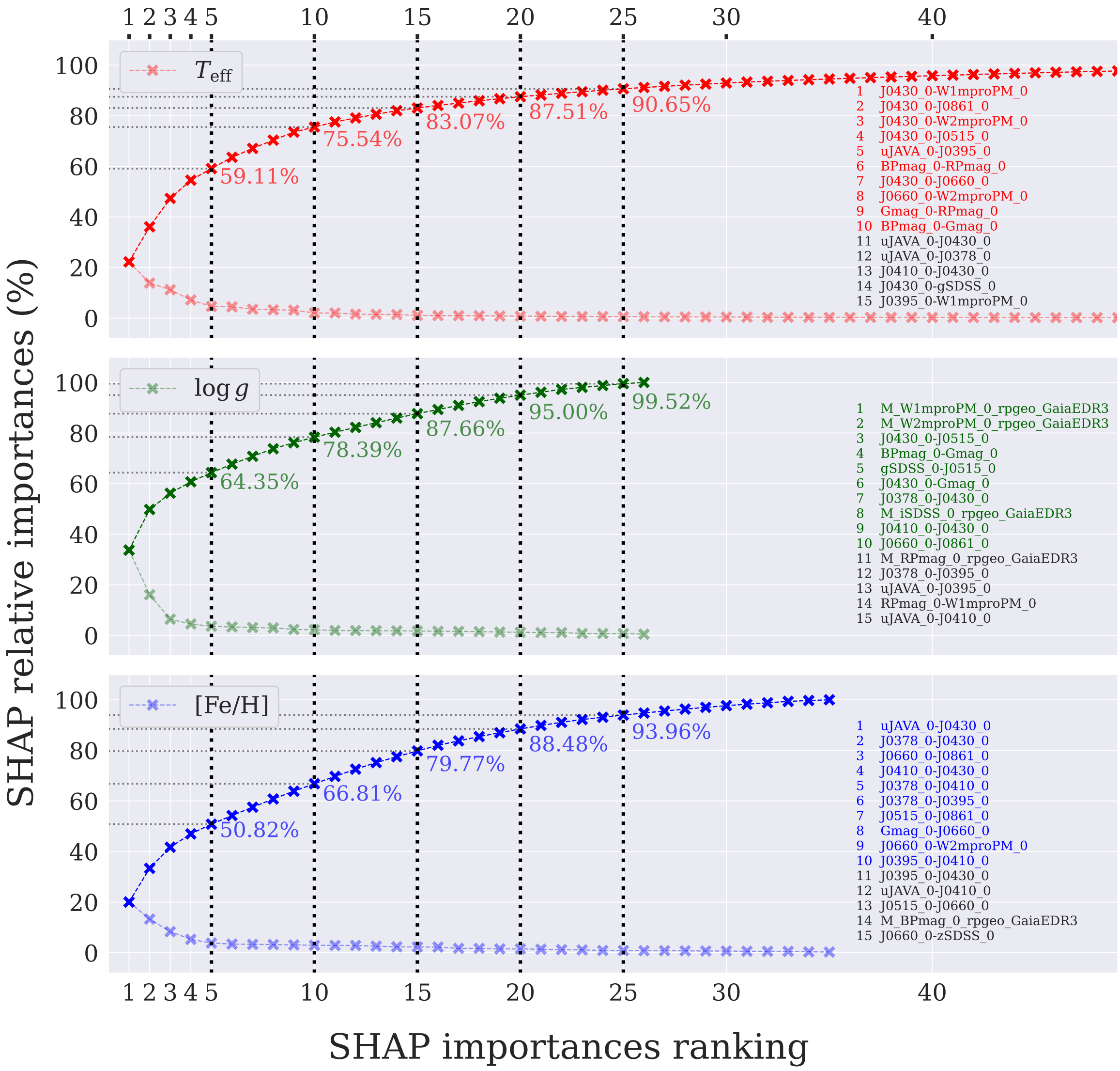

Figure 4.6: Relative SHAP importance as a function of feature ranking. Translucent crosses represent the ratio of each individual importance over the sum of all features' importances, while opaque crosses represent their cumulative sum, periodically indicated with dotted black lines for their values (x-axis) and their respective number of features (y-axis). The first 15 most important features are also shown, with the first 10 highlighted in the color of the respective data points (red for T_{eff}, green for $\log g$, and blue for [Fe/H]).

percentages represent the models' variability, i.e., to what extent the variations in the target predictions ($T_{\rm eff}$, log g, and [Fe/H]) can be assigned to variations in the features (colors and magnitudes). Hence, for instance, the top five most important features account for approximately 59.11%, 64.35%, and 50.82% of the variance in the predicted $T_{\rm eff}$, log g, and [Fe/H], respectively. These values increase to 75.54%, 78.39%, and 66.81% when considering the top ten features.

It is important to mention that, in this work, it was chosen to proceed with all selected best features, instead of selecting only a subset, given the following reasoning. If one chooses to use, for instance, only each parameter's three most important features, then only eight magnitudes (uJAVA, J0378, J0430, J0515, J0660, J0861, W1, and W2) would be necessary for predicting them, instead of all 17. However, this approach comes at the cost of constructing a model with much poorer performance, as possibly anticipated from Figure 4.6: barely half of the variations in the metallicity predictions, for instance, would be accounted for from the variations in the input photometry, hindering models' transparency.

In comparison, using the five or ten most important features improves the explaining power in [Fe/H] in terms of cumulative importances (respectively 50.82% and 66.81%). Nevertheless, at that point, the number of needed filters increases to at least 12. This mainly undermines the purpose of relying on fewer features, since the only absent ones would be Gaia RP and SDSS *gri*, and it is reasonable to expect that most stars should have available photometry for those. Hence, it is considered unjustified to let go of any of these adopted filters in this context and move on with the full set of best features.

Nevertheless, it should be recognized that the selection illustrated in Figure 4.4 carries a degree of stochasticity. The plateau observed from approximately 10 features (marked by dotted black lines) onward indicates that the differences in MAE among a wide range of feature counts are marginal, often at the percent level or below. In this context, the actual selected numbers of features (67, 26, and 35) reflect local optima, but not necessarily unique or sharply defined thresholds of performance.

This introduces a subtle but important inconsistency. If, by chance, the selection had landed on a larger number (for instance, 40 instead of 35 for the [Fe/H] model), the original choice (35) would then fall into the category of using "less than ideal", despite showing similar performance. Thus, the flatness of the curve undermines any rigid interpretation of the selected value as strictly optimal. At the same time, the analysis in Figure 4.6 discourages using fewer features than selected, citing the risk of degraded performance, especially if outside the plateau region. This tension highlights a conceptual limitation in over-interpreting the selected count as a hard boundary.

One might argue, by extrapolation, that using all 153 original features would be the safest choice, avoiding the ambiguity altogether. However, this line of reasoning is also misleading. Adding more features beyond the plateau contributes little to predictive performance, while potentially increasing model complexity, computation time, and susceptibility to overfitting. From a regularization standpoint, a more parsimonious model is often preferable when the gains are saturated, reflecting optimization to merely marginal improvements.

In view of these considerations, the approach adopted here follows a pragmatic path. The number of features selected by RFECV is used as a reference, not because it marks a uniquely ideal configuration, but because it offers a reproducible, data-driven point within a region of near-optimal solutions. This choice is made with full awareness of the surrounding ambiguity, yet avoids the pitfalls of both under- and over-selection.

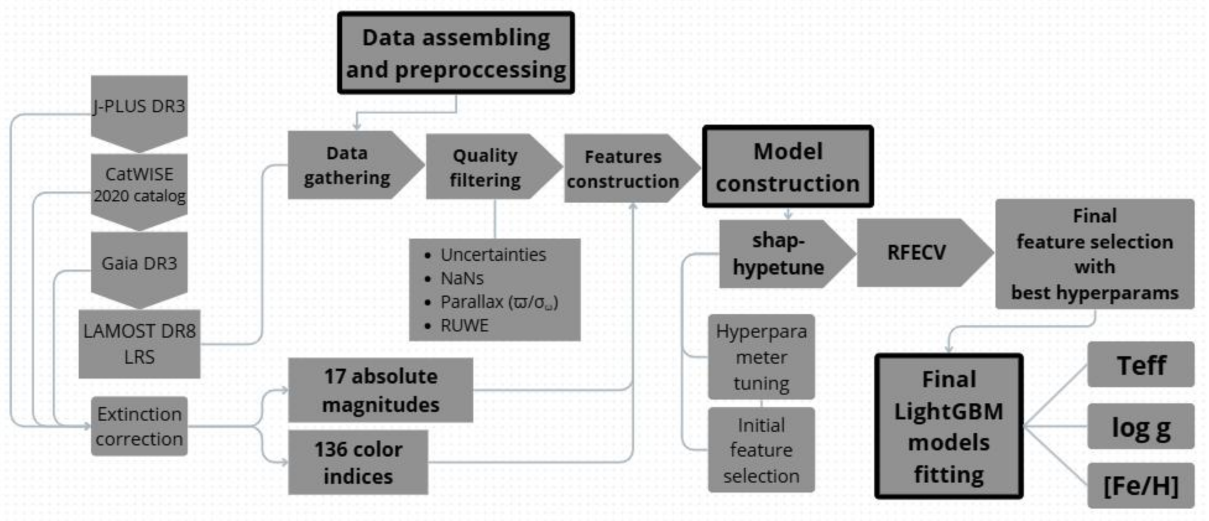

Figure 4.7: Overview flow chart depicting the processes for constructing models for stellar atmospheric parameter estimation using multiband photometry.

Finally, an overview of the steps taken to construct the models and predict the atmospheric parameters is exhibited in Figure 4.7. The process starts with data gathering from J-PLUS and cross-matched catalogs (CatWISE, Gaia, and LAMOST), followed by quality filtering and construction of extinction-corrected features (Section 4.1). Model construction included hyperparameter tuning and feature selection via `shap-hypetune` and RFECV (this section), with final LightGBM models fitted to predict T_{eff}, $\log g$, and [Fe/H]. Next, the results and performances of the final fitted models are presented and discussed.

4.2.2 Results and models performances

The results for the final LightGBM models applied to the test set (8842 stars, 10% of the full training sample) can be seen in Figure 4.8. In each panel, the horizontal axis shows the original values from LAMOST, while the vertical axis shows residuals (this work minus LAMOST), with MAE values exhibited and mean (μ) and standard deviation (σ) values indicated; marginal histograms represent the respective distributions. At first, outliers are easily spotted and moderate to weak trends can be identified, especially in $\log g$ space. On the other hand, MAE values (42 K, 0.055 dex, and 0.058 dex for T_{eff}, $\log g$, and [Fe/H], respectively) indicate a good predictive power, and relatively little scatter can be seen. Moreover, T_{eff} predictions are noticeably better constrained than the other two parameters.

The residual distributions in Figure 4.8 are approximately symmetric and centered around zero, with no strong indication of systematic bias across the parameter ranges. Slight increases in scatter appear toward the edges of the parameter space, where mild positive biases can be observed, particularly for metal-poor stars and low-gravity giants, reflecting limited coverage of those regimes in the training sample. The overall scatter is of the same order as the internal uncertainties derived from the LAMOST DR8 sample used earlier in this work, namely 54 K, 0.083 dex, and 0.054 dex for T_{eff}, $\log g$, and [Fe/H], respectively. The predicted MAE values lie within this range for T_{eff} and [Fe/H], and remain reasonably close for $\log g$. Moreover, the use of regularized gradient boosting and the observed generalization performance on the test set suggest that any risk of overfitting is well controlled.

Additionally, to further evaluate the impact of reducing the feature set, Figure 4.9 illustrates how models with different numbers of selected features perform in terms of MAE values. For each parameter, the MAE values for a reduced set of five features, for instance, were: $\text{MAE}_{T_{\text{eff}}} = 54$ K (compared to 43 K in the full model), $\text{MAE}_{\log g} = 0.067$ dex (compared to 0.055 dex), and $\text{MAE}_{[\text{Fe/H}]} = 0.079$ dex (compared to 0.058 dex). These correspond to percentage increases in error of 26%, 23%, and 37% for T_{eff}, $\log g$, and [Fe/H], respectively, affecting metallicities more aggressively.

Furthermore, implementing this shortened feature set still requires 13 of the 17 magnitudes from the full set[17], with the remaining four[18] being from broad-band filters that are unlikely to impose significant constraints on the sample.

Although the MAE curves flatten after a certain number of features, keeping the full set helps maintain robustness and avoid localized performance degradation. Taken together, these results support the decision to retain all features in the final models to ensure reliable predictions across all parameters.

In summary, the results presented in this chapter demonstrate that the constructed LightGBM models are well-suited to provide accurate and robust estimates of stellar atmospheric parameters. The residual distributions indicate good agreement with the LAMOST reference values and suggest that the models generalize well across the parameter space. Feature selection experiments further confirmed that although using fewer features than those retained in the final selection can yield similar MAE values, preserving the

---

[17] Namely, from J-PLUS: uJAVA, J0378, J0395, J0410, J0430, J0515, J0660, J0861, gSDSS; from *Gaia*: BPmag, Gmag; from CatWISE: W1mproPM and W2mproPM.

[18] These are RPmag (Gaia), and rSDSS, iSDSS, zSDSS (Sloan), all of which are broad-band filters.

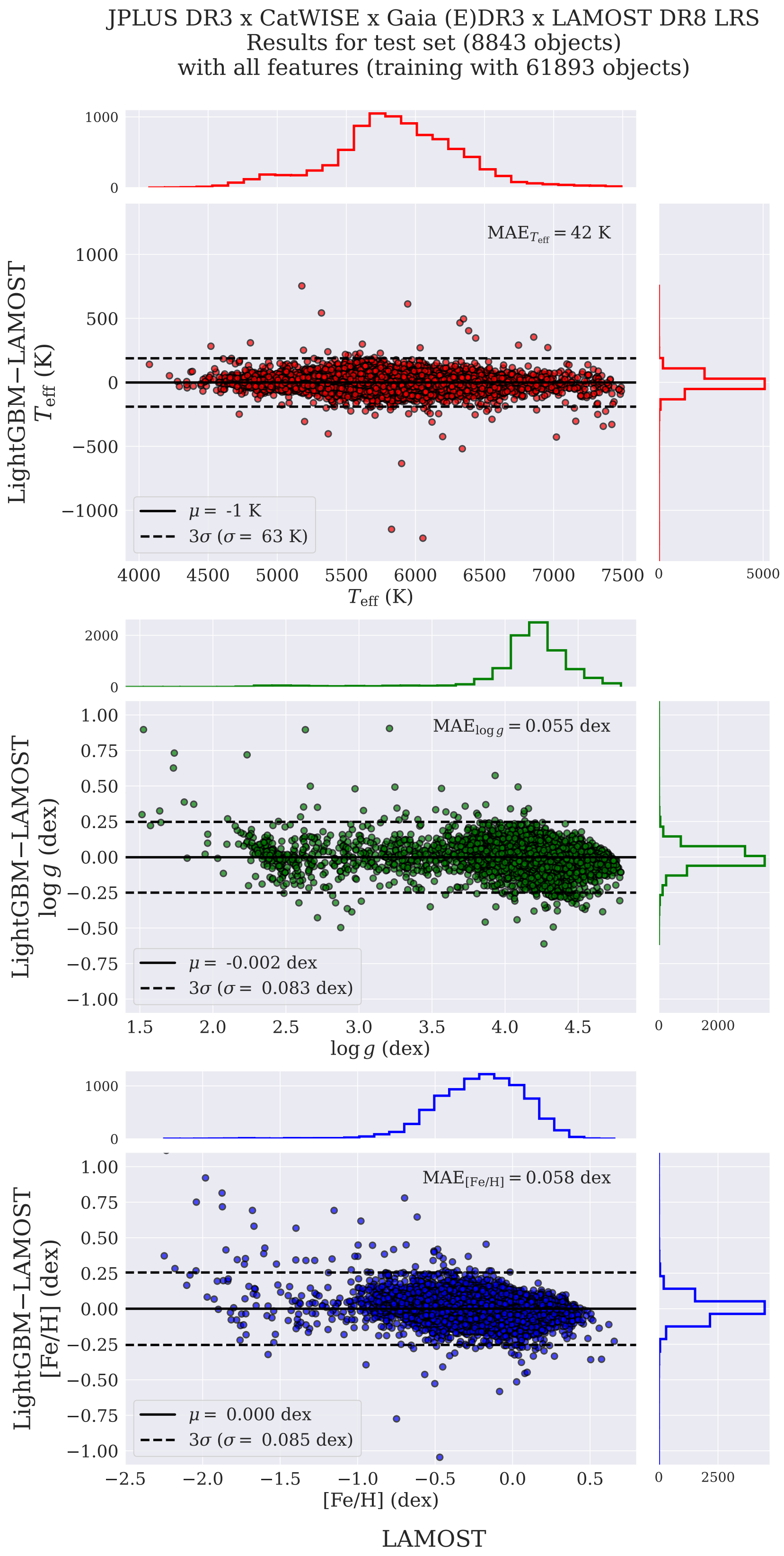

Figure 4.8: Residuals between predicted LightGBM parameters and original LAMOST values in the test set. MAE are also exhibited along with mean (solid black lines) and three standard deviations (dashed black lines) values. Marginal histograms show the distribution of values along each axis.

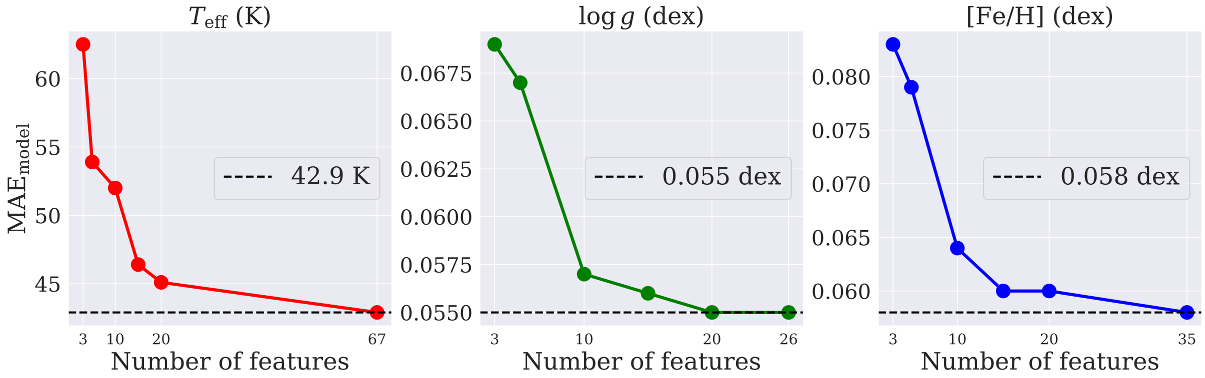

Figure 4.9: MAE values as a function of the number of features used in model fitting. The black dashed lines in each panel mark the best MAE achieved with the full feature set for each parameter. While similar in appearance, this figure differs from Figure 4.4, which refers to the RFECV step used to rank feature importance. In contrast, the present figure shows model performance when refitted using varying-sized subsets of the ranked features.

full selected set ensures maximal precision and stability by avoiding the loss of predictive information. It is important to emphasize that the strategy adopted here was to intentionally pursue the most efficient models possible, so that the results could be interpreted as representative of the current upper limit of what can be achieved using photometric and astrometric data alone within the scope of gradient boosting methods. The next chapter explores such an application, where the trained models are used to estimate metallicities of open clusters, assessing their reliability and potential for astrophysical inference.

Chapter 5

Science case: open clusters

The application of the models constructed here to real data is addressed in this chapter, along with some care taken in handling uncertainties, which were not directly digested into the model. Section 5.1 presents a brief contextualization for open clusters (OCs), which are the objects explored in the present work for science case application of these models, while the sample used here is presented in Section 5.1.1. Next, Section 5.2 discusses the metallicities of OC members within the J-PLUS footprint and explains how their uncertainties were estimated. The results are compared with spectroscopic values by directly drawing from the LAMOST DR8 catalog, while the metallicities derived in this work for OCs are contrasted with literature values.

5.1 Open clusters

Contextualization

Open clusters (OCs) are loosely bound groups of stars, predominantly young (typically less than 1 Gyr; FUJII & ZWART, 2016; ZHANG *et al.*, 2024) and numbering from a few tens to a few thousand, that formed from the same giant molecular cloud. Since their members roughly born together, these objects act as practical laboratories equipped with diverse examples, especially regarding a wide range stellar masses, constrained to roughly same age, distance and overall chemical composition.

As such, OCs are essential pieces to astrophysics for a number of reasons. Stellar ages are fundamentally hinged on clusters, which serve as benchmarks for calibrations of virtually every stellar age (see for instance SODERBLOM, 2010; ZHANG *et al.*, 2024, and references therein). They play a fundamentally informative role spanning from young and very young (on star formation mechanisms and regions) to intermediate and old (chemical enrichment of the galactic disc) cases. As a natural consequence, the stellar structure and evolution (probed via Hertzsprung-Russell diagrams and their variations, CMDs, and Kiel diagrams) of OCs provide observed, empirical isochrones for test beds and comparisons with theoretical models. In addition, the internal kinematics and potential mass segregations in OCs may offer additional insights on galactic structure by means of the evaporation of its stars into the field stellar population, highlighting the importance of studying cluster dynamics and evolution. Furthermore, resolving individual stars from galactic clusters is particularly beneficial as their fundamental parameters are accessible more reliably, e.g., age, distance, metallicity, and composition.

The naming coincidence hints on the fact that clustering is one of the most widely used ML techniques in the scope of OCs, especially when the goal is looking for new objects (HUNT & REFFERT, 2021). The very nature of clustering methods lies in seeking entities grouped in any given space (depending on the complexity of the task) by assessing their similarities, traditionally through unsupervised approaches. Several works in the literature have identified thousands of new OCs (e.g. BICA *et al.*, 2019; CANTAT-GAUDIN *et al.*, 2020; DIAS *et al.*, 2021; HUNT & REFFERT, 2021; HUNT & REFFERT, 2023; KHARCHENKO *et al.*, 2013) in recent years, as can be seen in Figure 5.1, largely due to the high-precision astrometry provided by Gaia and/or AI efforts (e.g. CANTAT-GAUDIN *et al.*, 2020; CASTRO-GINARD *et al.*, 2018; HUNT & REFFERT, 2021; PERREN *et al.*, 2023). Even though the currently known sample is far from complete (following e.g. BONATTO *et al.*, 2006, an upper limit is estimated to

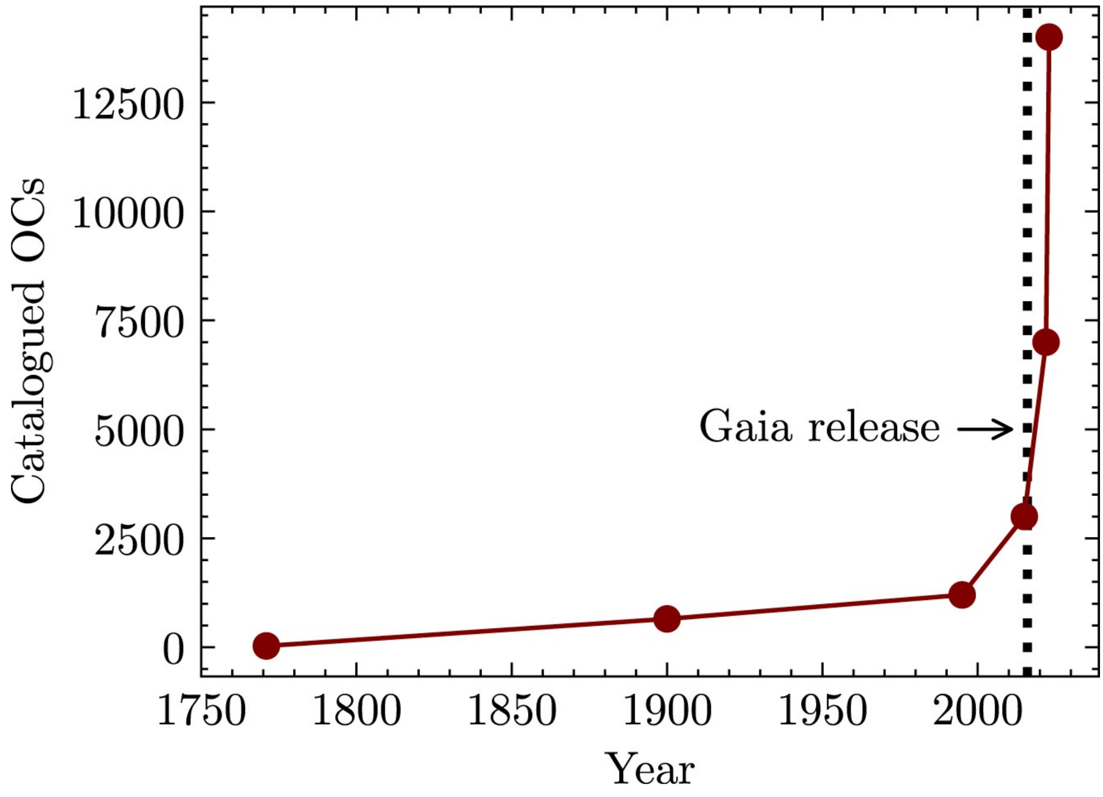

Figure 5.1: Growing number of catalogued OCs in the literature from the late 1700s to the present day, with the release date of Gaia's survey data indicated by the dotted line. Figure 1 from PERREN *et al.* (2023).

be over 10^5 OCs), these advances have drastically expanded the census of galactic OCs, bringing the number of OCs candidates to $\sim 14\,000$ (PERREN *et al.*, 2023) and confirmed ones to ~ 5600 (HUNT & REFFERT, 2024).

In the present work, the specific interest in OCs comes from the scarcity of studies on this topic using data from J-PLUS. Although a vast body of literature has emerged over the last decade using S-PLUS, J-PLUS, and (mini)J-PAS, OC studies within these surveys remain deficient. Multi-band approaches can be distinctly valuable for OC research, especially since narrow-band filters capture distinct astrophysical features, enhancing the characterization of both cluster properties and individual members in various aspects (e.g. BICA *et al.*, 2019; CAVALLO *et al.*, 2023; MAYO *et al.*, 2023; NARDIELLO *et al.*, 2023; SÁNCHEZ *et al.*, 2021). In particular, the work of SÁNCHEZ *et al.* (2021) also employs the same 12-band optical system at the JAST80 telescope for OC studies, independently from J-PLUS. By combining the multi-band approach with Gaia broad bands, their results highlight the benefits of relying on J-PLUS filters for cluster characterization, especially in SED fitting for extinction estimation.

Importantly, cluster members are virtually at the same distance, which allows individual stellar traits to be examined within a common environment. This becomes particularly evident and useful in visualizations such as the Hertzsprung–Russell (HR) diagram (HERTZSPRUNG, 1911; ROSENBERG, 1910; RUSSELL, 1914) and similar representations. Such spaces are primarily based on intrinsic stellar properties, originally displaying the relationship between absolute magnitude and spectral classification. These quantities delineate key fundamental aspects such as surface brightness, radius, and the hydrogen and helium content of the stellar core, leading to a distinct separation of the various stages of stellar evolution.

For illustration, Figure 5.2 shows several stellar groups, from various globular clusters, identifiable in an HR diagram, namely the main sequence, various giant branches, and the loci of white and red dwarfs. As generally old objects, globular clusters provide better visualization of the distinct evolutionary phases of stars, in contrast to typically young OCs.

Although less populated than globular clusters, HR diagrams of OCs may also exhibit these groups. Figure 5.3 shows an example for the young cluster Hyades (e.g. HUNT & REFFERT, 2023), with an age of 576^{+413}_{-223} Myr, displaying some of the stages highlighted in Figure 5.2, namely the main sequence, turnoff point, red giant branch, and white dwarf sequence.

The aforementioned other similar representations are built upon quantities related to those of HR diagrams, including luminosity, effective temperature, surface gravity, and combinations of magnitudes (color indices). The so-called Kiel diagram is a modern conception of the HR diagram, proposed following the advent of stellar atmosphere model calculations and enabled by the quantitative determination of stellar surface gravities

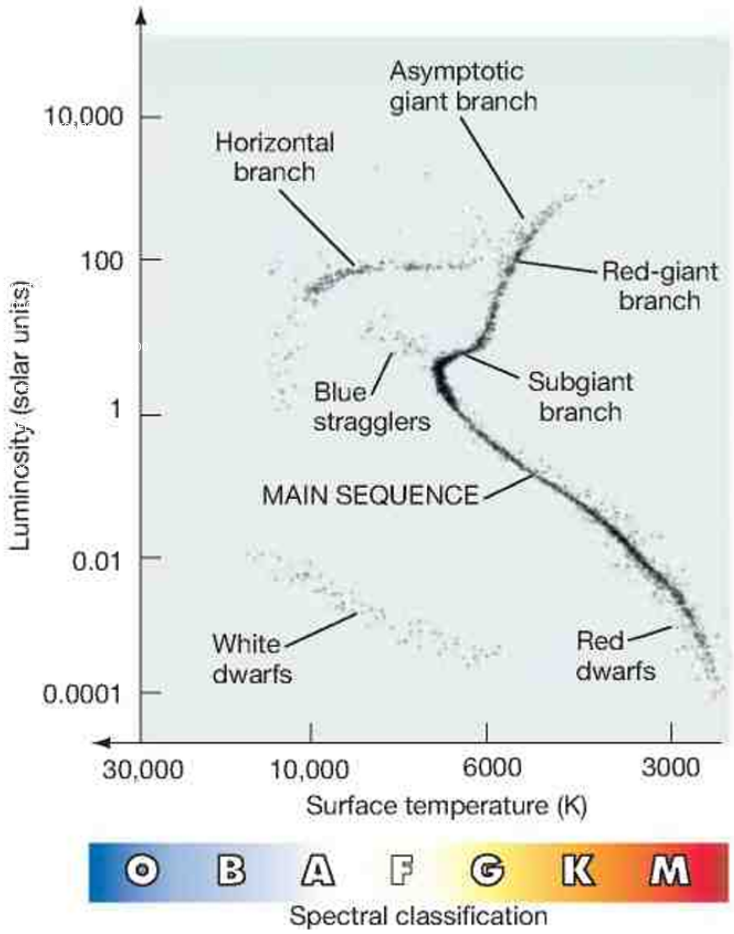

Figure 5.2: Combined HR diagram for several globular clusters with compositions similar to that if the M80 cluster, based on ground- and space-based observations. Major evolutionary stages and the population of blue stragglers are visible. The color bar indicates the spectral classification from hot O-type to cool M-type stars. Adapted from Figure 20.14 of CHAISSON & MCMILLAN (2005).

(see, e.g., LANGER & KUDRITZKI, 2014, and references therein). For stars of a given surface temperature, larger values of log g can imply a larger mass (GREENSTEIN & SARGENT, 1974; NEWELL, 1973), thereby serving as a proxy for stellar luminosity, with the advantage of not requiring distance measurements.

Perhaps the most commonly used space on this topic is the color–magnitude diagram (CMD), also known as the observational HR diagram, which, as the name suggests, is built with magnitudes and color indices. It arguably dominates in observational contexts due to ease of photometric data collection and relevance to stellar evolution and population studies. This dominance stems from the fact that color and magnitude serve as observational proxies for temperature and luminosity, respectively. Color indices reflect the spectral energy distribution and thus correlate with effective temperature, while absolute magnitudes (derived from apparent magnitudes and distances) approximate stellar

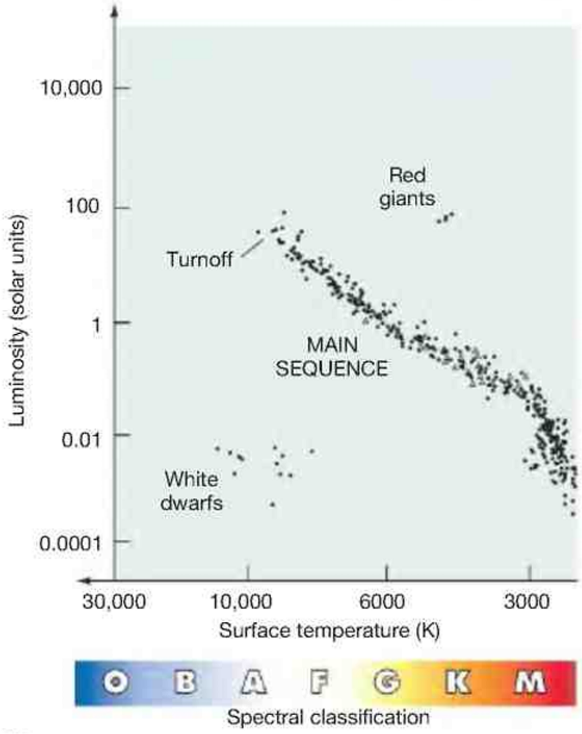

Figure 5.3: Hertzsprung–Russell diagram of the Hyades. The turnoff at about A spectral type implies an age of about 600 Myr, old enough for some massive stars to have become white dwarfs. The color bar indicates the spectral classification from hot O-type to cool M-type stars. Adapted from Figure 20.19 of CHAISSON & MCMILLAN (2005).

luminosity. As a result, CMDs closely resemble theoretical HR diagrams and effectively trace stellar evolution. Furthermore, color-color diagrams (constructed only with colors, instead of magnitudes) are typically employed for photometric calibration and outlier detection, but they can also yield valuable insights into open cluster properties.

In this context, in addition to providing the major grounds for estimating parameters in this work, J-PLUS photometry also carries notable value, especially in the scope of OCs studies. Its 12-filter system offers a powerful assembly of photometric data that can be used in various ways, which naturally includes constructing HR diagrams (or similar) for stellar populations. With all the above in mind, OCs are chosen as targets for application of the models constructed here, by predicting atmospheric parameters of individual members of clusters within the J-PLUS footprint. The main goal is to estimate metallicities for these objects, as further discussed after presenting the motivation and construction steps of the sample explored in next section.

5.1.1 Open clusters sample

The open cluster sample is based mainly on the work of HUNT & REFFERT (2023, hereafter HR23), who conducted the largest search for open clusters to date by employing the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDB-

SCAN) algorithm to the Gaia DR3 database. HR23 recovered 7167 clusters, 2387 of which were potential discoveries. However, the most reliable sample recommended by the authors includes 4105 clusters, 739 of which are new. Additionally, their catalog provides valuable information such as membership probability, cluster age, extinction, and distance.

During the development of this work, HUNT & REFFERT (2024) updated their previous OC census from HR23, reclassifying some of the systems selected here as moving groups. It was chosen to proceed with the original design based on HR23, with these objects serving as a representative astrophysical application of the method, since the primary focus is on presenting and discussing the developed tool. In this context, note that references to OCs throughout the text may include systems that were later reclassified as moving groups.

To identify which clusters fell within the J-PLUS DR3 footprint, a sample of J-PLUS stars was selected through a more relaxed query compared to that presented in Section 4.1[1] (Appendix B), aiming to retain the largest possible number of sources from J-PLUS. In practice, that one more restricted selection corresponds to a subsample of the broader one used now, which resulted in 8 633 441 stars. These were cross-matched with the most reliable cluster sample of HR23, composed by 529 004 probable members (with membership probability $P \geqslant 0.5$, a threshold chosen here also to favor a higher number of matches). This procedure returned 238 stars associated with 29 clusters, shown in Figure 5.4 as diamonds with black outlines plotted over the J-PLUS DR3 footprint, represented by gray regions. All markers are color-coded according to differential extinction as reported in HR23.

Such a low number could be somewhat anticipated considering the fact that stars (and pointings, in general) in J-PLUS are mainly in the halo, which reflects its extragalactic focus and is, in fact, in line with the previously mentioned scarcity of studies. Since the overall population of OCs is observed along the disk, few occurrences of clusters and members should be expected to be recovered: those 238 stars correspond to less than 0.05% of the probable members in the sample from HR23.

Next, that list was cross-matched with CatWISE and Gaia data, taking care to keep the same stars by using their Gaia DR3 source id, which resulted in 191 members. Then, those with obviously poor photometry (e.g., magnitudes and/or uncertainties set as 99) are filtered out, keeping only those with magnitude uncertainties smaller than 0.5 mag. This process resulted in 157 stars distributed in 23 systems: six of those have at least 10 members, and only two have at least 20 members (HSC 749 and NGC 1039, with 31 and 33, respectively). Additionally, 22 stars in the sample lacked Gaia extinction information. For these cases, the median extinction values in the G, BP, and RP bands were adopted, along

---

[1] `sglc_prob_star` > 0.5, now with no quality cuts imposed (i.e., no cuts on `mask_flags` or `norm_wmap_val`)

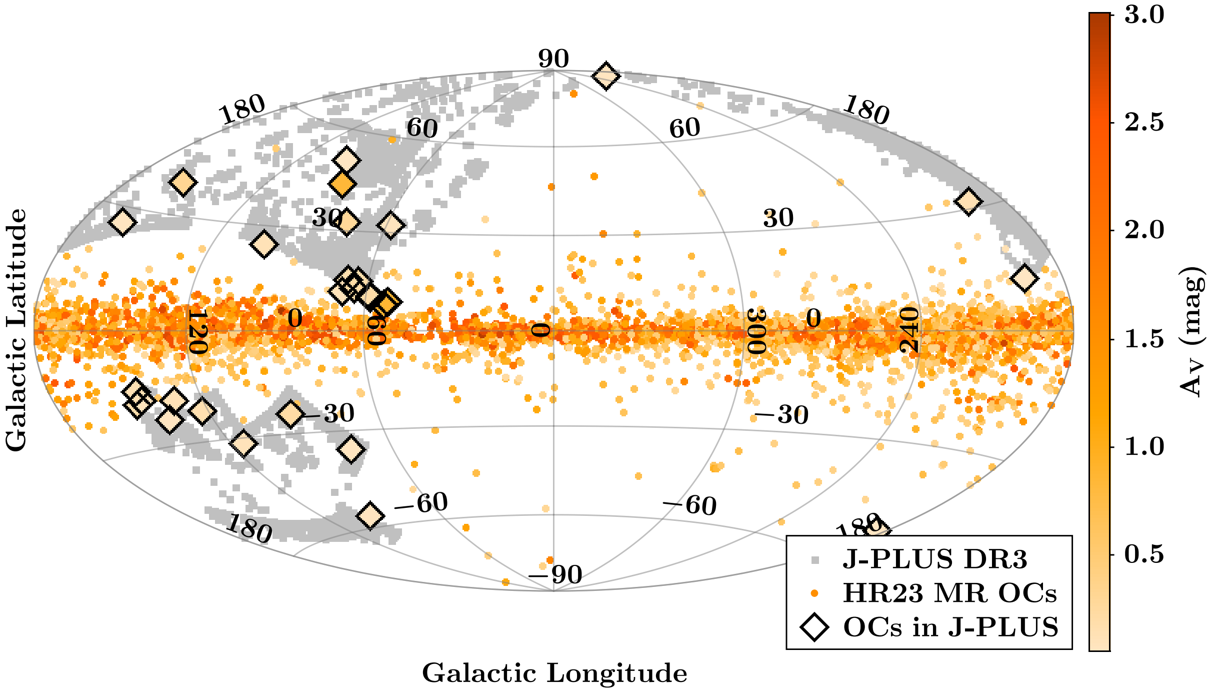

Figure 5.4: Sky Aitoff projection for the J-PLUS DR3 footprint (gray regions), the most reliable sample of OCs from HR23 (extinction-colored points) and the ones found in the J-PLUS DR3 footprint (extinction-colored diamonds with black outlines).

with their corresponding uncertainties. In light of the limited sample size, it was chosen to retain these stars in the analysis, especially since the extinctions involved are below 0.3 mag for all bands (specifically, maximum values of $A_G = 0.229$ mag, $A_{BP} = 0.298$ mag, and $A_{RP} = 0.176$ mag, as computed from the remaining 157 stars selected in the previous step) and thus are not expected to significantly impact the results.

Further filtering excluded stars with distances significantly deviating from the median value of their respective associated clusters, as compared to individual estimates from BAILER-JONES *et al.* (2021). Specifically, for each cluster, outliers were removed by comparing each star's distance to that cluster's median distance, which was derived from those individual estimates. This step was performed using the sigma clipping process provided by the SIGMA_CLIP function from the ASTROPY[2] Python library, with the median distance as the central value and a threshold of 3σ, with σ representing the robust median absolute deviation[3] (rMAD; see Appendix C) from the median. This selection process reduced the sample from 157 to 154 stars, with the three stars excluded belonging to the associations Alessi 62, HSC 759, and Theia 1188, and the total number of 23 systems remaining unchanged. Moreover, the median distances of the clusters, computed from individual stars using the values from BAILER-JONES *et al.* (2021), are generally consistent with those reported by HR23; differences exceed 5% only for cluster HSC 695.

---

[2] https://docs.astropy.org/en/stable/index.html

[3] https://docs.astropy.org/en/stable/api/astropy.stats.mad_std.html

The next section focuses on applying the [Fe/H] model to open cluster members and their parent clusters, presenting the resulting metallicity estimates, associated uncertainties, and comparisons with external references.

5.2 Application of the [Fe/H] model

This section applies the [Fe/H] model developed in this work to open cluster members and their parent clusters. Section 5.2.1 presents the procedure for estimating metallicities and associated uncertainties for individual stars, followed by comparisons with reference values from LAMOST DR8 LRS. Section 5.2.2 then describes the corresponding estimates for clusters, along with comparisons against values from open cluster catalogs.

5.2.1 Individual stars

This part addresses the model outputs for individual cluster members, covering both the estimation of uncertainties and the comparison of the resulting metallicities with external references.

Resampling for uncertainties estimation

Although the training step did not consider errors directly as input information, the uncertainties of the predicted parameters are provided according to the methodology described in this section. A resampling procedure is performed seeking to estimate uncertainties for the parameters. Importantly, this process also yielded the final [Fe/H] values adopted here.

In this method, multiple artificial datasets are generated by adding random noise to the original photometric and distance data, reflecting the observational errors. More specifically, each star goes through a sampling from a normal distribution where the imputed mean is the observed quantity (magnitudes and distances) and the dispersion is each measurement's error. For each resampled dataset, the model is reapplied to predict the atmospheric parameters. This process is repeated 1000 times, resulting in 1000 predicted values for each parameter of each star. From the distribution of these predicted values, the mean and standard deviation are adopted as the final values in this work. This resampling approach allows for capturing the variability inherent in the observed measurements and the model, thus offering a robust quantification of the statistical component of the parameter uncertainties.

Importantly, this procedure propagates the reported statistical uncertainties of the input photometry and distances under the assumption that measurement errors are independent and Gaussian. While correlations between bands and potential systematic effects in the photometric calibration, distance scale, extinction estimates, or the machine learning model are not explicitly modeled, these are expected to contribute primarily to the overall accuracy rather than to the internal precision quantified here. As a result, the quoted uncertainties should be interpreted as statistical errors and may be regarded as a lower limit on the total uncertainty budget.

Figure 5.5 illustrates the distributions of the final adopted parameters (top) and their respective errors (bottom). The average values of the errors, marked in dashed black lines, are 58 K for $T_{\rm eff}$, 0.047 dex for log g, and 0.169 dex for [Fe/H]. In particular, the

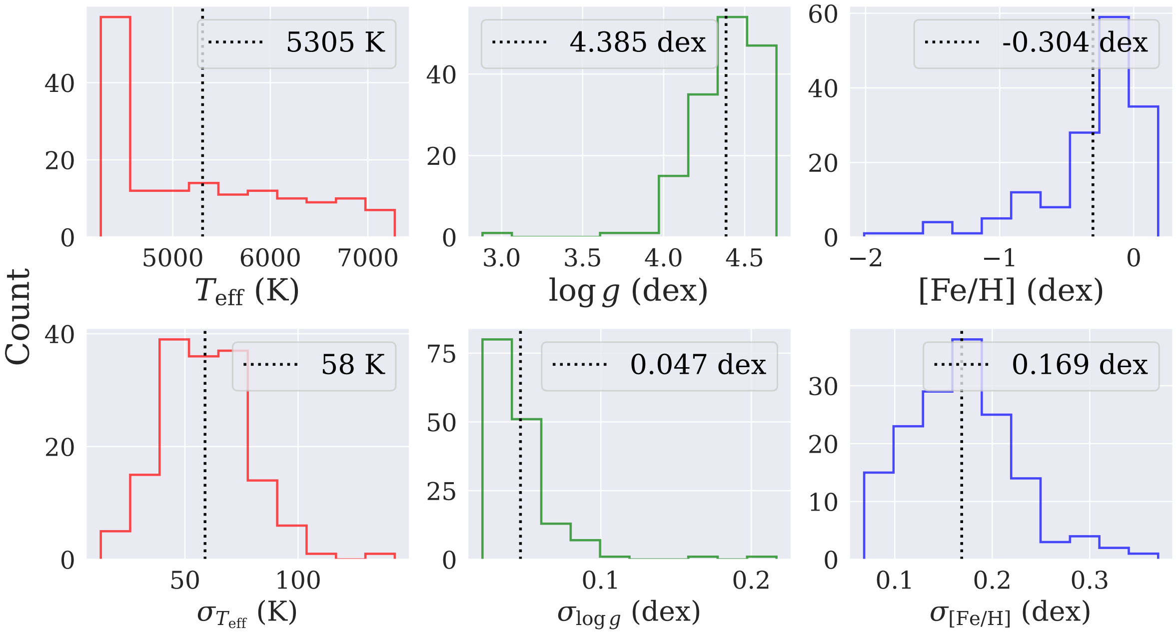

Figure 5.5: Parameters (top) and respective errors (bottom) obtained from the resampling process (discussed in the text) for the 154 cluster members. The derived mean and standard deviation are indicated. Dashed black lines indicate the average errors for each distribution.

peak in the leftmost portion of the temperature panel, around $T_{\rm eff} \lesssim 4600$ K, represents 38% of the sample (59 out of 154) and indicates that several stars are likely cooler than their predicted $T_{\rm eff}$, since temperature regimes below $T_{\rm eff} < 4000$ K are not covered in the training sample (see Fig. 4.2). Figure 5.6 shows the histograms of $(\rm J0430 - W1)_0$ for the full training sample (magenta) and the OCs members (cyan), overlaid by fitted Gaussian curves, and densograms indicated on top. It becomes evident that the sample of OC members covers a much wider range of $(\rm J0430 - W1)_0$ values. From the well-behaved trend of $T_{\rm eff}$ with $(\rm J0430 - W1)_0$, seen in Figure 5.7, it is reasonable to expect that if training data were available up until $(\rm J0430 - W1)_0 \gtrsim 7$ mag, most likely $T_{\rm eff}$ should keep monotonically decreasing. This reasoning is backed up by assuming that cooler stars consistently present smaller W1 values (i.e., brighter in redder regimes) as $T_{\rm eff}$ decreases, increasing the difference to J0430 and the value of $(\rm J0430 - W1)_0$ itself. Moreover, in this

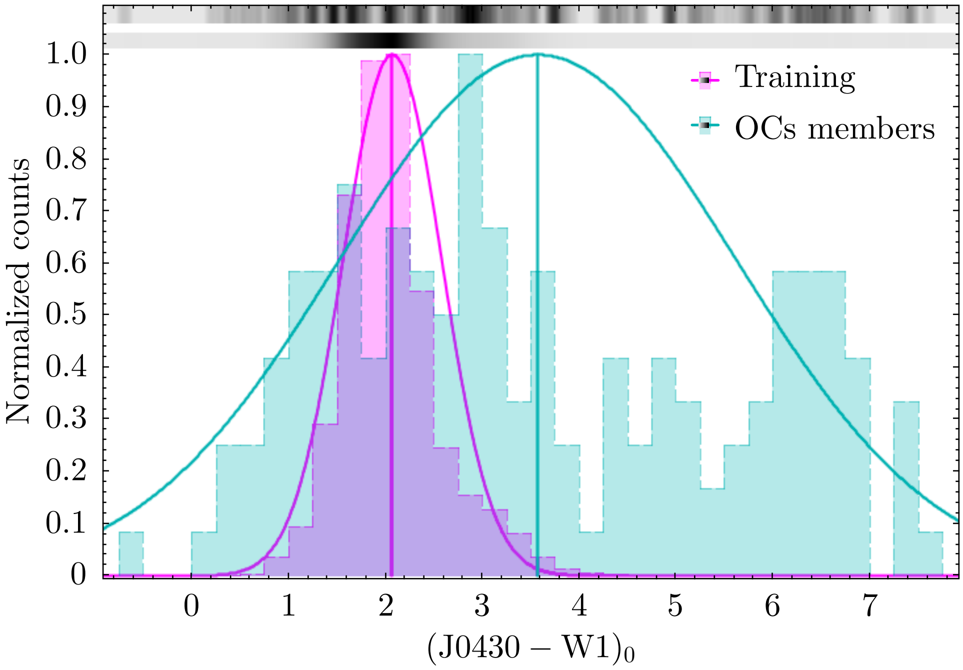

Figure 5.6: Histogram of $(\mathrm{J0430} - \mathrm{W1})_0$ comparing the ranges of the training sample (magenta) and the OCs members (cyan). Gaussian curves are overlaid and densograms are indicated above the bars. Counts are normalized to each distribution's respective maximum values.

