Characteristics, Classification and Challenges in Searching for CEMP Stars Postprint

Authors: Lefeng He, Xiangru Li, Xiaoming Kong, A-Li Luo and Haifeng Yang

Date: 2025-06-13T16:51:13+00:00

Abstract

The study of carbon-enhanced metal-poor (CEMP) stars is of great significance for understanding the chemical evolution of the early universe and stellar formation. CEMP stars are characterized by carbon overabundance and are classified into several subclasses based on the abundance patterns of neutron-capture elements,including CEMP-s,CEMP-no,CEMP-r,and CEMP-r/s. These subclasses provide important insights into the formation of the first stars,early stellar nucleosynthesis,and supernova explosions. However, one of the major challenges in CEMP star research is the relatively small sample size of identified stars, which limits statistical analyses and hinders a comprehensive understanding of their properties. Fortunately, a series of large-scale spectroscopic survey projects have been launched and developed in recent years, providing unprecedented opportunities and technical challenges for the search and study of CEMP stars. To this end, this paper draws on the progress and future prospects of existing methods in constructing large CEMP data sets and offers an in-depth discussion from a technical standpoint, focusing on the strengths and limitations. In addition,we review recent advancements in the identification of CEMP stars,emphasizing the growing role of machine learning in processing and analyzing the increasingly large data sets generated by modern astronomical surveys. Compared to traditional spectral analysis methods,machine learning offers greater efficiency in handling complex data,automatic extraction of stellar parameters,and improved prediction accuracy. Despite these advancements,the research faces persistent challenges,including the scarcity of labeled samples,limitations imposed by low-resolution spectra,and the lack of interpretability in machine learning models. To address these issues,the paper proposes potential solutions and future research directions aimed at advancing the study of CEMP stars and enhancing our understanding of their role in the chemical evolution of the universe.

Full Text

Preamble

Research in Astronomy and Astrophysics, 25:055012 (10pp), 2025 May https://doi.org/10.1088/1674-4527/adccf5 CSTR: 32081.14.RAA.adccf5

Characteristics, Classification and Challenges in Searching for CEMP Stars

Lefeng He[1], Xiangru Li[1], Xiaoming Kong[23], A-Li Luo[456], and Haifeng Yang[7]

[1] School of Computer Science, South China Normal University, Guangzhou 510631, China; xiangru.li@qq.com [2] School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai 264209, China [3] Shandong Key Laboratory of Intelligent Electronic Packaging Testing and Application, Shandong University, Weihai 264209, China [4] Key Lab of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China [5] University of Chinese Academy of Sciences, Nanjing 211135, China [7] School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China

Received 2025 March 12; revised 2025 April 8; accepted 2025 April 14; published 2025 May 9

Abstract

The study of carbon-enhanced metal-poor (CEMP) stars is of great significance for understanding the chemical evolution of the early universe and stellar formation. CEMP stars are characterized by carbon overabundance and are classified into several subclasses based on the abundance patterns of neutron-capture elements, including CEMP-s, CEMP-no, CEMP-r, and CEMP-r/s. These subclasses provide important insights into the formation of the first stars, early stellar nucleosynthesis, and supernova explosions. However, one of the major challenges in CEMP star research is the relatively small sample size of identified stars, which limits statistical analyses and hinders a comprehensive understanding of their properties. Fortunately, a series of large-scale spectroscopic survey projects have been launched and developed in recent years, providing unprecedented opportunities and technical challenges for the search and study of CEMP stars. To this end, this paper draws on the progress and future prospects of existing methods in constructing large CEMP data sets and offers an in-depth discussion from a technical standpoint, focusing on the strengths and limitations. In addition, we review recent advancements in the identification of CEMP stars, emphasizing the growing role of machine learning in processing and analyzing the increasingly large data sets generated by modern astronomical surveys. Compared to traditional spectral analysis methods, machine learning offers greater

efficiency in handling complex data, automatic extraction of stellar parameters, and improved prediction accuracy. Despite these advancements, the research faces persistent challenges, including the scarcity of labeled samples, limitations imposed by low-resolution spectra, and the lack of interpretability in machine learning models. To address these issues, the paper proposes potential solutions and future research directions aimed at advancing the study of CEMP stars and enhancing our understanding of their role in the chemical evolution of the universe.