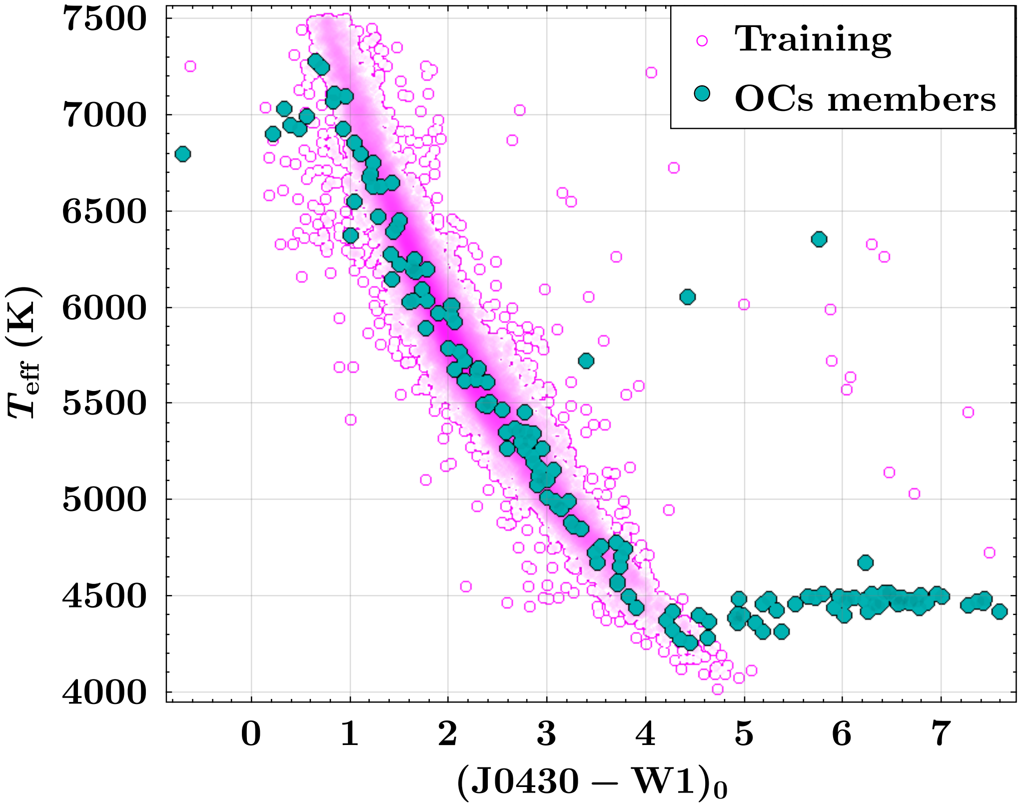

Figure 5.7: T_{eff} as a function of its most important feature, $(\mathrm{J0430} - \mathrm{W1})_0$, for the full training sample (magenta) and OC members (cyan). The training distribution defines the overall relation, while OC members keep on until $(\mathrm{uJAVA} - \mathrm{J0430})_0 \gtrsim 7$.

analysis, note that $(\mathrm{J0430}-\mathrm{W1})_0$ is compared to T_{eff} just for being its most important feature, hence serving as a proxy.

When the model struggles to map correlations between features and targets (because of high scatter, nonlinear relationships, or insufficient resolution in feature space, etc.), it minimizes error by regressing toward the mean value of the training distribution ($T_{\mathrm{eff}}^{\mathrm{mean}} = 5860$ K; again, see Figure 4.2), which pulls predictions away from its minimum value ($T_{\mathrm{eff}}^{\mathrm{min}} = 4020$ K). Because the model cannot predict $T_{\mathrm{eff}} < 4000$ K, this peak should be interpreted as a mixture of correctly predicted values in the range $4000 < T_{\mathrm{eff}}/\mathrm{K} \lesssim 4500$, together with most likely upper limits for the temperatures of these stars.

Comparisons against LAMOST

The resulting predictions from the application of the [Fe/H] model to the list of OCs members were compared to the original values of LAMOST. Only those stars with reference parameters within the training range for each parameter were kept, namely $[4000, 7500]$ K for T_{eff}, $[0, 5]$ dex for log g and $[-2.5, 1.0]$ dex for [Fe/H]. Since tree-based models, such as those from random forests and gradient boosting techniques, inherently preserve the limits of the training data and do not extrapolate beyond their range, it was not necessary to impose these selections for values in this work. For inputs falling outside the training domain, these models generally revert to predictions biased toward the mean response of the training data.

For comparisons with LAMOST, the cross-match was performed using the Gaia DR3 source ID, resulting in 30 stars from the initial 154. Of those, 25 fall within the comparison limits, namely $[4000, 7500]$ K for T_{eff}, $[0, 5]$ dex for log g and $[-2.5, 1.0]$ dex for [Fe/H]. Figure 5.8 shows the distribution of residuals for the predicted stellar parameters, where Δ represents this work (TW) minus LAMOST. The Gaussian curves depict distributions

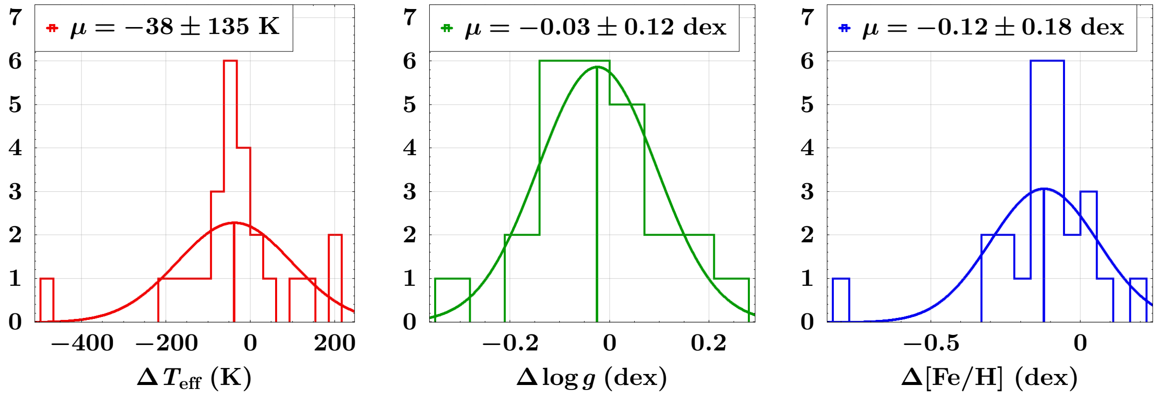

Figure 5.8: Distributions of residuals for T_{eff}, log g, and [Fe/H] of clusters members cross-matched with LAMOST. Each panel shows the histogram (in counts) of differences between model predictions and reference LAMOST values, with a Gaussian fit overlaid. The mean and standard deviation ($\mu \pm \sigma$) of each distribution are indicated.

generally centered around zero, with mean residuals and standard deviations ($\mu \pm \sigma$) of -38 ± 135 K for $T_{\rm eff}$ (left panel), -0.03 ± 0.12 dex for $\log g$ (center panel), and -0.12 ± 0.18 dex for [Fe/H] (right panel). The presence of outliers in the histogram tails suggests that larger deviations from the reference values occasionally occur. Additionally, the mean metallicity residual indicates a slight, systematic underestimation in the predictions from this work. Although this behavior could also be pointed in the $T_{\rm eff}$ panel, it is likely driven by the small size and composition of the constrained sample rather than by an intrinsic model bias.

Given the limited sample size, this analysis must be interpreted with caution. For example, note that relaxing the parameter-range constraints yields mean residuals of -9 ± 220 K for $T_{\rm eff}$ (a shift toward zero of about 76%), -0.02 ± 0.12 dex for $\log g$ (no significant change), and -0.08 ± 0.19 dex for [Fe/H] (a shift toward zero of about 25%). This smaller change for [Fe/H] leaves open the possibility of a mild systematic underestimation, although the limited comparison set precludes any firm conclusion.

Figure 5.9 compares the metallicities predicted by the model with those from LAMOST. The left panel color-codes data points by $T_{\rm eff}$ from this work, $T_{\rm eff}^{\rm TW}$, while the right panel uses $\Delta T_{\rm eff}$. The solid black lines in both panels represent the 1:1 relation, and the dashed gray lines denote one standard deviation of the residuals in [Fe/H], Δ[Fe/H] ≈ -0.18 dex. Large error bars indicate significant uncertainties in some cases, and a few outliers deviate substantially from the 1:1 relation. Notably, [Fe/H] of metal-rich stars in LAMOST ($[{\rm Fe/H}]_{\rm LAMOST} > 0$) are all underestimated, while metal-poor ones appear better distributed around the 1:1 line. In particular, the stars with underestimated metallicities (below the 1:1 line) span nearly the entire temperature range, and the color scale in the right panel indicates that these stars also tend to have slightly underestimated $T_{\rm eff}$ (yellow-toned points).

This behavior suggests some degeneracy between [Fe/H] and $T_{\rm eff}$, so [Fe/H] was plotted as a function of $T_{\rm eff}$ for various datasets in Figure 5.10. The gray density distribution and contour curves represent the full LAMOST dataset, while black dots indicate the 154 OC members in this study. The data points accumulated at $T_{\rm eff} \gtrapprox 4500$ K reflect the peak in the histogram of Figure 5.5. The colored data points correspond to cross-matched stars between the two datasets, where cyan diamonds denote values from LAMOST and magenta circles represent the predicted values obtained in this work. Each cyan–magenta pair is connected by a thin black line to illustrate the parameter differences for the same object. The overlaid straight lines are linear fits to the respective datasets, following the same color coding.

LAMOST metallicities (gray points) exhibit no significant trend with $T_{\rm eff}$ (dashed gray line), while this work OC members (black points) show a mild trend of increasing [Fe/H] as $T_{\rm eff}$ increases (full black line). A more pronounced decreasing trend can be seen for stars in the cross-match with LAMOST (full magenta line), further suggesting that hotter

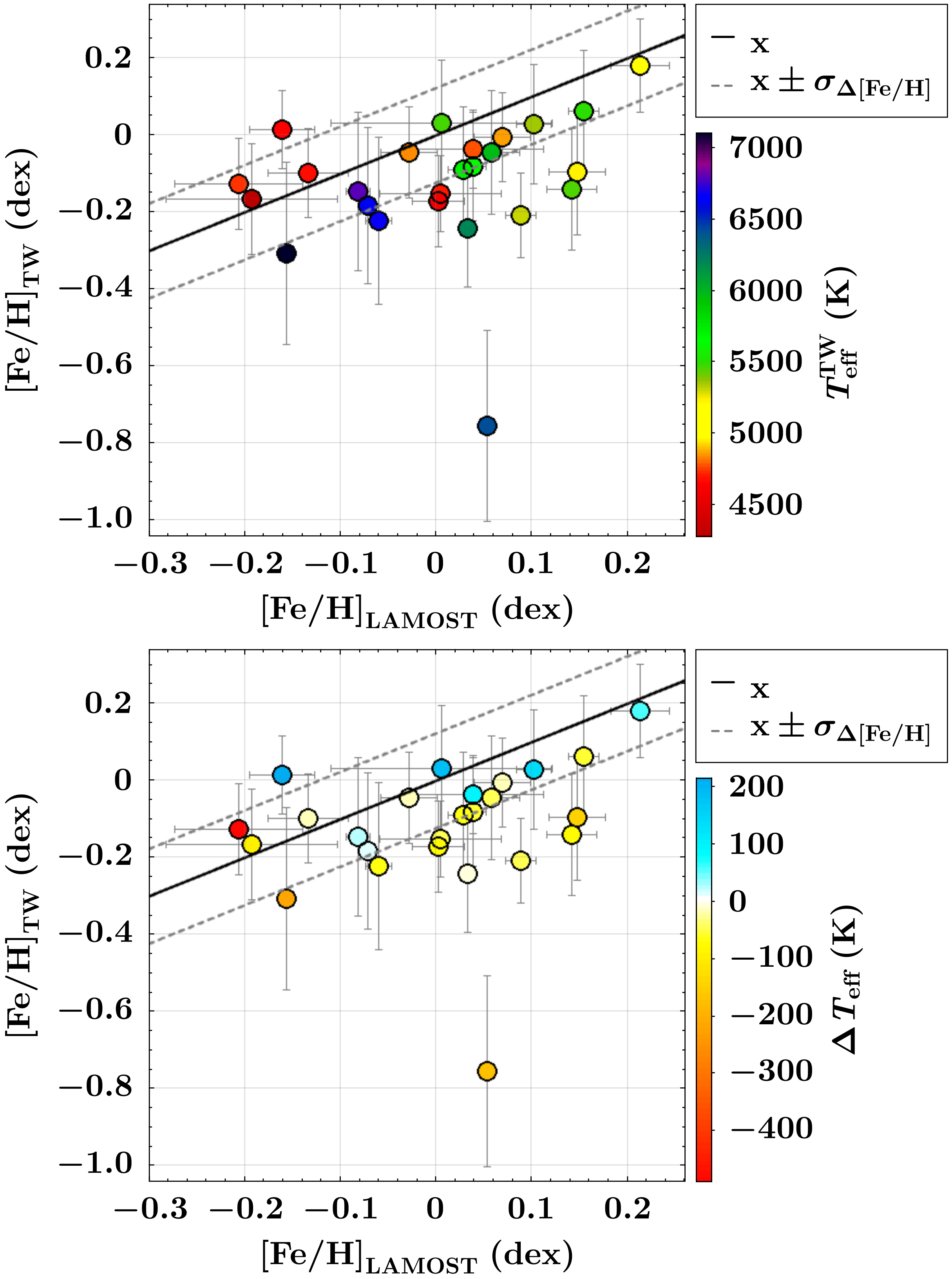

Figure 5.9: Comparison between metallicities from this work and from LAMOST. The left panel is color-coded by $T_{\text{eff}}^{\text{TW}}$, while the right panel uses ΔT_{eff}. The solid black line represents the 1:1 relation, and the dashed gray lines indicate one standard deviation of the residuals in [Fe/H].

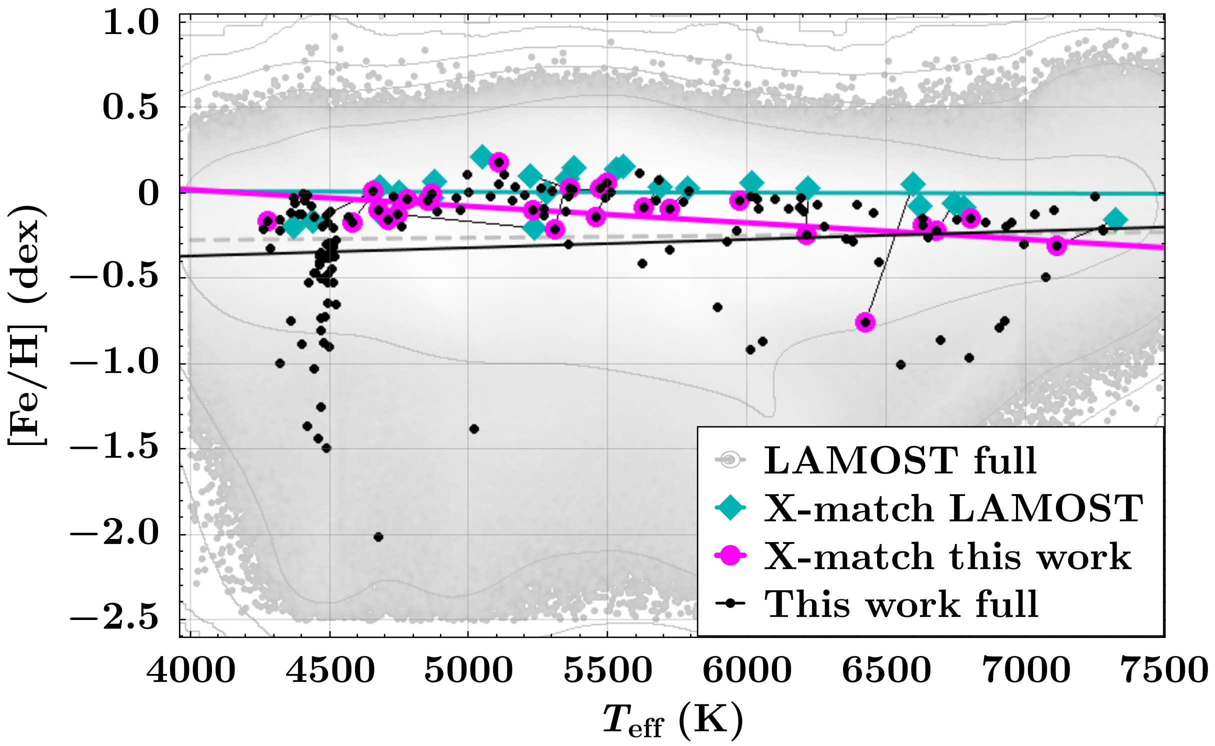

Figure 5.10: [Fe/H] values as a function of $T_{\rm eff}$ for different datasets. Gray contours show the full LAMOST dataset, while black dots represent OC members. Cyan diamonds and magenta circles denote cross-matched stars, linked to indicate the same object. Linear fits follow the same color coding.

stars may be systematically underestimated in [Fe/H]. However, since underestimations occur in both cooler and hotter stars (see Figure 5.9; another way to visualize this is by checking that most magenta points are placed below cyan ones), it is unlikely that temperature alone drives this effect. Instead, it may influence metallicity predictions indirectly through correlations with photometric features.

To further investigate this, Figure 5.11 displays [Fe/H] residuals, color-coded by $T_{\rm eff}$, as a function of the most important photometric feature for [Fe/H] $({\rm uJAVA} - {\rm J0430})_0$ in the left panel, and for $T_{\rm eff}$ $({\rm J0430} - {\rm W1})_0$ in the right panel. Two distinct regions emerge in the left panel: data points appear evenly distributed around zero for $({\rm uJAVA} - {\rm J0430})_0 > 1.22$ mag, while a clear depression is seen at smaller values. In addition to highlighting a trend of underestimated metallicities in the latter regime, the color bar shows that it spans almost the entire temperature range ($T_{\rm eff} \gtrsim 5000$ K), whereas the former contains only cooler stars ($4000 < T_{\rm eff}/{\rm K} < 5000$). Similarly, the same stars gathered close to $\Delta[{\rm Fe/H}] = 0$ in the left panel remain well-behaved in the right one for $({\rm J0430} - {\rm W1})_0 > 2.92$ mag, while those with underestimated metallicities are again grouped, in this case for $({\rm J0430} - {\rm W1})_0 < 2.92$ mag. Although the separation may not be as visually striking, the regime change is still evident in this space[4]. Altogether, although

---

[4] We highlight that the values 1.22 and 2.92 were chosen conveniently, purely to aid visualization of

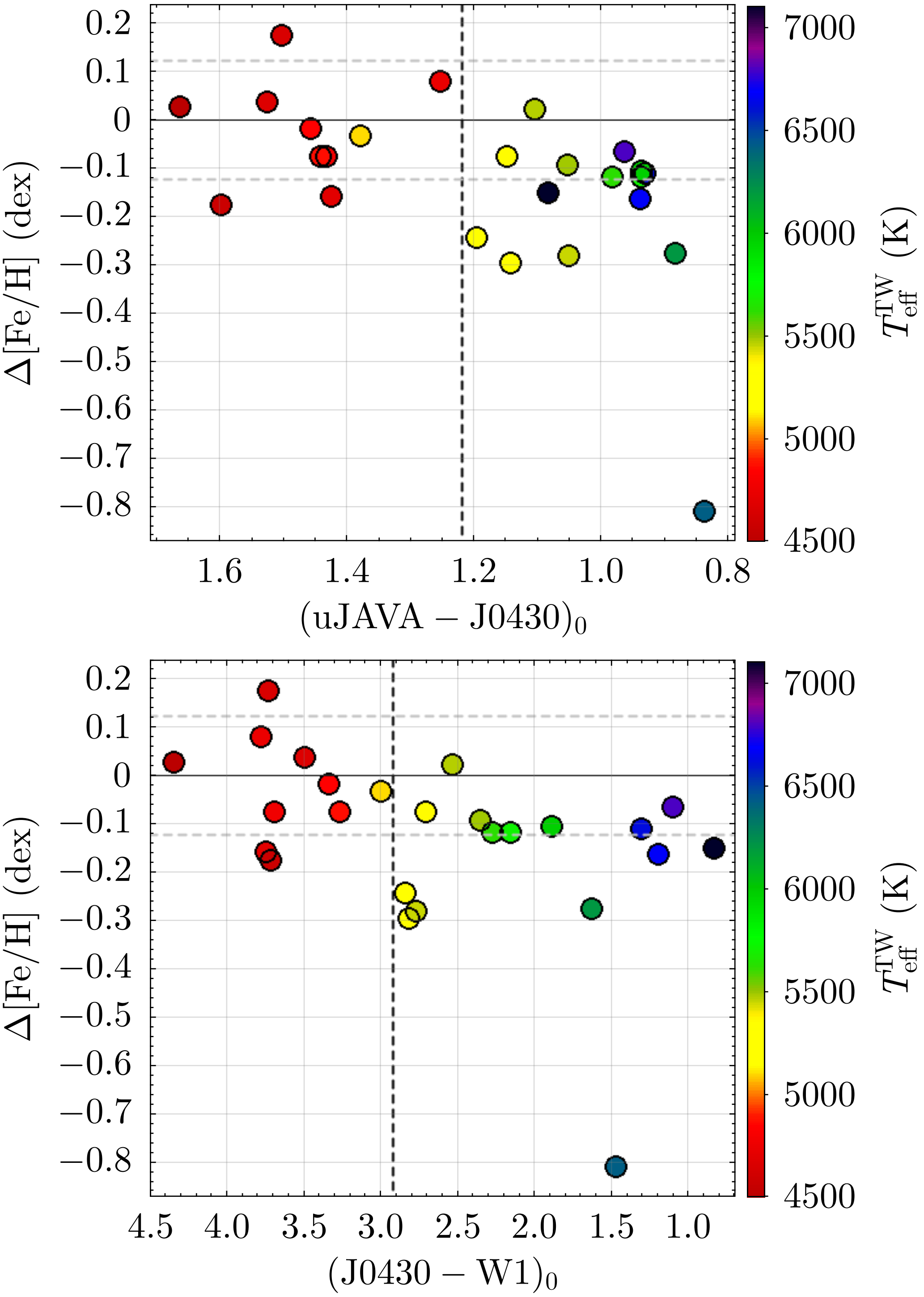

Figure 5.11: Residuals in [Fe/H] as a function of the most important feature for [Fe/H] (left) and $T_{\rm eff}$ (right), both color-coded by $T_{\rm eff}$. Solid black lines at (uJAVA − J0430)$_0$ = 1.22 mag and (J0430 − W1)$_0$ = 2.92 mag mark the division between distinct Δ[Fe/H] regimes.

based on few data points, these plots suggest that metallicity is reliably estimated only for the cooler stars.

The relation between T_{eff} and the most important feature of [Fe/H] is explicitly shown in Figure 5.12, where $(\text{uJAVA} - \text{J0430})_0$ and T_{eff} are plotted for the training sample, with [Fe/H] shown in the color bar. Larger circles indicate the 25 LAMOST matched stars, colored by their predicted metallicities. Dashed, black lines mark the color range of the cross-matched sample, and the solid black lines indicate the approximate temperature (vertical, $T_{\text{eff}} = 5000$ K) and color (horizontal, $(\text{uJAVA} - \text{J0430})_0 = 1.2$ mag) where the regime change occurs in Figure 5.11. In this space, the coolest stars $(4000 < T_{\text{eff}}/\text{K} < 5000)$ follow a spread but well-defined relation, which begins to break down beyond 5000 K, when $(\text{uJAVA} - \text{J0430})_0$ falls within the range $0.8 < ((\text{uJAVA} - \text{J0430})_0)/\text{mag} < 1.2$. While 88% of the sample (77 857/88 421) lies within this regime, these stars span a wide temperature range $(5000 < T_{\text{eff}}/\text{K} < 7500)$, practically saturating this region. Thus,

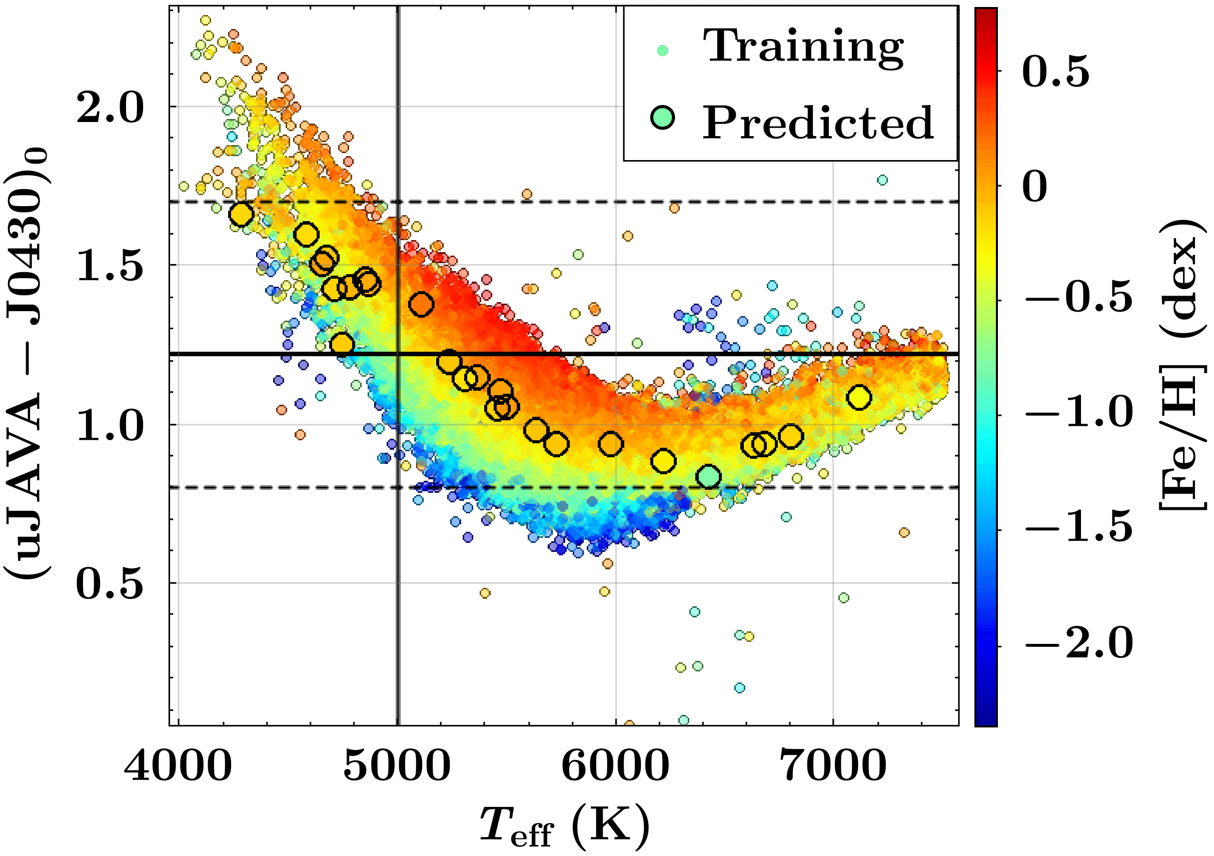

Figure 5.12: $(\text{uJAVA} - \text{J0430})_0$ as a function of LAMOST T_{eff}. Small circles represent the full training sample, color-coded by their LAMOST [Fe/H]. Larger symbols show the 25 open cluster members matched with LAMOST, color-coded by their predicted [Fe/H]. The solid, black lines mark $T_{\text{eff}} = 5000$ K (vertical) and $(\text{uJAVA} - \text{J0430})_0 = 1.22$ mag. The dashed, horizontal, black lines mark the limits of the $(\text{uJAVA} - \text{J0430})_0$ range for the 25 LAMOST matched stars, namely $0.80 < (\text{uJAVA} - \text{J0430})_0/\text{mag} < 1.70$.

---

the distinction.

within this $(\mathrm{uJAVA} - \mathrm{J0430})_0$ range, stars can have virtually any temperature (other than $T_{\mathrm{eff}} < 5000$ K). Moreover, the color layers in this figure also attest the relation between $(\mathrm{uJAVA} - \mathrm{J0430})_0$ and [Fe/H].

Again, the model struggles to map the correlations between features and targets in this case, hence minimizing error by regressing toward the central tendency (mean value) of the training distribution. Since the training sample has a mean metallicity of $\mu^{\mathrm{training}}_{[\mathrm{Fe/H}]} = -0.21$ dex (Figure 4.2 shows $\mu^{\mathrm{test}}_{[\mathrm{Fe/H}]} = -0.21$ dex for the test, blind set), this results in a systematic underestimation of metallicity in this temperature range. Moreover, recall that the models used to predict parameters in this work are independent of each other. The predicted values do not directly influence one another, meaning the [Fe/H] model does not have direct access to T_{eff}. When referring to the model's ability to distinguish temperatures, it is meant through photometric relations.

To further confirm that metallicity predictions are influenced by temperature-sensitive features, the most important ones are examined for both parameters. This is illustrated in Figure 5.13, where solid black lines mark the regime change values: $(\mathrm{uJAVA} - \mathrm{J0430})_0 = 1.22$ mag (horizontal) and $(\mathrm{J0430} - \mathrm{W1})_0 = 2.92$ mag (vertical)[5]. Larger circles indicate the 25 LAMOST matched stars, colored by their predicted metallicities. Dashed black lines indicate the respective ranges included in the cross-matched sample. The correlation between $(\mathrm{uJAVA} - \mathrm{J0430})_0$, the most important feature for [Fe/H], and $(\mathrm{J0430} - \mathrm{W1})_0$, the most important feature for T_{eff}, is expected and could already be inferred from Figure 5.12, since T_{eff} itself is well constrained.

It is also important to sort out why [Fe/H] is more strongly influenced by the behaviors discussed here, while T_{eff} is not. The observed degeneracy between T_{eff} and [Fe/H] appears to stem from a non-linear interaction between their most informative features. This degeneracy impacts [Fe/H] predictions more severely, since T_{eff} is primarily constrained by the overall continuum shape, which is broadly and smoothly reflected in photometric colors, making T_{eff} easier to recover. In contrast, [Fe/H] depends on narrower metallic lines and localized spectral depressions that are less effectively traced by photometry—at least in the form our features are built, as basic magnitude subtractions. Although a mild correlation exists between T_{eff} and [Fe/H], the most important T_{eff} feature is not highly ranked in the [Fe/H] model, likely because its inclusion would degrade performance, as it does not provide sufficient constraint on [Fe/H].

In fact, considering the most important feature for each of T_{eff} and [Fe/H], we find that reliable predictions are confined to specific color ranges, namely $1.2 < (\mathrm{uJAVA} - \mathrm{J0430})_0/\mathrm{mag} < 1.7$ and $2.9 < (\mathrm{J0430} - \mathrm{W1})_0/\mathrm{mag} < 4.4$. Although this restriction may raise concerns—since the most informative feature becomes less predictive outside these regimes—, it does not invalidate the model's overall performance. The main caveat we identify is a systematic effect known as attenuation bias (TING, 2024), which arises from

---

[5] Both values were set arbitrarily after simple visual inspection, merely for visualization.

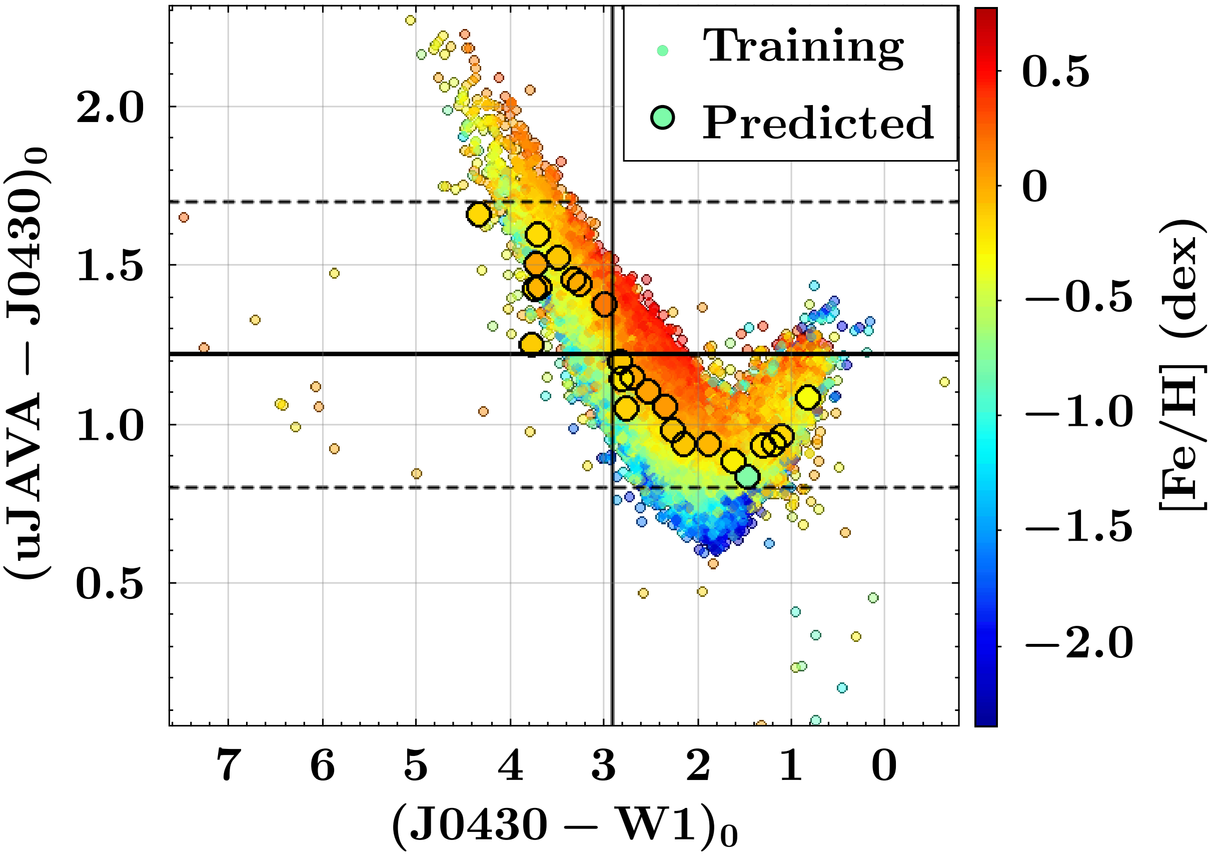

Figure 5.13: $(\mathrm{uJAVA} - \mathrm{J0430})_0$ as a function of LAMOST $(\mathrm{J0430} - \mathrm{W1})_0$Ṡmall circles represent the full training sample, color-coded by their LAMOST [Fe/H]. Larger symbols show the 25 open cluster members matched with LAMOST, color-coded by their predicted [Fe/H]. The solid black lines mark $(\mathrm{J0430} - \mathrm{W1})_0 = 2.92$ mag (vertical) and $(\mathrm{uJAVA} - \mathrm{J0430})_0 = 1.22$ mag, beyond which T_{eff} begins to be systematically underestimated. The dashed gray lines mark the limits of the respective ranges for the 25 LAMOST matched stars, namely $0.80 < (\mathrm{J0430} - \mathrm{W1})_0/\mathrm{mag} < 4.40$ and $0.80 < (\mathrm{uJAVA} - \mathrm{J0430})_0/\mathrm{mag} < 1.70$.

limited resolution in feature space. In regimes where the most informative feature loses sensitivity (in our case for stars hotter than 5000 K) this leads the model to underestimate metallicities by pulling predictions toward the central tendency (mean value) of the training set. In contrast, for stars with $4000 < T_{\mathrm{eff}}/\mathrm{K} < 5000$, the stronger correlation between $(\mathrm{uJAVA} - \mathrm{J0430})_0$ and metallicity allows the model to resolve the relationship more effectively and yield more accurate predictions.

Importantly, this systematic behavior is only partially captured by the reported uncertainties. Because they are derived from repeated model predictions under perturbed inputs (magnitudes and distances), regions affected by attenuation bias may appear to yield confident results, despite the predictions being systematically biased toward the central tendency. Furthermore, this additional contribution is not explicitly quantified here, as this work does not provide a well defined framework or external calibration to do so, and the reported uncertainties should therefore be regarded as lower limits in these regimes.

It is therefore important that we have explicitly identified and mapped this limitation, which will be more thoroughly addressed in future work. Nevertheless, the overall performance metrics of the [Fe/H] model indicate that it performs well despite this caveat. The mean uncertainty in adopted metallicity, $\langle \sigma_{[\mathrm{Fe/H}]} \rangle = 0.169$ dex, supports the reliability of the predictions, though it is important to emphasize that this uncertainty is focused in capturing the quality of the input photometry, rather than internal model biases arising from feature degeneracies. In this context, the ability to diagnose and map such limitations becomes a key outcome of this study. Together with the model's overall predictive performance, this characterization represents one of the most relevant contributions of the present work.

Comparisons against APOGEE DR17 were not performed due to even lower numbers than those from LAMOST comparisons: 14 stars in common, of which only 7 make the parameters range cut. Moreover, no stars were found in common with SEGUE, Gaia-ESO or GALAH DR3 (which is expected in this case, since covered hemispheres do not coincide).

5.2.2 Clusters

This part examines the results at the cluster level, summarizing the metallicities obtained from the member stars and the corresponding uncertainties. The adopted values are then compared with those reported in open cluster catalogs.

Adopted metallicities and errors

Given the context of this work, it is reasonable to expect that photometric results for metallicity may not fall within the narrow internal iron abundance spreads typically inferred from spectroscopic studies of open clusters. These spreads are generally on the order of $\lesssim 0.02$–0.03 dex (BOVY, 2016; POOVELIL *et al.*, 2020). In contrast, photometric metallicity estimates tend to show uncertainties around 0.05–0.10 dex (NETOPIL *et al.*, 2013, 2016), even when internal consistency is enforced. Therefore, the metallicity scatter observed in this work is expected to reflect the limitations of photometric precision rather than any true deviation from open cluster homogeneity.

The cluster metallicity was adopted as the sample median of member metallicities, with its uncertainty given by the bootstrap standard error of the median ($\mathrm{SE}_{\mathrm{med}}$). This provides a robust point estimate together with a data driven error scale, which we denote as

$$[\mathrm{Fe/H}]_{\mathrm{cluster}} = \tilde{x} \pm \mathrm{SE}_{\mathrm{med}},$$

where $\tilde{x}$ is the sample median of the members' metallicities. The $\mathrm{SE}_{\mathrm{med}}$ characterizes the dispersion of the bootstrap distribution of the median and thus the sampling uncertainty of the central estimate. Since no simple closed form expression exists for the SE of the median, it was determined by non parametric bootstrap resampling: $B = 1000$ new samples of size n (cluster size, in this context) were generated with replacement from the original data, the median was computed for each, and the standard deviation of these bootstrap medians was taken as $\mathrm{SE}_{\mathrm{med}}$. In practice, the standard deviation was computed with $B - 1$ in the denominator (`ddof=1`), since the mean of the bootstrap distribution is itself estimated from the same B realizations. With this correction, the sample variance is unbiased (IVEZIĆ *et al.*, 2020). Moreover, while the rMAD quantifies the variability of the data points themselves (their scatter; see Appendix C), the SE of a statistic measures the uncertainty associated with that statistic. Both rMAD and the $\mathrm{SE}_{\mathrm{med}}$ are reported in this work, the former for internal scatter and the latter for uncertainty estimation based on bootstrap resampling of the observed values.

These choices were motivated by the presence of clear outliers in some clusters, which deviate significantly from the bulk of the member distribution (as illustrated for NGC 1039 and HSC 749 in the histograms of Figure 5.14, described in the next paragraph). The use of bootstrap resampling, when applied to robust statistics such as the median and rMAD, provides a more reliable characterization of the typical cluster metallicity. In this context, the robustness comes from the statistics themselves rather than the bootstrap procedure, which simply quantifies their sampling variability. This makes the estimates less sensitive to outliers and therefore more representative of the bulk of the member distribution. This methodology is particularly suitable under the assumption that cluster members formed together from a chemically homogeneous molecular cloud. Moreover, when only a single member was available, neither rMAD nor $\mathrm{SE}_{\mathrm{med}}$ could be computed. In these cases, the individual metallicity and its associated uncertainty were adopted, with the caveat that the quoted uncertainties should be interpreted as lower limits, as is the case for the multi-member systems.

Figure 5.14 illustrates these results by showing the 23 associations distributed in a space of metallicity vs. the number of members retrieved, color-coded by metallicity uncertainty. Overall, the estimates are somewhat gathered around $[\mathrm{Fe/H}] \sim -0.2$ dex (more precisely between -0.1 and -0.3), which is consistent with the underestimation scenario described in Section 5.2.1, given that OCs are generally expected to present near solar metallicity. As previously discussed, stars hotter than 5000 K occupy a narrow region in the $(\mathrm{uJAVA} - \mathrm{J0430})_0$ vs. T_{eff} space (Figure 5.12), which limits the model's ability to capture variations in metallicity. Given that $(\mathrm{uJAVA} - \mathrm{J0430})_0$ is the most important feature for predicting [Fe/H], this compression in feature space hinders the model's capacity to learn meaningful correlations, even though evaluation metrics indicate generally good predictions. In such cases, as described earlier, the model tends to regress toward the mean of the training distribution (in this case, $\mu_{[\mathrm{Fe/H}]}^{\mathrm{training}} = -0.21$ dex), leading to a systematic bias in that direction. Furthermore, a few more metal-poor cases occur

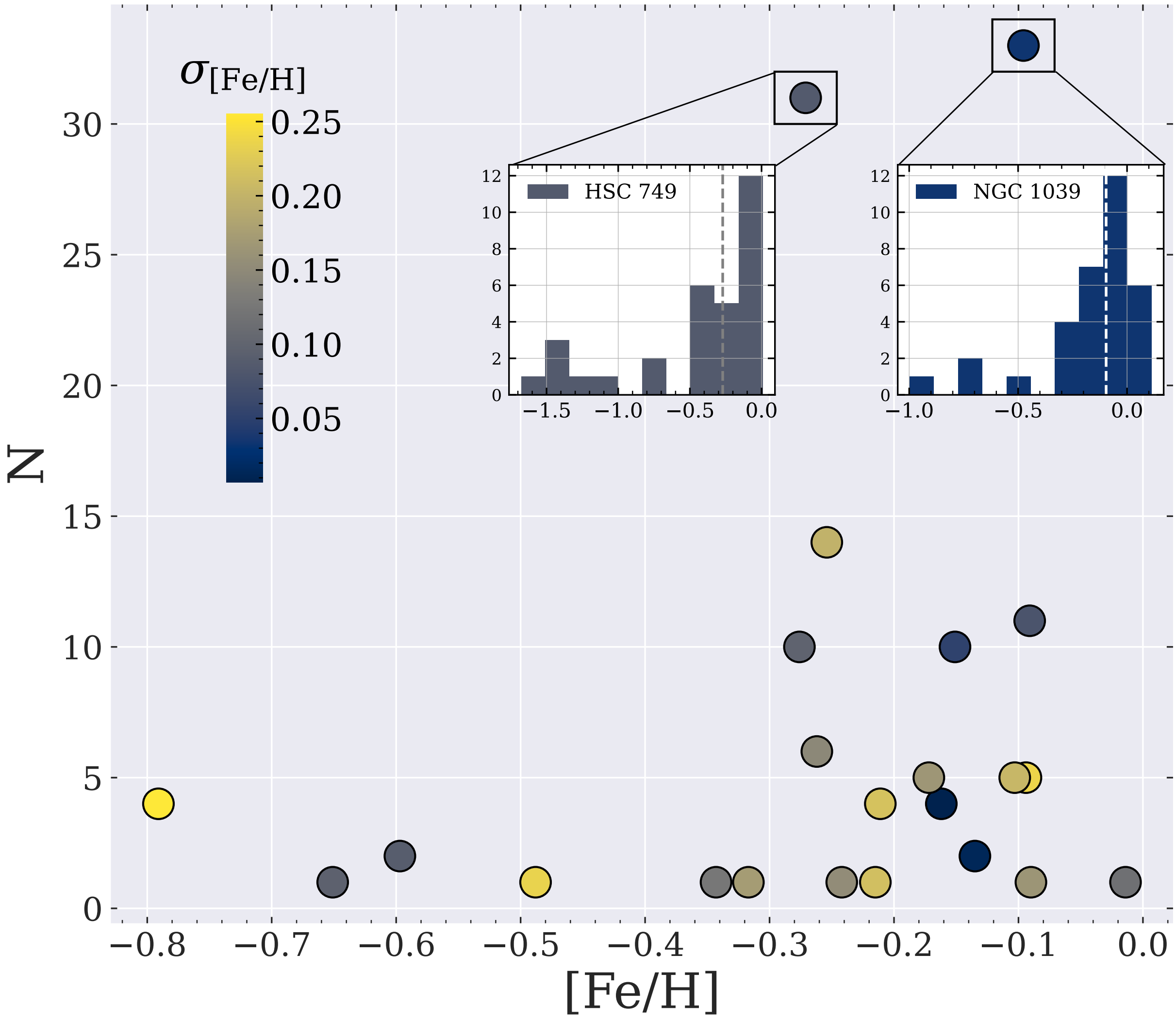

Figure 5.14: Number of members in each of the 23 clusters plotted against adopted metallicity, with circles color-coded by adopted uncertainty for the clusters. The gray dashed line marks the mean value for [Fe/H].

for objects with fewer than five members, but note that some of those are represented by only one object. In such cases, considering predictions of eventual outliers by the model (see Figure 5.8, these results can possibly be just not representative of the object.

As discussed in Section 5.1, being formed together make HR diagrams and similar spaces powerful tools in studying OCs and its members. Figure 5.15 presents Kiel diagrams displaying the parameters predicted in this study for the six most populated clusters. The vertical and horizontal axes represent log g and T_{eff}, respectively, while [Fe/H] is indicated by the color bar. Clusters names, number of members (N), [Fe/H], and uncertainty in [Fe/H] are indicated. The data points accumulated at the cool end ($T_{\mathrm{eff}} < 5000$ K) in Figure 5.15 merely reflect the behavior identified in Figure 5.5 and discussed in Section 5.2.1, that these stars are cooler than the model is capable of predicting. It is therefore not a coincidence that they also exhibit the largest uncertainties in T_{eff} as well as anomalous metallicities. In a way, this figure condenses the main results

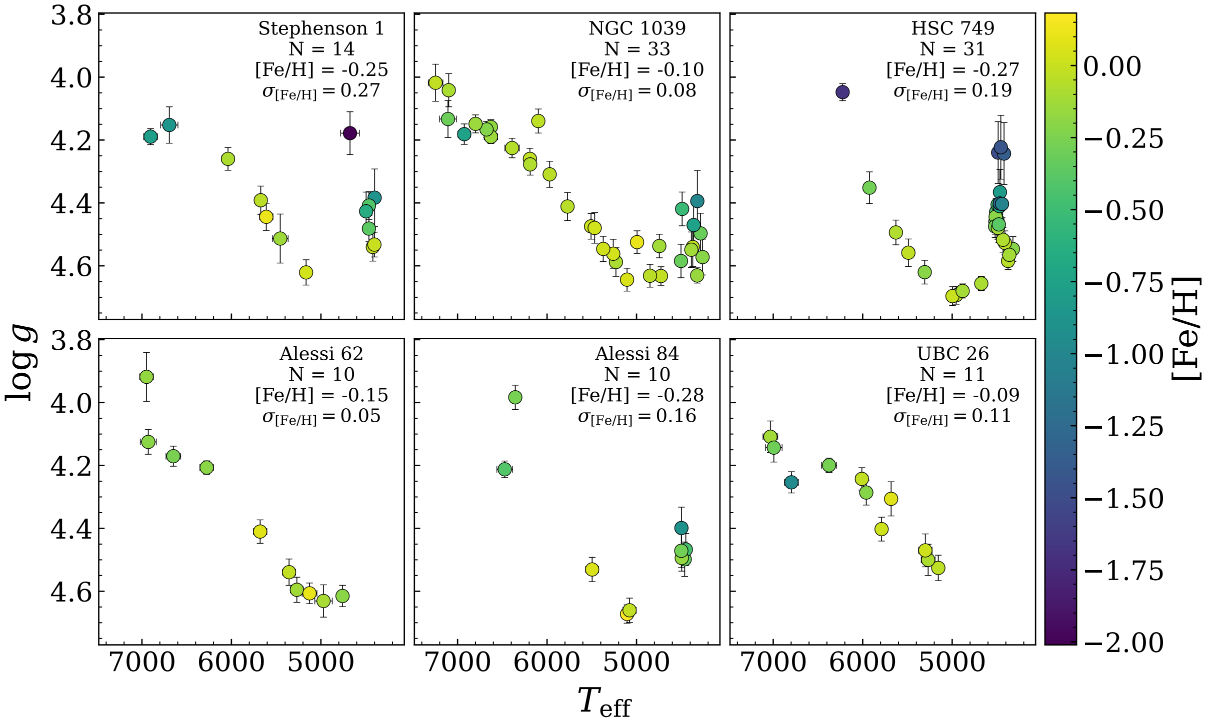

Figure 5.15: Kiel diagrams for stars in the six most populated systems in the present work, color-coded by [Fe/H]. Each panel indicates the cluster name, number of stars (N), adopted metallicity and respective uncertainty. The trends are overall consistent with the expected main-sequence morphology, with some dispersion in metallicity and data points accumulated in the cool end.

of the present work, presenting the three parameters estimated here applied to OCs, and more importantly, recovering the overall structure of their main sequences.

That structure is nicer behaved in a CMD space, for instance, as in Figure 5.16. The horizontal and vertical axes represent the most important features for T_{eff} and log g, respectively, namely the color $(\text{J0430}-\text{W1})_0$ and the absolute magnitude $W1_0^{\text{abs}}$. The main sequence here assumes a more typical form, particularly without data points clumping in the cold end, since T_{eff} is not explicitly invoked. As in Figure 5.15, data points are color coded by [Fe/H].

For further illustration, Figure 5.17 displays color-color diagrams composed only of narrow-band filters. Here, $(J0430-J0861)$ serves as a proxy for T_{eff} and $(J0430-J0515)$ for log g both among the most important features for each respective parameter. This space mimics OC layouts in CMDs, for instance, exhibiting well-delineated main sequences.

For a closing example, Figure 5.18 also employs only J-PLUS photometry, but introduces the broad-band filter uJAVA_0 in the vertical axis, while keeping $(J0430-J0861)$ in the horizontal axis. This combination resulted in a clearer stretching of the points, better resolving the data along the main-sequence structure, and potentially making it visually easier to distinguish evolutionary stages across the clusters.

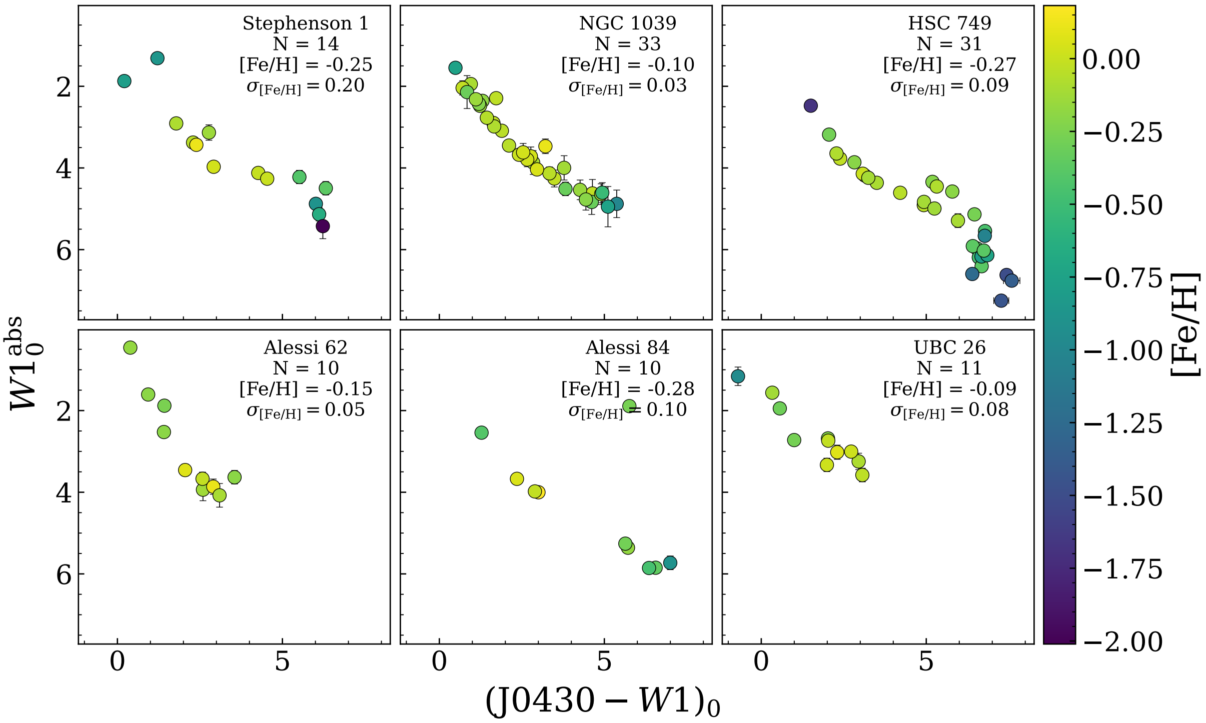

Figure 5.16: CMDs of the six most populated clusters in the present work, color-coded by [Fe/H]. Horizontal and vertical axes represent the most important features for $T_{\rm eff}$ ($(\rm J0430 - W1)_0$) and log $g(W1_0^{\rm abs})$, respectively. Data points are colored by [Fe/H].

In all these cases, the trends are consistent with the expected main-sequence morphology, with some dispersion in metallicity and data points accumulated in the cool end. Importantly, these examples also illustrate the power of the 12-filters system for exploring stellar populations, especially in stellar clusters (e.g. BONATTO *et al.*, 2019, for a globular cluster example).

Comparisons against OCs catalogs

The estimates of [Fe/H] for OCS derived in this work were compared against values from the literature, specifically from the works of CAVALLO *et al.* (2023) and ZHANG *et al.* (2024), as shown in Figure 5.19.

CAVALLO *et al.* (2023) uses an ANN approach with multi-band CMDs to estimate parameters, including metallicity, for approximately 5400 Gaia OCs. For error estimates, the ANN provides uncertainty based on 1σ confidence interval for [Fe/H] from a distribution of predictions, with the catalog providing 16th, 50th, and 84th percentiles to derive this. The average uncertainty for all parameters is approximately $\leqslant 0.2$ dex.

ZHANG *et al.* (2024) investigates the metallicity distribution of 1131 open clusters by combining Gaia DR3 data with LAMOST low-resolution spectra. Error estimates include a typical 0.07 dex error for individual stellar [Fe/H] from the LAMOST pipeline. After corrections and data cleaning, the adopted uncertainty for cluster mean metallicity values

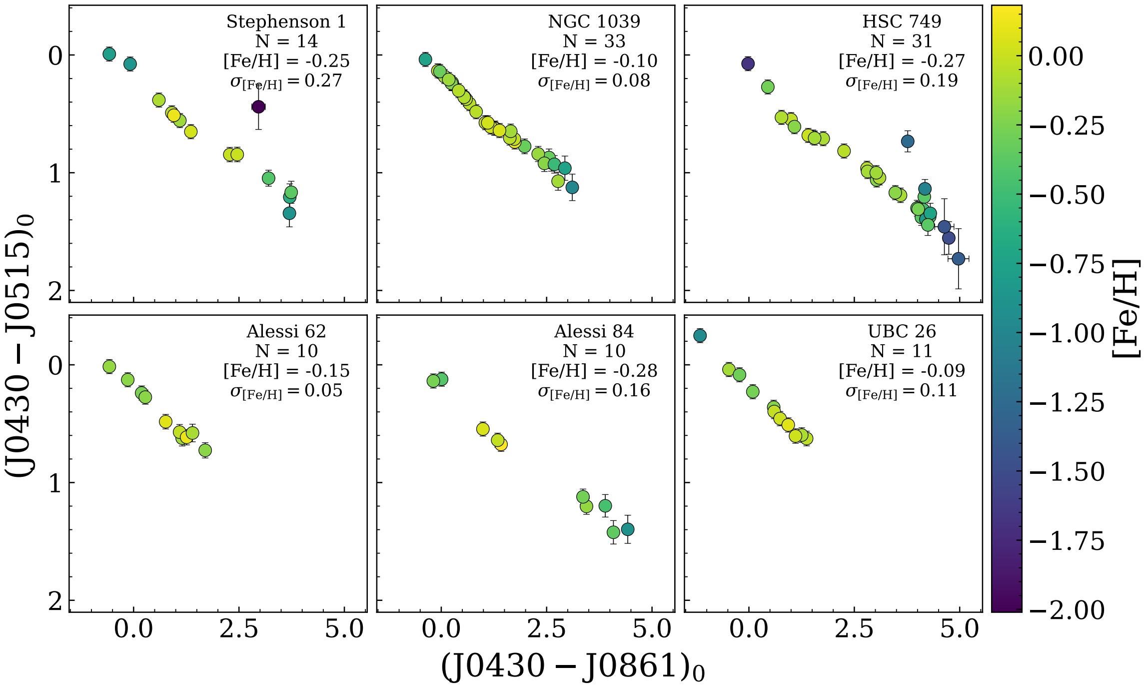

Figure 5.17: Color-color diagrams of the six most populated clusters in the present work. Horizontal and vertical axes represent the second most important feature for $T_{\rm eff}$ $(J0430-J0861)$, and the third most important one log g, $(J0430-J0515)$, respectively. Both are composed only of narrow-band filters. Although both axes are colors, data points mimic the distributions in CMDs.

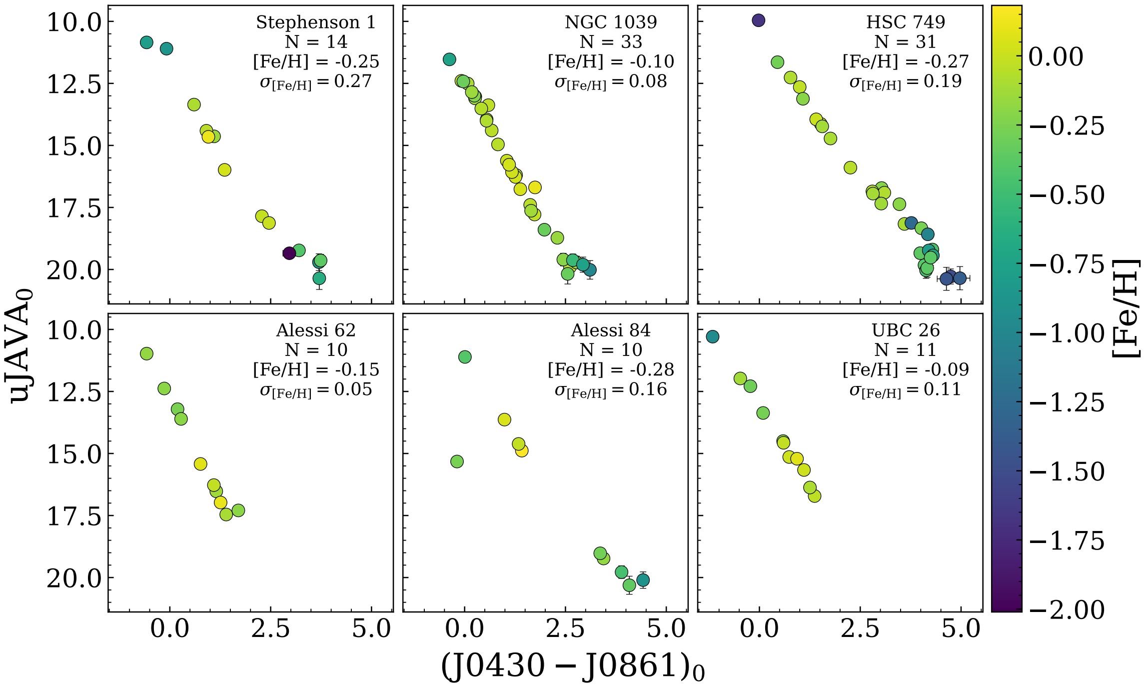

Figure 5.18: Color-color diagrams of the six most populated clusters in the present work. The horizontal axis again represents the second most important feature for $T_{\rm eff}$ $(J0430-J0861)$, while the vertical axis now shows the broad-band magnitude $uJAVA_0$, which visually stretches the main sequence structure.

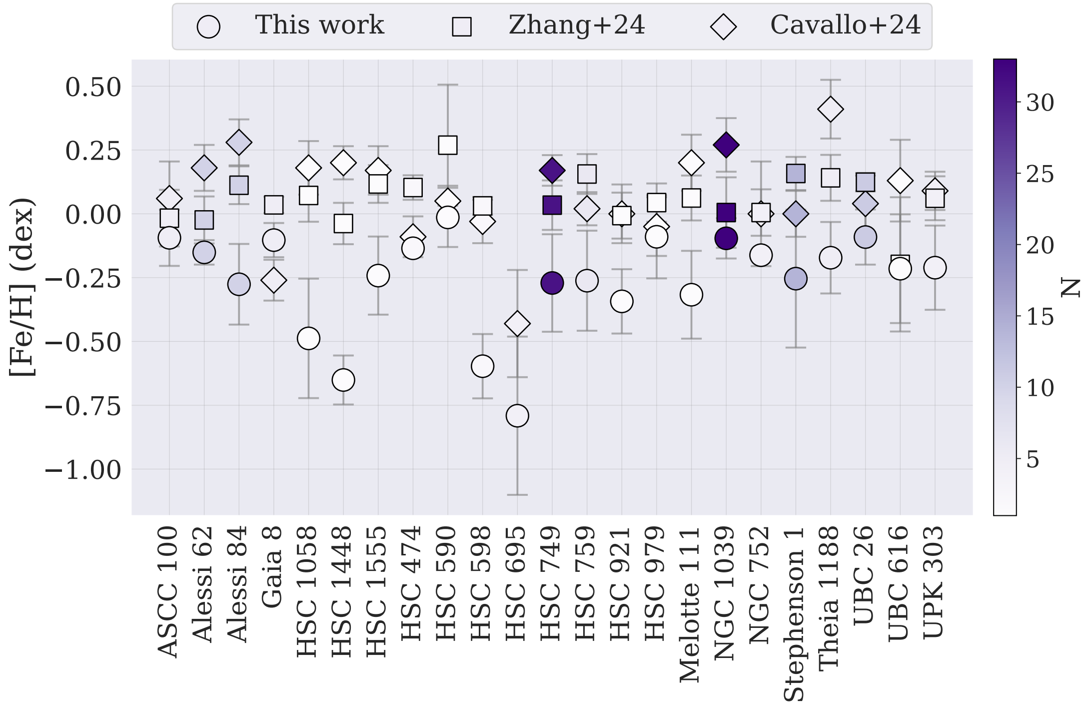

Figure 5.19: Literature comparison of [Fe/H] values for the 23 stellar systems analyzed here. Results are shown for this work (circles), CAVALLO *et al.* (2023, diamonds), and ZHANG *et al.* (2024, squares), colored by the number of members analyzed in this work. The systems are ordered alphabetically along the x-axis.

derived via Monte Carlo sampling is approximately 0.12 dex, with the standard deviation from this method serving as the final uncertainty.

Results from this work are shown alongside those from these two literature sources, with distinct markers indicating each source: circles for this work, diamonds for CAVALLO *et al.* (2023), and squares for ZHANG *et al.* (2024). Marker colors show the number of retrieved members in the present work, and soft gray error bars accompany all points, allowing for visual inspection of overall agreement within uncertainties.

The majority of systems display reasonable consistency across the three works, with some exceptions showing notable discrepancies, such as Alessi 84 (10 recovered members), Melotte 111 (1 member), and Theia 1188 (5 members). In such cases, the small sample sizes may reflect the non-representativeness discussed earlier. Beyond small-number statistics, these differences could also arise from real chemical inhomogeneity, differences in membership selection, or simply unreliable [Fe/H] estimates.

Importantly, all objects labeled with 'HSC' (candidate new systems reported in HR23) have been later updated as moving groups, and are therefore excluded from the list of the cited discrepant OC examples. Note also that [Fe/H] estimates in the present work tend to be systematically lower than those in the literature, as discussed in Section 5.2.1, reflecting the known underestimation bias of the method. Despite this bias, the relative trends across systems are preserved, and agreement remains satisfactory within uncertainties.

Furthermore, with an appropriately updated training sample containing more metal-poorer stars, this approach could be adapted for the analysis of multiple stellar populations in galactic globular clusters (see e.g. HARTMANN *et al.*, 2022, and references therein). This direction could also contribute to building metallicity distribution functions of unresolved extragalactic globular clusters. Such applications would be particularly relevant for studies of low-redshift galaxies.

Comparisons against recovered members in spectroscopic surveys

Further metallicities comparisons were performed for OCs using their retrieved members directly from spectroscopic samples. LAMOST DR8 and APOGEE D17 were cross matched with the list of probable members of the most reliable clusters sample of HR23. For both, exact match was required using Gaia DR3 source id, which resulted in 8267 stars from LAMOST and 6273 from APOGEE. Recall the parameters range considered in this work to filter out stars outside these intervals $4000 < T_{\mathrm{eff}} < 7500$ K, $0 < \log g < 5$ dex, $-2.5 < [\mathrm{Fe/H}] < 1$ dex), which retained 5388 in LAMOST (65%) and 2774 (44%) in APOGEE. In total, 328 stars distributed in 21 clusters were found in LAMOST, and 151 stars in 9 clusters in APOGEE.

The same procedure adopted for the cluster members analyzed in this work was applied to the stars from LAMOST and APOGEE to derive cluster metallicities and uncertainties. For each cluster, the adopted value is the median [Fe/H], and the formal uncertainty is the $\mathrm{SE}_{\mathrm{med}}$, which quantifies the precision of the central estimate, i.e., how reliably that particular sampled median estimate reflects the true population value. In addition, rMAD is used as a one sigma scale for the internal scatter.

In the comparisons with CAVALLO *et al.* (2023) and ZHANG *et al.* (2024) (Figure 5.19), the central values were consistent overall, and the differences, compatible with the reported errors (and with the rMAD spread as well, although these are not exhibited in the figure). As for comparisons against LAMOST and APOGEE, shown in Figure 5.20, the [Fe/H] values are consistent when looking at the rMAD scale but not with the smaller $\mathrm{SE}_{\mathrm{med}}$. The top panel shows error bars drawn from $\mathrm{SE}_{\mathrm{med}}$ values, which results in clear disagreements not seen in Figure 5.19. The bottom panel shows error bars drawn from rMAD, exhibiting a closer agreement to that of Figure 5.19. In both the color bar follows the number of retrieved stars to be analyzed.

In spite of the slight underestimation pattern identified for [Fe/H] predictions in the present work, general agreement is observed when using rMAD in Figure 5.20, with most of the discrepant cases corresponding to systems later reclassified from OC candidates to moving groups. While $\mathrm{SE}_{\mathrm{med}}$ quantifies the scatter across bootstrapped resamples of the median, rMAD instead reflects the star-to-star spread within a cluster, combining intrinsic dispersion with measurement noise. This way, although rMAD is not an error

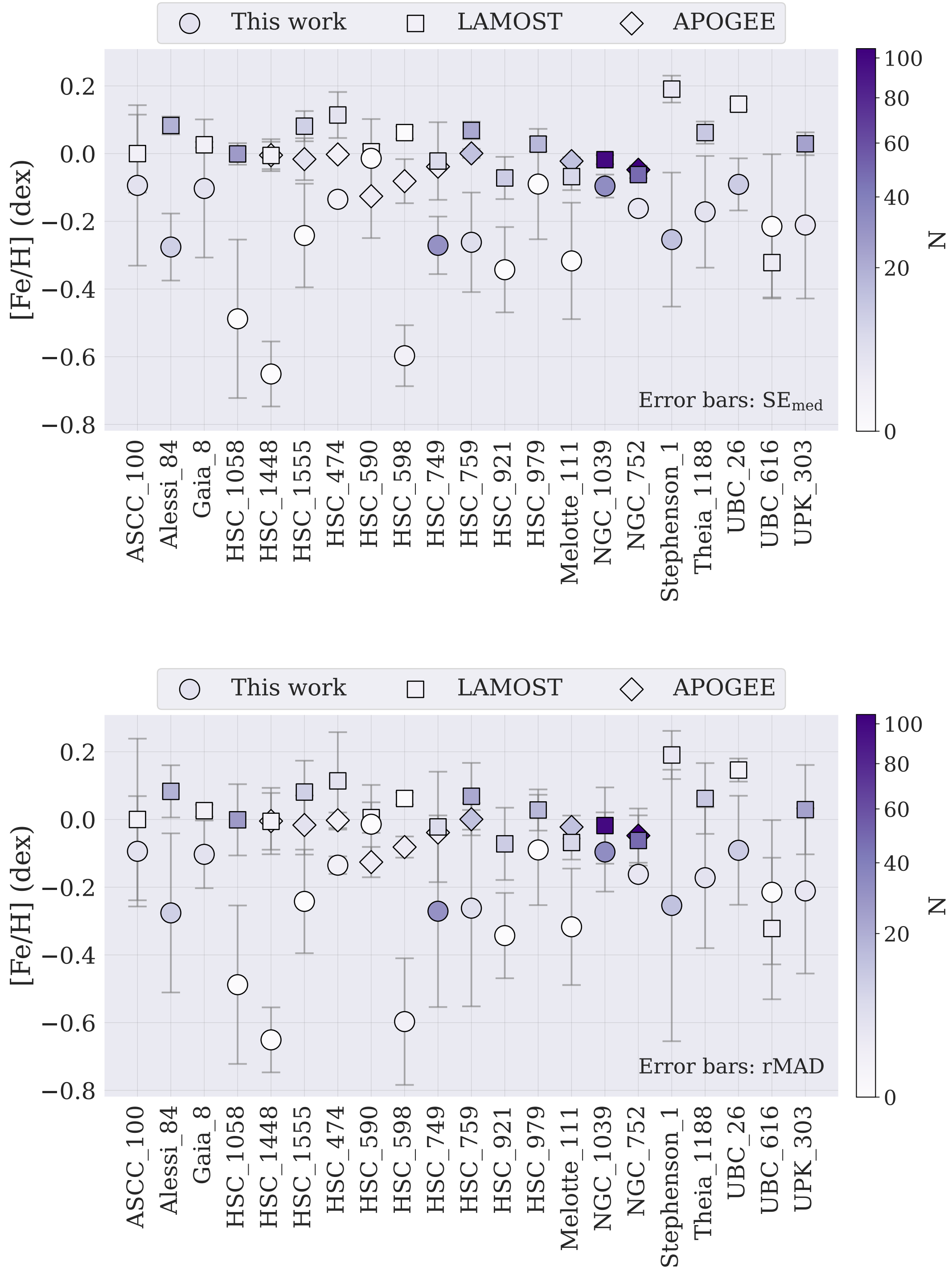

Figure 5.20: Comparison of [Fe/H] values for 21 of the 23 stellar systems with members retrieved either in LAMOST, APOGEE< or both. Results are shown for this work (circles), LAMOST (square), and diamonds (APOGEE), colored by the number of members retrieved in each case. The systems are ordered alphabetically along the x-axis. Error bars are given by SE$_{\rm med}$ (top) and rMAD (bottom), as indicated accordingly.

estimate in the strict sense, conformity of results within this scale provides a defensible basis to argue that the method delivers a valid first approximation, suitable for guiding subsequent detailed spectroscopic studies.

Chapter 6

Conclusions and perspectives

The work developed for this thesis was founded on the application of ML methods to photometric data in the context of large astronomical surveys. The unprecedented volume and diversity of data released over the last decade have driven a rapid growth of AI applications in astrophysics, motivating the search for scalable approaches to extract stellar parameters from photometry. Within this scenario, the methodology presented here demonstrates how gradient-boosted decision trees, carefully tuned and systematically validated, can be adapted to infer $T_{\rm eff}$, log g, and [Fe/H] from multi-band photometry. J-PLUS provided the main observational backbone, complemented by Gaia and CatWISE, while LAMOST spectroscopy served as the reference for supervised learning. The next sections summarize the main concluding remarks and future possible implementations.

6.1 Overall conclusions

A distinctive contribution of this work is the emphasis on uncertainty-aware predictions. By propagating input measurement errors through resampling, the models deliver reliable error estimates that remain tied to the quality of the photometry itself, rather than purely internal metrics of model variance. In addition, the process of feature selection proved enlightening. Using `shap-hypetune`, features and hyperparameters were optimized jointly, enabling the identification of the most informative combinations of colors and absolute magnitudes. Importantly, the results showed that beyond a certain point, aggressively cutting the number of features did not improve performance and could even be detrimental, as it in fact discarded substantial sets of lower-ranked features, which combined proved genuinely informative. This outcome underlines that interpretability and parsimony must be balanced with predictive accuracy, and that exhaustive pruning is not necessarily a path to more robust models. Moreover, the models achieved median absolute deviations of 42 K, 0.06 dex, 0.06 dex, relying on 67, 26, and 35 physically meaningful features, for $T_{\rm eff}$, log g, and [Fe/H], respectively.

The science case on open clusters further demonstrated the applicability of the method.

Photometric parameters were estimated for 154 members across 23 systems, for which comparisons with spectroscopic metallicities revealed overall agreement within uncertainties ($\mathrm{SE}_{\mathrm{med}}$), although reflecting an identified mild systematic underestimation, likely arising from attenuation bias. This bias also relates to compression effects in the training distribution, highlighting the importance of continuously expanding and diversifying reference samples and investing in assembling the best training sample possible. More broadly, the exercise underscored the role of photometric machine learning as a complement to spectroscopy: In clusters with extensive spectroscopic coverage, photometric results are best interpreted as corroborative context, reinforcing the reliability of spectroscopic determinations through an independent methodology; in clusters with sparse coverage, they serve as priors guiding efficient spectroscopic allocation. Together, these results attest to the viability of photometric ML pipelines as scalable tools for stellar population studies in the era of large surveys.

While [Fe/H] has been emphasized in this thesis as a key application, the method is designed to provide the full set of stellar parameters, T_{eff}, log g, and [Fe/H], ensuring broad applicability in contexts where spectroscopy is unavailable.. Comparisons with LAMOST and APOGEE show overall agreement within scattering (rMAD), whose role was contrasted against that of the standard error of the median, $\mathrm{SE}_{\mathrm{med}}$: while rMAD captures star-to-star dispersion within clusters, $\mathrm{SE}_{\mathrm{med}}$ quantifies the stability of median values under resampling. The analyses demonstrated the complementary roles of these different scatter estimators.

The approach developed here is not intended as a substitute for high-quality spectroscopy, but rather as a complement to it. More broadly, this framework contributes to stellar astrophysics by scaling atmospheric parameter estimation to large photometric samples. It reinforces the role of photometric surveys in stellar population studies, even if originally conceived with cosmological aims, and provides a scalable methodology readily applicable to S-PLUS and adaptable to J-PAS.

In clusters with extensive spectroscopic coverage, photometric results are best interpreted as corroborative context, while reinforcing the reliability of spectroscopic determinations through an independent methodology. In clusters with sparse coverage, where spectroscopy often concentrates on the brighter or more massive stars, photometric predictions contribute by extending parameter estimates to the fainter, lower-mass members, providing a more complete population-level view. In this sense, they act as practical priors that can guide efficient allocation of spectroscopic resources and enable uniform comparisons across different clusters. Together, these results attest to the viability of photometric ML pipelines as scalable tools for stellar population studies in the era of large surveys. In short, the box below outlines practical considerations and the intended-use guidelines for photometric metallicities, highlighting opportunities and limits of applying such models in different contexts.

INTENDED USE OF THE PHOTOMETRIC ML METALLICITIES

Use the model output when:

1. Few or no spectra exist for a cluster, particularly for its low-mass members that are typically underrepresented in spectroscopic surveys.
2. Available spectra come from mixed pipelines with unknown offsets.
3. A first-pass ranking is desired for prioritizing spectroscopic follow-up.
4. A uniform reference is required to compare many clusters at once.

Avoid relying on it when high-quality homogeneous spectroscopy is available for the specific objects of interest and a precise abundance scale is required. In such cases, use the photometric estimates only as ancillary context.

6.2 Future work

A long list of possibilities permeates the time ahead of this project, not only because of what it has to offer, but also given the broader context in which it is embedded. In the following sections, potential paths to follow are briefly exposed, with emphasis on applying this method to S-PLUS and making the data available to the community.

6.2.1 Overall prospects

In view of the results presented here, a major limitation to be addressed is the systematic underestimation observed in metallicity predictions. A first step toward mitigating this issue is to extend the application of the models, particularly for [Fe/H], to a much wider sample. Increasing the number of stars with predicted parameters will provide a firmer basis for statistical analyses. In parallel, an interesting and potentially simpler strategy is to implement ad hoc corrections to the current predictions, for example, by fitting polynomial functions to the residuals of the models.

A prospect initially intended to be implemented in the present work was to apply predicted metallicities to constrain the early identification of cluster members. For this, relevant fields must have their stars mapped for [Fe/H] so that the respective populations could stand out accordingly.

With regard to AI methods and large surveys, there is a very broad window to explore in future work. The most immediate effort seems to be implementing the methodology developed here in S-PLUS, since its filter system and main instrumentation are identical to those of J-PLUS. Recent efforts on both sides have provided valuable contributions in various fields (e.g. HUANG *et al.*, 2024; LOPES *et al.*, 2025; QUISPE-HUAYNASI *et al.*, 2023; YANG *et al.*, 2022), attesting that this exchange benefits the whole community.

A less straightforward but still closely related endeavor is adjusting this methodology to fit J-PAS. Its deeper coverage and richer filter system will enable major breakthroughs photometrically. However, to fully harness these assets, the present procedure must be adapted to meet the particular needs and seize the advantages of this tool set.

In addition to applying the method to S-PLUS, releasing the trained models and the products for the full J-PLUS catalog is among the short-term plans. This would not only make the framework more broadly accessible to the community but also facilitate independent applications to diverse areas. By releasing trained models and catalog products, the shared methods developed here enables for the community to engage in comparative studies across surveys, stimulating follow-up spectroscopic programs, and establishing a benchmark for future adaptations in the era of increasingly data-rich astronomy.

6.2.2 Further models

An obvious extension of the present framework is to experiment with additional color combinations beyond the traditional set of magnitudes and colors. Although this may appear inconsistent with the systematic feature selection strategy adopted here, incorporating more flexible or unconventional combinations could open windows unavailable to classical diagnostics. Compound or weighted color indices may uncover subtle patterns in stellar populations that standard two-band subtractions may not capture, potentially optimizing photometric relations.

A natural complement to this idea is the use of dimensionality reduction techniques. By projecting high-dimensional feature spaces into compact representations, methods such as principal component analysis (PCA) or t-distributed stochastic neighbor embedding (t-SNE) can compress information while retaining information relevant to stellar parameters. This approach would allow the model to explore richer sets of input features without suffering from overfitting or redundancy, balancing complexity with interpretability.

Another promising direction lies in the incorporation of Gaia XP spectra, which provide low-resolution, prism-based spectrophotometry for over 200 million stars. Generating synthetic photometry from these spectra, either by replicating J-PLUS/S-PLUS filters or designing optimized bandpasses, would enrich the feature space considerably. This could serve as a bridge between photometric and spectroscopic domains, offering a continuous framework to explore atmospheric parameters and, eventually, chemical abundances. In particular, integrating this idea to J-PAS photometry could provide an unprecedented homogeneous photometric data set for, for example, $[\alpha/\mathrm{Fe}]$.

Finally, models could be extended to make use of features based directly on observed fluxes rather than magnitudes. While magnitudes are convenient for their logarithmic scaling, they also compress dynamic range and can obscure fine differences in flux ratios.

Working with calibrated fluxes may therefore capture more linear correlations with stellar properties, at the cost of requiring stricter attention to calibration and extinction corrections. This path could prove valuable, especially when cross-matching heterogeneous datasets where flux-based representations may provide a more universal baseline.

6.2.3 Further parameters

A straightforward extension of the present work is to broaden the scope of predicted parameters beyond the classical trio of $T_{\rm eff}$, $\log g$, and [Fe/H]. Among these, the abundance of [α/Fe] stands out as a particularly relevant quantity. Because α-enhancement encodes information about the relative contributions of Type II and Type Ia supernovae, it is a cornerstone observable for disentangling the chemical enrichment history of the Milky Way. Photometric estimates of [α/Fe], even at modest precision, would open up a wide range of applications. These include mapping the Galactic metallicity gradients with added chemical sensitivity, refining age–metallicity relations, and tracing the assembly of different stellar populations such as the disk, halo, and even stellar streams. Extending the methodology developed here to include [α/Fe] predictions would therefore provide a powerful tool for chemodynamical studies, helping to bridge the gap between photometric surveys and high-resolution spectroscopic campaigns.

With a careful feature selection, but in a different approach, reframing some of the problems addressed here as classification or clustering tasks may be particularly fruitful. Besides traditional spectral classification, photometric models could be trained to classify stars into key loci of astrophysical interest, such as metal-poor regimes, carbon-enhanced metal-poor (CEMP) stars, chemically peculiar subgroups, or plain unresolved binaries flagging through inconsistent parameter combinations.

Importantly, these strategies could be carried out before or alongside the estimation of $T_{\rm eff}$, log g, and [Fe/H]: discrete categories capture broad stellar characteristics, while finer continuous parameters refine the description. As a result, these approaches could maximize the potential of photometric surveys to flag rare stellar populations across vast datasets and identify interesting systems for follow-up.

In a broader sense, these possibilities illustrate how the methodology developed in this thesis can evolve in the future from predicting the traditional set of stellar parameters into a more versatile framework that addresses both continuous and categorical tasks. By combining regression for atmospheric parameters with clustering and categorization, spectral classification, and eccentric objects identification, photometric ML can progressively expand its scientific reach. Such extensions would not only enhance the characterization of stars on an individual basis but also strengthen the role of large photometric surveys in Galactic archaeology and stellar populations, linking abundance studies with the panoramic view offered by all-sky imaging campaigns.

Bibliography

ADORF, H.-M., JOHNSTON, M. D., 1988, *Artificial neural nets in astronomy.* Arbeitspapier 329, Gesellschaft für Mathematik und Datenverarbeitung.

AGHAMOUSA, A., THE DESI COLLABORATION, 2016, “The DESI Experiment Part I: Science, Targeting, and Survey Design”, *arXiv e-prints*, (October). doi: 10.48550/arXiv.1611.00036. for the DESI Collaboration.

AHN, C. P., ALEXANDROFF, R., ALLENDE PRIETO, C., et al., 2012, “The Ninth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Baryon Oscillation Spectroscopic Survey”, Astrophys. J. Suppl., 203(2):21. doi: 10.1088/0067-0049/203/2/21.

AIHARA, H., ALLENDE PRIETO, C., AN, D., et al., 2011, “The Eighth Data Release of the Sloan Digital Sky Survey: First Data from SDSS-III”, Astrophys. J. Suppl., 193(2):29. doi: 10.1088/0067-0049/193/2/29.

ALMEIDA, A., ANDERSON, S. F., ARGUDO-FERNÁNDEZ, M., et al., 2023, “The Eighteenth Data Release of the Sloan Digital Sky Surveys: Targeting and First Spectra from SDSS-V”, Astrophys. J. Suppl., 267(2):44. doi: 10.3847/1538-4365/acda98.

ANDERS, F., CHIAPPINI, C., SANTIAGO, B. X., et al., 2018, “Dissecting stellar chemical abundance space with t-SNE”, *Astronomy & Astrophysics*, v. 619, pp. A125. doi: 10.1051/0004-6361/201833099. Disponível em: <`https://doi.org/10.1051/0004-6361/201833099`>.

ANDRAE, R., FOUESNEAU, C., HARRISON, C. S., et al., 2018, “Gaia Data Release 2: First stellar parameters from Apsis”, *Astronomy & Astrophysics*, v. 616, pp. A8. doi: 10.1051/0004-6361/201732516. Disponível em: <`https://doi.org/10.1051/0004-6361/201732516`>.

ANDRAE, R., FOUESNEAU, M., SORDO, R., et al., 2023, “Gaia Data Release 3: Analysis of the Gaia BP/RP spectra using the General Stellar Parameterizer from Photometry”, *Astronomy & Astrophysics*, v. 674, pp. A27. doi: 10.1051/

0004-6361/202243462. Disponível em: <https://www.aanda.org/articles/aa/full_html/2023/06/aa43462-22/aa43462-22.html>.

AVDEEVA, A. S., KOVALEVA, D. A., MALKOV, O. Y., et al., 2024, “Quality Flags for GSP-Phot Gaia DR3 Astrophysical Parameters with Machine Learning: Effective Temperatures Case Study”, *Monthly Notices of the Royal Astronomical Society*, v. 527, n. 3, pp. 7382–7393. doi: 10.1093/mnras/stad3601.

AVENI, A. F., 2001, *Skywatchers: A Revised and Updated Version of Skywatchers of Ancient Mexico*. Austin, University of Texas Press.

AYRES, T. R., JOHNSON, H. R., 1977, “The surface gravity and mass of Arcturus”, Astrophys. J., v. 214 (jun.), pp. 410–417. doi: 10.1086/155265. Disponível em: <https://ui.adsabs.harvard.edu/abs/1977ApJ...214..410A>.

BABUSIAUX, C., FABRICIUS, C., KHANNA, S., et al., 2023, “Gaia Data Release 3. Catalogue validation”, Astron. Astrophys., 674:A32. doi: 10.1051/0004-6361/202243790.

BAI, Y., LIU, J., BAI, Z., et al., 2019, “Machine-Learning Regression of Stellar Effective Temperatures in the Second Gaia Data Release”, *The Astronomical Journal*, v. 158, n. 2, pp. 93. doi: 10.3847/1538-3881/ab3048.

BAILER-JONES, C. A. L., RYBIZKI, J., FOUESNEAU, M., et al., 2021, “Estimating Distances from Parallaxes. V. Geometric and Photogeometric Distances to 1.47 Billion Stars in Gaia Early Data Release 3”, Astron. J., 161(3):147. doi: 10.3847/1538-3881/abd806.

BALL, N. M., BRUNNER, R. J., 2010, “Data mining and machine learning in astronomy”, *International Journal of Modern Physics D*, v. 19, n. 07, pp. 1049–1106. doi: 10.1142/S0218271810017160.

BALL, N. M., LOVEDAY, J., FUKUGITA, M., et al., 2004, “Galaxy types in the Sloan digital sky survey using supervised artificial neural networks”, *Monthly Notices of the Royal Astronomical Society*, v. 348, n. 3, pp. 1038–1046.

BANERJI, M., LAHAV, O., LINTOTT, C. J., et al., 2010, “Galaxy Zoo: reproducing galaxy morphologies via machine learning”, Mon. Not, Roy. Astron. Soc., v. 406, n. 1 (jul.), pp. 342–353. doi: 10.1111/j.1365-2966.2010.16713.x.

BANSAL, S., PHAN, P., RAHMAN, Z., 2024, “Enhancing Stellar Temperature Estimation through Machine Learning and Multifaceted Data Exploration”, *Research Notes of the AAS*, v. 8, n. 4, pp. 97. doi: 10.3847/2515-5172/ad3919.

BARKLEM, P. S., 2002, "Detailed analysis of Balmer lines in cool dwarf stars", *Astronomy & Astrophysics*, v. 385, n. 3, pp. 951–967. doi: 10.1051/0004-6361:20020163. Disponível em: <https://doi.org/10.1051/0004-6361:20020163>.

BARON, D., 2019, "Machine Learning in Astronomy: a practical overview", *arXiv e-prints*, art. arXiv:1904.07248. doi: 10.48550/arXiv.1904.07248.

BAYES, T., 1763, "LII. An essay towards solving a problem in the doctrine of chances", *Philosophical Transactions of the Royal Society of London*, v. 53 (jan.), pp. 370–418. doi: 10.1098/rstl.1763.0053. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Richard Price, in a letter to John Canton, A. M. F. R. S.

BAYO, A., RODRIGO, C., BARRADO Y NAVASCUÉS, D., et al., 2008, "VOSA: virtual observatory SED analyzer. An application to the Collinder 69 open cluster", Astron. Astrophys., v. 492, n. 1 (dez.), pp. 277–287. doi: 10.1051/0004-6361:200810395.

BAZELL, D., AHA, D. W., 2001, "Ensembles of Classifiers for Morphological Galaxy Classification", *The Astrophysical Journal*, v. 548, n. 1, pp. 219. doi: 10.1086/318696.

BELLM, E. C., KULKARNI, S. R., GRAHAM, M. J., et al., 2019, "The Zwicky Transient Facility: System Overview, Performance, and First Results", *Publications of the Astronomical Society of the Pacific*, v. 131, n. 995, pp. 018002. doi: 10.1088/1538-3873/aaecbe.

BENITEZ, N., DUPKE, R., MOLES, M., et al., 2014, "J-PAS: The Javalambre-Physics of the Accelerated Universe Astrophysical Survey", *arXiv e-prints*, art. arXiv:1403.5237. doi: 10.48550/arXiv.1403.5237.

BERGEMANN, M., LIND, K., COLLET, R., et al., 2012, "Non-LTE line formation of Fe in late-type stars. I. Standard stars with 1D and ⟨3D⟩ model atmospheres", Mon. Not, Roy. Astron. Soc., v. 427, n. 1 (nov.), pp. 27–49. doi: 10.1111/j.1365-2966.2012.21687.x. Disponível em: <https://doi.org/10.1111/j.1365-2966.2012.21687.x>.

BERGER-TAL, O., NATHAN, J., MERON, E., et al., 2014, "The Exploration-Exploitation Dilemma: A Multidisciplinary Framework", *PLOS ONE*, v. 9, n. 4, pp. e95693. doi: 10.1371/journal.pone.0095693. Disponível em: <https://doi.org/10.1371/journal.pone.0095693>.

BERTIN, E., ARNOUTS, S., 1996, “SExtractor: Software for source extraction.” Astron. Astrophys. Suppl., v. 117 (jun.), pp. 393–404. doi: 10.1051/aas:1996164.

BESSELL, M. S., 2008, “Photometric methods for stellar parameter determinations”, *Physica Scripta*, 2008(T133):014019. doi: 10.1088/0031-8949/2008/T133/014019. Disponível em: <https://doi.org/10.1088/0031-8949/2008/T133/014019>.

BICA, E., PAVANI, D. B., BONATTO, C. J., et al., 2019, “A Multi-band Catalog of 10978 Star Clusters, Associations, and Candidates in the Milky Way”, *The Astronomical Journal*, v. 157, n. 1, pp. 12. doi: 10.3847/1538-3881/aaef8d. Disponível em: <https://doi.org/10.3847/1538-3881/aaef8d>.

BONATTO, C., KERBER, L. O., BICA, E., et al., 2006, “Probing disk properties with open clusters”, *Astronomy & Astrophysics*, v. 446, pp. 121–135. doi: 10.1051/0004-6361:20053573. Disponível em: <https://doi.org/10.1051/0004-6361:20053573>.

BONATTO, C., CHIES-SANTOS, A. L., COELHO, P. R. T., et al., 2019, “J-PLUS: A wide-field multi-band study of the M 15 globular cluster. Evidence of multiple stellar populations in the RGB”, Astron. Astrophys., 622:A179. doi: 10.1051/0004-6361/201732441.

BONNOR, W. B., 1957, “Jeans’ Formula for Gravitational Instability”, *Monthly Notices of the Royal Astronomical Society*, v. 117, pp. 104–117. doi: 10.1093/mnras/117.1.104.

BONOLI, S., MARÍN-FRANCH, A., VARELA, J., et al., 2021, “The miniJPAS survey: A preview of the Universe in 56 colors”, Astron. Astrophys., 653:A31. doi: 10.1051/0004-6361/202038841.

BORISOV, V., LEEMANN, T., SESSLER, K., et al., 2024, “Deep Neural Networks and Tabular Data: A Survey”, *IEEE Transactions on Neural Networks and Learning Systems*, v. 35, n. 6, pp. 7499–7519. doi: 10.1109/TNNLS.2022.3229161.

BORUCKI, W. J., KOCH, D., BASRI, G., et al., 2010, “Kepler planet-detection mission: Introduction and first results”, *Science*, v. 327, n. 5968 (fev.), pp. 977–980. doi: 10.1126/science.1185402. Disponível em: <https://ui.adsabs.harvard.edu/abs/2010Sci...327..977B>.

BOVY, J., 2016, “THE CHEMICAL HOMOGENEITY OF OPEN CLUSTERS”, *The Astrophysical Journal*, v. 817, n. 1 (jan.), pp. 49. ISSN: 0004-637X. doi: 10.3847/0004-637X/817/1/49.

BREIMAN, L., 1996, "Bagging predictors", *Machine learning*, v. 24, n. 2, pp. 123–140. doi: https://doi.org/10.1007/BF00058655.

BREIMAN, L., 2001, "Random Forests", *Machine Learning*, v. 45, n. 1 (October), pp. 5–32. doi: 10.1023/A:1010933404324. Disponível em: <`https://doi.org/10.1023/A:1010933404324`>.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., et al., 1984, *Classification and Regression Trees.* 1st ed. New York, Chapman and Hall/CRC. doi: 10.1201/9781315139470. Disponível em: <`https://doi.org/10.1201/9781315139470`>.

BRESSAN, A., MARIGO, P., GIRARDI, L., et al., 2012, "PARSEC: Stellar tracks and isochrones with the PAdova and TRieste Stellar Evolution Code", Mon. Not, Roy. Astron. Soc., v. 427, n. 1 (nov.), pp. 127–145. doi: 10.1111/j.1365-2966.2012.21948.x. Disponível em: <`https://ui.adsabs.harvard.edu/abs/2012MNRAS.427..127B`>.

BRETON, S. N., SANTOS, A. R. G., BUGNET, L., et al., 2021, "ROOSTER: A Machine-Learning Analysis Tool for Kepler Stellar Rotation Periods", *Astronomy & Astrophysics*, v. 647, pp. A125. doi: 10.1051/0004-6361/202039947.

BUDER, S., KOS, J., WANG, E. X., et al., 2025. "The GALAH Survey: Data Release 4". Disponível em: <`https://arxiv.org/abs/2409.19858`>.

BUDER, S., BLAND-HAWTHORN, J., SHARMA, S., et al., 2021, "The GALAH+ survey: Third data release", *Monthly Notices of the Royal Astronomical Society*, v. 506, n. 1, pp. 150–174. doi: 10.1093/mnras/stab1242.

BÜHLMANN, P., HOTHORN, T., 2007, "Boosting Algorithms: Regularization, Prediction and Model Fitting", *Statistical Science*, v. 22, n. 4 (November), pp. 477–505. doi: 10.1214/07-STS242. Disponível em: <`https://doi.org/10.1214/07-STS242`>.

CANNON, A. J., 1901, "Preliminary Classification of 1 225 Stars", *Harvard College Observatory Circular*, v. 54, pp. 1–3.

CANNON, A. J., PICKERING, E. C., 1918, "The Henry Draper Catalogue", *Annals of Harvard College Observatory*, v. 91, pp. 1–473.

CANTAT-GAUDIN, T., ANDERS, F., 2020, "Clusters and mirages: cataloguing stellar aggregates in the Milky Way", Astron. Astrophys., 633:A99. doi: 10.1051/0004-6361/201936691.

CANTAT-GAUDIN, T., ANDERS, F., CASTRO-GINARD, A., et al., 2020, “Painting a portrait of the Galactic disc with its stellar clusters”, *Astronomy & Astrophysics*, v. 640, pp. A1. doi: 10.1051/0004-6361/202038192. Disponível em: <https://doi.org/10.1051/0004-6361/202038192>.

CARRARO, G., NG, Y. K., PORTINARI, L., 1998, “On the Galactic disc age-metallicity relation”, Mon. Not, Roy. Astron. Soc., v. 296, n. 4 (jun.), pp. 1045–1056. doi: 10.1046/j.1365-8711.1998.01460.x. Disponível em: <https://ui.adsabs.harvard.edu/abs/1998MNRAS.296.1045C>.

CARROLL, B. W., OSTLIE, D. A., 2017, *An Introduction to Modern Astrophysics*. 2 ed. Cambridge, Cambridge University Press. ISBN: 9781108422161.

CASTRO-GINARD, A., JORDI, C., LURI, X., et al., 2018, “A new method for unveiling open clusters in Gaia. New nearby open clusters confirmed by DR2”, Astron. Astrophys., 618:A59. doi: 10.1051/0004-6361/201833390. Disponível em: <https://ui.adsabs.harvard.edu/abs/2018A&A...618A..59C>.

CASTRO-GINARD, A., JORDI, C., LURI, X., et al., 2019, “Hunting for open clusters in Gaia DR2: the Galactic anticentre”, Astron. Astrophys., 627:A35. doi: 10.1051/0004-6361/201935531. Disponível em: <https://ui.adsabs.harvard.edu/abs/2019A&A...627A..35C>.

CASTRO-GINARD, A., PENOYRE, Z., CASEY, A. R., et al., 2024, “Gaia DR3 detectability of unresolved binary systems”, *Astronomy & Astrophysics*, v. 688, pp. A1. doi: 10.1051/0004-6361/202450172. Disponível em: <https://doi.org/10.1051/0004-6361/202450172>. Published online: 30 July 2024.

CATANZARO, G., LEONE, F., DALL, T. H., 2004, “Balmer lines as T_{eff} and log g indicators for non-solar composition atmospheres: An application to the extremely helium-weak star HR 6000”, Astron. Astrophys., v. 425 (out.), pp. 641–648. doi: 10.1051/0004-6361:20040558. Disponível em: <https://doi.org/10.1051/0004-6361:20040558>.

CAUCHY, A., 1847, “Méthode générale pour la résolution des systèmes d’équations simultanées”, *Comptes Rendus Hebdomadaires des Séances de l’Académie des Sciences*, v. 25, pp. 536–538.

CAVALLO, L., SPINA, L., CARRARO, G., et al., 2023, “Parameter Estimation for Open Clusters using an Artificial Neural Network with a QuadTree-based Feature Extractor”, *The Astronomical Journal*, v. 167, n. 1 (dez.), pp. 12. ISSN: 1538-3881. doi: 10.3847/1538-3881/ad07e5.

CAYREL, R., VAN 'T VEER-MENNERET, C., ALLARD, N. F., et al., 2011, "The Hα Balmer line as an effective temperature criterion. I. Calibration using 1D model stellar atmospheres", *Astronomy & Astrophysics*, v. 531, pp. A83. doi: 10.1051/0004-6361/201116911. Disponível em: <https://doi.org/10.1051/0004-6361/201116911>.

CENARRO, A. J., MOLES, M., CRISTÓBAL-HORNILLOS, D., et al., 2010, "The Javalambre Astrophysical Observatory project". In: Angeli, G. Z., Dierickx, P. (Eds.), *Modeling, Systems Engineering, and Project Management for Astronomy IV*, v. 7738, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, p. 77380V, jul. doi: 10.1117/12.861562.

CENARRO, A. J., MOLES, M., CRISTÓBAL-HORNILLOS, D., et al., 2012, "The Observatorio Astrofísico de Javalambre: goals and current status". In: Peck, A. B., Seaman, R. L., Comeron, F. (Eds.), *Observatory Operations: Strategies, Processes, and Systems IV*, v. 8448, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, p. 84481A, set. doi: 10.1117/12.925432.

CENARRO, A. J., MOLES, M., MARÍN-FRANCH, A., et al., 2014, "The Observatorio Astrofísico de Javalambre: current status, developments, operations, and strategies". In: Peck, A. B., Benn, C. R., Seaman, R. L. (Eds.), *Observatory Operations: Strategies, Processes, and Systems V*, v. 9149, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, p. 91491I, ago. doi: 10.1117/12.2055455.

CENARRO, A. J., MOLES, M., CRISTÓBAL-HORNILLOS, D., et al., 2019, "J-PLUS: The Javalambre Photometric Local Universe Survey", Astron. Astrophys., 622: A176. doi: 10.1051/0004-6361/201833036.

CHAISSON, E., MCMILLAN, S., 2005, *Astronomy Today*. 5 ed. Upper Saddle River, NJ, Prentice Hall. ISBN: 9780131445963.

CHEN, S.-X., SUN, W.-M., HE, Y., 2022, "Application of Random Forest Regressions on Stellar Parameters of A-type Stars and Feature Extraction", *Research in Astronomy and Astrophysics*, v. 22, n. 2, pp. 025017. doi: 10.1088/1674-4527/ac41c5.

CHEN, T., GUESTRIN, C., 2016, "XGBoost: A scalable tree boosting system", *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794.

CHIAPPINI, C., MATTEUCCI, F., GRATTON, R., 1997, "The chemical evolution of the Galaxy: The two-infall model", Astrophys. J., v. 477 (mar.), pp. 765–780.

doi: 10.1086/303726. Disponível em: <https://ui.adsabs.harvard.edu/abs/1997ApJ...477..765C>.

CHIERICHINI, S., LIU, J., KORSÓS, M. B., et al., 2024, "CME Arrival Modeling with Machine Learning", Astrophys. J., 963(2):121. doi: 10.3847/1538-4357/ad1cee.

CHOI, J., DOTTER, A., CONROY, C., et al., 2016, "Mesa Isochrones and Stellar Tracks (MIST). I. Solar-scaled models", Astrophys. J., 823(2):102. doi: 10.3847/0004-637X/823/2/102. Disponível em: <https://ui.adsabs.harvard.edu/abs/2016ApJ...823..102C>.

CHRISTENSEN-DALSGAARD, J., 2021, "Solar structure and evolution", *Living Reviews in Solar Physics*, v. 18, pp. 2. doi: 10.1007/s41116-020-00028-3. Disponível em: <https://doi.org/10.1007/s41116-020-00028-3>.

COLLISTER, A. A., LAHAV, O., 2004, "ANNz: Estimating photometric redshifts using artificial neural networks", *Publications of the Astronomical Society of the Pacific*, v. 116, n. 823, pp. 345–351.

CONDORCET, N. D., 2014, *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Cambridge Library Collection - Mathematics. Cambridge, Cambridge University Press. ISBN: 9781139923972. doi: 10.1017/CBO9781139923972. Disponível em: <https://doi.org/10.1017/CBO9781139923972>. First published in 1785.

CORDEIRO DA SILVA, M. V. E., BORGES, M., 2023, *IDENTIFICAÇÃO DE SUBANÃS QUENTES EM LEVANTAMENTOS ASTRONÔMICOS*. Tese de Mestrado, Observatorio Nacional, Rio de Janeiro, Brazil. Disponível em: <http://localhost:8080/tede/handle/tede/154>. Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Astronomia, Divisão de Programas de Pós-Graduação - DIPPG.

CRAWFORD, D. L., 1958, "Two-Dimensional Spectral Classification by Narrow-Band Photometry for B Stats in Clusters and Associations." Astrophys. J., v. 128 (set.), pp. 185. doi: 10.1086/146536.

CREEVEY, O. L., SORDO, R., PAILLER, F., et al., 2023, "Gaia Data Release 3. Astrophysical parameters inference system (Apsis). I. Methods and content overview", Astron. Astrophys., 674:A26. doi: 10.1051/0004-6361/202243688.

CREVIER, D., 1993, *AI: the tumultuous history of the search for artificial intelligence*. USA, Basic Books, Inc. ISBN: 0465029973.

CROPPER, M., KATZ, D., SARTORETTI, P., et al., 2018, "Gaia Data Release 2. Gaia Radial Velocity Spectrometer", Astron. Astrophys., 616:A5. doi: 10.1051/0004-6361/201832763.

CUI, X.-Q., ZHAO, Y.-H., CHU, Y.-Q., et al., 2012, "The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST)", *Research in Astronomy and Astrophysics*, v. 12, pp. 1197.

CUNTZ, M., WANG, Z., 2018, "The Mass–Luminosity Relation for a Refined Set of Late-K/M Dwarfs", *Research Notes of the AAS*, v. 2, n. 1, pp. 19. doi: 10.3847/2515-5172/aaaa67.

CUTRI, R. M., WRIGHT, E. L., CONROW, T., et al., 2012. "VizieR Online Data Catalog: WISE All-Sky Data Release (Cutri+ 2012)". VizieR On-line Data Catalog: II/311. Originally published in: 2012wise.rept....1C, abr. Authors from "Explanatory Supplement to the WISE All-Sky Data Release Products" in https://wise2.ipac.caltech.edu/docs/release/allsky/expsup/index.html.

CUTRI, R. M., WRIGHT, E. L., CONROW, T., et al., 2014. "VizieR Online Data Catalog: AllWISE Data Release (Cutri+ 2013)". VizieR On-line Data Catalog: II/328. Originally published in: IPAC/Caltech (2013), fev. updated version, 16-Feb-2021.

DA SILVA, A. R., SMILJANIC, R., 2025, "Discovery of a pair of very metal-poor stars enriched in neutron-capture elements: The proto-disk r-II star BPS CS 29529-0089 and the Gaia-Sausage-Enceladus r-I star TYC 9219-2422-1", *Astronomy & Astrophysics*, v. 696, pp. A122. doi: 10.1051/0004-6361/202453295. Disponível em: <https://doi.org/10.1051/0004-6361/202453295>.

DALTON, G., THE WEAVE CONSORTIUM, 2012, "WEAVE: the next generation wide-field spectroscopy facility for the William Herschel Telescope". In: *Ground-based and Airborne Instrumentation for Astronomy IV*, v. 8446, p. 84460P, September. doi: 10.1117/12.925950. for the WEAVE Consortium.

DAVIES, B., KUDRITZKI, R.-P., PLEZ, B., et al., 2013, "The Temperatures of Red Supergiants", *The Astrophysical Journal*, v. 767, n. 1, pp. 3. doi: 10.1088/0004-637X/767/1/3. Disponível em: <https://doi.org/10.1088/0004-637X/767/1/3>.

DAVIS, A. B., CISEWSKI, J., DUMUSQUE, X., et al., 2017, "Insights on the Spectral Signatures of Stellar Activity and Planets from PCA", *The Astrophysical Journal*, v. 846, n. 1, pp. 59. doi: 10.3847/1538-4357/aa8303. Disponível em: <https://doi.org/10.3847/1538-4357/aa8303>.

DAVIS, M., GUHATHAKURTA, P., KONIDARIS, N. P., et al., 2007, "The All-Wavelength Extended Groth Strip International Survey (AEGIS) Data Sets", Astrophys. J. Lett., v. 660, n. 1 (maio), pp. L1–L6. doi: 10.1086/517931.

DE JONG, R., 4MOST CONSORTIUM, 2019, "4MOST: Project overview and information for the First Call for Proposals", *The Messenger*, v. 175, pp. 3–11. doi: 10.18727/0722-6691/5117. for the 4MOST Consortium.

DE LA CALLEJA, J., FUENTES, O., 2004, "Machine learning and image analysis for morphological galaxy classification", *Monthly Notices of the Royal Astronomical Society*, v. 349, n. 1, pp. 87–93. doi: 10.1111/j.1365-2966.2004.07442.x.

DEEMING, T. J., 1964, "Stellar spectral classification, I." Mon. Not, Roy. Astron. Soc., v. 127 (jan.), pp. 493. doi: 10.1093/mnras/127.6.493.

DEPUYDT, L., 1997, *Civil Calendar and Lunar Calendar in Ancient Egypt*. Leuven, Orientalia Lovaniensia Analecta.

DIAS, W. S., MONTEIRO, H., MOITINHO, A., et al., 2021, "Updated parameters of 1743 open clusters based on Gaia DR2", *Monthly Notices of the Royal Astronomical Society*, v. 504, pp. 356–371. doi: 10.1093/mnras/stab770. Disponível em: <https://doi.org/10.1093/mnras/stab770>.

DIELEMAN, S., WILLETT, K. W., DAMBRE, J., 2015, "Rotation-invariant convolutional neural networks for galaxy morphology prediction", *Monthly Notices of the Royal Astronomical Society*, v. 450, n. 2, pp. 1441–1459. doi: 10.1093/mnras/stv632.

DIETTERICH, T. G., 2000, "Ensemble Methods in Machine Learning". In: Kittler, J., Roli, F. (Eds.), *Multiple Classifier Systems*, v. 1857, *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, pp. 1–15. ISBN: 978-3-540-67704-8. doi: 10.1007/3-540-45014-9_1.

DING, M.-Y., SHI, J.-R., WU, Y., et al., 2022, "Stellar Atmospheric Parameters of M-type Stars from LAMOST DR8", *The Astrophysical Journal Supplement Series*, v. 260, n. 2, pp. 45. doi: 10.3847/1538-4365/ac6754. Disponível em: <https://doi.org/10.3847/1538-4365/ac6754>.

DJORGOVSKI, S. G., MAHABAL, A., DRAKE, A., et al., 2013, "Sky Surveys". In: Oswalt, T. D., Bond, H. E. (Eds.), *Planets, Stars and Stellar Systems*, Springer, Dordrecht. doi: 10.1007/978-94-007-5618-2_5.

DOMÍNGUEZ SÁNCHEZ, H., HUERTAS-COMPANY, M., BERNARDI, M., et al., 2018, "Improving Galaxy Morphologies for SDSS with Deep Learning", *Monthly*

Notices of the Royal Astronomical Society, v. 476, n. 3, pp. 3661–3676. doi: 10.1093/mnras/sty338.

D'ISANTO, A., POLSTERER, K. L., 2018, "Photometric redshift estimation via deep learning: Generalized and pre-classification-less, image based, fully probabilistic redshifts", *Astronomy & Astrophysics*, v. 609, pp. A111. doi: 10.1051/0004-6361/201731326.

EDVARDSSON, B., 1988, "Spectroscopic surface gravities and chemical compositions for 8 nearby single sub-giants", Astron. Astrophys., v. 190 (jan.), pp. 148–166. Disponível em: <https://ui.adsabs.harvard.edu/abs/1988A&A...190..148E>.

EDVARDSSON, B., ANDERSEN, J., GUSTAFSSON, B., et al., 1993, "The chemical evolution of the Galactic disk. I. Analysis and results", Astron. Astrophys., v. 275 (ago.), pp. 101–152. Disponível em: <https://ui.adsabs.harvard.edu/abs/1993A&A...275..101E>.

EISENHARDT, P. R. M., MAROCCO, F., FOWLER, J. W., et al., 2020, "The CatWISE Preliminary Catalog: Motions from WISE and NEOWISE Data", Astrophys. J. Suppl., 247(2):69. doi: 10.3847/1538-4365/ab7f2a.

EL-BADRY, K., RIX, H.-W., CENDES, Y., et al., 2023, "A red giant orbiting a black hole", Mon. Not, Roy. Astron. Soc., v. 521, n. 3 (maio), pp. 4323–4348. doi: 10.1093/mnras/stad799.

EL-KHOLY, R. I., HAYMAN, Z. M., 2025, "Optimised sampling of SDSS-IV MaStar spectra for stellar classification using supervised models", *Astronomy & Astrophysics*, v. 693, pp. A300. doi: 10.1051/0004-6361/202451309. Disponível em: <https://doi.org/10.1051/0004-6361/202451309>.

FISHER, R. A., 1922, "On the Mathematical Foundations of Theoretical Statistics", *Philosophical Transactions of the Royal Society of London. Series A*, v. 222, n. 594–604 (jan.), pp. 309–368. doi: 10.1098/rsta.1922.0009. Disponível em: <http://dx.doi.org/10.1098/rsta.1922.0009>.

FLEMING, W. P., 1890, "Stars Having Peculiar Spectra", *Annals of the Astronomical Observatory of Harvard College*, v. 18, pp. 1–19.

FLUKE, C. J., JACOBS, C., 2020, "Surveying the reach and maturity of machine learning and artificial intelligence in astronomy", *WIREs Data Mining and Knowledge Discovery*, 10(2):e1349. doi: 10.1002/widm.1349.

FOLKES, S. R., LAHAV, O., MADDOX, S. J., 1996, "An artificial neural network approach to the classification of galaxy spectra." Mon. Not, Roy. Astron. Soc., v. 283, n. 2 (dez.), pp. 651–665. doi: 10.1093/mnras/283.2.651.

FOTOPOULOU, S., 2024, "A review of unsupervised learning in astronomy", *Astronomy and Computing*, v. 48 (July), pp. 100851. doi: 10.1016/j.ascom.2024.100851. Disponível em: <`https://doi.org/10.1016/j.ascom.2024.100851`>.

FREUND, Y., SCHAPIRE, R. E., 1997, "A decision-theoretic generalization of on-line learning and an application to boosting", *Journal of computer and system sciences*, v. 55, n. 1, pp. 119–139.

FRIEDMAN, J. H., 2001, "Greedy function approximation: A gradient boosting machine", *Annals of Statistics*, v. 29, n. 5, pp. 1189–1232. doi: 10.1214/aos/1013203451.

FUHRMANN, K., 1998, "Nearby stars of the Galactic disk and halo", Astron. Astrophys., v. 338 (out.), pp. 161–183. Disponível em: <`https://ui.adsabs.harvard.edu/abs/1998A&A...338..161F`>.

FUHRMANN, K., AXER, M., GEHREN, T., 1993, "Balmer lines in cool dwarf stars. I. Basic influence of atmospheric models", *Astronomy & Astrophysics*, v. 271, pp. 451–462. Disponível em: <`https://ui.adsabs.harvard.edu/abs/1993A&A...271..451F`>. Provided by the SAO/NASA Astrophysics Data System.

FUHRMANN, K., PFEIFFER, M., FRANK, C., et al., 1997, "The surface gravities of cool dwarf stars revisited", Astron. Astrophys., v. 323 (jul.), pp. 909–922. Disponível em: <`https://ui.adsabs.harvard.edu/abs/1997A&A...323..909F`>.

FUJII, M. S., ZWART, S. P., 2016, "The formation and dynamical evolution of young star clusters", Astrophys. J., 817(1):4. doi: 10.3847/0004-637X/817/1/4. Disponível em: <`https://ui.adsabs.harvard.edu/abs/2016ApJ...817....4F`>.

GAIA COLLABORATION, 2022, "VizieR Online Data Catalog: Gaia DR3 Part 1. Main source (Gaia Collaboration, 2022)", *VizieR Online Data Catalog*, art. I/355. doi: 10.26093/cds/vizier.1355.

GAIA COLLABORATION, BROWN, A. G. A., VALLENARI, A., et al., 2016a, "Gaia Data Release 1. Summary of the astrometric, photometric, and survey properties", Astron. Astrophys., 595:A2. doi: 10.1051/0004-6361/201629512.

GAIA COLLABORATION, PRUSTI, T., DE BRUIJNE, J. H. J., et al., 2016b, "The Gaia mission", Astron. Astrophys., 595:A1. doi: 10.1051/0004-6361/201629272.

GAIA COLLABORATION, BROWN, A. G. A., VALLENARI, A., et al., 2018a, "Gaia Data Release 2. Summary of the contents and survey properties", Astron. Astrophys., 616:A1. doi: 10.1051/0004-6361/201833051.

GAIA COLLABORATION, MIGNARD, F., KLIONER, S. A., et al., 2018b, "Gaia Data Release 2. The celestial reference frame (Gaia-CRF2)", Astron. Astrophys., 616: A14. doi: 10.1051/0004-6361/201832916.

GAIA COLLABORATION, BROWN, A. G. A., VALLENARI, A., et al., 2021, "Gaia Early Data Release 3. Summary of the contents and survey properties", Astron. Astrophys., 649:A1. doi: 10.1051/0004-6361/202039657.

GAIA COLLABORATION, VALLENARI, A., BROWN, A. G. A., et al., 2023, "Gaia Data Release 3. Summary of the content and survey properties", Astron. Astrophys., 674:A1. doi: 10.1051/0004-6361/202243940.

GALARZA, C. A., DAFLON, S., PLACCO, V. M., et al., 2022, "J-PLUS: Searching for very metal-poor star candidates using the SPEEM pipeline", Astron. Astrophys., 657:A35. doi: 10.1051/0004-6361/202141717.

GARS, S. L., 2015, "Spectroscopy, History of". In: Gargaud, M., others (Eds.), *Encyclopedia of Astrobiology*, Springer, Berlin, Heidelberg. ISBN: 978-3-662-44184-8. doi: 10.1007/978-3-662-44185-5_1490. Disponível em: <https://doi.org/10.1007/978-3-662-44185-5_1490>.

GAVEL, A., ANDRAE, R., FOUESNEAU, M., et al., 2021, "Estimating [α/Fe] from Gaia low-resolution BP/RP spectra using the ExtraTrees algorithm", *Astronomy & Astrophysics*, v. 656, pp. A93. doi: 10.1051/0004-6361/202141589. Disponível em: <https://doi.org/10.1051/0004-6361/202141589>.

GHOJOGH, B., CROWLEY, M., 2019, "The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial", *arXiv e-prints*, art. arXiv:1905.12787. doi: 10.48550/arXiv.1905.12787.

GILMORE, G., RANDICH, S., ASPLUND, M., et al., 2012, "The Gaia-ESO Public Spectroscopic Survey", *The Messenger*, v. 147 (mar.), pp. 25–31.

GILMORE, G., RANDICH, S., WORLEY, C. C., et al., 2022, "The Gaia-ESO Public Spectroscopic Survey", *Astronomy & Astrophysics*, v. 666, pp. A120. doi: 10.1051/0004-6361/202243134. Disponível em: <https://www.aanda.org/articles/aa/full_html/2022/10/aa43134-22/aa43134-22.html>.

GORLOVA, N. I., MEYER, M. R., RIEKE, G. H., et al., 2003, "Gravity indicators in the near-infrared spectra of brown dwarfs", Astrophys. J., v. 593, n. 2 (ago.), pp. 1074–1092. doi: 10.1086/376730. Disponível em: <https://doi.org/10.1086/376730>.

GRASSHOFF, G., 1990, *The History of Ptolemy's Star Catalogue*, v. 14, *Studies in the History of Mathematics and Physical Sciences*. New York, Springer. doi: 10.1007/978-1-4612-3468-6.

GRAY, D. F., 2005, *The Observation and Analysis of Stellar Photospheres*. 3 ed. Cambridge, Cambridge University Press. ISBN: 9780521851863. doi: 10.1017/CBO9781316036570. Disponível em: <https://doi.org/10.1017/CBO9781316036570>.

GRAY, D. F., JOHANSON, H. L., 1991, "Precise Measurement of Stellar Temperatures Using Line-Depth Ratios", *Publications of the Astronomical Society of the Pacific*, v. 103, n. 663, pp. 439–447. doi: 10.1086/132839. Disponível em: <https://doi.org/10.1086/132839>.

GREENSTEIN, J. L., SARGENT, A. I., 1974, "Photoelectric observations of faint blue stars. IV", *Astrophysical Journal Supplement Series*, v. 28, pp. 157. doi: 10.1086/190318.

GUTIÉRREZ-SOTO, L. A., DE OLIVEIRA, R. L., AKRAS, S., et al., 2025, "Mapping Hα excess candidate point sources in the southern hemisphere using S-PLUS data", *Astronomy & Astrophysics*, v. 695, pp. A104. doi: 10.1051/0004-6361/202453167. Disponível em: <https://doi.org/10.1051/0004-6361/202453167>.

HAFEZ, I., 2010, *The Influence of Islamic Astronomy on the European Renaissance*. Phd thesis, James Cook University. Disponível em: <https://researchonline.jcu.edu.au/28854/>.

HAIGH, T., 2023, "There Was No 'First AI Winter"', *Commun. ACM*, v. 66, n. 12 (nov.), pp. 35–39. ISSN: 0001-0782. doi: 10.1145/3625833. Disponível em: <https://doi.org/10.1145/3625833>.

HARTMANN, E. A., BONATTO, C. J., CHIES-SANTOS, A. L., et al., 2022, "S-PLUS: exploring wide field properties of multiple populations in galactic globular clusters at different metallicities", Mon. Not, Roy. Astron. Soc., v. 515, n. 3 (set.), pp. 4191–4200. doi: 10.1093/mnras/stac1411.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J., 2009, *The elements of statistical learning*. Springer Series in Statistics. 2nd ed. New York, NY, Springer. ISBN: 978-0-387-84857-0. doi: 10.1007/978-0-387-84858-7.

HE, X.-J., LUO, A. L., CHEN, Y.-Q., 2022, “Identification, mass, and age of primary red clump stars from spectral features derived with the LAMOST DR7”, Mon. Not, Roy. Astron. Soc., v. 512, n. 2 (maio), pp. 1710–1721. doi: 10.1093/mnras/stac484.

HERTZSPRUNG, E., 1911, “On the Use of Photographic Effective Wavelengths for the Determination of Color Equivalents”, *Publications of the Astrophysical Observatory in Potsdam*, v. 22, n. 63, pp. 1–.

HEZAVEH, Y. D., LEVASSEUR, L. P., MARSHALL, P. J., 2017, “Fast Automated Analysis of Strong Gravitational Lenses with Convolutional Neural Networks”, *Nature*, v. 548, n. 7669, pp. 555–557. doi: 10.1038/nature23463.

HIDALGO, S. L., PIETRINFERNI, A., CASSISI, S., et al., 2018, “The updated BaSTI stellar evolution models and isochrones. I. Solar-scaled calculations”, Astrophys. J., 856(2):125. doi: 10.3847/1538-4357/aab158. Disponível em: <`https://ui.adsabs.harvard.edu/abs/2018ApJ...856..125H`>.

HO, T. K., 1995, “Random decision forests”. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*, v. 1, pp. 278–282 vol.1, August. doi: 10.1109/ICDAR.1995.598994.

HOGG, D. W., CASEY, A. R., NESS, M., et al., 2016, “Chemical Tagging Can Work: Identification of Stellar Phase-space Structures Purely by Chemical-abundance Similarity”, *The Astrophysical Journal*, v. 833, n. 2, pp. 262. doi: 10.3847/1538-4357/833/2/262. Disponível em: <`https://doi.org/10.3847/1538-4357/833/2/262`>.

HOLTZMAN, J. A., NIDEVER, D. L., SHETRONE, M., et al., 2018, “Abundances, Stellar Parameters, and Spectra from the SDSS-IV APOGEE-2 Survey”, *The Astronomical Journal*, v. 156, n. 3, pp. 125. doi: 10.3847/1538-3881/aad4f9.

HOURIHANE, A., FRANÇOIS, P., WORLEY, C. C., et al., 2023, “The Gaia-ESO Survey: Homogenisation of stellar parameters and elemental abundances”, Astron. Astrophys., 676:A129. doi: 10.1051/0004-6361/202345910.

HOWELL, S. B., SOBECK, C., HAAS, M., et al., 2014, “The K2 Mission: Characterization and Early Results”, Publ. Astron. Soc. Pacific, v. 126, n. 938 (abr.), pp. 398. doi: 10.1086/676406.

HUANG, Y., BEERS, T. C., XIAO, K., et al., 2024, “J-PLUS: Beyond Spectroscopy. III. Stellar Parameters and Elemental-abundance Ratios for Five Million Stars from DR3”, Astrophys. J., 974(2):192. doi: 10.3847/1538-4357/ad6b94.

HUNT, E. L., REFFERT, S., 2021, “Improving the open cluster census. I. Comparison of clustering algorithms applied to Gaia DR2 data”, *Astronomy & Astrophysics*, v. 646, pp. A104. doi: 10.1051/0004-6361/202039341. Disponível em: <https://doi.org/10.1051/0004-6361/202039341>.

HUNT, E. L., REFFERT, S., 2023, “Improving the open cluster census. II. An all-sky cluster catalogue with Gaia DR3”, Astron. Astrophys., 673:A114. doi: 10.1051/0004-6361/202346285.

HUNT, E. L., REFFERT, S., 2024, “Improving the open cluster census - III. Using cluster masses, radii, and dynamics to create a cleaned open cluster catalogue”, *A&A*, v. 686, pp. A42. doi: 10.1051/0004-6361/202348662. Disponível em: <https://doi.org/10.1051/0004-6361/202348662>.

ISHIDA, E. E. O., DE SOUZA, R. S., 2013, “Kernel PCA for Type Ia supernovae photometric classification”, *Monthly Notices of the Royal Astronomical Society*, v. 430, n. 1, pp. 509–532. doi: 10.1093/mnras/sts650. Disponível em: <https://doi.org/10.1093/mnras/sts650>.

IVEZIĆ, Î, THE LSST COLLABORATION, 2019, “LSST: From Science Drivers to Reference Design and Anticipated Data Products”, *The Astrophysical Journal*, v. 873, n. 2, pp. 111. doi: 10.3847/1538-4357/ab042c. for the LSST Collaboration.

IVEZIĆ, Î, CONNOLLY, A. J., VANDERPLAS, J. T., et al., 2020, *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data, Updated Edition*. Princeton Series in Modern Observational Astronomy. Revised ed. Princeton, NJ, Princeton University Press. doi: 10.2307/j.ctvrxk1hs. Disponível em: <https://www.jstor.org/stable/j.ctvrxk1hs>.

JEANS, J. H., 1902, “The Stability of a Spherical Nebula”, *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, v. 199, pp. 1–53. doi: 10.1098/rsta.1902.0012.

JEANS, J. H., 1928, *Astronomy and Cosmogony*. Cambridge, Cambridge University Press.

JOHNSON, H. L., 1964, “The colors, Bolometric corrections and effective temperatures of the bright stars”, *Boletin de los Observatorios Tonantzintla y Tacubaya*, v. 3 (jun.), pp. 305–324.

JOHNSON, H. L., MORGAN, W. W., 1953, "Fundamental stellar photometry for standards of spectral type on the Revised System of the Yerkes Spectral Atlas." Astrophys. J., v. 117 (maio), pp. 313. doi: 10.1086/145697.

JOHNSON, H. L., MITCHELL, R. I., IRIARTE, B., et al., 1966, "UBVRIJKL Photometry of the Bright Stars", *Communications of the Lunar and Planetary Laboratory*, v. 4 (jan.), pp. 99–110.

JORDI, C., GEBRAN, M., CARRASCO, J. M., et al., 2010, "Gaia broad band photometry", Astron. Astrophys., 523:A48. doi: 10.1051/0004-6361/201015441.

JR., M. W. H., KUDRITZKI, R.-P., BRESOLIN, F., et al., 2014, "Quantitative spectroscopy of blue supergiants in metal-poor dwarf galaxy NGC 3109", Astrophys. J., 785(2):151. doi: 10.1088/0004-637X/785/2/151. Disponível em: <https://ui.adsabs.harvard.edu/abs/2014ApJ...785..151H>.

JÖNSSON, H., NIDEVER, D. L., SHETRONE, M., et al., 2020, "Stellar Abundances in APOGEE DR16: New Insights into the Chemical Evolution of the Milky Way", *The Astronomical Journal*, v. 160, n. 3, pp. 120. doi: 10.3847/1538-3881/aba592.

KAMULALI, J., NSAMBA, B., ADIBEKYAN, V., et al., 2025, "MAISTEP: A new grid-based machine learning tool for inferring stellar parameters. I. Ages of giant planet host stars", *Astronomy & Astrophysics*, v. 695, pp. A57. doi: 10.1051/0004-6361/202453268. Disponível em: <https://doi.org/10.1051/0004-6361/202453268>.

KARATAŞ, Y., SCHUSTER, W. J., 2006, "Metallicity and absolute magnitude calibrations for UBV photometry", Mon. Not, Roy. Astron. Soc., v. 371, n. 4 (out.), pp. 1793–1812. doi: 10.1111/j.1365-2966.2006.10800.x. Disponível em: <https://doi.org/10.1111/j.1365-2966.2006.10800.x>.

KE, G., MENG, Q., FINLEY, T., et al., 2017, "LightGBM: A highly efficient gradient boosting decision tree", *Advances in neural information processing systems*, v. 30, pp. 3146–3154.

KEARNS, M., 1988, "Thoughts on hypothesis boosting", Unpublished manuscript, circulated in ML community.

KEARNS, M., VALIANT, L. G., 1989, "Cryptographic limitations on learning Boolean formulae and finite automata", *Proceedings of the 21st annual ACM symposium on Theory of computing*, pp. 433–444.

KELLER, S. C., SCHMIDT, B. P., BESSELL, M. S., et al., 2007, "The SkyMapper Telescope and The Southern Sky Survey", Publ. Astron. Soc. Australia, v. 24, n. 1 (maio), pp. 1–12. doi: 10.1071/AS07001.

KHAN, A., HUERTA, E. A., WANG, S., et al., 2019, "Deep Learning at Scale for the Construction of Galaxy Catalogs in the Dark Energy Survey", *Physics Letters B*, v. 795, pp. 248–258. doi: 10.1016/j.physletb.2019.06.009.

KHARCHENKO, N. V., PISKUNOV, A. E., SCHILBACH, E., et al., 2013, "Global survey of star clusters in the Milky Way. II. The catalogue of basic parameters", *Astronomy & Astrophysics*, v. 558 (out.), pp. A53. doi: 10.1051/0004-6361/201322302. Disponível em: <https://doi.org/10.1051/0004-6361/201322302>. Section: Catalogs and data, Published online: 03 October 2013.

KONG, Z., JIANG, J. H., BURN, R., et al., 2022, "Analyzing the Habitable Zones of Circumbinary Planets Using Machine Learning", *The Astrophysical Journal*, v. 929, n. 2, pp. 187. doi: 10.3847/1538-4357/ac5c5a.

KORDOPATIS, G., SCHULTHEIS, M., MCMILLAN, P. J., et al., 2023, "Stellar ages, masses, extinctions, and orbital parameters based on spectroscopic parameters of Gaia DR3", Astron. Astrophys., 669:A104. doi: 10.1051/0004-6361/202244283.

KREMER, J., STENSBO-SMIDT, K., GIESEKE, F., et al., 2017, "Big universe, big data: Machine learning and image analysis for astronomy", *IEEE Intelligent Systems*, v. 32, n. 2, pp. 16–22. doi: 10.1109/MIS.2017.40.

KRON, G. E., WHITE, H. S., GASCOIGNE, S. C. B., 1953, "Red and Infrared Magnitudes for 138 Stars Observed as Photometric Standards." Astrophys. J., v. 118 (nov.), pp. 502. doi: 10.1086/145778.

KRTÍCKA, J., KRTÍCKOVÁ, I., JANÍK, J., et al., 2024, "Hot subdwarf wind models with accurate abundances. II. Helium-dominated merger products CD–46 8926 and CD–51 11879", Astron. Astrophys., 683:A80. doi: 10.1051/0004-6361/202347978. Disponível em: <https://doi.org/10.1051/0004-6361/202347978>.

KUDRITZKI, R.-P., URBANEJA, M. A., BRESOLIN, F., et al., 2008, "Quantitative spectroscopy of 24 A supergiants in the Sculptor galaxy NGC 300: Flux-weighted gravity-luminosity relationship, metallicity, and metallicity gradient", Astrophys. J., v. 681, n. 1 (jul.), pp. 269–289. doi: 10.1086/588647. Disponível em: <https://ui.adsabs.harvard.edu/abs/2008ApJ...681..269K>.

KURTZ, D. W., 2022, "Asteroseismology across the Hertzsprung–Russell diagram", Ann. Rev. Astron. Astrohys., v. 60, pp. 31–71. doi: 10.1146/annurev-astro-052920-094232. Disponível em: <https://doi.org/10.1146/annurev-astro-052920-094232>.

LAHAV, O., NAIM, A., SODRÉ, JR., L., et al., 1996, "Neural computation as a tool for galaxy classification: methods and examples", Mon. Not, Roy. Astron. Soc., v. 283 (nov.), pp. 207. doi: 10.1093/mnras/283.1.207.

LANDEMORE, H., ELSTER, J., 2012, *Collective Wisdom: Principles and Mechanisms*. Cambridge, Cambridge University Press. ISBN: 9780511846427. Disponível em: <https://www.cambridge.org/core/books/collective-wisdom/B5BCD789235A8BA125F86752EB11A7E9>.

LANGER, N., KUDRITZKI, R.-P., 2014, "The spectroscopic Hertzsprung–Russell diagram", *Astronomy & Astrophysics*, v. 564 (abr.), pp. A52. doi: 10.1051/0004-6361/201423374. Disponível em: <https://doi.org/10.1051/0004-6361/201423374>. Section: Stellar structure and evolution.

LANUSSE, F., MA, Q., LI, N., et al., 2018, "CMU DeepLens: Deep Learning for Automatic Image-based Galaxy–Galaxy Strong Lens Finding", *Monthly Notices of the Royal Astronomical Society*, v. 473, n. 3, pp. 3895–3906. doi: 10.1093/mnras/stx1665.

LAPLACE, P.-S., 1812, *Théorie Analytique des Probabilités*. Paris, Courcier. Disponível em: <https://archive.org/details/theorieanaldepro00laplrich/page/n7/mode/2up>.

LASTENNET, E., VALLS-GABAUD, D., 2002, "Detached double-lined eclipsing binaries as critical tests of stellar evolution: Age and metallicity determinations from the HR diagram", Astron. Astrophys., v. 396 (dez.), pp. 551–580. doi: 10.1051/0004-6361:20021312. Disponível em: <https://ui.adsabs.harvard.edu/abs/2002A&A...396..551L>.

LAYCOCK, S., TANG, S., GRINDLAY, J., et al., 2010, "Digital Access to a Sky Century at Harvard: Initial Photometry and Astrometry", Astron. J., v. 140, n. 4 (out.), pp. 1062–1077. doi: 10.1088/0004-6256/140/4/1062.

LEAMER, E. E., 1978, *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley Series in Probability and Statistics – Applied Probability and Statistics Section. 1st ed. New York, John Wiley & Sons.

LEE, Y. S., BEERS, T. C., SIVARANI, T., et al., 2008a, “The SEGUE Stellar Parameter Pipeline. II. Validation with Galactic globular and open clusters”, Astron. J., v. 136, n. 5 (nov.), pp. 2050–2069. doi: 10.1088/0004-6256/136/5/2050. Disponível em: <`https://ui.adsabs.harvard.edu/abs/2008AJ....136.2050L`>.

LEE, Y. S., BEERS, T. C., SIVARANI, T., et al., 2008b, “THE SEGUE STELLAR PARAMETER PIPELINE. I. DESCRIPTION AND COMPARISON OF INDIVIDUAL METHODS”, *The Astronomical Journal*, v. 136, n. 5 (oct), pp. 2022. doi: 10.1088/0004-6256/136/5/2022. Disponível em: <`https://dx.doi.org/10.1088/0004-6256/136/5/2022`>.

LEMOS, P., CRANMER, M., ABIDI, M., et al., 2023, “Robust simulation-based inference in cosmology with Bayesian neural networks”, *Machine Learning: Science and Technology*, v. 4, n. 1, pp. 01LT01. doi: 10.1088/2632-2153/acbb53.

LI, Q., 2000, “Astronomy in Ancient China”, *Proceedings of the International Astronomical Union*, v. 2000, pp. 23–34.

LI, Z., ZHAO, G., CHEN, Y., et al., 2022, “Stellar Atmospheric Parameters of M-type Stars from LAMOST DR8”, *arXiv preprint*.

LIANG, J., BU, Y., TAN, K., et al., 2022, “Estimation of Stellar Atmospheric Parameters with Light Gradient Boosting Machine Algorithm and Principal Component Analysis”, *The Astronomical Journal*, v. 163, n. 4, pp. 153. doi: 10.3847/1538-3881/ac4d97.

LIND, K., AMARSI, A. M., 2024, “Three-dimensional nonlocal thermodynamic equilibrium abundance analyses of late-type stars”, *Annual Review of Astronomy and Astrophysics*, v. 62, pp. 475–527. doi: 10.1146/annurev-astro-052722-103557. Disponível em: <`https://doi.org/10.1146/annurev-astro-052722-103557`>.

LIND, K., BERGEMANN, M., ASPLUND, M., 2012, “Non-LTE line formation of Fe in late-type stars. II. 1D spectroscopic stellar parameters”, Mon. Not, Roy. Astron. Soc., v. 427, n. 1 (nov.), pp. 50–60. doi: 10.1111/j.1365-2966.2012.21686.x. Disponível em: <`https://doi.org/10.1111/j.1365-2966.2012.21686.x`>.

LIU, C., FU, J., SHI, J., et al., 2020. “LAMOST Medium-Resolution Spectroscopic Survey (LAMOST-MRS): Scientific goals and survey plan”. Disponível em: <`https://arxiv.org/abs/2005.07210`>.

LOH, W.-Y., 2014, "Fifty Years of Classification and Regression Trees", *International Statistical Review*, v. 82, n. 3, pp. 329–348. doi: https://doi.org/10.1111/insr.12016. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12016>.

LONGO, G., MERÉNYI, E., TINO, P., 2019, "Foreword to the focus issue on machine intelligence in astronomy and astrophysics", *Publications of the Astronomical Society of the Pacific*, v. 131, n. 1001, pp. 100101. doi: 10.1088/1538-3873/ab2743.

LOPES, C. E. F., GUTIÉRREZ-SOTO, L. A., ALBERICE, V. S. F., et al., 2025, "Stellar atmospheric parameters and chemical abundances of ∼5 million stars from S-PLUS multiband photometry", Astron. Astrophys., 693:A306. doi: 10.1051/0004-6361/202451491. Disponível em: <https://ui.adsabs.harvard.edu/abs/2025A&A...693A.306F>.

LÓPEZ-SANJUAN, C., VÁZQUEZ RAMIÓ, H., VARELA, J., et al., 2019, "J-PLUS: Morphological star/galaxy classification by PDF analysis", Astron. Astrophys., 622:A177. doi: 10.1051/0004-6361/201732480.

LÓPEZ-SANJUAN, C., YUAN, H., VÁZQUEZ RAMIÓ, H., et al., 2021, "J-PLUS: Systematic impact of metallicity on photometric calibration with the stellar locus", Astron. Astrophys., 654:A61. doi: 10.1051/0004-6361/202140444.

LU, H., ZHOU, H., WANG, J., et al., 2006, "Ensemble Learning for Independent Component Analysis of Normal Galaxy Spectra", *The Astronomical Journal*, v. 131, n. 2, pp. 790. doi: 10.1086/498711.

LUNDBERG, S. M., LEE, S.-I., 2017, "A unified approach to interpreting model predictions", *Advances in neural information processing systems*, v. 30.

LÓPEZ-SANJUAN, C., RAMIÓ, H. V., XIAO, K., et al., 2024, "J-PLUS: Toward a homogeneous photometric calibration using Gaia BP/RP low-resolution spectra", *Astronomy & Astrophysics*, v. 683, pp. A29. doi: 10.1051/0004-6361/202346012. Disponível em: <https://doi.org/10.1051/0004-6361/202346012>.

MACHADO-PEREIRA, E., DAFLON, S., PLACCO, V. M., et al., 2026, "J-PLUS: stellar parameter determination of members of open clusters using machine learning", *Monthly Notices of the Royal Astronomical Society*. doi: 10.1093/mnras/stag132. Early View article, published 21 January 2026.

MACIEL, W. J., 2016, *Introduction to Stellar Structure.* Springer Praxis Books. 1 ed. Cham, Springer Cham. ISBN: 978-3-319-16141-9. doi: 10.1007/978-3-319-16142-6. Originally published as *Introdução à Estrutura e Evolução Estelar*, Editora da Universidade de São Paulo.

MAINZER, A., BAUER, J., GRAV, T., et al., 2011, "Preliminary Results from NEOWISE: An Enhancement to the Wide-field Infrared Survey Explorer for Solar System Science", Astrophys. J., 731(1):53. doi: 10.1088/0004-637X/731/1/53.

MAJEWSKI, S. R., ZASOWSKI, G., SCHLEGEL, D. G., et al., 2017, "The Apache Point Observatory Galactic Evolution Experiment (APOGEE)", *The Astronomical Journal*, v. 154, n. 3, pp. 94. doi: 10.3847/1538-3881/aa784d.

MAROCCO, F., EISENHARDT, P. R. M., FOWLER, J. W., et al., 2021, "The CatWISE2020 Catalog", Astrophys. J. Suppl., 253(1):8. doi: 10.3847/1538-4365/abd805.

MAS-BUITRAGO, P., SOLANO, E., GONZÁLEZ-MARCOS, A., et al., 2022, "J-PLUS: Discovery and characterisation of ultracool dwarfs using Virtual Observatory tools. II. Second data release and machine learning methodology", Astron. Astrophys., 666:A147. doi: 10.1051/0004-6361/202243895.

MASON, L., BAXTER, J., BARTLETT, P. L., et al., 1999, "Boosting algorithms as gradient descent", *Advances in Neural Information Processing Systems*, v. 12.

MATSUNO, T., STARKENBURG, E., BALBINOT, E., et al., 2024, "Improving metallicity estimates for very metal-poor stars in the Gaia DR3 GSP-Spec catalog", Astron. Astrophys., 685:A59. doi: 10.1051/0004-6361/202245762. Disponível em: <`https://ui.adsabs.harvard.edu/abs/2024A&A...685A..59M`>.

MAURY, A. C., 1897, "Spectra of Bright Stars Photographed with the 11-Inch Draper Telescope as Part of the Henry Draper Memorial", *Annals of the Astronomical Observatory of Harvard College*, v. 28, pp. 1–128.

MAYO, A. W., DRESSING, C. D., VANDERBURG, A., et al., 2023, "Hyades Member K2-136c: The Smallest Planet in an Open Cluster with a Precisely Measured Mass", *The Astronomical Journal*, v. 165, n. 6, pp. 235. doi: 10.3847/1538-3881/acca1c. Disponível em: <`https://doi.org/10.3847/1538-3881/acca1c`>.

MCCULLOCH, W. S., PITTS, W., 1943, "A Logical Calculus of the Ideas Immanent in Nervous Activity", *The Bulletin of Mathematical Biophysics*, v. 5, pp. 115–133. doi: 10.1007/BF02478259.

MCGIBBON, R. J., KHOCHFAR, S., 2023, “Multi-epoch machine learning 2: identifying physical drivers of galaxy properties in simulations”, *Monthly Notices of the Royal Astronomical Society*, v. 523, n. 4, pp. 5583–5597. doi: 10.1093/mnras/stad1811.

MCGRAYNE, S. B., 2011, *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy.* New Haven, CT, Yale University Press. ISBN: 9780300169690. Disponível em: <`http://www.jstor.org/stable/j.ctt1np76s`>.

MELÉNDEZ, J., ASPLUND, M., GUSTAFSSON, B., et al., 2009, “The peculiar solar composition and its possible relation to planet formation”, Astrophys. J. Lett., v. 704, n. 1 (out.), pp. L66–L70. doi: 10.1088/0004-637X/704/1/L66. Disponível em: <`https://ui.adsabs.harvard.edu/abs/2009ApJ...704L..66M`>.

MENDES DE OLIVEIRA, C., RIBEIRO, T., SCHOENELL, W., et al., 2019, “The Southern Photometric Local Universe Survey (S-PLUS): improved SEDs, morphologies, and redshifts with 12 optical filters”, Mon. Not, Roy. Astron. Soc., v. 489, n. 1 (out.), pp. 241–267. doi: 10.1093/mnras/stz1985.

MIENYE, I. D., SUN, Y., 2022, “A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects”, *IEEE Access*, v. 10, pp. 99129–99149. doi: 10.1109/ACCESS.2022.3207287.

MIHALAS, D., 1978, *Stellar atmospheres.* 2nd ed. San Francisco, W. H. Freeman.

MILLER, A. A., BLOOM, J. S., RICHARDS, J. W., et al., 2015, “A Machine-Learning Method to Infer Fundamental Stellar Parameters from Photometric Light Curves”, *The Astrophysical Journal*, v. 798, n. 2, pp. 122. doi: 10.1088/0004-637X/798/2/122.

MILLER, A. S., 1993, “A review of neural network applications in astronomy”, *Vistas in Astronomy*, v. 36, pp. 141–161. doi: 10.1016/0083-6656(93)90118-4.

MINSKY, M., 1954, *Neural Nets and the Brain-Model Problem.* Tese de Doutorado, Princeton University. Disponível em: <`https://searchworks.stanford.edu/view/11643293`>. Discusses SNARC's design (joint work with Edmonds) in Chapter 3.

MINSKY, M., EDMONDS, D., 1951, “The SNARC Machine: First Neural Network Hardware”, Unpublished work; described in MINSKY (1954) and RUSSELL

& NORVIG (2016). Hardware constructed at Harvard/Princeton; no surviving technical report.

MINSKY, M., PAPERT, S., 1969, *Perceptrons: An Introduction to Computational Geometry.* Cambridge, MA, MIT Press. First edition.

MINSKY, M., PAPERT, S., 1988, *Perceptrons: An Introduction to Computational Geometry.* Expanded edition ed. Cambridge, MA, MIT Press. ISBN: 0-262-63111-3. Originally published in 1969; expanded edition includes handwritten alterations from 1972 and a new preface and epilogue.

MONTEGRIFFO, P., ANGELI, F. D., ANDRAE, R., et al., 2023, "Gaia Data Release 3: External calibration of BP/RP low-resolution spectroscopic data", *Astronomy & Astrophysics*, v. 666, pp. A121. doi: 10.1051/0004-6361/202243880. Disponível em: <`https://www.aanda.org/articles/aa/full_html/2023/06/aa43880-22/aa43880-22.html`>.

MORGAN, J. N., SONQUIST, J. A., 1963, "Problems in the Analysis of Survey Data, and a Proposal", *Journal of the American Statistical Association*, v. 58, n. 302, pp. 415–434. doi: 10.1080/01621459.1963.10500855. Disponível em: <`https://doi.org/10.1080/01621459.1963.10500855`>.

MOSBY, G., THE ROMAN SPACE TELESCOPE DETECTOR TEAM, 2020, "Properties and characteristics of the Nancy Grace Roman Space Telescope H4RG-10 detectors", *Journal of Astronomical Telescopes, Instruments, and Systems*, v. 6, n. 4, pp. 046001. doi: 10.1117/1.JATIS.6.4.046001. for the Roman Space Telescope Detector Team.

MOYA, A., LÓPEZ-SASTRE, R. J., 2022, "Stellar Mass and Radius Estimation Using Artificial Intelligence", *Astronomy & Astrophysics*, v. 663, pp. A112. doi: 10.1051/0004-6361/202142930.

MYEONG, G. C., BELOKUROV, V., AGUADO, D. S., et al., 2022, "Milky Way's Eccentric Constituents with Gaia, APOGEE, and GALAH", Astrophys. J., 938 (1):21. doi: 10.3847/1538-4357/ac8d68.

NARAYAN, G., ZAIDI, T., SORAISAM, M. D., et al., 2018, "Machine-learning-based Brokers for Real-time Classification of the LSST Alert Stream", *The Astrophysical Journal Supplement Series*, v. 236, n. 1, pp. 9. doi: 10.3847/1538-4365/aab781.

NARDIELLO, D., BEDIN, L. R., GRIGGIO, M., et al., 2023, "Photometry and astrometry with JWST – III. A NIRCam-Gaia DR3 analysis of the open cluster NGC 2506", *Monthly Notices of the Royal Astronomical Society*, v. 525,

n. 2, pp. 2585–2604. doi: 10.1093/mnras/stad2445. Disponível em: <https://doi.org/10.1093/mnras/stad2445>.

NESS, M., HOGG, D. W., RIX, H.-W., et al., 2015, "The Cannon: A Data-driven Approach to Stellar Label Determination", *The Astrophysical Journal*, v. 808, n. 1, pp. 16. doi: 10.1088/0004-637X/808/1/16.

NETOPIL, M., PAUNZEN, E., HEITER, U., et al., 2013, "Towards a photometric metallicity scale for open clusters", *A&A*, v. 556, pp. A26. doi: 10.1051/0004-6361/201321118.

NETOPIL, M., PAUNZEN, E., HEITER, U., et al., 2016, "On the metallicity of open clusters - III. Homogenised sample", *A&A*, v. 585, pp. A150. doi: 10.1051/0004-6361/201527939.

NEWBERG, H. J., YANNY, B., ROCKOSI, C., et al., 2002, "The Ghost of Sagittarius and Lumps in the Halo of the Milky Way", *The Astrophysical Journal*, v. 569, n. 1 (apr), pp. 245. doi: 10.1086/338983. Disponível em: <https://dx.doi.org/10.1086/338983>.

NEWELL, E. B., 1973, "Photoelectric observations of faint blue stars. III", *Astrophysical Journal Supplement Series*, v. 26, pp. 37. doi: 10.1086/190276.

ODEWAHN, S. C., STOCKWELL, E. B., PENNINGTON, R. L., et al., 1992, "Automated Star/Galaxy Discrimination With Neural Networks", Astron. J., v. 103 (jan.), pp. 318. doi: 10.1086/116063.

OESTREICHER, M. O., SCHMIDT-KALER, T., 1998, "Red supergiants in the LMC. II. Spectrophotometry and model atmospheres", *Monthly Notices of the Royal Astronomical Society*, v. 299, n. 3, pp. 625–636. doi: 10.1046/j.1365-8711.1998.01501.x. Disponível em: <https://doi.org/10.1046/j.1365-8711.1998.01501.x>.

OGILVIE, M. B., 2000, "Obligatory Amateurs: Annie Maunder (1868–1947) and British Women Astronomers at the Dawn of Professional Astronomy", *British Journal for the History of Science*, v. 33, n. 1, pp. 67–84.

OLIVEIRA SCHWARZ, G. B., 2022. "Southern Photometric Local Universe Survey Cloud". jan. Disponível em: <https://doi.org/10.5281/zenodo.10980447>.

OSUNA, P., ORTIZ, I., LUSTED, J., et al., 2008. "IVOA Astronomical Data Query Language Version 2.00". IVOA Recommendation 30 October 2008, out.

OWENS, E. A., GRIFFITHS, R. E., RATNATUNGA, K. U., 1996, "Using oblique decision trees for the morphological classification of galaxies", *Monthly Notices of the Royal Astronomical Society*, v. 281, n. 1 (jul), pp. 153–157. doi: 10.1093/mnras/281.1.153. Disponível em: <https://doi.org/10.1093/mnras/281.1.153>. Published: 01 July 1996.

PAGEL, B. E. J., 2009, *Nucleosynthesis and Chemical Evolution of Galaxies*. 2nd ed. Cambridge, UK, Cambridge University Press. ISBN: 9780521840309.

PAUNZEN, E., 2015, "A new catalogue of Strömgren-Crawford $uvby\beta$ photometry", Astron. Astrophys., 580:A23. doi: 10.1051/0004-6361/201526413. Disponível em: <https://doi.org/10.1051/0004-6361/201526413>.

PEROTTONI, H. D., PLACCO, V. M., ALMEIDA-FERNANDES, F., et al., 2024, "The S-PLUS Ultra-Short Survey: First data release", Astron. Astrophys., 691:A138. doi: 10.1051/0004-6361/202348801.

PERREN, G. I., PERA, M. S., NAVONE, H. D., et al., 2023, "The Unified Cluster Catalogue: towards a comprehensive and homogeneous data base of stellar clusters", *Monthly Notices of the Royal Astronomical Society*, v. 526, n. 3 (dez.), pp. 4107–4119. ISSN: 0035-8711. doi: 10.1093/mnras/stad2826.

PIETRINFERNI, A., CASSISI, S., SALARIS, M., et al., 2004, "A large stellar evolution database for population synthesis studies. I. Scaled-solar models and isochrones", Astrophys. J., v. 612, n. 1 (set.), pp. 168–190. doi: 10.1086/422498. Disponível em: <https://ui.adsabs.harvard.edu/abs/2004ApJ...612..168P>.

POGSON, N. R., 1856, "Magnitudes of Thirty-six of the Minor Planets for the First Day of Each Month of the Year 1857", *Monthly Notices of the Royal Astronomical Society*, v. 17, n. 1, pp. 12–15. Disponível em: <https://doi.org/10.1093/mnras/17.1.12>. Contains the original definition of the modern logarithmic magnitude scale.

POOVELIL, V. J., ZASOWSKI, G., HASSELQUIST, S., et al., 2020, "Open Cluster Chemical Homogeneity throughout the Milky Way", *The Astrophysical Journal*, v. 903, n. 1 (nov.), pp. 55. ISSN: 0004-637X. doi: 10.3847/1538-4357/abb93e.

PRIETO, C. A., BEERS, T. C., WILHELM, D. E., et al., 2006, "FERRE: A Code for Efficient Parameter Estimation in Large Stellar Spectroscopic Surveys", *The Astrophysical Journal*, v. 636, n. 2, pp. 804–820. doi: 10.1086/498208.

PÉREZ, A. E. G., MÉSZÁROS, S., PRIETO, C. A., et al., 2016, "The APOGEE Stellar Parameter and Chemical Abundances Pipeline", *The Astronomical Journal*, v. 151, n. 6, pp. 144. doi: 10.3847/0004-6256/151/6/144.

QAHWAJI, R., COLAK, T., 2007, "Automatic short-term solar flare prediction using machine learning and sunspot associations", *Solar Physics*, v. 241, n. 1, pp. 195–211.

QU, C.-X., LUO, A.-L., WANG, R., et al., 2024, "Stellar Atmospheric Parameters for Cool Dwarfs in Gaia Data Release 3", *The Astrophysical Journal Supplement Series*, v. 270, n. 2, pp. 32. doi: 10.3847/1538-4365/ad103c.

QUISPE-HUAYNASI, F., ROIG, F., DAFLON, S., et al., 2023, "J-PLUS: characterization of high-velocity stars in the second data release", Mon. Not, Roy. Astron. Soc., v. 522, n. 3 (jul.), pp. 3898–3911. doi: 10.1093/mnras/stad1230.

QUISPE-HUAYNASI, F., ROIG, F., PLACCO, V. M., et al., 2024, "Characterization of high-velocity stars in the S-PLUS internal fourth data release", Mon. Not, Roy. Astron. Soc., v. 527, n. 3 (jan.), pp. 6173–6188. doi: 10.1093/mnras/stad3610. Disponível em: <`https://ui.adsabs.harvard.edu/abs/2024MNRAS.527.6173Q`>.

RAMÍREZ, I., MELÉNDEZ, J., 2005, "The Effective Temperature Scale of FGK Stars. II. T_{eff}:Color:[Fe/H] Calibrations", *The Astrophysical Journal*, v. 626, n. 1, pp. 465–485. doi: 10.1086/430102. Disponível em: <`https://doi.org/10.1086/430102`>.

RANDICH, S., GILMORE, G., GAIA-ESO CONSORTIUM, 2013, "The Gaia-ESO Large Public Spectroscopic Survey", *The Messenger*, v. 154 (dez.), pp. 47–49.

RANDICH, S., GILMORE, G., MAGRINI, L., et al., 2022, "The Gaia-ESO Public Spectroscopic Survey: Implementation, data products, open cluster survey, science, and legacy", *Astronomy & Astrophysics*, v. 666, pp. A121. doi: 10.1051/0004-6361/202243141. Disponível em: <`https://www.aanda.org/articles/aa/full_html/2022/10/aa43141-22/aa43141-22.html`>.

RAPPAPORT, B., ANDERSON, K., 1988, "Automated galaxy recognition". In: *Astronomy from Large Databases: Scientific Objectives and Methodological Approaches*, pp. 233–238, Garching, Federal Republic of Germany, October. European Southern Observatory. Proceedings of the Conference, Garching, Germany, October 12–14, 1987. (A89-27176 10-82).

RICHARDS, J. W., STARR, D. L., BUTLER, N. R., et al., 2011, "On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data", *The Astrophysical Journal*, v. 733, n. 1, pp. 10. doi: 10.1088/0004-637X/733/1/10.

RICKER, G. R., WINN, J. N., VANDERSPEK, R., et al., 2015, "Transiting Exoplanet Survey Satellite (TESS)", *Journal of Astronomical Telescopes, Instruments, and Systems*, 1:014003. doi: 10.1117/1.JATIS.1.1.014003.

ROBINSON, D., AVESTRUZ, C., GNEDIN, N. Y., 2024, "Exploring the dependence of gas cooling and heating functions on the incident radiation field with machine learning", Mon. Not, Roy. Astron. Soc., v. 528, n. 1 (fev.), pp. 255–269. doi: 10.1093/mnras/stad3880.

ROCKOSI, C. M., LEE, Y. S., MORRISON, H. L., et al., 2022, "SEGUE-2: Old Milky Way Stars Near and Far", Astrophys. J. Suppl., 259(2):60. doi: 10.3847/1538-4365/ac5323.

RODRIGO, C., SOLANO, E., 2020, "The SVO Filter Profile Service". In: *XIV.0 Scientific Meeting (virtual) of the Spanish Astronomical Society*, p. 182, July.

RODRIGO, C., SOLANO, E., BAYO, A., 2012. "SVO Filter Profile Service Version 1.0". IVOA Working Draft 15 October 2012, October.

RODRIGO, C., CRUZ, P., AGUILAR, J. F., et al., 2024, "Photometric segregation of dwarf and giant FGK stars using the SVO Filter Profile Service and photometric tools", *Astronomy & Astrophysics*, v. 689, pp. A93. doi: 10.1051/0004-6361/202449998. Published online: 06 September 2024.

RODRÍGUEZ, J.-V., RODRÍGUEZ-RODRÍGUEZ, I., WOO, W. L., 2022, "On the application of machine learning in astronomy and astrophysics: A text-mining-based scientometric analysis", *WIREs Data Mining and Knowledge Discovery*, v. 12, n. 5, pp. e1476. doi: https://doi.org/10.1002/widm.1476. Disponível em: <`https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1476`>.

ROMAN, N. G., 1954, "A group of high velocity F-type stars", Astron. J., v. 59 (jan.), pp. 307–312. doi: 10.1086/107017. Disponível em: <`https://ui.adsabs.harvard.edu/abs/1954AJ.....59..307R`>.

ROSENBERG, H., 1910, "Über den Zusammenhang von Helligkeit und Spektraltypus in den Plejaden", *Astronomische Nachrichten*, v. 186, pp. 71–78. doi: 10.1002/asna.19101860503.

ROSENBLATT, F., 1957, *The Perceptron, a Perceiving and Recognizing Automaton (Project Para)*. Relatório Técnico 85, Cornell Aeronautical Laboratory, Buffalo, NY.

ROUSSEEUW, P. J., CROUX, C., 1993, "Alternatives to the Median Absolute Deviation", *Journal of the American Statistical Association*, v. 88, n. 424, pp. 1273–1283. doi: 10.1080/01621459.1993.10476408. Disponível em: <https://doi.org/10.1080/01621459.1993.10476408>.

RUMELHART, D. E., HINTON, G. E., WILLIAMS, R. J., 1986a, "Learning Representations by Back-Propagating Errors", *Nature*, v. 323, n. 6088 (October), pp. 533–536. doi: 10.1038/323533a0. Disponível em: <https://doi.org/10.1038/323533a0>.

RUMELHART, D. E., MCCLELLAND, J. L., THE PDP RESEARCH GROUP, 1986b, *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*. Cambridge, MA, The MIT Press. ISBN: 9780262291408. doi: 10.7551/mitpress/5236.001.0001. Disponível em: <https://doi.org/10.7551/mitpress/5236.001.0001>.

RUSSELL, H. N., 1914, "Relations Between the Spectra and Other Characteristics of the Stars", *Popular Astronomy*, v. 22, pp. 275–294.

RUSSELL, S. J., NORVIG, P., 2016, *Artificial Intelligence: A Modern Approach*. 3rd ed. New Jersey, Pearson. ISBN: 978-1-292-15396-4. Added here for brief historical note on SNARC.

SADEH, I., ABDALLA, F. B., LAHAV, O., 2016, "ANNz2: Photometric Redshift and Probability Distribution Function Estimation using Machine Learning", *Publications of the Astronomical Society of the Pacific*, v. 128, n. 968, pp. 104502. doi: 10.1088/1538-3873/128/968/104502.

SAGI, O., ROKACH, L., 2018, "Ensemble learning: A survey", *WIREs Data Mining and Knowledge Discovery*, v. 8, n. 4, pp. e1249. doi: https://doi.org/10.1002/widm.1249. Disponível em: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1249>.

SAMUEL, A. L., 1959, "Some Studies in Machine Learning Using the Game of Checkers", *IBM Journal of Research and Development*, v. 3, pp. 210–229. doi: 10.1147/rd.33.0210.

SARRO, L. M., ORDIERES-MERÉ, J., BELLO-GARCÍA, A., et al., 2018, "Estimates of the Atmospheric Parameters of M-type Stars: A Machine-Learning Perspective”, *Monthly Notices of the Royal Astronomical Society*, v. 476, n. 1, pp. 1120–1139. doi: 10.1093/mnras/sty165.

SCHILD, R., PETERSON, D. M., OKE, J. B., 1971, “Effective Temperatures of B- and A-Type Stars”, *The Astrophysical Journal*, v. 166, pp. 95–110. doi: 10.1086/150944. Disponível em: <https://ui.adsabs.harvard.edu/abs/1971ApJ...166...95S>. Provided by the SAO/NASA Astrophysics Data System.

SCHLAFLY, E. F., MEISNER, A. M., GREEN, G. M., 2019, “The unWISE Catalog: Two Billion Infrared Sources from Five Years of WISE Imaging”, Astrophys. J. Suppl., 240(2):30. doi: 10.3847/1538-4365/aafbea.

SEN, S., AGARWAL, S., CHAKRABORTY, P., et al., 2022, “Astronomical big data processing using machine learning: A comprehensive review”, *Experimental Astronomy*, v. 53 (fev.), pp. 1–43. doi: 10.1007/s10686-021-09827-4. Disponível em: <https://doi.org/10.1007/s10686-021-09827-4>. Published online: 14 January 2022.

SETO, H., OYAMA, A., KITORA, S., et al., 2022, “Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data”, *Scientific reports*, v. 12, n. 1, pp. 15889.

SHALLUE, C. J., VANDERBURG, A., 2018, “Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90”, *The Astronomical Journal*, v. 155, n. 2, pp. 94. doi: 10.3847/1538-3881/aa9e09.

SHANKAR, S., GULLY-SANTIAGO, M. A., MORLEY, C. V., 2025, “A New Hybrid Machine Learning Method for Stellar Parameter Inference”, *The Astrophysical Journal*, v. 985, n. 1, pp. 68. doi: 10.3847/1538-4357/adcb47.

SHAPLEY, L. S., 1951, *Notes on the N-Person Game — II: The Value of an N-Person Game*. Santa Monica, CA, RAND Corporation. doi: 10.7249/RM0670.

SKRUTSKIE, M. F., CUTRI, R. M., STIENING, R., et al., 2006, “The Two Micron All Sky Survey (2MASS)”, Astron. J., v. 131, n. 2 (fev.), pp. 1163–1183. doi: 10.1086/498708.

SMALLEY, B., 2005, “$T_{\rm eff}$ and log g determinations”, *Memorie della Società Astronomica Italiana Supplementi*, v. 8 (jan.), pp. 130–139. doi: 10.48550/arXiv.astro-ph/0509535. Disponível em: <https://ui.adsabs.harvard.edu/abs/2005MSAIS...8..130S>.

SMITH, M. J., GEACH, J. E., 2023, “Astronomia ex machina: a history, primer and outlook on neural networks in astronomy”, *Royal Society Open Science*, v. 10, n. 2, pp. 221454. doi: 10.1098/rsos.221454. Disponível em: <`http://doi.org/10.1098/rsos.221454`>.

SNEDEN, C., 1973, “The nitrogen abundance of the very metal-poor star HD 122563”, Astrophys. J., v. 184 (set.), pp. 839–849. doi: 10.1086/152374. Disponível em: <`https://ui.adsabs.harvard.edu/abs/1973ApJ...184..839S`>.

SNEDEN, C., BEAN, J., IVANS, I., et al., 2012. “MOOG: LTE line analysis and spectrum synthesis”. Astrophysics Source Code Library, record ascl:1202.009, fev. Disponível em: <`https://ui.adsabs.harvard.edu/abs/2012ascl.soft02009S`>.

SOBEL, D., 2016, *The Glass Universe: How the Ladies of the Harvard Observatory Took the Measure of the Stars.* Viking.

SODERBLOM, D. R., 2010, “The Ages of Stars”, *Annual Review of Astronomy and Astrophysics*, v. 48 (set.), pp. 581–629. doi: 10.1146/annurev-astro-081309-130806. Disponível em: <`https://doi.org/10.1146/annurev-astro-081309-130806`>. First published as a Review in Advance on May 25, 2010.

SOKOLOV, N. A., 1998, “Effective temperatures of Ap stars”, *Astronomy & Astrophysics Supplement Series*, v. 130, pp. 215–222. doi: 10.1051/aas:1998226. Disponível em: <`https://ui.adsabs.harvard.edu/abs/1998A&AS..130..215S`>. Provided by the SAO/NASA Astrophysics Data System.

SOLANO, E., MARTÍN, E. L., CABALLERO, J. A., et al., 2019, “J-PLUS: Discovery and characterisation of ultracool dwarfs using Virtual Observatory tools”, Astron. Astrophys., 627:A29. doi: 10.1051/0004-6361/201935256.

SOUBIRAN, C., LE CAMPION, J.-F., BROUILLET, N., et al., 2016, “The PASTEL catalogue: 2016 version”, Astron. Astrophys., 591:A118. doi: 10.1051/0004-6361/201628497.

SPARKE, L. S., GALLAGHER, III, J. S., 2007, *Galaxies in the Universe: An Introduction.* 2nd ed. , Cambridge University Press.

STARKENBURG, E., MARTIN, N., YOUAKIM, K., et al., 2017, “The Pristine survey - I. Mining the Galaxy for the most metal-poor stars”, Mon. Not, Roy. Astron. Soc., v. 471, n. 3 (nov.), pp. 2587–2604. doi: 10.1093/mnras/stx1068.

STEPHENS, Z. D., LEE, S. Y., FAGHRI, F., et al., 2015, "Big Data: Astronomical or Genomical?" *PLoS Biology*, v. 13, n. 7, pp. e1002195. doi: 10.1371/journal.pbio.1002195.

STEPHENSON, F. R., 1997, *Historical Eclipses and Earth's Rotation*. Cambridge, Cambridge University Press.

STORRIE-LOMBARDI, M. C., LAHAV, O., SODRE, JR., L., et al., 1992, "Morphological Classification of Galaxies by Artificial Neural Networks", Mon. Not, Roy. Astron. Soc., v. 259 (nov.), pp. 8P. doi: 10.1093/mnras/259.1.8P.

STRÖMGREN, B., 1956, "Two-dimensional spectral classification of F stars through photoelectric photometry with interference filters", *Vistas in Astronomy*, v. 2, n. 1 (jan.), pp. 1336–1346. doi: 10.1016/0083-6656(56)90060-5.

STRÖMGREN, B., 1966, "Spectral Classification Through Photoelectric Narrow-Band Photometry", Ann. Rev. Astron. Astrohys., v. 4 (jan.), pp. 433. doi: 10.1146/annurev.aa.04.090166.002245.

SUN, Y., CHENG, Z., YE, S., et al., 2021, "A Catalog of 323 Cataclysmic Variables from LAMOST DR6", *The Astrophysical Journal Supplement Series*, v. 257, n. 2, pp. 65. doi: 10.3847/1538-4365/ac283a. Disponível em: <`https://doi.org/10.3847/1538-4365/ac283a`>.

SUÁREZ GÓMEZ, S. L., GONZÁLEZ-GUTIÉRREZ, C., DÍEZ ALONSO, E., et al., 2019, "Experience with Artificial Neural Networks Applied in Multi-object Adaptive Optics", *Publications of the Astronomical Society of the Pacific*, v. 131, n. 1004, pp. 108012. doi: 10.1088/1538-3873/ab1ebb.

SÁNCHEZ, N., LÓPEZ-MARTÍNEZ, F., OCANDO, S., et al., 2021, "Study of the open cluster Alessi-Teutsch 9 (ASCC 10) using multiband photometry and Gaia EDR3", *Astrophysics and Space Science*, v. 366, pp. 92. doi: 10.1007/s10509-021-03999-3. Disponível em: <`https://doi.org/10.1007/s10509-021-03999-3`>.

TARSITANO, F., BRUDERER, C., SCHAWINSKI, K., et al., 2022, "Image feature extraction and galaxy classification: a novel and efficient approach with automated machine learning", *Monthly Notices of the Royal Astronomical Society*, v. 511, n. 3, pp. 3330–3338. doi: 10.1093/mnras/stac233.

TAYLOR, M. B., 2005, "TOPCAT & STIL: Starlink Table/VOTable Processing Software". In: Shopbell, P., Britton, M., Ebert, R. (Eds.), *Astronomical Data Analysis Software and Systems XIV*, v. 347, *Astronomical Society of the Pacific Conference Series*, p. 29, dez.

THE MULTIMODAL UNIVERSE COLLABORATION, AUDENAERT, J., BOWLES, M., et al., 2024. "The Multimodal Universe: Enabling Large-Scale Machine Learning with 100TB of Astronomical Scientific Data". Disponível em: <https://arxiv.org/abs/2412.02527>.

TING, Y.-S., 2024, "Why Machine Learning Models Systematically Underestimate Extreme Values", *arXiv e-prints*, art. arXiv:2412.05806. doi: 10.48550/arXiv.2412.05806.

TING, Y.-S., CONROY, C., RIX, H.-W., et al., 2019, "The Payne: Self-consistent Ab Initio Fitting of Stellar Spectra", *The Astrophysical Journal*, v. 879, n. 2, pp. 69. doi: 10.3847/1538-4357/ab2331.

TINSLEY, B. M., 1979, "Stellar lifetimes and abundance ratios in chemical evolution", Astrophys. J., v. 229 (maio), pp. 1046–1056. doi: 10.1086/157039. Disponível em: <https://ui.adsabs.harvard.edu/abs/1979ApJ...229.1046T>.

TOOSI, A., BOTTINO, A. G., SABOURY, B., et al., 2021, "A Brief History of AI: How to Prevent Another Winter (A Critical Review)", *PET Clinics*, v. 16, n. 4 (out.), pp. 449–469. ISSN: 1556-8598. doi: 10.1016/j.cpet.2021.07.001. Disponível em: <http://dx.doi.org/10.1016/j.cpet.2021.07.001>.

TURING, A. M., 1950, "Computing Machinery and Intelligence", *Mind*, v. 59, pp. 433–460. doi: 10.1093/mind/LIX.236.433. Disponível em: <https://doi.org/10.1093/mind/LIX.236.433>.

TWAROG, B. A., 1980a, "The chemical evolution of the solar neighborhood. I. A bias-free reduction technique and data sample", Astrophys. J. Suppl., v. 44 (set.), pp. 1–29. doi: 10.1086/190683. Disponível em: <https://ui.adsabs.harvard.edu/abs/1980ApJS...44....1T>.

TWAROG, B. A., 1980b, "The chemical evolution of the solar neighborhood. II. The age-metallicity relation and the history of star formation in the galactic disk", Astrophys. J., v. 242 (nov.), pp. 242–259. doi: 10.1086/158460. Disponível em: <https://ui.adsabs.harvard.edu/abs/1980ApJ...242..242T>.

VICENTE, B., ABAD, C., GARZÓN, F., 2007, "Astrometry with Carte du Ciel plates, San Fernando zone. I. Digitization and measurement using a flatbed scanner", Astron. Astrophys., v. 471, n. 3 (set.), pp. 1077–1089. doi: 10.1051/0004-6361:20066843.

VOGEL, H. C., 1900, "On the Spectroscopic Work at the Astrophysical Observatory of Potsdam", *Publications of the Astronomical Society of the Pacific*, v. 12, pp. 223–229. doi: 10.1086/121942.

WADADEKAR, Y., 2005, "Estimating Photometric Redshifts Using Support Vector Machines", Publ. Astron. Soc. Pacific, v. 117, n. 827 (jan.), pp. 79–85. doi: 10.1086/427710.

WALLERSTEIN, G., 1962, "Abundances in G dwarfs. VI. A survey of field stars", Astrophys. J. Suppl., v. 6 (fev.), pp. 407–435. doi: 10.1086/190067. Disponível em: <`https://ui.adsabs.harvard.edu/abs/1962ApJS....6..407W`>.

WALLERSTEIN, G., CARLSON, M., 1960, "Letter to the editor: On the ultraviolet excess in G dwarfs", Astrophys. J., v. 132 (set.), pp. 276–281. doi: 10.1086/146926. Disponível em: <`https://ui.adsabs.harvard.edu/abs/1960ApJ...132..276W`>.

WANG, C., BAI, Y., YUAN, H., et al., 2022, "J-PLUS: Support vector regression to measure stellar parameters", Astron. Astrophys., 664:A38. doi: 10.1051/0004-6361/202243130.

WANG, H.-F., CARRARO, G., LI, X., et al., 2024a, "Age Determination of LAMOST Red Giant Branch Stars Based on the Gradient Boosting Decision Tree Method", *The Astrophysical Journal*, v. 967, n. 1, pp. 37. doi: 10.3847/1538-4357/ad3b90.

WANG, J., ZHONG, Z., 2018, "Revisiting the mass-luminosity relation with an effective temperature modifier", *Astronomy & Astrophysics*, v. 619, pp. L1. doi: 10.1051/0004-6361/201834109.

WANG, J., DING, X., LI, J., et al., 2024b, "A Method of Rapidly Deriving Late-type Contact Binary Parameters and Its Application in the Catalina Sky Survey", *The Astrophysical Journal Supplement Series*, v. 273, n. 2, pp. 31. doi: 10.3847/1538-4365/ad5953.

WANG, J., 2025, "Training a convolutional neural network for exoplanet classification with transit photometry data", *Scientific Reports*, v. 15, pp. 15408. doi: 10.1038/s41598-025-98935-8.

WANG, K., GUO, P., YU, F., et al., 2018, "Computational intelligence in astronomy: A survey", *International Journal of Computational Intelligence Systems*, v. 11, n. 1, pp. 575. doi: 10.2991/ijcis.11.1.43.

WEBB, S., 1999, *Measuring the Universe: The Cosmological Distance Ladder*. Springer.

WEIR, N., DJORGOVSKI, S., FAYYAD, U. M., 1995, "Initial Galaxy Counts from Digitized Poss-II", Astron. J., v. 110 (jul.), pp. 1. doi: 10.1086/117493.

WEIR, N., FAYYAD, U. M., DJORGOVSKI, S. G., et al., 1995, "The SKICAT System for Processing and Analyzing Digital Imaging Sky Surveys", *Publications of the Astronomical Society of the Pacific*, v. 107, n. 718, pp. 1243–1254. doi: 10.1086/133683. Disponível em: <https://doi.org/10.1086/133683>.

WHITTEN, D. D., PLACCO, V. M., BEERS, T. C., et al., 2019, "J-PLUS: Identification of low-metallicity stars with artificial neural networks using SPHINX", Astron. Astrophys., 622:A182. doi: 10.1051/0004-6361/201833368.

WHITTEN, D. D., PLACCO, V. M., BEERS, T. C., et al., 2021, "The photometric metallicity and carbon distributions of the Milky Way's halo and solar neighborhood from S-PLUS observations of SDSS Stripe 82", Astrophys. J., 912(2):147. doi: 10.3847/1538-4357/abee7e. Disponível em: <https://ui.adsabs.harvard.edu/abs/2021ApJ...912..147W>.

WHITTEN, D. D., BEERS, T. C., PLACCO, V. M., et al., 2019, "Constraints on the Galactic Inner Halo Assembly History from the Age Gradient of Blue Horizontal-branch Stars", *The Astrophysical Journal*, v. 884, n. 1, pp. 67. doi: 10.3847/1538-4357/ab4269. Disponível em: <https://doi.org/10.3847/1538-4357/ab4269>.

WOLFF, S. C., KUHI, L. V., HAYES, D., 1968, "The Effective Temperatures of A and B Stars", *The Astrophysical Journal*, v. 152, pp. 871–880. doi: 10.1086/149603. Disponível em: <https://ui.adsabs.harvard.edu/abs/1968ApJ...152..871W>. Provided by the SAO/NASA Astrophysics Data System.

WOLPERT, D. H., 1992, "Stacked generalization", *Neural networks*, v. 5, n. 2, pp. 241–259. doi: https://doi.org/10.1016/S0893-6080(05)80023-1.

WORTHEY, G., 1999, "The age-metallicity degeneracy". In: Hubeny, I., Heap, S., Cornett, R. (Eds.), *Spectrophotometric Dating of Stars and Galaxies*, v. 192, *Astronomical Society of the Pacific Conference Series*, pp. 283–295, jan. Disponível em: <https://ui.adsabs.harvard.edu/abs/1999ASPC..192..283W>.

WOŹNIAK, P. R., WILLIAMS, S. J., VESTRAND, W. T., et al., 2004, "Identifying red variables in the northern sky variability survey", *The Astronomical Journal*, v. 128, n. 6, pp. 2965–2976.

WRIGHT, E. L., EISENHARDT, P. R. M., MAINZER, A. K., et al., 2010, "The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance", Astron. J., v. 140, n. 6 (dez.), pp. 1868–1881. doi: 10.1088/0004-6256/140/6/1868.

WRIGHT, E. L., EISENHARDT, P. R. M., MAINZER, A. K., et al., 2019. “AllWISE Source Catalog”. NASA IPAC DataSet, IRSA1, jan.

WU, Y., DU, B., LUO, A., et al., 2014, “Automatic Stellar Spectral Parameterization Pipeline for LAMOST Survey”. In: *Proceedings of the International Astronomical Union, Volume 10, Symposium S306*, pp. 340–342, may. doi: 10.1017/S1743921314010825.

XIAO, K., YUAN, H., LÓPEZ-SANJUAN, C., et al., 2023, “J-PLUS: Photometric Recalibration with the Stellar Color Regression Method and an Improved Gaia XP Synthetic Photometry Method”, Astrophys. J. Suppl., 269(2):58. doi: 10.3847/1538-4365/ad0645.

YANG, L., YUAN, H., XIANG, M., et al., 2022, “J-PLUS: Stellar parameters, C, N, Mg, Ca, and [α/Fe] abundances for two million stars from DR1”, Astron. Astrophys., 659:A181. doi: 10.1051/0004-6361/202142724.

YANNY, B., ROCKOSI, C., NEWBERG, H. J., et al., 2009, “SEGUE: A SPECTROSCOPIC SURVEY OF 240,000 STARS WITH g = 14–20”, *The Astronomical Journal*, v. 137, n. 5 (apr), pp. 4377. doi: 10.1088/0004-6256/137/5/4377. Disponível em: <https://dx.doi.org/10.1088/0004-6256/137/5/4377>.

YI, S., DEMARQUE, P., KIM, Y.-C., et al., 2001, “Toward better age estimates for stellar populations: The Y^2 isochrones for solar mixture”, Astrophys. J. Suppl., v. 136, n. 2 (out.), pp. 417–437. doi: 10.1086/321795. Disponível em: <https://ui.adsabs.harvard.edu/abs/2001ApJS..136..417Y>.

YORK, D. G., ADELMAN, J., ANDERSON, JR., J. E., et al., 2000, “The Sloan Digital Sky Survey: Technical Summary”, Astron. J., v. 120, n. 3 (set.), pp. 1579–1587. doi: 10.1086/301513.

YUAN, H. B., LIU, X. W., XIANG, M. S., 2013, “Empirical extinction coefficients for the GALEX, SDSS, 2MASS and WISE passbands”, Mon. Not, Roy. Astron. Soc., v. 430, n. 3 (abr.), pp. 2188–2199. doi: 10.1093/mnras/stt039.

YUAN, H. B., YANG, L., CRUZ, P., et al., 2023, “The miniJPAS survey: stellar atmospheric parameters from 56 optical filters”, Mon. Not, Roy. Astron. Soc., v. 518, n. 2 (jan.), pp. 2018–2033. doi: 10.1093/mnras/stac3155.

Zhang, C., Ma, Y. (Eds.), 2012, *Ensemble Machine Learning: Methods and Applications*. Engineering. 1 ed. New York, NY, Springer. ISBN: 978-1-4419-9325-0. doi: 10.1007/978-1-4419-9326-7. Disponível em: <https://doi.org/10.1007/978-1-4419-9326-7>.

ZHANG, R., WANG, G.-J., LU, Y. L., et al., 2024, "When LAMOST meets Gaia DR3: Exploring the metallicity of open clusters", *Astronomy & Astrophysics*, v. 692 (dez.), pp. A212. doi: 10.1051/0004-6361/202450726. Disponível em: <https://doi.org/10.1051/0004-6361/202450726>. Published online: 13 December 2024, Section: Galactic structure, stellar clusters and populations.

ZHANG, X., GREEN, G. M., RIX, H.-W., 2023, "Parameters of 220 million stars from Gaia BP/RP spectra", *Monthly Notices of the Royal Astronomical Society*, v. 524, n. 2, pp. 1855–1876. doi: 10.1093/mnras/stad188. Disponível em: <https://academic.oup.com/mnras/article/524/2/1855/7209172>.

ZHAO, T., LYU, R., WANG, H., et al., 2023, "Space-based gravitational wave signal detection and extraction with deep neural network", *Communications Physics*, v. 6, pp. 212. doi: 10.1038/s42005-023-01334-6.

ZHOU, Z.-H., 2012, *Ensemble Methods: Foundations and Algorithms*. 1st ed. New York, Chapman and Hall/CRC. ISBN: 9780429151095. doi: 10.1201/b12207. Disponível em: <https://doi.org/10.1201/b12207>.

ZHOU, Z.-H., 2025, *Ensemble Methods: Foundations and Algorithms*. 2nd ed. New York, Chapman and Hall/CRC. ISBN: 9781003587774. doi: 10.1201/9781003587774. Disponível em: <https://doi.org/10.1201/9781003587774>.

ZOU, H., GAO, J., ZHOU, X., et al., 2019, "Photometric Redshifts and Stellar Masses for Galaxies from the DESI Legacy Imaging Surveys", *The Astrophysical Journal Supplement Series*, v. 242, n. 1, pp. 8. doi: 10.3847/1538-4365/ab1847. Disponível em: <https://doi.org/10.3847/1538-4365/ab1847>.

ZUCKER, S., GIRYES, R., 2018, "Shallow Transits—Deep Learning. I. Feasibility Study of Deep Learning to Detect Periodic Transits of Exoplanets", *The Astronomical Journal*, v. 155, n. 4, pp. 147. doi: 10.3847/1538-3881/aaae05.

Appendix A

All filters in (mini)J-PAS

In Table A.1, detailed information is provided on all 57 filters used in J-PAS and the full set of 60 employed in miniJPAS. As discussed in the main text, miniJPAS includes three additional broad-band filters: uJPAS, gSDSS, and rSDSS, which are not present in the standard J-PAS configuration. Filter central wavelengths, and widths were retrieved from the Javalambre services using the Table Access Protocol (TAP) via the TOPCAT software. The table lists each filter's identifier, name, central wavelength (λ_{central}), bandwidth, a coefficient k_x relevant to calibration, and the assigned color code in hexadecimal format.

Table A.1: Detailed properties of all filters employed in J-PAS and miniJPAS surveys.

ID	Name	λ_{central} (Å)	Width (Å)	k_{x}	Color (hex)
1	uJAVA	3497	509	4.916	#CC00FF
2	J0378	3782	157	4.567	#9900FF
3	J0390	3904	150	4.442	#6F008C
4	J0400	3996	148	4.341	#6900B0
5	J0410	4110	144	4.241	#4300DE
6	J0420	4203	147	4.143	#0000FF
7	J0430	4303	148	4.048	#0069FF
8	J0440	4403	152	3.953	#00B8FF
9	J0450	4503	150	3.861	#00FFFF
10	J0460	4603	148	3.771	#00FFB8
11	J0470	4701	148	3.682	#00FF69
12	J0480	4799	142	3.596	#00FF10
13	J0490	4902	154	3.511	#2BFF00
14	J0500	5002	152	3.428	#4CFF00
15	J0510	5097	148	3.347	#67FF00
16	J0520	5202	150	3.268	#85FF00

ID	Name	$\lambda_{\mathrm{central}}$ (Å)	Width (Å)	k_{x}	Color (hex)
17	J0530	5296	150	3.190	#9CFF00
18	J0540	5389	152	3.115	#B3FF00
19	J0550	5498	149	3.041	#CEFF00
20	J0560	5596	150	2.969	#E5FF00
21	J0570	5701	150	2.900	#FFFF00
22	J0580	5803	148	2.832	#FFE300
23	J0590	5917	152	2.765	#FFC400
24	J0600	6010	150	2.701	#FFA600
25	J0610	6107	150	2.639	#FF8A00
26	J0620	6206	148	2.578	#FF6900
27	J0630	6309	150	2.519	#FF4600
28	J0640	6408	150	2.462	#FF1D00
29	J0650	6506	146	2.407	#FC0000
30	J0660	6607	151	2.354	#F70000
31	J0670	6710	146	2.302	#F10000
32	J0680	6812	152	2.251	#EC0000
33	J0690	6912	148	2.203	#E70000
34	J0700	7007	148	2.155	#E30000
35	J0710	7119	148	2.109	#DD0000
36	J0720	7207	146	2.064	#D80000
37	J0730	7307	150	2.020	#D30000
38	J0740	7414	148	1.977	#CD0000
39	J0750	7502	142	1.935	#C90000
40	J0760	7602	144	1.894	#C30000
41	J0770	7719	146	1.854	#BE0000
42	J0780	7811	145	1.815	#B80000
43	J0790	7907	142	1.776	#B30000
44	J0800	8009	140	1.739	#AE0000
45	J0810	8124	144	1.702	#A70053
46	J0820	8226	143	1.666	#A2004B
47	J0830	8329	148	1.630	#9C0043
48	J0840	8429	148	1.595	#97003B
49	J0850	8523	146	1.561	#910034
50	J0860	8620	148	1.528	#8B002D
51	J0870	8716	146	1.495	#860026
52	J0880	8810	146	1.462	#800020
53	J0890	8912	150	1.431	#7A0019
54	J0900	9000	154	1.400	#750014

ID	Name	$\lambda_{\mathrm{central}}$ (Å)	Width (Å)	k_{x}	Color (hex)
55	J0910	9107	152	1.370	#6F000E
56	J1007	9316	635	1.119	#610001
57	uJPAS	3623	736	4.884	#B400FF
58	gSDSS	4750	1250	3.629	#006600
59	rSDSS	6250	1250	2.527	#FF0000
60	iSDSS	7725	1250	1.825	#990033

Appendix B

Query for retrieving stars in J-PLUS DR3

The following ADQL query was implemented using Python to get stars from JPLUS DR3 using restricted criteria regarding flags, uncertainties and probability of being a star. This approach is necessary given the limitation in J-PLUS's online data access services of returning up to 1 million rows per query. The conditions to which this query is restricted as listed below, and the query is exhibited on the next page.

- `norm_wmap_val` > 0.8: adequate exposure,
- `mask_flags` $= 0$: avoiding any known issues affecting the quality of the photometry,
- `mag_err_aper_cor_6_0` > 0.5: filtering mostly obvious bad quality sources,
- `mag_aper_cor_3_0` < 22: limiting magnitude, with some small tolerance, as presented in LÓPEZ-SANJUAN *et al.* (2024),
- `sglc_prob_star` > 0.7: 70% or higher probability of being a star.

```sql
1  SELECT
2  dualobj.number, dualobj.tile_id,
3  dualobj.alpha_j2000, dualobj.delta_j2000,
4  sgc.sglc_prob_star, dualobj.class_star,
5  ps.mag_aper_cor_6_0, ps.mag_err_aper_cor_6_0,
6  ext.ax, ext.ax_err, ext.ebv, ext.ebv_err,
7  dualobj.mag_aper_3_0, dualobj.mag_err_aper_3_0,
8  dualobj.mag_aper_6_0, dualobj.mag_err_aper_6_0,
9  dualobj.mag_auto, dualobj.mag_err_auto,
10 array_sub(dualobj.mag_auto, ext.ax) AS mag_auto_0,
11 dualobj.norm_wmap_val,
12 ps.mag_aper_cor_3_0, ps.mag_err_aper_cor_3_0,
13 ti.aper_cor_3_0, ti.aper_cor_6_0, ti.aper_cor_err_6_0
14
15 FROM
16 jplus.MagABDualObj AS dualobj
17 INNER JOIN jplus.MagABDualPointSources AS ps ON ps.number =
      dualobj.number AND ps.tile_id = dualobj.tile_id
18 INNER JOIN jplus.MWExtinction AS ext ON ext.number = dualobj.
      number AND ext.tile_id = dualobj.tile_id
19 INNER JOIN jplus.StarGalClass AS sgc ON sgc.number = dualobj.
      number AND sgc.tile_id = dualobj.tile_id
20 INNER JOIN jplus.TileImage AS ti ON ti.tile_id = dualobj.
      tile_id
21
22 WHERE
23 (array_min_float(dualobj.norm_wmap_val) > 0.8) AND
24 (array_max_int(dualobj.mask_flags) = 0) AND
25 (array_max_float(ps.mag_err_aper_cor_6_0) < 0.5) AND
26 (array_max_float(dualobj.mag_aper_3_0) < 22) AND
27 sgc.sglc_prob_star >= 0.7 AND
28 dualobj.tile_id IN {0}
```

The typeset 0 at line 28 is responsible for looping over 83 blocks files containing the merged 1642 reference tiles.

Appendix C

Median Absolute Deviation (MAD)

The median absolute deviation is a measure of dispersion defined as the median over the absolute deviations from the sample median:

$$\mathrm{MAD} = \mathrm{median}\left(|x_i - \mathrm{median}(x)|\right). \tag{C.1}$$

Despite similarities, unlike the standard deviation, MAD is more robust to outliers, making it well-suited for astrophysical data with potential contamination. However, without proper scaling, the MAD underestimates the true variability if interpreted as a standard deviation measure. Hence, a scaling factor, k, is necessary because the raw MAD and the traditional standard deviation measure variability on different numerical scales due to their definitions and sensitivity to data values. The former is based on squared deviations from the mean, which emphasizes larger deviations, while the latter uses median absolute deviations from the median, making it more robust to outliers but systematically smaller in magnitude for the same data. Applying the scaling factor adjusts the MAD so that it enables meaningful comparison and interpretation of dispersion across methods. Ignoring that factor means ignoring these differences, leading to inconsistent or biased estimates of spread. This way, a consistent estimator for that standard deviation is written as:

$$\sigma = k \times \mathrm{MAD}, \tag{C.2}$$

where k is the scaling factor that corrects for the biased tendency of the raw MAD.

To derive the scaling factor for normally distributed data, which is typical in that context (ROUSSEEUW & CROUX, 1993), note that the MAD corresponds to the middle 50% of the data around the median. Formally, the probability that a normally distributed value lies within ±MAD of the median is 50%:

$$P\big(-\mathrm{MAD} \leqslant X - \mu \leqslant \mathrm{MAD}\big) = 0.5, \tag{C.3}$$

where μ is the mean (same as median for normal data). Using the standardized variable

$$Z = \frac{X-\mu}{\sigma} \quad \Rightarrow \quad P(-c \leqslant Z \leqslant c) = 0.5, \tag{C.4}$$

with

$$c = \frac{\text{MAD}}{\sigma}. \tag{C.5}$$

This means that the standardized variable, Z, holds 50% probability of lying between $-c$ and c. Using the cumulative distribution function (CDF), Φ, of the standard normal distribution[1], this probability is expressed as:

$$\Phi(c) - \Phi(-c) = 0.5, \tag{C.6}$$

By symmetry of the normal distribution, $\Phi(-c) = 1 - \Phi(c)$, so

$$\Phi(c) - (1 - \Phi(c)) = 0.5 \quad \Rightarrow \quad 2\Phi(c) - 1 = 0.5 \quad \Rightarrow \quad \Phi(c) = 0.75. \tag{C.7}$$

Thus, c is the value whose cumulative probability is 0.75, i.e. the 75th percentile of the standard normal distribution. From standard normal tables (or computational evaluation, since the CDF itself does not have a closed-form solution with elementary functions):

$$c = \Phi^{-1}(0.75) \approx 0.67449. \tag{C.8}$$

Since $c = \text{MAD}/\sigma$, the scale factor k is

$$k = \frac{1}{c} = \frac{1}{0.67449} \approx 1.4826, \tag{C.9}$$

and the standard deviation estimator is given by:

$$\sigma = 1.4826 \text{ MAD}. \tag{C.10}$$

which represents a robust and unbiased estimate of the standard deviation using MAD.

This scaled MAD (robust MAD, referred here as rMAD) was used in Chapter 4.2 through the `mad_std`[2] function from the `astropy` package to exclude stars with distances significantly deviating from cluster medians and to estimate internal scatter on cluster metallicities with reduced influence of outliers. Same results are found using the `scipy.stats.median_abs_deviation`[3] function.

---

[1] Calculated by integrating the normal probability density function (PDF) from negative infinity up to x: $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$

[2] https://docs.astropy.org/en/stable/api/astropy.stats.mad_std.html

[3] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.median_abs_deviation.html

Appendix D

General information on OCs

Table D.1 shows general information for the 23 clusters recovered in this work: metallicities from this work and useful OCs information, namely distances (both individually computed from BJ21 based on GaiaEDR3 and provided by HR23), relative difference between distances (in relation to distance of HR23, ages, and type. Here, Δ means the difference (HR23 − BJ21).

Table D.1: Metallicities from this work and useful OCs information, namely distances (both individually computed from GaiaEDR3 (BAILER-JONES *et al.*, 2021) and provided by HR23), relative difference between distances, ages, and type.

Cluster -	N -	$[\mathrm{Fe/H}]_{\mathrm{TW}}$ (dex)	$d^{\mathrm{median}}_{\mathrm{rpgeo}}$ (pc)	$d^{\mathrm{median}}_{\mathrm{HR23}}$ (pc)	$\frac{\Delta}{d_{\mathrm{HR23}}}$ -	Age (log yr)	Type -
ASCC 100	5	-0.094 ± 0.237	$367.0^{+4.3}_{-4.3}$	$360.4^{+0.3}_{-0.3}$	0.018	$7.812^{+0.299}_{-0.303}$	o → m
Alessi 62	10	-0.151 ± 0.052	$614.6^{+4.3}_{-17.3}$	$606.3^{+0.6}_{-0.6}$	0.014	$8.629^{+0.253}_{-0.206}$	o
Alessi 84	10	-0.276 ± 0.099	$196.8^{+3.2}_{-2.2}$	$198.3^{+0.1}_{-0.1}$	0.008	$8.068^{+0.248}_{-0.252}$	o → m
Gaia 8	5	-0.103 ± 0.204	$286.4^{+8.5}_{-12.9}$	$289.2^{+0.2}_{-0.2}$	0.010	$7.324^{+0.254}_{-0.256}$	o
HSC 1058	1	-0.488 ± 0.234	$485.5^{+5.4}_{-5.2}$	$501.6^{+0.7}_{-0.7}$	0.032	$8.069^{+0.239}_{-0.276}$	m
HSC 1448	1	-0.651 ± 0.096	$115.8^{+0.6}_{-0.5}$	$118.0^{+0.1}_{-0.1}$	0.019	$8.318^{+0.330}_{-0.298}$	m
HSC 1555	1	-0.242 ± 0.153	$171.4^{+2.1}_{-1.8}$	$172.9^{+0.2}_{-0.2}$	0.009	$8.107^{+0.270}_{-0.271}$	o → m
HSC 474	2	-0.135 ± 0.013	$328.4^{+11.2}_{-11.2}$	$315.4^{+0.3}_{-0.3}$	0.041	$7.491^{+0.257}_{-0.244}$	o → m
HSC 590	1	-0.014 ± 0.116	$115.1^{+0.1}_{-0.1}$	$118.1^{+0.0}_{-0.0}$	0.026	$8.504^{+0.313}_{-0.285}$	o → m
HSC 598	2	-0.597 ± 0.090	$193.1^{+1.6}_{-1.6}$	$184.3^{+0.1}_{-0.1}$	0.048	$8.212^{+0.342}_{-0.324}$	o → m
HSC 695	4	-0.791 ± 0.255	$6975.5^{+1577.1}_{-1448.0}$	$9550.3^{+827.3}_{-705.1}$	0.270	$8.826^{+0.532}_{-0.398}$	m
HSC 749	31	-0.271 ± 0.085	$118.4^{+5.1}_{-4.1}$	$121.2^{+0.0}_{-0.0}$	0.023	$8.354^{+0.301}_{-0.408}$	o → m
HSC 759	6	-0.262 ± 0.147	$97.7^{+5.1}_{-4.3}$	$96.0^{+0.0}_{-0.1}$	0.019	$8.174^{+0.269}_{-0.294}$	o → m
HSC 921	1	-0.343 ± 0.126	$166.9^{+0.8}_{-0.9}$	$172.9^{+0.1}_{-0.1}$	0.035	$7.635^{+0.280}_{-0.236}$	o → m
HSC 979	1	-0.090 ± 0.163	$314.1^{+2.5}_{-1.8}$	$317.2^{+0.4}_{-0.4}$	0.010	$8.263^{+0.266}_{-0.358}$	m
Melotte 111	1	-0.317 ± 0.172	$87.7^{+0.2}_{-0.3}$	$85.3^{+0.0}_{-0.0}$	0.029	$8.816^{+0.221}_{-0.195}$	o
NGC 1039	33	-0.096 ± 0.034	$492.2^{+15.1}_{-12.3}$	$489.7^{+0.3}_{-0.3}$	0.005	$8.087^{+0.280}_{-0.232}$	o
NGC 752	4	-0.162 ± 0.007	$435.8^{+4.8}_{-11.2}$	$433.6^{+0.3}_{-0.3}$	0.005	$9.152^{+0.216}_{-0.182}$	o
Stephenson 1	14	-0.254 ± 0.198	$353.0^{+4.4}_{-9.0}$	$354.7^{+0.2}_{-0.2}$	0.005	$7.425^{+0.275}_{-0.262}$	o
Theia 1188	5	-0.172 ± 0.165	$324.2^{+4.6}_{-2.4}$	$325.2^{+0.3}_{-0.3}$	0.003	$8.756^{+0.236}_{-0.237}$	o → m
UBC 26	11	-0.091 ± 0.077	$600.7^{+16.5}_{-9.9}$	$593.5^{+0.5}_{-0.5}$	0.012	$7.401^{+0.208}_{-0.211}$	o
UBC 616	1	-0.215 ± 0.213	$2823.7^{+242.0}_{-160.2}$	$2758.4^{+44.5}_{-43.2}$	0.024	$9.050^{+0.211}_{-0.171}$	o
UPK 303	4	-0.211 ± 0.217	$215.8^{+3.8}_{-2.2}$	$211.2^{+0.1}_{-0.1}$	0.022	$8.045^{+0.266}_{-0.314}$	o → m

Appendix E

Comparison with literature

Tables E.1 and E.2 show the data used for comparison between metallicities of this work and from literature, namely CAVALLO *et al.* (2023, hereafter C23) and ZHANG *et al.* (2024, hereafter Z24) in Table E.1, and LAMOST and APOGEE in Table E.2. Number of used stars to estimate [Fe/H] are represented by N and internal scatter by robust MAD (rMAD)

Table E.1: Metallicities from this work and from C23 and Z24.

Cluster	N_{TW}	$[\mathrm{Fe/H}]_{\mathrm{TW}}$	$\mathrm{rMAD}^{[\mathrm{Fe/H}]}_{\mathrm{TW}}$	$[\mathrm{Fe/H}]_{\mathrm{C23}}$	$[\mathrm{Fe/H}]_{\mathrm{Z24}}$
-	-	(dex)	(dex)	(dex)	(dex)
ASCC 100	5	-0.094 ± 0.237	0.163	0.060 ± 0.145	-0.017 ± 0.111
Alessi 84	10	-0.276 ± 0.099	0.235	0.280 ± 0.090	0.112 ± 0.074
Gaia 8	5	-0.103 ± 0.204	0.100	-0.260 ± 0.080	0.035 ± 0.032
HSC 1058	1	-0.488 ± 0.234	0.234	0.180 ± 0.105	0.072 ± 0.103
HSC 1448	1	-0.651 ± 0.096	0.096	0.200 ± 0.065	-0.038 ± 0.081
HSC 1555	1	-0.242 ± 0.153	0.153	0.170 ± 0.095	0.117 ± 0.074
HSC 474	2	-0.135 ± 0.013	0.026	-0.090 ± 0.080	0.103 ± 0.048
HSC 590	1	-0.014 ± 0.116	0.116	0.050 ± 0.060	0.269 ± 0.237
HSC 598	2	-0.597 ± 0.090	0.187	-0.030 ± 0.085	0.031 ± 0.035
HSC 749	31	-0.271 ± 0.085	0.283	0.170 ± 0.060	0.034 ± 0.097
HSC 759	6	-0.262 ± 0.147	0.290	0.020 ± 0.065	0.156 ± 0.078
HSC 921	1	-0.343 ± 0.126	0.126	0.000 ± 0.115	-0.007 ± 0.090
HSC 979	1	-0.090 ± 0.163	0.163	-0.050 ± 0.115	0.044 ± 0.075
Melotte 111	1	-0.317 ± 0.172	0.172	0.200 ± 0.110	0.062 ± 0.088
NGC 1039	33	-0.096 ± 0.034	0.117	0.270 ± 0.105	0.005 ± 0.138
NGC 752	4	-0.162 ± 0.007	0.007	0.000 ± 0.205	0.005 ± 0.091
Stephenson 1	14	-0.254 ± 0.198	0.401	0.000 ± 0.090	0.158 ± 0.065
Theia 1188	5	-0.172 ± 0.165	0.208	0.410 ± 0.115	0.141 ± 0.090
UBC 26	11	-0.091 ± 0.077	0.161	0.040 ± 0.110	0.124 ± 0.036
UBC 616	1	-0.215 ± 0.213	0.213	0.130 ± 0.160	-0.198 ± 0.263
UPK 303	4	-0.211 ± 0.217	0.244	0.090 ± 0.075	0.061 ± 0.086

Table E.2: Metallicities from this work and from LAMOST and APOGEE surveys. Number of objects retrieved and scattering are also reported, represented by N and the rMAD, respectively.

Cluster	N_{TW}	$[\mathrm{Fe/H}]_{\mathrm{TW}}$	$\mathrm{rMAD}_{\mathrm{TW}}^{[\mathrm{Fe/H}]}$	N_{LAMOST}	$[\mathrm{Fe/H}]_{\mathrm{LAMOST}}$	$\mathrm{rMAD}_{\mathrm{LAMOST}}^{[\mathrm{Fe/H}]}$	N_{APOGEE}	$[\mathrm{Fe/H}]_{\mathrm{APOGEE}}$	$\mathrm{rMAD}_{\mathrm{APOGEE}}^{[\mathrm{Fe/H}]}$
-	-	(dex)	(dex)	-	(dex)	(dex)	-	(dex)	(dex)
ASCC 100	5	-0.094 ± 0.237	0.163	2	0.000 ± 0.115	0.239	-	-	-
Alessi 84	10	-0.276 ± 0.099	0.235	19	0.083 ± 0.026	0.077	-	-	-
Gaia 8	5	-0.103 ± 0.204	0.100	2	0.026 ± 0.009	0.018	-	-	-
HSC 1058	1	-0.488 ± 0.234	0.234	27	-0.001 ± 0.032	0.105	-	-	-
HSC 1448	1	-0.651 ± 0.096	0.096	2	-0.006 ± 0.040	0.084	2	-0.005 ± 0.047	0.098
HSC 1555	1	-0.242 ± 0.153	0.153	10	0.081 ± 0.045	0.093	5	-0.017 ± 0.062	0.087
HSC 474	2	-0.135 ± 0.013	0.026	5	0.114 ± 0.068	0.144	2	-0.002 ± 0.011	0.023
HSC 590	1	-0.014 ± 0.116	0.116	2	0.006 ± 0.022	0.045	3	-0.126 ± 0.124	0.045
HSC 598	2	-0.597 ± 0.090	0.187	1	0.062 ± 0.018	0.018	3	-0.082 ± 0.065	0.031
HSC 749	31	-0.271 ± 0.085	0.283	7	-0.022 ± 0.115	0.163	3	-0.039 ± 0.008	0.006
HSC 759	6	-0.262 ± 0.147	0.290	22	0.069 ± 0.026	0.099	14	0.001 ± 0.013	0.048
HSC 921	1	-0.343 ± 0.126	0.126	11	-0.072 ± 0.062	0.107	-	-	-
HSC 979	1	-0.090 ± 0.163	0.163	17	0.028 ± 0.022	0.061	-	-	-
Melotte 111	1	-0.317 ± 0.172	0.172	8	-0.068 ± 0.040	0.050	14	-0.022 ± 0.011	0.034
NGC 1039	33	-0.096 ± 0.034	0.117	99	-0.018 ± 0.013	0.113	-	-	-
NGC 752	4	-0.162 ± 0.007	0.007	49	-0.062 ± 0.011	0.074	105	-0.048 ± 0.012	0.080
Stephenson 1	14	-0.254 ± 0.198	0.401	4	0.191 ± 0.040	0.071	-	-	-
Theia 1188	5	-0.172 ± 0.165	0.208	12	0.062 ± 0.033	0.105	-	-	-
UBC 26	11	-0.091 ± 0.077	0.161	2	0.146 ± 0.016	0.034	-	-	-
UBC 616	1	-0.215 ± 0.213	0.213	3	-0.322 ± 0.103	0.209	-	-	-
UPK 303	4	-0.211 ± 0.217	0.244	24	0.029 ± 0.034	0.132	-	-	-

Monthly Notices
of the Royal Astronomical Society

J-PLUS: metallicity determination in open clusters using machine learning

Journal:	*Monthly Notices of the Royal Astronomical Society*
Manuscript ID	MN-25-1285-MJ
Manuscript type:	Main Journal
Date Submitted by the Author:	06-Jun-2025
Complete List of Authors:	Machado-Pereira, Eduardo; Observatório Nacional; NSF’s NOIRLab Daflon, S.; Observatório Nacional Placco, Vinicius; NSF’s NOIRLab, Cruz, Patricia; Centro de Astrobiología, Departamento de Astrofísica Borges Fernandes, Marcelo; Observatorio Nacional, Jiménez-Esteban, Francisco; Centro de Astrobiología (INTA-CSIC), Departamento de Astrofísica Caballero, José; Centro de Astrobiologia, Astrophysics Anguiano, Borja; CEFCA; University of Virginia, Department of Astronomy Yuan, H.-B.; Beijing Normal University; Beijing Normal University Bonatto, Charles; UFRGS, Departamento de Astronomia Alvarez-Candal, Alvaro; Instituto de Astrofísica de Andalucía Cenarro, Javier; CEFCA, Hornillos, David Hernández-Monteagudo, C.; Instituto de Astrofísica de Canarias, Department of Astrophysics Research; Universidad de La Laguna, Departamento de Astrofísica López-Sanjuan, Carlos; CEFCA, Galaxy Evolution Marín-Franch, Antonio; CEFCA, Moles, Mariano; Instituto de Astrofisica de Andalusia (C.S.I.C.), Apartado 2004, 1 8080 Granada, Varela, Jesús; CEFCA, Galaxy evolution Vázquez Ramió, Héctor; Centro de Estudios de Física del Cosmos de Aragón, Scientific Operation Alcaniz, Jailson; Observatorio Nacional, Departamento de Astronomia Dupke, Renato de Alencar; Observatorio Nacional, Astronomy Ederoclite, Alessandro Sodre Junior, Laerte; IAG, Astronomy Angulo, Raul; Donostia International Physics Center, ;
Keywords:	stars: fundamental parameters < Stars, methods: data analysis < Astronomical instrumentation, methods, and techniques, software: machine learning < Software, (Galaxy:) open clusters and associations: general < The Galaxy, techniques: photometric < Astronomical instrumentation, methods, and techniques, surveys < Astronomical Data bases

MNRAS **000**, 1–19 (2025) Preprint 6 June 2025 Compiled using MNRAS LATEX style file v3.0

J-PLUS: metallicity determination in open clusters using machine learning

E. Machado-Pereira[1,2★], S. Daflon[1], V. M. Placco[2], P. Cruz[3], M. Borges Fernandes[1], F. Jiménez-Esteban[3], J. A. Caballero[3], B. Anguiano[4,5], H. Yuan (苑海波)[6,7], C. Bonatto[8], A. Alvarez-Candal[9], J. Cenarro[4], D. Cristóbal-Hornillos[4], C. Hernández-Monteagudo[4], C. López-Sanjuan[4], A. Marín-Franch[4], M. Moles[4], J. Varela[4], H. Vázquez Ramió[4], J. Alcaniz[1], R. Dupke[1], A. Ederoclite[4], L. Sodré Jr.[10], R. E. Angulo[11,12]

[1] *Observatório Nacional - MCTI (ON), Rua Gal. José Cristino 77, São Cristóvão, 20921-400 Rio de Janeiro, Brazil*
[2] *NSF NOIRLab, Tucson, AZ 85719, USA*
[3] *Centro de Astrobiología (CAB), CSIC-INTA, Camino Bajo del Castillo s/n, E-28692, Villanueva de la Cañada, Madrid, Spain.*
[4] *Centro de Estudios de Física del Cosmos de Aragón (CEFCA), Plaza San Juan 1, E-44001 Teruel, Spain*
[5] *Department of Astronomy, University of Virginia, Charlottesville, VA, 22904, USA*
[6] *Institute for Frontiers in Astronomy and Astrophysics, Beijing Normal University, Beijing 102206, People's Republic of China*
[7] *Department of Astronomy, Beijing Normal University, Beijing 100875, People's Republic of China*
[8] *Universidade Federal do Rio Grande do Sul, Departamento de Astronomia, CP 15051, RS, Porto Alegre 91501-970, Brazil*
[9] *Instituto de Astrofísica de Andalucía, Glorieta de la Astronomía S/N, 18008 Granada, Spain*
[10] *Universidade de São Paulo, Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Departamento de Astronomia, SP 05508-090, São Paulo, Brazil*
[11] *Donostia International Physics Center (DIPC), Manuel Lardizabal Ibilbidea, 4, E-20018 Donostia, San Sebastián, Spain*
[12] *Ikerbasque, Basque Foundation for Science, E-48013 Bilbao, Spain*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We present a machine learning approach to estimate stellar atmospheric parameters ($T_{\rm eff}$, log g, and [Fe/H]) for stars in the J-PLUS DR3 footprint using photometric and astrometric data from J-PLUS, Gaia, and CatWISE. Our models are based on the gradient boosting algorithm LightGBM and were trained using extinction-corrected photometry and absolute magnitudes derived from Gaia parallaxes. Through a systematic feature selection process, we identified the most informative features for each parameter, showing that using the full set of selected features yields the best overall performance. Evaluation against spectroscopic values from LAMOST reveals that our models achieve competitive mean absolute errors of 42 K, 0.06 dex and 0.06 dex for $T_{\rm eff}$, log g, and [Fe/H], respectively. We applied the trained models to a curated sample of 154 stars belonging to open clusters and moving groups. While our metallicity estimates are mild systematically underestimated by -0.13 dex, we attribute this to the compression of training data in a compact and non-linear region of the feature space. Despite this bias, the relative metallicity patterns among clusters remain consistent with literature values, demonstrating the robustness of our predictions for comparative studies. Our method provides a reliable alternative for estimating stellar parameters using multi-band photometry combined with astrometric information, offering further applications to surveys such as S-PLUS and J-PAS. Future improvements for parameters predictions will focus on additional magnitudes combinations, while on color–magnitude diagram decontamination techniques for cluster member selection.

Key words: stars: fundamental parameters – methods: data analysis – software: machine learning – Galaxy: open clusters and associations: general – techniques: photometric – surveys

1 INTRODUCTION

Over the past decades, repeated and extensive coverage of large sky regions has enabled numerous large-scale astronomical surveys. The systematic collection and processing of vast data volumes have paved the way for increasingly ambitious spectroscopic and photometric efforts. Precise determinations of stellar atmospheric parameters, typically provided by large spectroscopic surveys, are critical for addressing a wide range of questions in stellar astrophysics.

One of the most extensive efforts is the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST), which operates from the Xinglong Station, China. LAMOST is a 4-meter telescope that carries observations of up to 4000 objects via a fiber-fed spectrograph, covering 3700–9000 Å at R ≈ 1800 (Cui et al. 2012; Zhao et al. 2012). The LAMOST Stellar Parameter Pipeline (LASP) estimates $T_{\rm eff}$, log g, and [Fe/H] using template matching and machine learning techniques (Wu et al. 2014). From data release 6 (DR6) on, more recent data releases include higher-resolution spectra (R ≈ 7500; Liu et al. 2020), expanding its capability for chemical abundance studies. Other prominent spectroscopic efforts further include the Apache Point Observatory Galactic Evolution Experiment (APOGEE; Majewski et al. 2017), the Sloan Extension for Galactic Understanding and Exploration (SEGUE; Yanny et al. 2009), and the GALactic

---

★ E-mail: eduardopereira@on.br

Archaeology with HERMES (GALAH; De Silva et al. 2015; Buder et al. 2021).

Despite their precision, spectroscopic surveys are limited by high costs and complex data reduction. In contrast, photometric surveys offer a more efficient way to observe large stellar samples over wide sky areas. The combination of foundational results from spectroscopy with vast datasets from photometry enable the estimation of fundamental stellar parameters that serve as a powerful alternative for wide-area stellar characterization.

Recent wide-field photometric optical surveys have significantly enhanced this potential by delivering extensive and high-quality data for stellar studies. The Sloan Digital Sky Survey (SDSS; York et al. 2000) has been one of the most influential, serving as a foundation for numerous stellar (e.g. Belokurov et al. 2006; Ivezić et al. 2008; Bochanski et al. 2010) and extragalactic studies (e.g. Kauffmann et al. 2003; Lietzen et al. 2016). Other notable surveys include the SkyMapper Southern Survey (SMSS; Keller et al. 2007; Onken et al. 2024), the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS; Chambers et al. 2016) and the Pristine survey (Pristine; Starkenburg et al. 2017).

The *Gaia* mission (Gaia Collaboration et al. 2016; Gaia Collaboration 2022) is arguably the most transformative endeavor and has revolutionized Milky Way studies by delivering microarcsecond-precision astrometry and broad-band photometry in three optical filters for nearly two billion sources (Gaia Collaboration et al. 2023; Creevey et al. 2023; Babusiaux et al. 2023). Its photometric system enables accurate estimates of stellar temperatures, luminosities, and variability, making it critical for stellar population research.

The all-sky Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010) has provided mid-infrared photometry essential for stellar and extragalactic studies. By utilizing the full-depth imaging over multiple epochs and intermediate versions (NEOWISE, AllWISE, unWISE; Mainzer et al. 2014; Cutri et al. 2013; Schlafly et al. 2019, respectively), the CatWISE2020 catalog (CatWISE; Marocco et al. 2021) extends WISE data by delivering a pre-extracted almost 2 billion source list with W1 and W2 photometry.

Dedicated surveys relying on multi-band photometric systems have also been developed to serve a wide range of scientific fields. One such effort is the Javalambre Photometric Local Universe Survey (J-PLUS; Cenarro et al. 2019), carried out with the Javalambre Auxiliary Survey Telescope (JAST/T80), an 83 cm telescope mounted at the Observatorio Astrofísico de Javalambre (OAJ, Teruel, Spain). J-PLUS, currently on its third data release (DR3), was initially planned to provide photometric calibration of the Javalambre Physics of the Accelerating Universe Survey (J-PAS; Benitez et al. 2014), conducted with the 250 cm Javalambre Survey Telescope (JST250) also from OAJ. J-PAS is currently in its early data release (EDR[1]) stage after an initial, 1 deg² wide approach with miniJPAS[2] (Bonoli et al. 2021). At the Cerro Tololo Inter-American Observatory (CTIO, Chile), a southern hemisphere effort complementary to J-PLUS is being conducted with an identical optical system and telescope, T80-South: the Southern Photometric Local Universe Survey (S-PLUS; Mendes de Oliveira et al. 2019), which as of now provides the internal fifth data release (iDR5) for collaboration members and DR4 publicly available (Herpich et al. 2024).

Such wealth of data provided by recent large surveys has enabled astronomy to enter the era of data-rich science, creating ideal conditions for applying deep and machine learning methods. While recent efforts increasingly use these techniques to extract insights from both spectroscopic and photometric vast survey datasets, we highlight a few studies which predict stellar atmospheric parameters.

Using Gaia XP spectra, Andrae et al. (2023) trained an XGBoost model on APOGEE and metal-poor stars to estimate metallicity, $T_{\rm eff}$, and log g for 175 million stars, achieving typical uncertainties of ~ 0.1 dex, ~ 120 K, and ~ 0.2 dex, respectively. Zhang et al. (2023) applied a neural-network-based forward model, trained on LAMOST and incorporating 2MASS (Skrutskie et al. 2006) and WISE photometry, to derive parameters for 220 million stars with improved precision (~ 90 K in $T_{\rm eff}$, ~ 0.15 dex in [Fe/H] and log g). More recently, Huang et al. (2025) used synthetic Gaia colors and the stellar locus method to estimate [Fe/H] for 100 million stars, with typical uncertainties ranging from 0.05 dex at solar metallicity to 0.5 dex near $[{\rm Fe/H}] \approx -4$.

Several works using data from J-PLUS also employed machine learning techniques. Whitten et al. (2019) applied their artifical neural network, SPHINX, to J-PLUS photometry to estimate stellar effective temperatures and metallicities, achieving typical uncertainties of ~ 91 K in $T_{\rm eff}$ and ~ 0.25 dex in [Fe/H]. Yang et al. (2022) applied a cost-sensitive neural network (CSNet) to J-PLUS DR1 photometry to estimate stellar parameters and elemental abundances for two million stars. Their method achieved precisions of ~ 55 K in $T_{\rm eff}$, ~ 0.15 dex in log g, and ~ 0.07 dex in [Fe/H], as well as uncertainties of 0.04–0.08 dex for [C/Fe], [N/Fe], [Mg/Fe] and $[\alpha/{\rm Fe}]$. In a complementary approach, Wang et al. (2022) used support vector regression (SVR) on J-PLUS DR1 to estimate stellar parameters for 2.5 million stars, obtaining root mean square errors of ~ 160 K in $T_{\rm eff}$, ~ 0.35 dex in log g, and ~ 0.25 dex in [Fe/H].

Galarza et al. (2022) developed the SPEEM pipeline, using Random Forest and XGBoost to estimate $T_{\rm eff}$, log g, and [Fe/H] from J-PLUS photometry. Applied to J-PLUS DR2, it achieved typical errors of 41 K for $T_{\rm eff}$, 0.11 dex for log g, and 0.09 dex for [Fe/H] successfully identifying a number of very metal-poor (VMP) and extremely metal-poor (EMP) stars with a 64% success rate in spectroscopic validation. Quispe-Huaynasi et al. (2023) identified 28 stars with Galactocentric velocities exceeding 400 km s^{-1} and inferred stellar parameters using a convolutional neural network (CNN) trained on LAMOST data, achieving typical uncertainties of ~ 100 K in $T_{\rm eff}$, ~ 0.12 dex in [Fe/H], and ~ 0.18 dex in log g. More recently, Huang et al. (2024) extended these efforts to J-PLUS DR3, deriving [Fe/H], [C/Fe], [Mg/Fe], and $[\alpha/{\rm Fe}]$ for five million stars using kernel principal component analysis. Their results achieved typical precisions of 0.10–0.20 dex for [Fe/H] and [C/Fe], and ~ 0.05 dex for [Mg/Fe] and $[\alpha/{\rm Fe}]$. Furthermore, given the resemblance in filter systems, S-PLUS has also been similarly used for stellar parameter estimation in the southern sky (Whitten et al. 2021; Quispe-Huaynasi et al. 2024; Lopes et al. 2025; Quispe-Huaynasi et al. 2025).

Although a vast body of literature has emerged over the last decade using S-PLUS, J-PLUS, and (mini)J-PAS, studies on open clusters (OCs) within these surveys remain scarce. Several works in literature have identified thousands of new OCs (e.g. Bica et al. 2019; Cantat-Gaudin et al. 2020; Dias et al. 2021; Hunt & Reffert 2021; Hunt & Reffert 2023) in recent years, largely due to the high-precision astrometry provided by Gaia (Hunt & Reffert 2021; Perren et al. 2023). Although the known sample is far from complete (following e.g. Bonatto et al. 2006, an upper limit is estimated to be over 10^5 OCs), these advances have drastically expanded the census of galactic OCs, bringing the number of OCs candidates to $\sim$ 14 000

---

[1] `https://www.j-pas.org/datareleases/jpas_early_data_release`

[2] `https://www.j-pas.org/datareleases/minijpas_public_data_release_pdr201912`

(Perren et al. 2023) and confirmed ones to ∼ 5600 (Hunt & Reffert 2024).

In this context, multi-band approaches can be particularly valuable for OC research, especially since narrow-band filters capture distinct astrophysical features, enhancing the characterization of both cluster properties and individual members in various aspects (e.g. Bica et al. 2019; Fragkou et al. 2019; Sánchez et al. 2021; Mayo et al. 2023; Nardiello et al. 2023; Cavallo et al. 2023). Independently from J-PLUS, the work of Sánchez et al. (2021) also employs the same 12-band optical system at the JAST80 telescope for OC studies. By combining the multi-band approach with Gaia broad-bands, their results highlight the benefits of relying on J-PLUS filters for cluster characterization, particularly in SED fitting for extinction estimation.

The primary goal of this work is to develop a novel tool for estimating stellar atmospheric parameters, particularly addressing this gap in OCs studies within J-PLUS. In this initial approach, we derive parameters for individual member stars of OCs, identified from the literature, within the J-PLUS DR3 footprint. These parameters were determined using machine learning models trained on both photometric and spectroscopic samples. In addition to the twelve J-PLUS filters, we incorporate photometry from CatWISE ($W1$, $W2$) and Gaia (G, B_P and R_P), totaling 17 distinct photometric bands to construct our models. Spectroscopic data from LAMOST DR8 (low resolution) was also utilized.

Spectroscopically determined stellar atmospheric parameters (effective temperature - T_eff, surface gravity - $\log g$, and metallicity - [Fe/H]) constitute the input *targets* we want the model to reproduce using the photometric information as input *features*. In other words, the model must map the patterns and relations between the targets and the features seeking to minimize the difference between predicted and true values.

This paper is structured as follows: Section 2 details the data retrieval process, followed by the construction of the models in Section 3. Results and comparisons for individual OC members are presented in Section 4, while results for clusters are presented in Section 5. Finally, our conclusions and perspectives for future work are given in Section 6.

2 DATA

In this section, we describe the datasets used to train and validate the machine learning algorithms used in this work. We elaborate on the cross-matches performed for sample construction and the preprocessing steps undertaken to enhance its quality and suitability. Then, we present the sample of open clusters members to which our models will be applied as a science case.

2.1 Training sample

2.1.1 J-PLUS DR3

Covering a 2 deg² field of view with $0.55''$ pixel^{-1} resolution, J-PLUS is conducted using the JAST80 telescope and the T80Cam wide-field camera. It is designed to image ∼8500 deg²of the northern sky in 12 optical bands: five broad (Sloan-like *ugriz*) and seven narrow, optimized for the detection of interesting astrophysical features.

In particular, its third data release (DR3; López-Sanjuan et al. 2024) includes 47.4 million sources across 3192 deg², delivering a unique multicolor dataset of the local universe with improved photometric calibration. Since July 2022, J-PLUS DR3 is fully available for the collaboration members. It comprises 1642 J-PLUS fields observed in twelve optical bands amounting to 3192 deg². J-PLUS DR3 is based on images collected from November 2015 to February 2022 by the JAST80 telescope and includes ≈ 30 million stars brighter than $r = 21$.

Data from the J-PLUS DR3 and previous data releases can be accessed from its website[3] in various ways, including Virtual Observatory (VO) services such as Table Access Protocol (TAP), e.g. through Astronomical Data Query Language (ADQL, Osuna et al. 2008) queries. However, the limit of rows that the TAP service returns is 1 million. To overcome this limitation and avoid the risk of not accessing all J-PLUS DR3 stars, our ADQL query (see Appendix A) was attached into a Python code to acquire the data in the form of blocks of reference tiles, so that the full query is split into several minor queries, as indicated by the user manual for the CEFCA' s Catalogues Portal[4] (Section 4.4.2). The script applies quality cuts and retrieves useful information such as sky positions, magnitudes, and their respective uncertainties.

In particular, J-PLUS photometry offers various kinds of magnitudes, e.g. measured from aperture or point spread function techniques. The standard 6" aperture limits flux contamination from nearby sources and prevents dominance by background noise, but it is insufficient to capture the total flux of the stars. In this work, following López-Sanjuan et al. (2019), we used the magnitude from a 6" diameter aperture after correction for aperture effects (column *APER_COR_6_0*).

2.1.2 J-PLUS DR3 stars cross matched with other catalogs

In this section, we outline the process of constructing the training sample after gathering data for the J-PLUS stars. All cross-matching steps were performed using TOPCAT[5] (Taylor 2005), with a maximum angular separation of 3" between matched sources.

First, for J-PLUS stars selection, the constraints applied to the DR3 database include the following (see Appendix A).

- Star/galaxy separation:
  `sglc_prob_star` > 0.7 to select objects with more than 70% probability of being a star.
- Magnitude error:
  `array_max_float(ps.mag_err_aper_cor_6_0)` < 0.5 to exclude stars (leniently) with too low S/N in the 6" diameter in any of the twelve filters.
- Adequate exposure:
  `array_min_float(dualobj.norm_wmap_val)` > 0.8 to avoid sky regions not observed at full depth.
- Quality flags:
  `array_max_int(dualobj.mask_flags)` = 0 to select stars without any warning flags, meaning with good photometry.

This query returned 5 114 494 sources, which are used for cross-matching with complementary catalogs, as described below. This resulting list was cross-matched with CatWISE (Marocco et al. 2021) to include *W1* and *W2* magnitudes in the model construction, since infrared photometry information from these bands helps reducing

---

[3] `https://www.j-plus.es/datareleases/data_release_dr3`

[4] `https://archive.cefca.es/doc/manuals/catalogues_portal_users_manual.pdf`, version 1.29 was consulted.

[5] Tool for OPerations on Catalogues And Tables: `https://www.star.bris.ac.uk/~mbt/topcat/`; version 4.9 was used.

Table 1. Summary of steps in the construction of our training sample, with each row consecutively describing the cross matches performed (within a 3 arcsec search radius from the J-PLUS RA and DEC).

Catalog	Sample size	Fraction	Notes
J-PLUS DR3	5 114 494	1.00000	Appendix A
CatWISE	5 027 068	0.98291	For $W1$ and $W2$
Gaia DR3	5 026 927	0.98288	For G, B_P, R_P
Gaia EDR3^(a)	5 004 860	0.97856	For distances^(a)
Gaia AP	5 004 738	0.97854	For extinctions
LAMOST DR8	529 565	0.10354	For spectroscopic targets

Notes. (a) Values from the work of Bailer-Jones et al. (2021).

temperature–extinction degeneracy (Andrae et al. 2023). The provided 5 027 068 stars were then cross-matched with Gaia data, more specifically with i) DR3 main source catalog for the usual Gaia magnitudes G, B_P and R_P (Gaia Collaboration 2022; Babusiaux et al. 2023; Gaia Collaboration et al. 2023), which returned 5 026 927 stars; ii) with EDR3 distances from Bailer-Jones et al. (2021), which returned 5 004 860 stars; and iii) with DR3 astrophysical parameters (Gaia Collaboration 2022; Creevey et al. 2023) for information on extinction corrections, which returned 5 004 738 stars.

For spectroscopic information, LAMOST DR8 was chosen due to its large data volume, which is beneficial for machine learning since larger datasets enable more detailed pattern recognition. During the development of the present study, newer releases became available (DR9 with parameters for approximately 6.92 million AFGK stars, and DR10 with 7.45 million), but we opted to retain the originally assembled sample for consistency and convenience, under the assumption that the increase in sample size or potential refinements in parameter estimates would not substantially impact model performance or the conclusions of this work. In addition, the range of stellar atmospheric parameters given by the LAMOST stellar database aligns well with the target sample of this study. From its database[6], other than identification columns, we selected $T_{\rm eff}$, $\log\,g$, [Fe/H], their corresponding uncertainties ($\sigma_{T_{\rm eff}}, \sigma_{\log\,g}, \sigma_{\rm [Fe/H]}$), and signal-to-noise ratios (S/N) in spectral regions associated with the photometric filters from the *ugriz* system. Keeping only stars with valid uncertainties ($\sigma_{T_{\rm eff}}, \sigma_{\log\,g}$ and $\sigma_{\rm [Fe/H]} > 0$) in the LAMOST DR8 sample resulted in 6 684 026 stars, and cross-matching these with the list obtained from the previous step cross-match step leaves our sample with 529 565 stars, which are further refined as detailed in Section 2.1.3. Table 1 summarizes the selecting steps described above.

For illustration, Figure 1 shows the transmission curves for each filter in the three photometric systems used in this work, i.e. from J-PLUS, CatWISE, and Gaia. Also shown are the theoretical spectral energy distributions (SEDs) for three example stars, highlighting absorption features relevant to stellar astrophysics, including Ca II H and K, Hα and the Ca II infrared triplet. The synthetic data has a wavelength sampling of $\Delta\log(\lambda/\text{Å}) = 8\times10^{-4}$ (R =?) and was taken from Coelho (2014) by using Virtual Observatory Sed Analyzer (VOSA; Bayo, A. et al. 2008). The SEDs (black lines) represent stars with $T_{\rm eff}$ values of 4000 K (dotted), 6000 K (solid) and 7500 K (dashed), with $\log\,g = 4.5$ dex, and solar values for [Fe/H] and [α/Fe].

Table 2. Summary of final training sample with 88 421 objects

Parameter	Mean	SD	Minimum	Maximum
$T_{\rm eff}$ (K)	5857	465	4020	7499
$\log\,g$ (dex)	4.134	0.393	0.665	4.844
[Fe/H] (dex)	-0.216	0.306	-2.350	0.777

2.1.3 Preparing data for training

From the selection presented in Section 2.1.2, we kept stars brighter than 21 mag in any of the 12 J-PLUS filters, with uncertainties smaller than 0.1 mag. The same uncertainty constraint was applied for CatWISE and Gaia photometry. The sample was also further constrained by the individual S/N retrieved from LAMOST (*ugriz*) and CatWISE ($W1$ and $W2$) databases: $\mathrm{SNR}_g^{\rm min} = \mathrm{SNR}_r^{\rm min} > 50$ and $\mathrm{SNR}_{W1}^{\rm min} = \mathrm{SNR}_{W2}^{\rm min} = \mathrm{SNR}_u^{\rm min} = \mathrm{SNR}_i^{\rm min} = \mathrm{SNR}_z^{\rm min} > 10$. These limits discard low-quality data from the sample whilst keeping it large enough for training. While no restriction was made to the values of $\log\,g$ or [Fe/H], the selected $T_{\rm eff}$ range was [4000 K, 7500 K], keeping spectral types F, G, and K, for which uncertainties were limited to less than 100 K in $T_{\rm eff}$ and less than 0.1 dex in both $\log\,g$ and [Fe/H]. Additionally, constraints from Gaia DR3 on the parallax-over-error ratio (ϖ/σ_ϖ) and the re-normalized unit weight error (RUWE) were respectively set to $\varpi/\sigma_\varpi > 10$ and RUWE < 1.4, avoiding likely poor quality astrometric solutions.

For these stars, interstellar extinction corrections for each of the 12 J-PLUS filters were performed using values provided in J-PLUS catalog. For Gaia ($A_{\rm G}, A_{\rm BP}, A_{\rm RP}$; Gaia Collaboration 2022; Creevey et al. 2023) and CatWISE ($A_{\rm W1}, A_{\rm W2}$; Yuan et al. 2013), extinction values were directly retrieved from the literature. All cases assume an extinction factor of $R_{\rm V} = 3.1$. After these cuts, 88 421 stars remained in our list for training. Summarizing statistics of this sample, including mean, standard deviation (SD) and maximum and minimum values, are presented in Table 2, where we can see that $\log\,g$ and [Fe/H] were kept limited to [0.665, 4.844] dex and [-2.350, 0.777] dex, respectively.

The following step was to calculate all 136 possible colors using the 17 magnitudes of all three systems used in this work, i.e. 12 from J-PLUS, three from Gaia, and two from CatWISE. Absolute magnitudes were calculated using photogeometric distances from Bailer-Jones et al. (2021), which are computed from a statistical model that assumes both photometric and direction-dependent priors. Altogether, the possible colors and the absolute magnitudes amount to a total of 153 features.

In this study, the full dataset is randomly partitioned into three subsets according to their role in the fitting process: 70% for training, 20% for validation, and 10% for testing. The training set enables the model to capture underlying patterns within the data; the validation set is specifically employed for hyper-parameter tuning, ensuring better generalization to new data by preventing over-fitting; the test set serves as the final evaluation benchmark, assessing the model's performance on data that is not part of the training or validation sets, providing a reliable internal measure of its present generalization capabilities to previously unseen portions of the dataset (Ivezić et al. 2020). Figure 2 presents the final distributions of the train (dotted lines), validation (dashed lines), and test (solid lines) samples within our full dataset. The mean and standard deviation values for the test sample are indicated in the legend and closely match those of the entire dataset, as shown in Table 2.

---

[6] `https://www.lamost.org/dr8/`

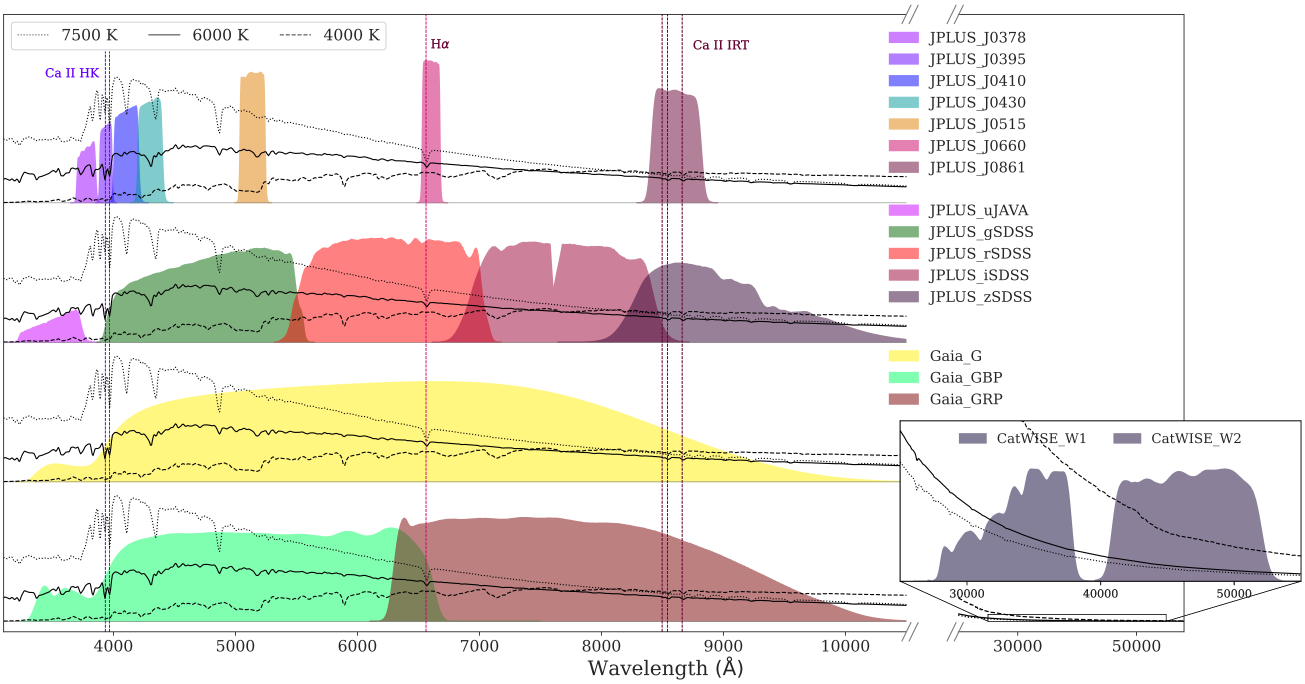

Figure 1. SEDs (black lines) based on synthetic spectra from Coelho (2014), shown for stars with $T_{\rm eff}$ = 7500 K (dashed), $T_{\rm eff}$ = 6000 K (solid), and $T_{\rm eff}$ = 4000 K (dotted; visible only in the inset), all with log g = 4.5 dex, [Fe/H] = 0 dex, and [α/Fe] = 0 dex. For visual clarity, the fluxes of the $T_{\rm eff}$ = 6000 K and $T_{\rm eff}$ = 4000 K SEDs were multiplied by arbitrary factors of 1.6 and 9, respectively. These SEDs are overplotted with the filled transmission curves of the filter sets from the three surveys used in this work: J-PLUS DR3, Gaia DR3, and CatWISE (shown in the inset). Colors and corresponding temperatures are indicated in the legend. Colored vertical dashed lines mark prominent absorption lines of Ca II and Hα.

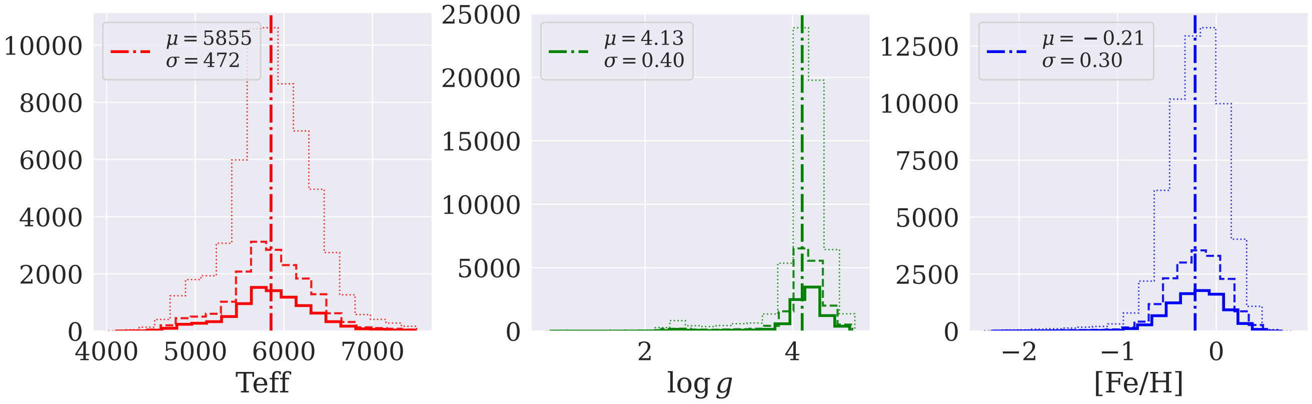

Figure 2. Distributions of the train (dotted lines), validation (dashed lines), and test (solid lines) sets within our full training sample. The mean and standard deviation of the test set are indicated by the legend and are as similar as expected to those of train and validation sets, even though the latter are not explicitly shown.

2.2 Open clusters sample

Our open cluster sample is based on the work of Hunt & Reffert (2023, hereafter HR23), who conducted the largest search for open clusters to date by employing the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm to the Gaia DR3 database. HR23 recovered 7167 clusters, 2387 of which were potential new discoveries. However, the most reliable sample recommended by the authors includes 4105 clusters, 739 of which are new. Additionally, their catalog provides valuable information such as membership probability, cluster age, extinction, and distance.

To identify which clusters fell within the J-PLUS DR3 footprint, we used a sample of J-PLUS stars selected through a more relaxed query compared to that presented in Section 2.1.2 and shown in Appendix A[7], aiming to retain the largest possible number of sources

---

[7] `sglc_prob_star > 0.5`, with no quality cuts imposed (i.e., no cuts on `mask_flags` or `norm_wmap_val`)

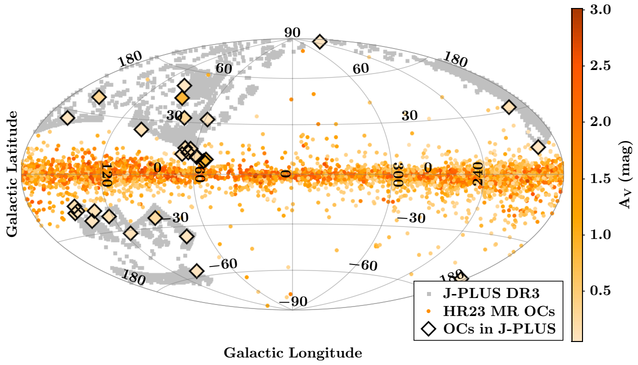

Figure 3. Sky Aitoff projection for the J-PLUS DR3 footprint (gray regions), the most reliable sample of OCs from HR23 (extinction-colored points) and the ones found in the J-PLUS DR3 footprint (extinction-colored diamonds with black outlines).

from J-PLUS. In practice, the somewhat more restricted selection described in Section 2.1.2 constitutes a subsample of the broader one used here, which resulted in 8 633 441 stars. These were cross-matched with the 529 004 probable members (with membership probability $P \geq 0.5$, also to favor a higher number of matches) from the most reliable cluster sample of HR23. This procedure returned 238 stars associated with 29 clusters, shown in Figure 3 as diamonds with black outlines plotted over the J-PLUS DR3 footprint, represented by gray regions. All markers are color-coded according to differential extinction as reported in HR23.

During the development of this work, Hunt & Reffert (2024) updated their previous OC census from HR23, reclassifying some of the systems selected here as moving groups. We chose to proceed with the original design of our work based on HR23, with these objects serving as a representative astrophysical application of the method, since our primary focus is on presenting and discussing the developed tool. In this context, we note that references to OCs throughout the text may include systems that were later reclassified as moving groups.

3 CONSTRUCTION OF MODELS

In this section, we provide details on the construction of the gradient boosting models used to estimate our stellar parameters, which are presented later. Gradient boosting is a powerful machine learning ensemble technique that combines the predictive strength of multiple weak learners, typically decision trees, to create a robust and accurate model (Friedman 2001).

In this work, we have chosen to use the machine learning Python framework Light Gradient-Boosting Machine (LightGBM[8]; Ke et al. 2017), which operates similarly to the widely used algorithm *XGBoost* (Chen & Guestrin 2016). Both algorithms are highly regarded for their performance, but LightGBM can be particularly advantageous in scenarios involving large datasets, high-dimensional data, and requirements for speed or lower memory consumption, as discussed in Ke et al. (2017).

3.1 Search for best features and parameters

Hyper-parameter tuning is a key step in any machine-learning pipeline, involving the adjustment of model settings and the selection of input features. These tasks are often performed separately, which can lead to sub-optimal performance and longer processing times.

In this work, our choice to address this challenge is SHAP-HYPETUNE[9], a Python package designed for simultaneous hyper-parameters tuning and features selection. This tool combines both tasks, optimizing the ideal number of features and parameter configurations concurrently, ensuring heightened efficiency for gradient-boosting models. We used the SHAP-HYPETUNE package to perform Bayesian optimization, which employs a probabilistic model to intelligently guide the search for optimal hyper-parameters based on past evaluations.

For feature selection, SHAP-HYPETUNE implements a recursive feature elimination (RFE) approach, which iteratively removes features that contribute less to the model. The feature importances are calculated using the Python software SHAP[10], which yields the so-called SHapley Additive exPlanation (SHAP) values (Lundberg & Lee 2017).

SHAP values are an explainability technique that breaks down a model's output into contributions from each feature, using Shapley values from (Shapley 1951). By evaluating local and global impacts on predictions, SHAP values explain feature importances while preserving the original units of the input features. Mathematically, SHAP values are computed using Shapley values from cooperative game theory. The SHAP value for each feature represents its average contribution to the prediction by considering all possible ways in which the feature can combine with other features. In other words, this method helps clarify how features influence specific predictions and their overall effect on the model' s performance. Moreover, it has recently started being used in astronomical contexts as machine learning techniques have become more accessible (e.g. He et al. 2022; Robinson et al. 2024; Chierichini et al. 2024).

Having obtained the optimal hyper-parameters and features from SHAP-HYPETUNE, we further performed another RFE step, but this time with cross-validation (RFECV) using the RFECV function from the Python library SCIKIT-LEARN[11] (Pedregosa et al. 2011). RFECV executes feature selection through an iterative 5-fold cross-validation process, which enhances model robustness compared to standard RFE, since cross-validation splits the sample into 5 groups, interchangeably using 4 groups for training and 1 group for validation in each iteration. This allows RFECV to evaluate feature importance across multiple subsets of the data, resulting in a more reliable estimation of feature significance and ensuring that the selected features better generalize to unseen data.

The results of this step for $T_{\rm eff}$, log g, [Fe/H], are shown in Figure 4, where the y-axis is the mean absolute error (MAE) and the number of features can be seen in the x-axis. The number of points in each panel (82, 83, 100 features for $T_{\rm eff}$, log g, and [Fe/H], respectively) are the results from the RFE approach previously implemented by SHAP-HYPETUNE before our RFECV step. One can easily notice that, although the optimal number of features varies for each stellar parameter as indicated (black dashed lines mark 62, 65, 76 features and indicate MAE values of 43 K, 0.056 dex and 0.062 dex for $T_{\rm eff}$, log g, and [Fe/H], respectively), in all cases a visual plateau

---

[8] `https://lightgbm.readthedocs.io/`

[9] `https://github.com/cerlymarco/shap-hypetune`

[10] `https://shap.readthedocs.io/en/latest/`

[11] `https://scikit-learn.org/stable/index.html`

is reached above around 10 features (dotted, vertical, gray lines). The horizontal black dashed line is plotted to mark the best mean RFECV MAE value in each panel, but visually demonstrates this behavior.

3.2 Training and test sets

In essence, the initial 153 features underwent two selection steps: first, SHAP-HYPETUNE used RFE to find the best features simultaneously with best hyper-parameters; then, we additionally executed RFECV seeking to improve reliability when applying the models to other datasets. The results for the RFECV process are summarized in the top panel of Figure 5, where relative (i.e. divided by their sum) feature importances from SHAP values are shown on the vertical axis and the feature positions on the horizontal axis. The first 15 most important features are indicated on each respective panel, with the first 10 highlighted in the color of the respective data points (red for $T_{\rm eff}$, green for log g, and blue for [Fe/H]).

Importantly, we present some standards regarding the names of the features shown in Figure 5: J-PLUS filters are the ones with J as first letter followed by four characters representing either the central wavelength of filter in nanometers (for the narrow bands—e.g. J0430) or an indication of which broad band (uJAVA or *griz*SDSS); CatWISE magnitudes are W1mproPM and W2mproPM, while Gaia magnitudes are BPmag, RPmag and Gmag; the majority of them are simple magnitudes subtractions; all magnitudes exhibit the subscript `_0` to indicate extinction correction; those with the prefix `M_` represent absolute magnitudes, which were calculated using photogeometric distances, *rpgeo* (indicated in the feature name), from Bailer-Jones et al. (2021, using Gaia EDR3, also indicated in the feature name).

More importantly, the top panel of Figure 5 shows the information on individual importances (translucent curves and crosses) and how their relative values accumulate (solid). The percentages represent the models' variability, i.e. to which extent the variations in the target predictions ($T_{\rm eff}$, log g, and [Fe/H]) can be assigned to variations in the features (colors and magnitudes). Hence, for instance, the top five most important features account for approximately 61.70%, 66.32%, and 55.64% of the variance in the predicted $T_{\rm eff}$, log g, and [Fe/H], respectively. These values increase to 75.34%, 76.43%, and 69.75% when considering the top ten features.

The bottom panel of Figure 5 shows importance values for the top ten most important features (same as exhibited in colors on the top panel) for each parameter's model. The leading features, particularly the top three, exhibit physically meaningful quantities, lending confidence to the model's internal consistency. For $T_{\rm eff}$, the dominance of the $J0430 - W1$ and $J0430 - J0861$ colors, for instance, reflects the strong sensitivity of these filters to the shape of the stellar continuum for the spectral types of interest to this work, which is largely driven by temperature. In the case of log g, the two most relevant features are the absolute magnitudes M_{W1} and M_{W2}, which are expected to correlate with surface gravity due to their dependence on stellar luminosity and radius. The third feature, $J0430 - J0515$, involves the intermediate-band filter $J0515$ that partially covers the Mg triplet region, which has some surface gravity sensitivity. For [Fe/H], the dominance of features such as $uJAVA - J0430$ and $J0378 - J0430$ highlights the role of the blue part of the spectrum, where many metal absorption lines are concentrated. Additionally, colors combining $J0660$ (centered on Hα) and $J0861$ (covering the Ca II triplet) further support the model's ability to capture metallicity effects, given the known sensitivity of these features to metal content.

It is important to mention that in this work we choose to proceed with all available filters instead of selecting only a subset. If we naively chose to use, for instance, only the top three most important features we found for each parameter, then only 8 magnitudes (uJAVA, J0378, J0430, J0515, J0660, J0861, W1, and W2) would be necessary for estimating all three parameters we seek in this work, instead of all 17. However, this approach comes at the cost of constructing a model with much poorer performance, as one could anticipate from the top panel of Figure 5: barely half of the variations in the metallicity predictions would be accounted for the variations in the input photometry. Fitting the models using the top five or top ten most important features improves the performance in [Fe/H] in terms of cumulative importances (respectively 55.64% and 69.75%). Nevertheless, at that point, the number of needed filters increases to 12. This mainly undermines the purpose of relying on fewer features, since the only absent ones are the Sloan filters *griz* and the Gaia filter *RP*, and we can expect that most stars should have available photometry for those. Hence, we consider it unjustified to exclude any of the available magnitudes in this context, even at the cost of having fewer stars available.

The results for the final LightGBM models applied to the test set (8842 stars, 10% of our training sample) can be seen in Figure 6. The horizontal axis shows the original values from LAMOST, while the vertical axis shows residuals (this work minus LAMOST), with MAE values exhibited and mean and standard deviation values indicated. Marginal histograms represent the respective distributions in each axis. At first, outliers are easily spotted and moderate trends can be identified, especially for log $g < 3$ dex and for $\rm [Fe/H] < -1$ dex. On the other hand, MAE values (42 K, 0.055 dex, and 0.061 dex for $T_{\rm eff}$, log g, and [Fe/H], respectively) indicate a good predictive power, and relatively little scatter can be seen outside of those just mentioned regions. Moreover, $T_{\rm eff}$ predictions seem to be better constrained than the other two parameters.

To further evaluate the impact of reducing the feature set, Figure 7 illustrates how models with different numbers of selected features perform in terms of MAE values. For each parameter, the MAE values for a reduced set of five features, for instance, were: $T_{\rm eff} = 55$ K (compared to 43 K in the full model), $\log g = 0.061$ dex (compared to 0.055 dex), and $\rm [Fe/H] = 0.075$ dex (compared to 0.061 dex). These correspond to percentage increases in error of 28%, 11%, and 23% for $T_{\rm eff}$, $\log g$, and [Fe/H], respectively. However, as highlighted in Figure 5, the analysis of relative importances indicates that less than two-thirds of the variance in the target predictions can be explained by the variations in the selected features, with the best-case scenario observed for $\log g$ at 66%. This implies that the reduced models' MAE values become increasingly less reliable in capturing parameter variations as the feature set is narrowed. Furthermore, implementing this reduced feature set still requires 12 of the 17 magnitudes from the full set, with the five omitted magnitudes being broad-band filters that are unlikely to limit the sample size significantly. This discussion, together with Figure 7, underscores the benefit in retaining the full feature set to ensure reliable predictions across all parameters.

4 SCIENCE CASE: PARAMETERS FOR INDIVIDUAL MEMBERS OF CLUSTERS

Here we present in Section 4.1 the results of applying the models discussed in the last section. Then we explored directly the LAMOST DR8 catalog for comparisons with spectroscopic results (Section 4.2). For this, we retrieved from LAMOST only those stars with reference parameters within the training range for each parameter, namely [4000, 7500] K for $T_{\rm eff}$, [0, 5] dex for log g and [−2.5, 1.0] dex for [Fe/H].

Since tree-based models, such as from random forests and gradi-

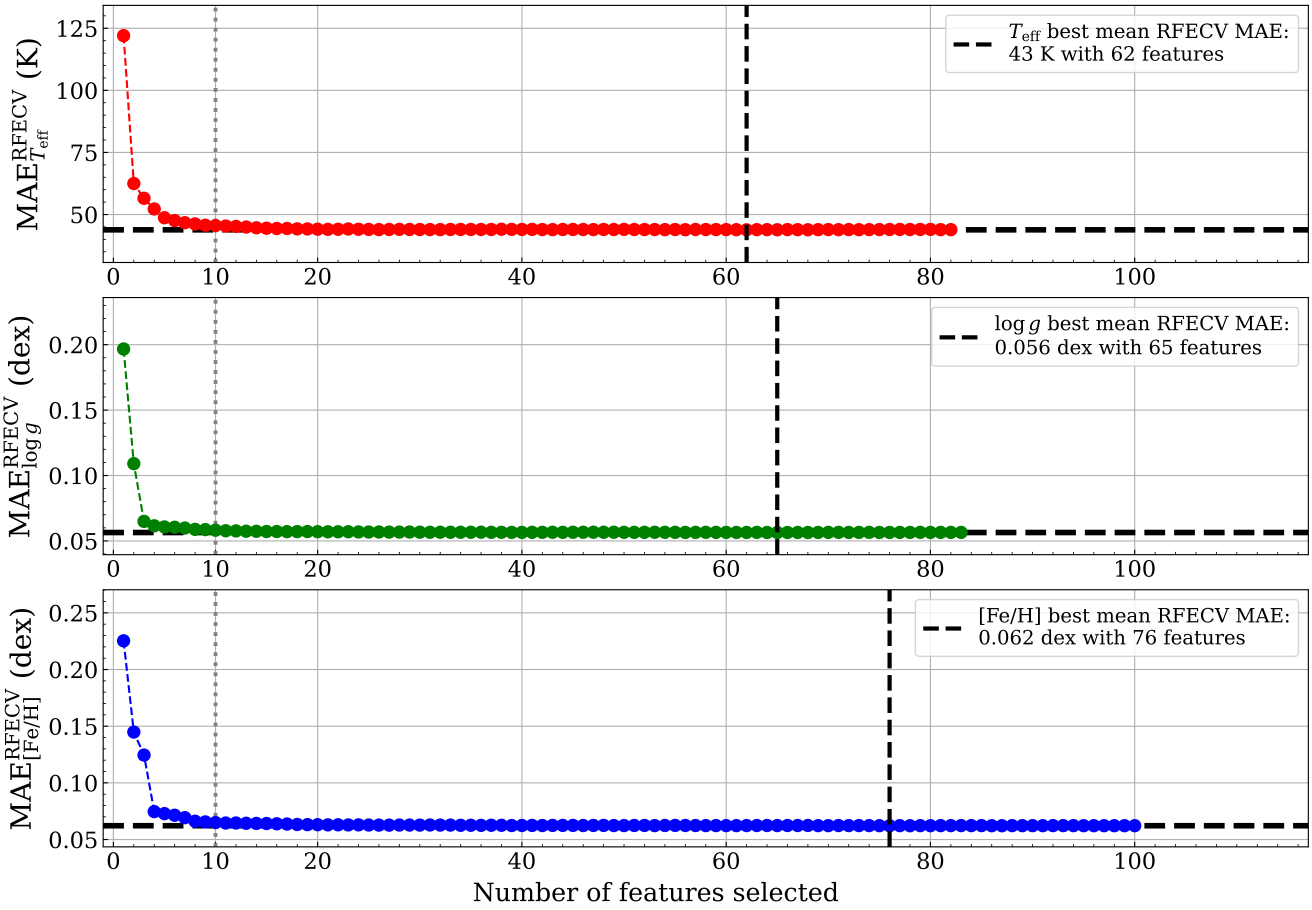

Figure 4. Mean RFECV MAE values as a function of the selected number of features. The black dashed lines in each panel mark the best mean RFECV MAE and the associated number of features.

ent boosting techniques, inherently preserve the limits of the training data and do not extrapolate beyond its range, it was not necessary to impose these selections for values in this work. For inputs falling outside the training domain, these models generally revert to predictions biased toward the mean response of the training data.

4.1 Application of our models

The first step is fetching, for members of our selected OCs, the photometry required by the models (see in Section 2.1.2). The list of 238 OCs members introduced in Section 2.2 was cross-matched with CatWISE and Gaia data, taking care of cross matching same stars using their Gaia DR3 source id, which resulted in 191 stars. Afterwards, we eliminate those with obvious poor photometry (e.g. magnitudes and/or uncertainties set as 99), keeping only those with magnitude uncertainties smaller than 0.5 mag. This process resulted in 157 stars distributed in 23 systems; six of those have at least 10 members and only two (HSC 749 and NGC 1039) have at least 20 members (31 and 33, respectively). Additionally, 22 stars in our sample lacked Gaia extinction information. For these cases, we adopted the median extinction values in the G, BP, and RP bands, along with their corresponding uncertainties. In light of our limited sample size, we chose to retain these stars in the analysis, especially since the extinction values involved are modest—below 0.3 mag for all bands (specifically, A_G = 0.229 mag, A_{BP} = 0.298 mag, and A_{RP} = 0.176 mag at most, computed from the remaining 157 stars selected in the previous step)—and thus are not expected to significantly impact the results.

Next, we perform a direct resampling procedure to estimate uncertainties for the parameters. In this method, we generate multiple artificial datasets by adding random noise to the original photometric data, reflecting the observational errors. More specifically, each star goes through a sampling from a normal distribution where the mean is the observed quantity (magnitude or distance) and the dispersion is the measurement error. For each resampled dataset, we reapply our model to predict the atmospheric parameters. This process is repeated 10 000 times, resulting in 10 000 predicted values for each parameter of each star. From the distribution of these predicted values, we derive the mean and standard deviation, which are adopted as the final values in this work, while the median values and specific percentiles (16th and 84th) are also provided. This resampling approach allows us to capture the variability inherent in the observed measurements and the model, thus offering a comprehensive and robust quantification of the parameter uncertainties. Figure 8 illustrates the distributions of the final adopted parameters (top) and their respective errors (bottom). The average values of the errors, marked in dashed black lines, are 57 K for $T_{\rm eff}$, 0.041 dex for $\log g$, and 0.158 dex for [Fe/H].

In particular, the peak in $T_{\rm eff}$ $\sim$ 4500 K indicates that several

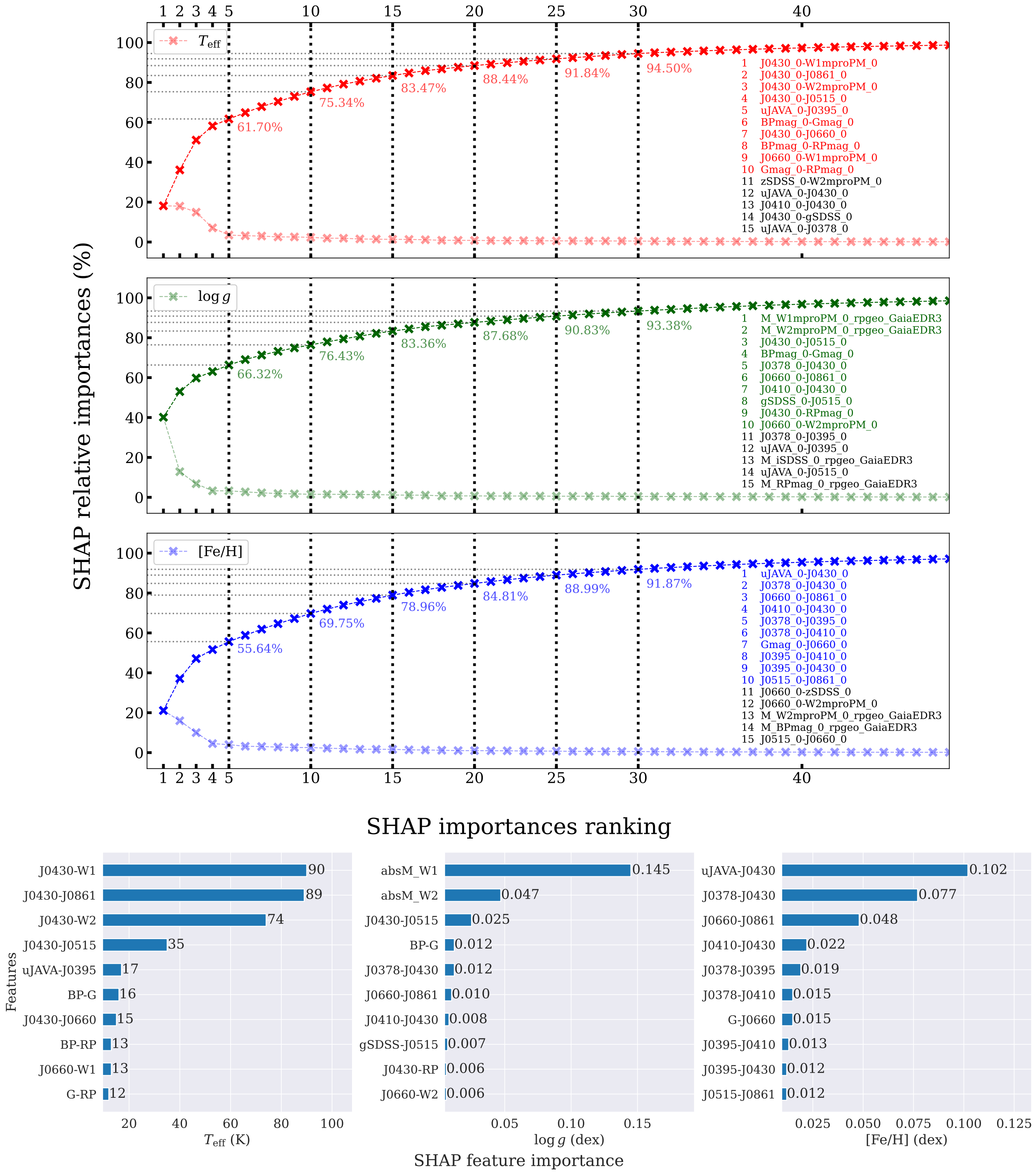

Figure 5. Top: relative SHAP importance as a function of feature ranking. Translucent crosses represent the ratio of each individual importance over the sum of all features' importances, while opaque crosses represent their cumulative sum, periodically indicated with dotted black lines for their values (x-axis) and their respective number of features (y-axis). The first 15 most important features are also shown, with the first 10 highlighted in the color of the respective data points (red for $T_{\rm eff}$, green for log g, and blue for [Fe/H]). Bottom: bar plot of the top 10 most important features (same as those shown in color in the top panel), with their corresponding importance values labeled.

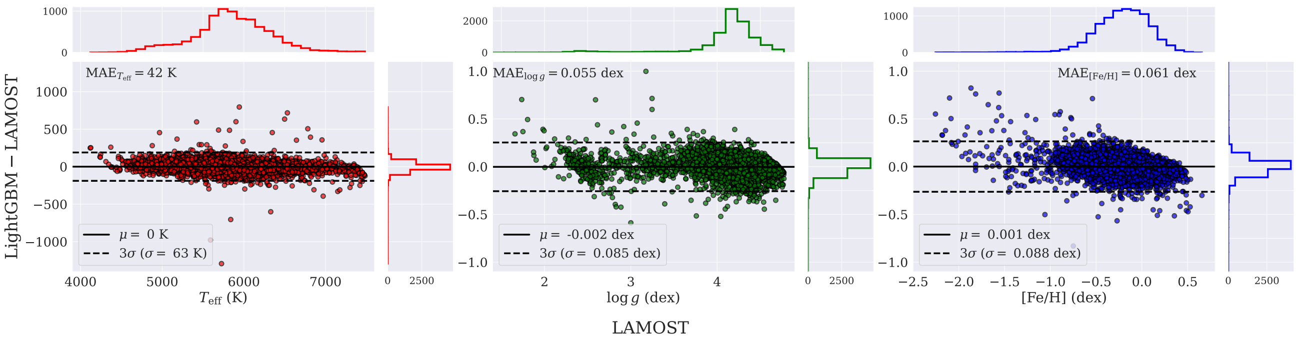

Figure 6. Residual between predicted LightGBM parameters and original LAMOST values in the test set. MAE are also exhibited along with mean (solid black lines) and three standard deviations (dashed black lines) values. Marginal histograms show the distribution of values along each axis.

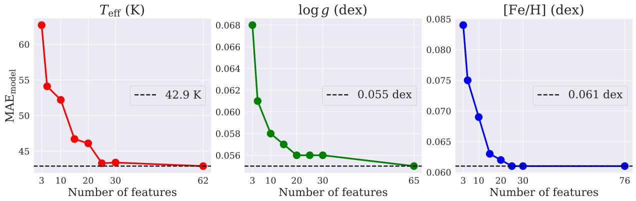

Figure 7. MAE values as a function of the number of features used in model fitting. The black dashed lines in each panel mark the best MAE values for each parameter. Despite similarities, note that Figure 4 differs from this one in the sense that the former depicts the RFECV step, from which we can identify the most important features, while the latter uses precisely these now uncovered features for fitting the final models.

stars that are likely cooler than their predicted T_eff. However, temperatures as low as $T_\mathrm{eff} < 4000$ K are not covered in the training sample (see Fig 2). When the model struggles to map the correlations between features and targets—due to high scatter, nonlinear relationships, or insufficient resolution in feature space—it minimizes error by regressing toward the mean of the training distribution ($T_\mathrm{eff} = 5857$ K), bringing these predictions away from the minimum value ($T_\mathrm{eff} = 4020$ K). Since our T_eff model cannot reach $T_\mathrm{eff} < 4000$ K, then we only have a probable upper limit for the temperatures of these stars.

We further filtered out stars with distances significantly deviating from the median value of their respective associated clusters as compared to individual estimates from Bailer-Jones et al. (2021). Specifically, for each cluster in our list, outliers were removed by comparing each star's distance to the cluster's median distance, which was derived from those individual estimates. This step was performed using the sigma clipping process provided by the SIGMA_CLIP function from the ASTROPY[12] Python library, with the median distance as the central value and a threshold of 3σ, with σ representing median absolute deviations[13] (MAD) from the median. This selection process reduced our sample from 157 to 154 stars, while preserving distributions similar to those shown in Figure 8. The three stars excluded belong to the associations Alessi 62, HSC 759, and Theia 1188, so the total number of 23 systems remains unchanged. Moreover, the median distances of the clusters, computed from individual stars using the values from Bailer-Jones et al. (2021), are generally consistent with those reported by HR23; differences exceed 5% only for cluster HSC 695 (see Appendix B).

4.2 Comparisons with LAMOST

For comparisons with LAMOST, the cross-match was performed using the Gaia DR3 source ID, resulting in 30 stars from the initial 154. Of those, 25 fall within the comparison limits, which we recall to be [4000, 7500] K for T_eff, [0, 5] dex for log g and [−2.5, 1.0] dex for [Fe/H]. Figure 9 shows the distribution of residuals for the predicted stellar parameters, where Δ represents this work (TW) minus LAMOST. The Gaussian curves depict distributions generally centered around zero, with mean residuals and standard deviations ($\mu \pm \sigma$) of -30 ± 130 K for T_eff, -0.02 ± 0.13 dex for log g, and -0.13 ± 0.20 dex for [Fe/H]. The presence of outliers in the histogram tails suggests that larger deviations from the reference values occasionally occur. Additionally, the mean metallicity residual indicates a slight, systematic underestimation in our predictions. Furthermore, given our limited sample size, this analysis must be interpreted with caution.

Figure 10 compares the metallicities predicted by our model with those from LAMOST. The left panel color-codes data points by T_eff from this work, $T_\mathrm{eff}^\mathrm{TW}$, while the right panel uses ΔT_eff. The solid black lines in both panels represents the 1:1 relation, and the dashed gray lines denote one standard deviation of the residuals in [Fe/H], Δ[Fe/H] $= -0.20$ dex. Large error bars indicate significant uncertainties in some cases, and a few outliers deviate substantially from the 1:1 relation. Notably, [Fe/H] of metal-rich stars in LAMOST ($\mathrm{[Fe/H]}_\mathrm{LAMOST} > 0$) are all underestimated, while metal-poor ones appear better distributed around the 1:1 line. In particular, the stars with underestimated metallicities (below the 1:1 line) span nearly the entire temperature range, and the color scale in the right panel indicates that these stars also tend to have slightly underestimated T_eff (yellow-toned points).

This behavior suggests some degeneracy between [Fe/H] and T_eff, so we plotted [Fe/H] as a function of T_eff for various datasets in Figure 11. The gray density distribution and contour curves represent the full LAMOST dataset, while black dots indicate the 154 OC members in this study. The data points accumulated at $T_\mathrm{eff} \gtrapprox 4500$ K reflects the peak in the histogram of Figure 8 (see Section 4.1). The colored data points correspond to cross-matched stars between the two datasets, where cyan diamonds denote values from LAMOST

---

[12] `https://docs.astropy.org/en/stable/index.html`
[13] `https://docs.astropy.org/en/stable/api/astropy.stats.mad_std.html`

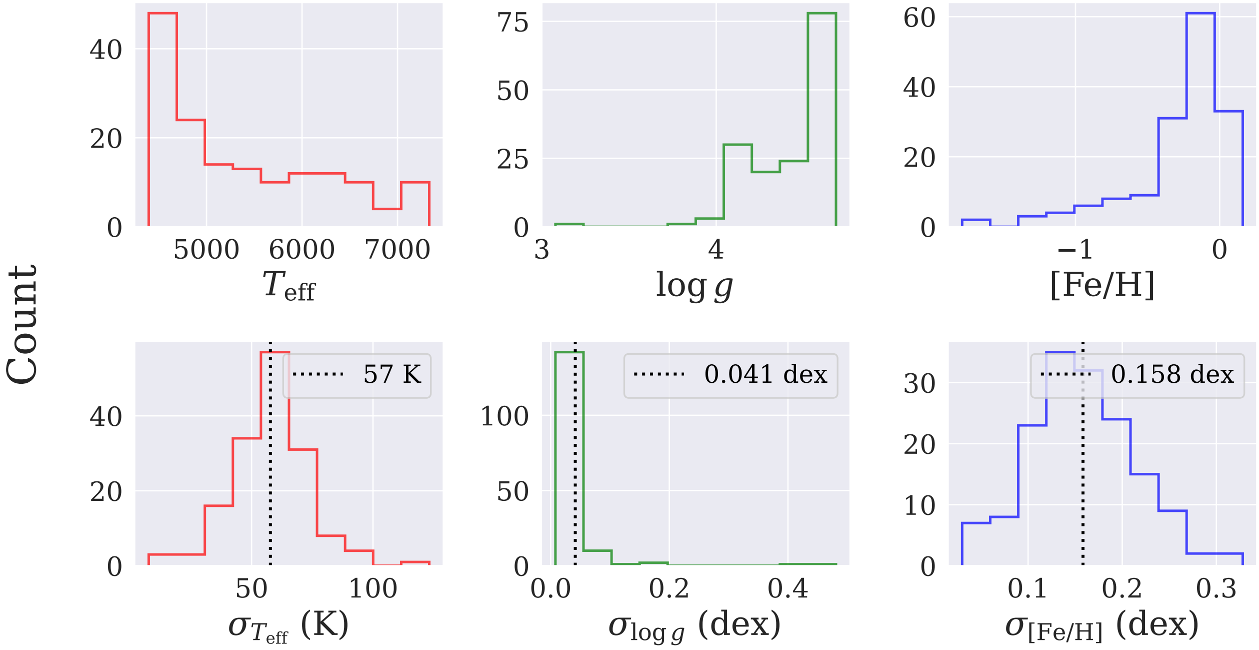

Figure 8. Distributions of the adopted values for 157 cluster members. Parameters (top) and respective errors (bottom) were both obtained from our resampling process (discussed in the text) as the derived mean and standard deviation, respectively. In the bottom panels, dashed black lines indicate the average errors for each distribution.

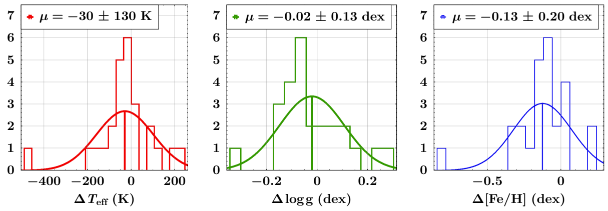

Figure 9. Distribution of residuals for $T_{\rm eff}$, log g, and [Fe/H] of clusters members cross matched with LAMOST. Each panel shows the histogram (in counts) of differences between our model predictions and reference LAMOST values, with a Gaussian fit overlaid. The mean and standard deviation ($\mu \pm \sigma$) of each distribution are indicated.

and magenta circles represent the predicted values obtained in this work. Each cyan–magenta pair is connected by a thin black line to illustrate the parameter differences for the same object. The overlaid straight lines are linear fits to the respective datasets, following the same color coding.

LAMOST metallicities (gray points) exhibit no significant trend with $T_{\rm eff}$ (dashed gray line), while our OC members (black points) show a mild trend of decreasing [Fe/H] as $T_{\rm eff}$ increases (full black line). This behavior is more pronounced for stars in the cross match with LAMOST (full magenta line), further suggesting that hotter stars may be systematically underestimated in [Fe/H]. However, since underestimations occur in both cooler and hotter stars (see Figure 10; another way to visualize this is by checking that most magenta points are placed below cyan ones, in both cooler and hotter regimes), it is unlikely that temperature alone drives this effect. Instead, it may influence metallicity predictions indirectly through correlations with photometric features.

To further investigate this, Figure 12 displays [Fe/H] residuals, color-coded by $T_{\rm eff}$, as a function of the most important photometric feature for [Fe/H]–(uJAVA − J0430)$_0$ in the left panel—and for $T_{\rm eff}$-–(J0430 − W1)$_0$ in the right panel. Two distinct regions emerge in the left panel: data points appear evenly distributed around zero for (uJAVA − J0430)$_0$ > 1.22 mag, while a clear depression is seen at smaller values. In addition to highlighting a trend of underestimated metallicities in the latter regime, the color bar shows that it spans almost the entire temperature range ($T_{\rm eff} \gtrsim 5000$ K), whereas the former contains only cooler stars ($4000 < T_{\rm eff}/{\rm K} < 5000$). Similarly, the same stars gathered close to Δ[Fe/H] = 0 in the left panel remain well-behaved in the right one for (J0430 − W1)$_0$ > 2.92 mag, while those with underestimated metallicities are again grouped, in this case for (J0430 − W1)$_0$ < 2.92 mag. Although the separation may not be as visually striking, the regime change is still evident in this space[14]. Altogether, although based on few data points, these plots suggest that metallicity is reliably estimated only for the cooler stars.

The relation between $T_{\rm eff}$ and the most important feature of [Fe/H] is explicitly shown in Figure 13, where (uJAVA − J0430)$_0$ and $T_{\rm eff}$ are plotted for our training sample, with [Fe/H] shown in the color bar. Dashed, black lines mark the color range of the cross matched sample, and the solid black lines indicate the approximate temperature (vertical, $T_{\rm eff}$ = 5000 K) and color (horizontal, (uJAVA − J0430)$_0$ = 1.2 mag) where the regime change occurs in Figure 12. In this space,

---

[14] We highlight that the values 1.22 and 2.92 were chosen conveniently, purely to aid visualization of the distinction.

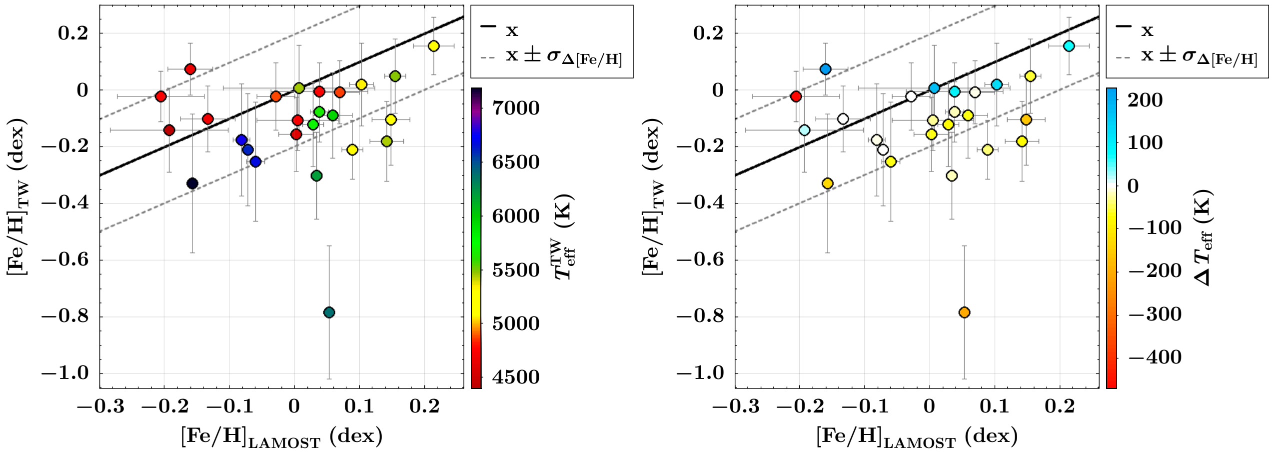

Figure 10. Comparison between metallicities predicted by our model and those from LAMOST. The left panel is color-coded by $T_{\mathrm{eff}}^{\mathrm{TW}}$, while the right panel uses ΔT_{eff}. The solid black line represents the 1:1 relation, and the dashed gray lines indicate one standard deviation of the residuals in [Fe/H].

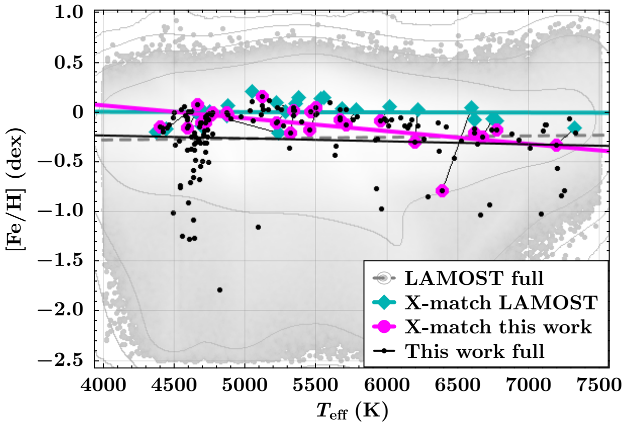

Figure 11. [Fe/H] values as a function of T_{eff} for different datasets. Gray contours show the full LAMOST dataset, while black dots represent OC members. Cyan diamonds and magenta circles denote cross-matched stars, linked to indicate the same object. Linear fits follow the same color coding.

the coolest stars ($4000 < T_{\mathrm{eff}}/\mathrm{K} < 5000$) follow a spread but well-defined relation, which begins to break down beyond 5000 K, when $(\mathrm{uJAVA} - \mathrm{J0430})_0$ falls within the range $0.8 < ((\mathrm{uJAVA} - \mathrm{J0430})_0)/\mathrm{mag} < 1.2$. While 88% of the sample (77 857/88 421) lies within this regime, these stars span a wide temperature range ($5000 < T_{\mathrm{eff}}/\mathrm{K} < 7500$), practically saturating this region. Thus, within this $(\mathrm{uJAVA} - \mathrm{J0430})_0$ range, stars can have virtually any temperature (other than $T_{\mathrm{eff}} < 5000$ K). Moreover, the color layers in this figure also attest the relation between $(\mathrm{uJAVA} - \mathrm{J0430})_0$ and [Fe/H].

Again, the model struggles to map the correlations between features and targets in this case, hence minimizing error by regressing toward the central tendency (mean value) of the training distribution. Since the training sample has a mean metallicity of $\mu_{[\mathrm{Fe/H}]}^{\mathrm{training}} = -0.22$ dex (in Figure 2, we show $\mu_{[\mathrm{Fe/H}]}^{\mathrm{test}} = -0.21$ dex for the test, blind set), this results in a systematic underestimation of metallicity in this temperature range.

We recall that the models used to predict parameters in this work are independent of each other. The predicted values do not directly influence one another, meaning the [Fe/H] model does not have direct access to T_{eff}. When referring to the model's ability to distinguish temperatures, we mean through photometric relations.

To further confirm that metallicity predictions are influenced by temperature-sensitive features, we examine the most important ones for both parameters. This is illustrated in Figure 14, where solid black lines mark the regime change values: $(\mathrm{uJAVA} - \mathrm{J0430})_0 = 1.2$ mag (horizontal) and $(\mathrm{J0430} - \mathrm{W1})_0 = 2.9$ mag (vertical). Dashed black lines indicate the respective ranges included in the cross-matched sample. The correlation between $(\mathrm{uJAVA} - \mathrm{J0430})_0$, the most important feature for [Fe/H], and $(\mathrm{J0430} - \mathrm{W1})_0$, the most important feature for T_{eff}, is expected and could already be inferred from Figure 13, since T_{eff} itself is well constrained.

It is also important to sort out why [Fe/H] is more strongly influenced by the behaviors discussed here, while T_{eff} is not. The observed degeneracy between T_{eff} and [Fe/H] appears to stem from a non-linear interaction between their most informative features. This degeneracy impacts [Fe/H] predictions more severely, since T_{eff} is primarily constrained by the overall continuum shape, which is broadly and smoothly reflected in photometric colors, making T_{eff} easier to recover. In contrast, [Fe/H] depends on narrower metallic lines and localized spectral depressions that are less effectively traced by photometry—at least in the form our features are built, as basic magnitude subtractions. Although a mild correlation exists between T_{eff} and [Fe/H], the most important T_{eff} feature is not highly ranked in the [Fe/H] model, likely because its inclusion would degrade performance, as it does not provide sufficient constraint on [Fe/H].

In fact, considering the most important feature for each of T_{eff} and [Fe/H], we find that reliable predictions are confined to specific color ranges, namely $1.2 < (\mathrm{uJAVA} - \mathrm{J0430})_0/\mathrm{mag} < 1.7$ and $2.9 < (\mathrm{J0430} - \mathrm{W1})_0/\mathrm{mag} < 4.4$. Although this restriction may raise concerns—since the most informative feature becomes less predictive outside these regimes—, it does not invalidate the model's overall performance. The main caveat we identify is a systematic effect known as attenuation bias (Ting 2024), which arises from limited resolution in feature space. In regimes where the most informative feature loses sensitivity (in our case for stars hotter than 5000 K) this leads the model to underestimate metallicities by pulling predictions toward the central tendency (mean value) of the training set. In contrast, for stars with $4000 < T_{\mathrm{eff}}/\mathrm{K} < 5000$, the stronger correlation between $(\mathrm{uJAVA} - \mathrm{J0430})_0$ and metallicity allows the model

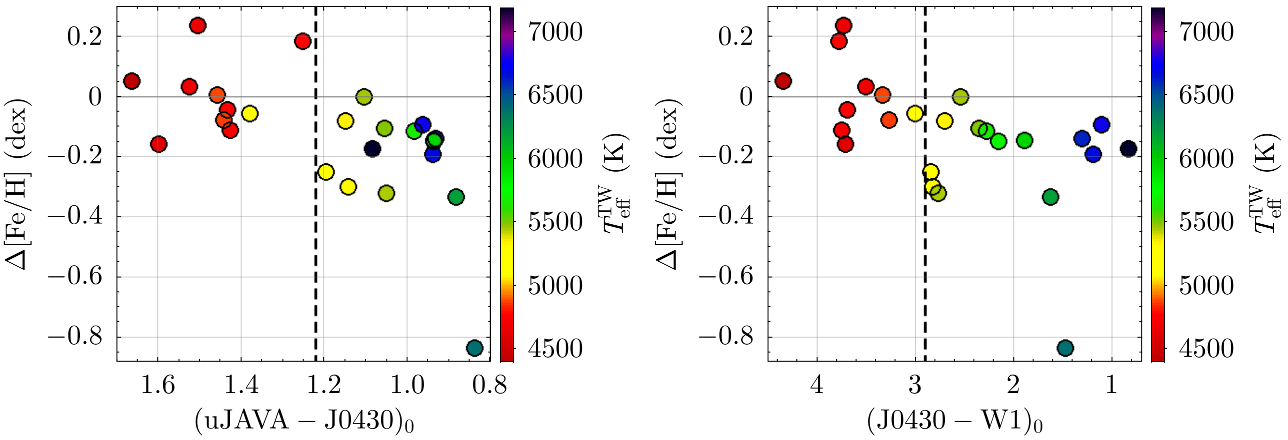

Figure 12. Residuals in [Fe/H] as a function of the most important feature for [Fe/H] (left) and $T_{\rm eff}$ (right), both color-coded by $T_{\rm eff}$. Solid black lines at (uJAVA − J0430)$_0$ = 1.22 mag and (J0430 − W1)$_0$ = 2.92 mag mark the division between distinct Δ[Fe/H] regimes.

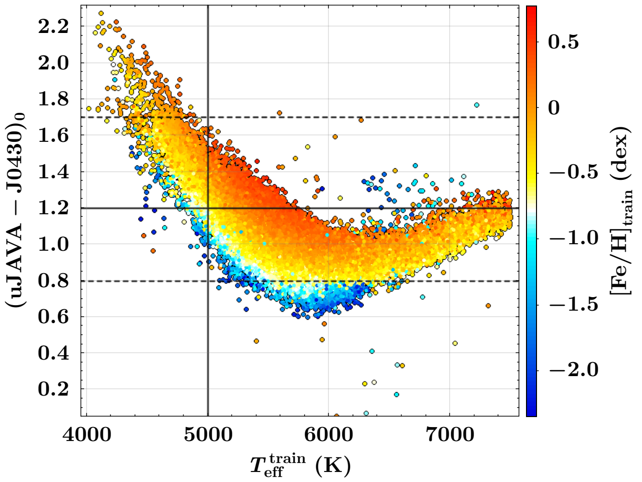

Figure 13. (uJAVA − J0430)$_0$ as a function of LAMOST $T_{\rm eff}$ and color coded by LAMOST [Fe/H] for our full training sample. The solid, black lines mark $T_{\rm eff}$ = 5000 K (vertical) and (uJAVA − J0430)$_0$ = 1.2 mag. The dashed, horizontal, black lines mark the limits of the (uJAVA − J0430)$_0$ range for the 25 LAMOST matched stars, namely 0.8 < (uJAVA − J0430)$_0$/mag < 1.7.

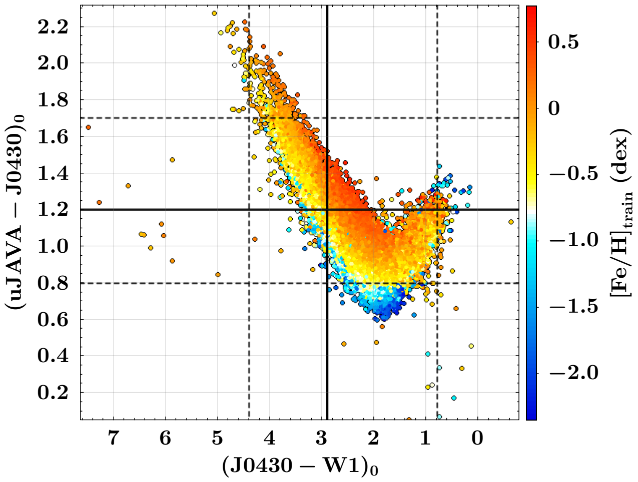

Figure 14. (uJAVA − J0430)$_0$ as a function of (J0430 − W1)$_0$ and color coded by LAMOST [Fe/H] for our full training sample. The solid black lines mark (J0430 − W1)$_0$ = 2.9 mag (vertical) and (uJAVA − J0430)$_0$ = 1.2 mag, beyond which $T_{\rm eff}$ begins to be systematically underestimated. The dashed gray lines mark the limits of the respective ranges for the 25 LAMOST matched stars, namely 0.8 < (J0430 − W1)$_0$/mag < 4.4 and 0.8 < (uJAVA − J0430)$_0$/mag < 1.7.

to resolve the relationship more effectively and yield more accurate predictions.

Importantly, this systematic behavior is only partially captured by the reported uncertainties. Because they are derived from repeated model predictions under perturbed inputs (magnitudes and distances), regions affected by attenuation bias may appear to yield confident results, despite the predictions being systematically biased toward the central tendency. As a result, uncertainties in such regimes are likely underestimated.

It is therefore important that we have explicitly identified and mapped this limitation, which will be more thoroughly addressed in future work. Nevertheless, the overall performance metrics of the [Fe/H] model indicate that it performs well despite this caveat. The mean uncertainty in adopted metallicity, $\langle\sigma_{\rm [Fe/H]}\rangle$ = 0.158 dex, supports the reliability of the predictions, though it is important to emphasize that this uncertainty is focused in capturing the quality of the input photometry, rather than internal model biases arising from feature degeneracies. In this context, the ability to diagnose and map such limitations becomes a key outcome of this study. Together with the model's overall predictive performance, this characterization represents one of the most relevant contributions of the present work.

5 CLUSTERS METALLICITIES

In the context of this work, it is reasonable to expect that photometric results for metallicity may not fall within the narrow internal iron abundance spreads typically inferred from spectroscopic studies of open clusters. These spreads are generally on the order of $\lesssim$ 0.02–0.03 dex (Bovy 2016; Poovelil et al. 2020). In contrast, photometric metallicity estimates tend to show uncertainties around

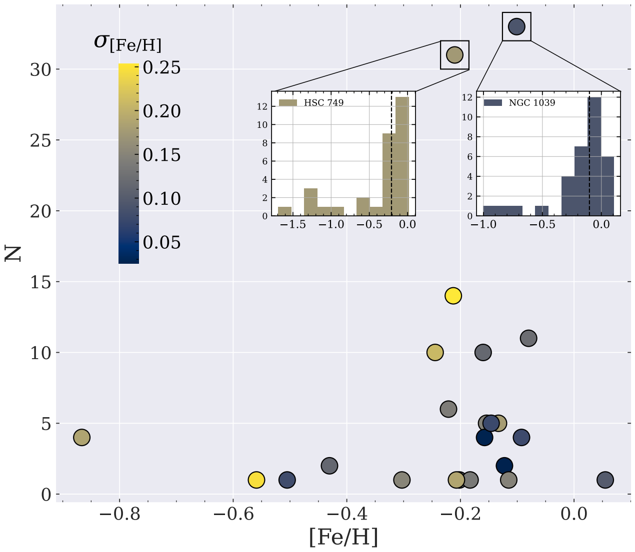

Figure 15. Number of members in each of our 23 clusters plotted against adopted metallicity, with circles color coded by adopted uncertainty for the clusters. The gray dashed line marks the mean value for [Fe/H].

0.05–0.10 dex (Netopil et al. 2013, 2016), even when internal consistency is enforced. Therefore, the metallicity scatter observed in this work is expected to reflect the limitations of photometric precision rather than any true deviation from open cluster homogeneity.

We adopted the median of individual members' metallicities to estimate each system's metallicity, and the MAD as the associated uncertainty. This choice was motivated by the presence of clear outliers in some clusters significantly deviating from the bulk of the member distribution, such as those illustrated for NGC 1039 and HSC 749 in the histograms in Figure 15. Robust statistics like the median and MAD are less sensitive to such outliers and thus provide a more reliable characterization of the typical cluster metallicity. This is particularly appropriate given the expectation that cluster members formed together from a chemically homogeneous molecular cloud. In cases where only one member was available, we adopted the individual metallicity and its associated uncertainty (estimated in the present work) for that star.

Figure 15 illustrates these results by showing our 23 associations distributed in a space of metallicity vs. the number of members we retrieved, color coded by metallicity uncertainty. Furthermore, a few more metal-poor cases occur for objects with less than 5 members, and we recall that some of those are represented by only one object. In such cases, these results must be interpreted with caution.

Our estimates are gathered around [Fe/H] ~ -0.2, which is consistent with the underestimation scenario described in Section 4.2. As previously discussed, stars hotter than 5000 K occupy a narrow region in the $(\mathrm{uJAVA} - \mathrm{J0430})_0$ vs. T_{eff} space (Figure 13), which limits the model's ability to capture variations in metallicity. Given that $(\mathrm{uJAVA} - \mathrm{J0430})_0$ is the most important feature for predicting [Fe/H], this compression in feature space hinders the model's capacity to learn meaningful correlations. In such cases, as described earlier, the model tends to regress toward the mean of the training distribution—in this case, $\mu^{\mathrm{training}}_{[\mathrm{Fe/H}]} = -0.21$ dex—leading to a systematic bias in that direction.

More specifically, Figure 16 presents Kiel diagrams displaying the parameters predicted in this study for our six most populated clusters. The vertical and horizontal axes represent log g and T_{eff}, respectively, while [Fe/H]is indicated by the color bar. Figure 17 shows color-magnitude diagrams (CMDs) also color-coded by metallicity, with the horizontal axis representing the feature that is most relevant to T_{eff} and composed only by narrow bands. In the left panel, absolute $W1$ magnitude, $W1^{\mathrm{abs}}_0$, was selected as it is the most important feature for log g, while in the right panel, uJAVA_0 was used to illustrate a CMD based solely on J-PLUS photometry. It is evident that the main sequences are better recovered in the latter case. Overall, the trends are consistent with the expected main-sequence morphology, with some dispersion in metallicity and data points accumulated in the cool end.

The data points accumulated at the cool end ($T_{\mathrm{eff}} < 5000$ K) in Figure 16 are the same as those identified and discussed in Figure 11. In these cases, this behavior is merely a reflection of our stars being cooler than the model is capable of predicting. It is therefore not a coincidence that they also exhibit the largest uncertainties in T_{eff} as well as anomalous metallicities.

We performed brief comparisons of our [Fe/H] estimates for OCs with values from the literature, specifically from the works of Cavallo et al. (2023) and Zhang et al. (2024). Figure 18 presents a direct comparison of the metallicity estimates for all 23 stellar systems for which we estimated [Fe/H]. Our results are shown alongside those from the two literature sources, with distinct markers indicating each source: circles for this work, diamonds for Cavallo et al. (2023), and squares for Zhang et al. (2024). Marker colors encode the associated uncertainties, and soft gray error bars accompany all points, allowing for visual inspection of agreement within uncertainties.

The majority of systems display reasonable consistency across the three works. A few objects—such as HSC 1058 and Theia 1188—show notable discrepancies, which may reflect chemical inhomogeneity, differences in membership selection, or small-number statistics. We also note that our [Fe/H] estimates tend to be systematically lower than those from the literature. This is expected given the underestimation bias inherent to our method. Despite this bias, the relative trends across systems are preserved, and the agreement remains satisfactory within uncertainties.

Furthermore, with an appropriately updated training sample directed at containing more metal-poorer stars, our approach may be a valuable tool in the analysis of multiple stellar populations in Galactic Globular Clusters (see e.g. Hartmann et al. 2022, and references therein). In particular, it could contribute in building the metallicity distribution function of unresolved extragalactic globular clusters belonging to low-redshift galaxies.

6 CONCLUSIONS

In this work, we developed machine learning models based on the gradient boosting framework LightGBM to estimate atmospheric stellar parameters for stars in J-PLUS DR3. Photometric data was also taken from CatWISE and Gaia DR3, from which we retrieved distances estimated in Bailer-Jones et al. (2021) to compute absolute magnitudes. After performing feature engineering on extinction-corrected data, we obtained solid results, as indicated by the mean absolute errors used to evaluate model performance. The models were then applied to a sample of 157 stars—later reduced to 154 after a distance-based filtering—identified as members of open clusters and moving groups within the J-PLUS DR3 footprint (Appendix B), selected based on the works of Hunt & Reffert (2023) and Hunt & Reffert (2024). Our results have demonstrated that this approach is

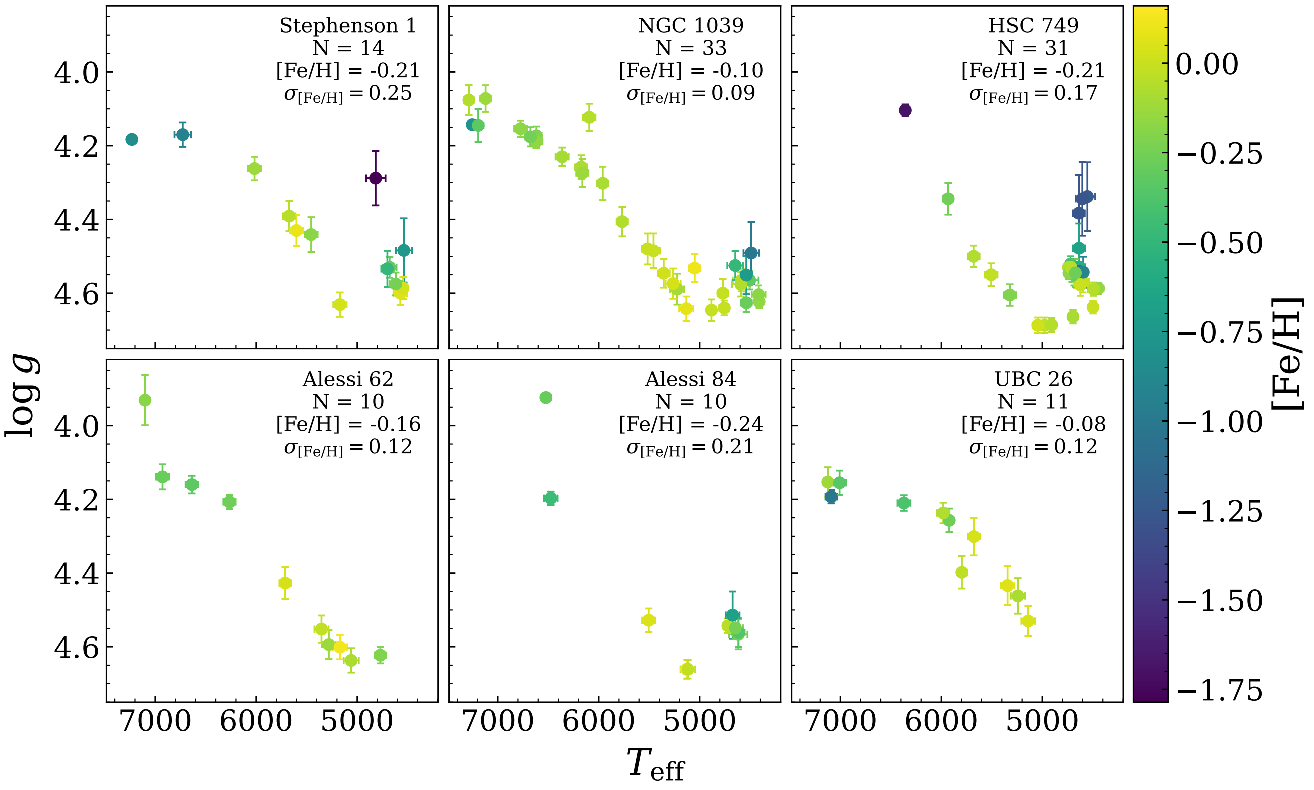

Figure 16. Kiel diagrams for stars in the six most populated systems in our work, color-coded by [Fe/H]. Each panel indicates the cluster name, number of stars (N), adopted metallicity and respective uncertainty. The trends are consistent with the expected main-sequence morphology, with some dispersion in metallicity and data points accumulated in the cool end.

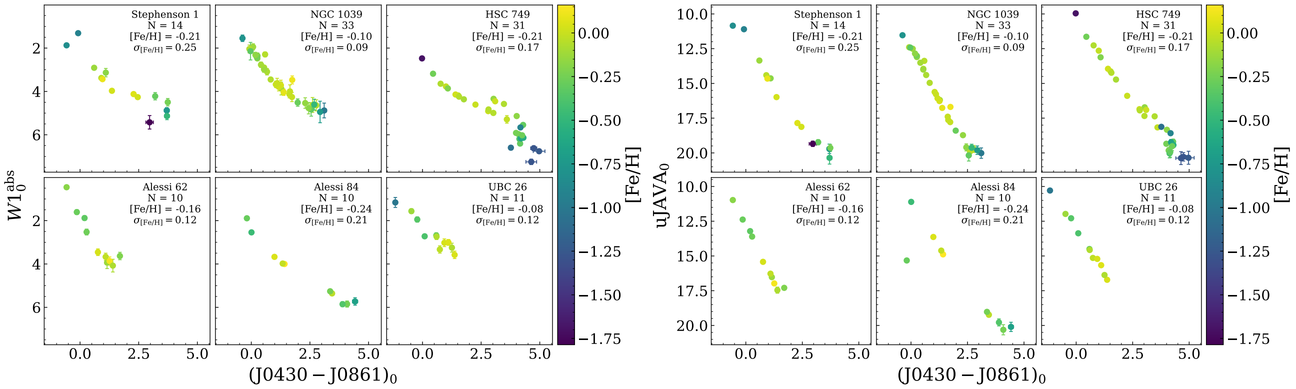

Figure 17. Same as in Figure 16, but here with color magnitude diagrams instead of Kiel diagrams. In the left panel, vertical axis represent extinction-corrected, absolute $W1$ magnitude, $W1_0^{\rm abs}$, the most important feature for $\log\,g$; in the right panel, vertical axis represent the broad band uJAVA$_0$ to illustrate J-PLUS photometry. Horizontal axes in both panels represent extinction-corrected $(\mathrm{J0430} - \mathrm{J0861})_0$, an important feature for $T_{\rm eff}$ that relies only J-PLUS narrow band photometry.

not only viable but also competitive for estimating stellar parameters from photometric data.

In this process, we showed that using all the available selected best features, instead of only a subset of those, is the best choice in this case. One of the main advantages of selecting fewer features would be to enlarge sample size, since less magnitudes would be necessary and quality cuts would filter out less stars. Nevertheless, in addition to evaluation metrics becoming less reliable, among the necessary magnitudes are mainly those that limit the sample size the most, e.g. J-PLUS filters J0395, J0378, especially for [Fe/H].

Our results for open cluster members reveal a mild degeneracy between $T_{\rm eff}$ and [Fe/H], leading to a systematic underestimation of metallicity at higher temperatures. This effect stems from a concentration of training examples in a compact region of feature space, where non-linear relationships likely further degrade model performance.

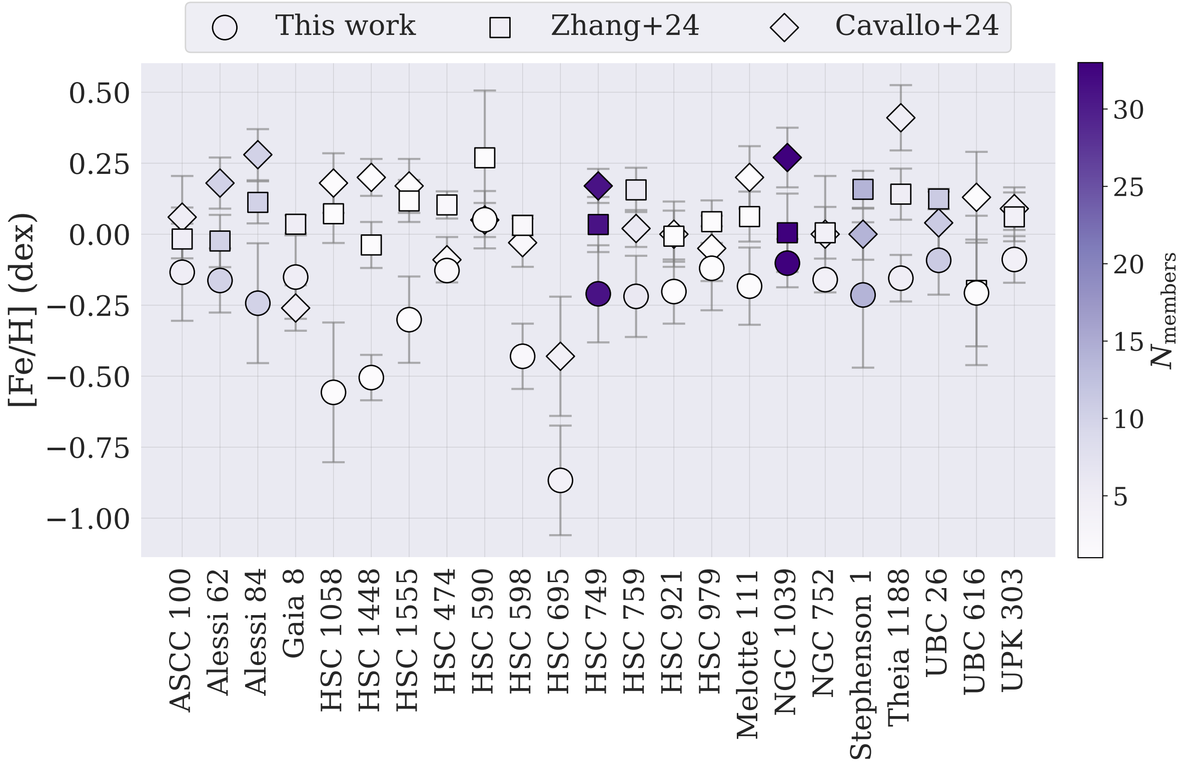

Figure 18. Comparison of [Fe/H] values for our 23 stellar systems. Results are shown for this work (circles), Cavallo et al. (2023, diamonds), and Zhang et al. (2024, squares), colored by the number of members analyzed in this work. The systems are ordered alphabetically along the x-axis.

These findings underscore that, although the models predict each parameter independently, correlations among photometric features introduce indirect dependencies. The underestimation of [Fe/H] in hotter stars appears to arise from the complex interplay between temperature-sensitive features and metallicity indicators, compounded by imbalances in the training data. While not severe, this bias is noticeable and highlights the difficulty of disentangling intrinsic stellar properties from their observational signatures. It also reinforces the importance of careful feature selection and model interpretation when applying machine learning to stellar parameter estimation.

When compared to OC literature estimates, metallicities predicted by our models tend to be systematically lower. This offset is consistent with the underestimation bias observed in our internal validation tests (see Section 4.2). Despite this, the relative trends among clusters are well preserved, and the estimates generally agree with literature values within the associated uncertainties. These results support the robustness of our metallicity predictions for comparative analyses, even if their absolute calibration is affected by a mild systematic shift.

An obvious next step is to apply the methodology presented in this paper to S-PLUS data, leveraging the fact that both surveys share the same photometric system. This is currently a work in progress (Machado-Pereira et al., in prep.) that will enable comparisons to the results found in the comprehensive work of Lopes et al. (2025).

A further possibility is to take advantage of the behavior observed in Figure 4 to seek a compromise between the number of features and model performance. That way, fewer filters are required and the training sample may be larger, making room for improvement in terms of accuracy. As discussed in this work, this strategy that may not be ideal depending on the goal. For instance, reliable metallicity estimates depend on bluer filters (see discussion in Section 3 and Figure 5, and stars for which these magnitudes are available will probably have measures on most of the other filters as well. On the other hand, log g estimates require less challenging magnitudes, allowing for further efforts on giant–dwarf separation, for example.

In future work, we also plan to apply photometric decontamination techniques to color-magnitude diagrams (CMDs) for membership selection (e.g., Bonatto & Bica 2007, 2010; Bica & Bonatto 2011). Our aim is to increase both the number of clusters detected in the J-PLUS footprint and the number of stars identified as cluster members, further leveraging J-PLUS's unique capabilities, in particular exploring the forthcoming DR4.

ACKNOWLEDGEMENTS

E. M.-P. acknowledges CAPES (procs. 88887.605761/2021-00 and 88881.846754/2023-01). The work of E.M.-P. and V.M.P. is supported by NOIRLab, which is managed by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the U.S. National Science Foundation. S.D. acknowledges CNPq/MCTI for grant 306859/2022-0 and FAPERJ for grant 210.688/2024. P.C. acknowledges financial support from the Spanish Virtual Observatory project funded by the Spanish Ministry of Science and Innovation/State Agency of Research

MCIN/AEI/10.13039/501100011033 through grant PID2020-112949GB-I00. M.B.F. acknowledges financial support from the National Council for Scientific and Technological Development (CNPq) Brazil (grant number: 307711/2022-6). This research has partially funded by MICIU/AEI/10.13039/501100011033/ through grant PID2023-146210NB-I00.

This research uses services or data provided by the Astro Data Lab at NSF's National Optical-Infrared Astronomy Research Laboratory. NOIRLab is operated by the Association of Universities for Research in Astronomy (AURA), Inc. under a cooperative agreement with the National Science Foundation. Based on observations made with the JAST/T80 telescope at the Observatorio Astrofísico de Javalambre (OAJ), in Teruel, owned, managed, and operated by the Centro de Estudios de Física del Cosmos de Aragón (CEFCA). We acknowledge the OAJ Data Processing and Archiving Unit (UPAD) for reducing and calibrating the OAJ data used in this work. Funding for the J-PLUS Project has been provided by the Governments of Spain and Aragón through the Fondo de Inversiones de Teruel; the Aragón Government through the Research Groups E96, E103, and E16_17R; the Spanish Ministry of Science, Innovation and Universities (MCIU/AEI/FEDER, UE) with grants PGC2018-097585-B-C21 and PGC2018-097585-B-C22; the Spanish Ministry of Economy and Competitiveness (MINECO) under AYA2015-66211-C2-1-P, AYA2015-66211-C2-2, AYA2012-30789, and ICTS-2009-14; and European FEDER funding (FCDD10-4E-867, FCDD13-4E-2685). The Brazilian agencies FINEP, FAPESP, and the National Observatory of Brazil have also contributed to this project. This work has made use of data from the European Space Agency (ESA) mission *Gaia* `https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/gaia/dpac/consortium`). Funding for the DPAC has been provided by national institutions, in particular, the institutions participating in the *Gaia* Multilateral Agreement. This paper also made use of the NASA ' s Astrophysics Data System (ADS) bibliographic services; the arXiv preprint server operated by Cornell University; the open-source Python packages Astropy (Astropy Collaboration et al. 2013, 2018, 2022), NumPy (Harris et al. 2020), pandas (Wes McKinney 2010; pandas development team 2023), Matplotlib (Hunter 2007), scikit-learn (Pedregosa et al. 2011); the SIMBAD database and the VizieR catalogue access tool, CDS, Strasbourg, France (DOI : 10.26093/cds/vizier). The original description of the VizieR service was published in 2000, A&AS 143, 23; and VOSA, developed under the Spanish Virtual Observatory (`https://svo.cab.inta-csic.es`) project funded by MCIN/AEI/10.13039/501100011033/ through grant PID2020-112949GB-I00. VOSA has been partially updated by using funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement nº 776403 (EXOPLANETS-A)

DATA AVAILABILITY

The J-PLUS DR3 Data Release (J-PLUS-DR3, July 2022) is based on images collected from November 2015 to February 2022 by the JAST80 telescope. J-PLUS DR3 data are available at `https://archive.cefca.es/catalogues/jplus-dr3`.

REFERENCES

Andrae R., Rix H.-W., Chandra V., 2023, ApJS, 267, 8
Astropy Collaboration et al., 2013, A&A, 558, A33
Astropy Collaboration et al., 2018, AJ, 156, 123
Astropy Collaboration et al., 2022, ApJ, 935, 167
Babusiaux C., et al., 2023, A&A, 674, A32
Bailer-Jones C. A. L., Rybizki J., Fouesneau M., Demleitner M., Andrae R., 2021, AJ, 161, 147
Bayo, A. Rodrigo, C. Barrado y Navascués, D. Solano, E. Gutiérrez, R. Morales-Calderón, M. Allard, F. 2008, A&A, 492, 277
Belokurov V., Zucker D. B., Evans N. W. e. a., 2006, The Astrophysical Journal, 642, L137
Benitez N., et al., 2014, arXiv e-prints, p. arXiv:1403.5237
Bica E., Bonatto C., 2011, A&A, 530, A32
Bica E., Pavani D. B., Bonatto C. J., Lima E. F., 2019, The Astronomical Journal, 157, 12
Bochanski J. J., Hawley S. L., Covey K. R. e. a., 2010, The Astronomical Journal, 139, 2679
Bonatto C., Bica E., 2007, Monthly Notices of the Royal Astronomical Society, 377, 1301
Bonatto C., Bica E., 2010, A&A, 516, A81
Bonatto C., Kerber L. O., Bica E., Santiago B. X., 2006, Astronomy & Astrophysics, 446, 121
Bonoli S., et al., 2021, Astronomy & Astrophysics, 653, A31
Bovy J., 2016, The Astrophysical Journal, 817, 49
Buder S., Asplund M., Duong L. e. a., 2021, Monthly Notices of the Royal Astronomical Society, 506, 150
Cantat-Gaudin T., Anders F., Castro-Ginard A., et al., 2020, Astronomy & Astrophysics, 640, A1
Cavallo L., et al., 2023, The Astronomical Journal, 167, 12
Cenarro A. J., et al., 2019, A&A, 622, A176
Chambers K. C., Magnier E. A., Metcalfe N. e. a., 2016, arXiv e-prints
Chen T., Guestrin C., 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. Association for Computing Machinery, New York, NY, USA, p. 785–794, doi:10.1145/2939672.2939785, `https://doi.org/10.1145/2939672.2939785`
Chierichini S., Liu J., Korsós M. B., Del Moro D., Erdélyi R., 2024, ApJ, 963, 121
Coelho P. R. T., 2014, Monthly Notices of the Royal Astronomical Society, 440, 1027
Creevey O. L., et al., 2023, A&A, 674, A26
Cui X.-Q., Zhao Y.-H., Chu Y.-Q. e. a., 2012, Research in Astronomy and Astrophysics, 12, 1197
Cutri R. M., Wright E. L., Conrow T. e. a., 2013, Explanatory Supplement to the AllWISE Data Release Products, `https://wise2.ipac.caltech.edu/docs/release/allwise/expsup/`
De Silva G. M., Freeman K. C., Bland-Hawthorn J. e. a., 2015, Monthly Notices of the Royal Astronomical Society, 449, 2604
Dias W. S., Monteiro H., Moitinho A., et al., 2021, Monthly Notices of the Royal Astronomical Society, 504, 356–371
Fragkou V., Parker Q. A., Zijlstra A., Shaw R., Lykou F., 2019, Monthly Notices of the Royal Astronomical Society, 484, 3078–3092
Friedman J. H., 2001, Annals of statistics, pp 1189–1232
Gaia Collaboration 2022, VizieR Online Data Catalog, p. I/355
Gaia Collaboration et al., 2016, A&A, 595, A1
Gaia Collaboration et al., 2023, A&A, 674, A1
Galarza C. A., et al., 2022, A&A, 657, A35
Harris C. R., et al., 2020, Nature, 585, 357
Hartmann E. A., et al., 2022, MNRAS, 515, 4191
He X.-J., Luo A. L., Chen Y.-Q., 2022, MNRAS, 512, 1710
Herpich F. R., et al., 2024, A&A, 689, A249
Huang Y., et al., 2024, The Astrophysical Journal, 974, 192
Huang B., Yuan H., Xu S., Xiao K., Xiang M., Huang Y., Beers T. C., 2025, The Astrophysical Journal Supplement Series, 277, 7
Hunt E. L., Reffert S., 2021, Astronomy & Astrophysics, 646, A104
Hunt E. L., Reffert S., 2023, A&A, 673, A114
Hunt E. L., Reffert S., 2024, A&A, 686, A42
Hunter J. D., 2007, Computing in Science & Engineering, 9, 90
Ivezić Ž., Sesar B., Jurić M. e. a., 2008, The Astrophysical Journal, 684, 287
Ivezić Ž., Connolly A. J., VanderPlas J. T., Gray A., 2020, Statistics, Data Mining, and Machine Learning in Astronomy. A Practical Python Guide

for the Analysis of Survey Data, Updated Edition. Princeton University Press

Kauffmann G., Heckman T. M., White S. D. M. e. a., 2003, Monthly Notices of the Royal Astronomical Society, 341, 33

Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y., 2017, in Advances in Neural Information Processing Systems. Curran Associates, Inc., `https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html`

Keller S. C., Schmidt B. P., Bessell M. S. e. a., 2007, Publications of the Astronomical Society of Australia, 24, 1

Lietzen H., Tempel E., Liivamägi L. J. e. a., 2016, Astronomy & Astrophysics, 588, L4

Liu C., et al., 2020, arXiv e-prints, p. arXiv:2005.07210

Lopes C. E. F., et al., 2025, Astronomy & Astrophysics, 693, A306

López-Sanjuan C., et al., 2019, A&A, 622, A177

López-Sanjuan C., et al., 2024, A&A, 683, A29

Lundberg S. M., Lee S.-I., 2017, Advances in neural information processing systems, 30

Mainzer A., Bauer J., Cutri R. M. e. a., 2014, The Astrophysical Journal, 792, 30

Majewski S. R., Schiavon R. P., Frinchaboy P. M. e. a., 2017, The Astronomical Journal, 154, 94

Marocco F., et al., 2021, ApJS, 253, 8

Mayo A. W., Dressing C. D., Vanderburg A., Fortenbach C. D., et al., 2023, The Astronomical Journal, 165, 235

Mendes de Oliveira C., et al., 2019, MNRAS, 489, 241

Nardiello D., Bedin L. R., Griggio M., Salaris M., Scalco M., Cassisi S., 2023, Monthly Notices of the Royal Astronomical Society, 525, 2585–2604

Netopil M., Paunzen E., Heiter U., Soubiran C., 2013, A&A, 556, A26

Netopil M., Paunzen E., Heiter U., Soubiran C., 2016, A&A, 585, A150

Onken C. A., Wolf C., Bessell M. S. e. a., 2024, Publications of the Astronomical Society of Australia, 41, e001

Osuna P., et al., 2008, IVOA Astronomical Data Query Language Version 2.00, IVOA Recommendation 30 October 2008 (`arXiv:1110.0503`), doi:10.5479/ADS/bib/2008ivoa.spec.1030O

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Perren G. I., Pera M. S., Navone H. D., Vázquez R. A., 2023, Monthly Notices of the Royal Astronomical Society, 526, 4107–4119

Poovelil V. J., et al., 2020, The Astrophysical Journal, 903, 55

Quispe-Huaynasi F., et al., 2023, Monthly Notices of the Royal Astronomical Society, 522, 3898

Quispe-Huaynasi F., et al., 2024, Monthly Notices of the Royal Astronomical Society, 527, 6173

Quispe-Huaynasi F., et al., 2025, arXiv e-prints, p. arXiv:2504.18491

Robinson D., Avestruz C., Gnedin N. Y., 2024, MNRAS, 528, 255

Schlafly E. F., Meisner A. M., Green G. M. e. a., 2019, The Astrophysical Journal Supplement Series, 240, 30

Shapley L. S., 1951, Notes on the N-Person Game —II: The Value of an N-Person Game. RAND Corporation, Santa Monica, CA, doi:10.7249/RM0670

Skrutskie M. F., et al., 2006, The Astronomical Journal, 131, 1163

Starkenburg E., Martin N. F., Youakim K. e. a., 2017, Monthly Notices of the Royal Astronomical Society, 471, 2587

Sánchez N., López-Martínez F., Ocando S., Blay P., 2021, Astrophysics and Space Science, 366, 92

Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, Astronomical Society of the Pacific Conference Series Vol. 347, Astronomical Data Analysis Software and Systems XIV. p. 29

Ting Y.-S., 2024, arXiv e-prints, p. arXiv:2412.05806

Wang C., et al., 2022, Astronomy & Astrophysics, 664, A38

Wes McKinney 2010, in Stéfan van der Walt Jarrod Millman eds, Proceedings of the 9th Python in Science Conference. pp 56 – 61, doi:10.25080/Majora-92bf1922-00a

Whitten D. D., et al., 2019, Astronomy & Astrophysics, 622, A182

Whitten D. D., et al., 2021, The Astrophysical Journal, 912, 147

Wright E. L., Eisenhardt P. R. M., Mainzer A. K. e. a., 2010, The Astronomical Journal, 140, 1868

Wu Y., Du B., Luo A.-L. e. a., 2014, Research in Astronomy and Astrophysics, 14, 719

Yang L., et al., 2022, Astronomy & Astrophysics, 659, A181

Yanny B., Rockosi C., Newberg H. J. e. a., 2009, The Astronomical Journal, 137, 4377

York D. G., Adelman J., Anderson J. E. e. a., 2000, The Astronomical Journal, 120, 1579

Yuan H. B., Liu X. W., Xiang M. S., 2013, MNRAS, 430, 2188

Zhang X., Green G. M., Rix H.-W., 2023, Monthly Notices of the Royal Astronomical Society, 524, 1855

Zhang R., et al., 2024, Astronomy & Astrophysics, 692, A212

Zhao G., Zhao Y.-H., Chu Y.-Q. e. a., 2012, Research in Astronomy and Astrophysics, 12, 723

pandas development team T., 2023, pandas-dev/pandas: Pandas, doi:10.5281/zenodo.10304236, `https://doi.org/10.5281/zenodo.10304236`

APPENDIX A: QUERY FOR RETRIEVING STARS IN J-PLUS DR3

The following ADQL query was implemented using Python to get stars from JPLUS DR3 using restricted criteria regarding flags, uncertainties and probability of being a star. This approach is necessary given the limitation in J-PLUS' online data access services of returning up to 1 million rows per query.

The conditions to which this query is restricted include the following.
`norm_wmap_val` > 0.8: adequate exposure,
`mask_flags` = 0: avoiding any known issues affecting the quality of the photometry,
`mag_err_aper_cor_6_0` > 0.5: filtering mostly obvious bad quality sources,
`mag_aper_cor_3_0` < 22: limiting magnitude described in López-Sanjuan et al. (2024, with some small tolerance),
`sglc_prob_star` > 0.7: 70% or higher probability of being a star.
The typeset 0 at line 28 is responsible for looping over 83 blocks files containing the merged 1642 reference tiles.

```
1  SELECT
2  dualobj.number, dualobj.tile_id,
3  dualobj.alpha_j2000, dualobj.delta_j2000,
4  sgc.sglc_prob_star, dualobj.class_star,
5  ps.mag_aper_cor_6_0, ps.mag_err_aper_cor_6_0,
6  ext.ax, ext.ax_err, ext.ebv, ext.ebv_err,
7  dualobj.mag_aper_3_0, dualobj.
       mag_err_aper_3_0,
8  dualobj.mag_aper_6_0, dualobj.
       mag_err_aper_6_0,
9  dualobj.mag_auto, dualobj.mag_err_auto,
10 array_sub(dualobj.mag_auto, ext.ax) AS
       mag_auto_0,
11 dualobj.norm_wmap_val,
12 ps.mag_aper_cor_3_0, ps.mag_err_aper_cor_3_0,
13 ti.aper_cor_3_0, ti.aper_cor_6_0, ti.
       aper_cor_err_6_0
14
15 FROM
16 jplus.MagABDualObj AS dualobj
17 INNER JOIN jplus.MagABDualPointSources AS ps
       ON ps.number = dualobj.number AND ps.
       tile_id = dualobj.tile_id
18 INNER JOIN jplus.MWExtinction AS ext ON ext.
       number = dualobj.number AND ext.tile_id =
        dualobj.tile_id
19 INNER JOIN jplus.StarGalClass AS sgc ON sgc.
       number = dualobj.number AND sgc.tile_id =
        dualobj.tile_id
20 INNER JOIN jplus.TileImage AS ti ON ti.
       tile_id = dualobj.tile_id
21
22 WHERE
23 (array_min_float(dualobj.norm_wmap_val) >
       0.8)
24 AND (array_max_int(dualobj.mask_flags) = 0)
25 AND (array_max_float(ps.mag_err_aper_cor_6_0)
        < 0.5)
26 AND (array_max_float(dualobj.mag_aper_3_0) <
       22)
27 AND sgc.sglc_prob_star >= 0.7
28 AND dualobj.tile_id IN {0}
```

APPENDIX B: PARAMETERS DERIVED IN THIS WORK

In Table B1, we provide parameters for the systems analyzed in this work.

This paper has been typeset from a TeX/LaTeX file prepared by the author.

Table B1. Cluster parameters including metallicity estimates, Gaia and HR23 distances, and age. Distances and age values are expressed with asymmetric uncertainties based on the 16th and 84th percentiles.

Cluster	$N\,(*)$	[Fe/H] (TW)	$[\mathrm{Fe/H}]_{\mathrm{C23}}$	$[\mathrm{Fe/H}]_{\mathrm{Z24}}$	$d^{\mathrm{rpgeo}}_{\mathrm{BJ+21}}$ (pc)	d_{HR23} (pc)	$\log(t/\mathrm{Gyr})_{\mathrm{HR23}}$	Type (**)
ASCC 100	5	-0.134 ± 0.171	0.06 ± 0.145	-0.017 ± 0.111	$367.0^{+4.3}_{-4.3}$	$360.4^{+0.3}_{-0.3}$	$7.812^{+0.299}_{-0.303}$	o → m
Alessi 62	10 (11)	-0.1625 ± 0.113	0.18 ± 0.09	-0.024 ± 0.092	$614.3^{+4.2}_{-25.2}$	$606.3^{+0.6}_{-0.5}$	$8.629^{+0.253}_{-0.206}$	o
Alessi 84	10	-0.2425 ± 0.2105	0.28 ± 0.09	0.112 ± 0.074	$196.8^{+3.2}_{-2.2}$	$198.3^{+0.1}_{-0.1}$	$8.068^{+0.248}_{-0.252}$	o → m
Gaia 8	5	-0.151 ± 0.147	-0.26 ± 0.08	0.035 ± 0.032	$286.4^{+8.5}_{-12.9}$	$289.2^{+0.2}_{-0.2}$	$7.324^{+0.254}_{-0.256}$	o
HSC 1058	1	-0.557 ± 0.246	0.18 ± 0.105	0.072 ± 0.103	$485.5^{+0.0}_{-0.0}$	$501.6^{+0.7}_{-0.7}$	$8.069^{+0.239}_{-0.276}$	m
HSC 1448	1	-0.505 ± 0.080	0.2 ± 0.065	-0.038 ± 0.081	$115.8^{+0.0}_{-0.0}$	$118.0^{+0.1}_{-0.1}$	$8.318^{+0.330}_{-0.298}$	m
HSC 1555	1	-0.301 ± 0.152	0.17 ± 0.095	0.117 ± 0.074	$171.4^{+0.0}_{-0.0}$	$172.9^{+0.2}_{-0.2}$	$8.107^{+0.270}_{-0.271}$	o → m
HSC 474	2	-0.1275 ± 0.0205	-0.09 ± 0.08	0.103 ± 0.048	$328.4^{+11.2}_{-11.3}$	$315.4^{+0.3}_{-0.3}$	$7.491^{+0.257}_{-0.244}$	o → m
HSC 590	1	0.051 ± 0.101	0.05 ± 0.06	0.269 ± 0.237	$115.1^{+0.0}_{-0.0}$	$118.1^{+0.1}_{-0.1}$	$8.504^{+0.313}_{-0.285}$	o → m
HSC 598	2	-0.4295 ± 0.1145	-0.03 ± 0.085	0.031 ± 0.035	$193.1^{+1.6}_{-1.7}$	$184.3^{+0.1}_{-0.1}$	$8.212^{+0.342}_{-0.324}$	o → m
HSC 695	4	-0.8665 ± 0.1925	-0.43 ± 0.21	—	$6975.5^{+1577.0}_{-1448.0}$	$9550.3^{+827.3}_{-705.1}$	$8.826^{+0.532}_{-0.398}$	m
HSC 749	31	-0.210 ± 0.171	0.17 ± 0.06	0.034 ± 0.097	$118.4^{+5.1}_{-4.1}$	$121.2^{+0.0}_{-0.0}$	$8.354^{+0.301}_{-0.408}$	o → m
HSC 759	6 (7)	-0.219 ± 0.143	0.02 ± 0.065	0.156 ± 0.078	$98.2^{+10.0}_{-4.0}$	$96.0^{+0.1}_{-0.1}$	$8.174^{+0.269}_{-0.294}$	o → m
HSC 921	1	-0.202 ± 0.113	0.0 ± 0.115	-0.007 ± 0.09	$166.9^{+0.0}_{-0.0}$	$172.9^{+0.1}_{-0.1}$	$7.635^{+0.280}_{-0.236}$	o → m
HSC 979	1	-0.120 ± 0.148	-0.05 ± 0.115	0.044 ± 0.075	$314.1^{+0.0}_{-0.0}$	$317.2^{+0.4}_{-0.4}$	$8.263^{+0.266}_{-0.358}$	m
Melotte 111	1	-0.183 ± 0.136	0.2 ± 0.11	0.062 ± 0.088	$87.7^{+0.0}_{-0.0}$	$85.3^{+0.0}_{-0.0}$	$8.816^{+0.221}_{-0.195}$	o
NGC 1039	33	-0.102 ± 0.085	0.27 ± 0.105	0.005 ± 0.138	$492.2^{+15.1}_{-12.3}$	$489.7^{+0.3}_{-0.3}$	$8.087^{+0.280}_{-0.232}$	o
NGC 752	4	-0.160 ± 0.022	0.0 ± 0.205	0.005 ± 0.091	$435.8^{+4.8}_{-11.2}$	$433.6^{+0.3}_{-0.3}$	$9.152^{+0.216}_{-0.182}$	o
Stephenson 1	14	-0.2135 ± 0.2555	0.0 ± 0.09	0.158 ± 0.065	$353.0^{+4.4}_{-9.0}$	$354.7^{+0.2}_{-0.2}$	$7.425^{+0.275}_{-0.262}$	o
Theia 1188	5 (6)	-0.155 ± 0.082	0.41 ± 0.115	0.141 ± 0.09	$325.7^{+9.8}_{-3.4}$	$325.2^{+0.3}_{-0.3}$	$8.756^{+0.236}_{-0.237}$	o → m
UBC 26	11	-0.092 ± 0.121	0.04 ± 0.11	0.124 ± 0.036	$600.7^{+16.5}_{-9.9}$	$593.5^{+0.5}_{-0.5}$	$7.401^{+0.208}_{-0.211}$	o
UBC 616	1	-0.207 ± 0.188	0.13 ± 0.16	-0.198 ± 0.263	$2823.7^{+0.0}_{-0.0}$	$2758.4^{+44.6}_{-43.2}$	$9.050^{+0.211}_{-0.171}$	o
UPK 303	4	-0.0885 ± 0.082	0.09 ± 0.075	0.061 ± 0.086	$215.8^{+3.8}_{-2.2}$	$211.2^{+0.1}_{-0.1}$	$8.045^{+0.266}_{-0.314}$	o → m

Notes:
(*) In parenthesis, the number of members before distance cut;
(**) type of object update from HR23 to HR24 (HR23→HR24, with *o* and *m* meaning open cluster and moving group, respectively).