Key words: stars: carbon – stars: abundances – methods: data analysis – methods: observational – surveys

1. Introduction

Metal-poor stars ([Fe/H] < -1.0) are stars with low metal abundance in their chemical composition (Beers & Christlieb 2005). They formed in the early universe when heavy elements had not yet been widely enriched, serving as chemical relics of the early universe. Their elemental abundance patterns record the characteristics of early stellar nucleosynthesis and supernova explosions, providing valuable clues for studying the chemical evolution and dynamic history of the Milky Way (Frebel & Norris 2015). In their study of very metal-poor (VMP) stars, Beers and Christlieb (2005) discovered that approximately 20% of stars with [Fe/H] ≤ -2.0 exhibit carbon overabundances. These stars are defined as carbon-enhanced metal-poor (CEMP) stars, characterized by low metallicity and relatively high carbon abundance (Beers & Christlieb 2005; Aoki et al. 2007). Subsequent studies have found that as metallicity decreases, the fraction of CEMP stars increases: approximately 20% in VMP ([Fe/H] ≤ -2.0) stars, around 40% in extremely metal-poor (EMP, [Fe/H] ≤ -3.0) stars, and about 80% in ultra metal-poor (UMP, [Fe/H] ≤ -4.0) stars (Lucatello et al. 2005; Placco et al. 2014; Banerjee et al. 2018; Yoon et al. 2018). Because CEMP stars play an important role in the formation and evolution of the early Milky Way, studying their abundance patterns and origins is crucial for understanding the chemical evolution of the early universe (Bonifacio et al. 2012).

Based on the overabundance characteristics of neutron-capture elements, CEMP stars can be divided into different types, such as CEMP-no (no significant neutron-capture element enrichment), CEMP-r (r-process enriched), CEMP-s (s-process enriched), and CEMP-r/s (enriched in both r-process and s-process elements) (Beers & Christlieb 2005). This classification reflects the different nucleosynthesis processes and environmental influences stars undergo during their evolution. It helps us understand the evolutionary history of the stars themselves and provides crucial information about the chemical history of the early universe (Bonifacio et al. 2012; Norris et al. 2012b; Hansen et al. 2016). For example, Shejeelammal & Goswami (2023) conducted a study on two CEMP-no stars discovered in the Hamburg/ESO Survey and performed chemical and kinematic analyses using high-resolution spectra. This work explored the origin of these two stars by determining fundamental stellar parameters (including ra-

dial velocity and atmospheric parameters such as effective temperature, surface gravity, and metallicity), combined with analysis of the abundances of various elements. These investigations provide important insights into the formation mechanisms and evolutionary history of such stars.

In recent years, with the development of large-scale survey projects such as the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST, also called the Guo Shou Jing Telescope) (Cui et al. 2012; Luo et al. 2015), astronomers have obtained vast amounts of photometric and spectral data. On this basis, significant progress has been made in the search for CEMP stars using both photometric and spectroscopic data.

In the search for CEMP stars based on photometric data, the application of filter systems has played a key role, significantly improving the ability to identify CEMP stars and estimate their carbon abundance ([C/Fe]). For example, Whitten et al. (2021) used the mixed-bandwidth (combining narrow-band and wide-band) photometric technique provided by S-PLUS Data Release 2 to estimate stellar parameters, including Teff, [Fe/H], [C/Fe], and absolute carbon abundances (A(C)), identifying 364 CEMP candidate stars. In addition, Perottoni Hélio et al. (2024) utilized the S-PLUS Ultra-Short Survey (USS), which collects multi-band photometric data through narrow, medium, and broad-band filters, targeting prominent stellar spectral features with the aim of identifying bright EMP and UMP stars. Huang et al. (2024) used J-PLUS DR3 and Gaia EDR3 data to determine the parameters and chemical abundances of over 5 million stars, providing an important photometric foundation for exploring the structure and evolution of the Milky Way. Although Perottoni Hélio et al. (2024) and Huang et al. (2024) did not directly involve the search for CEMP stars, they fully demonstrate the broad application potential of photometric stellar parameters and chemical abundance data in the study of stellar populations. Alternatively, an approach is to initially select metal-poor star candidates using photometric data, followed by subsequent spectral follow-up observations. Jacobson et al. (2015) utilized the unique photometric filter system of the SkyMapper Southern Sky Survey to select metal-poor star candidates in the Galaxy by focusing on the Ca II K line at 3933 Å. They conducted detailed studies of 122 metal-poor stars using high-resolution spectroscopy. Based on the 1D Local Thermodynamic Equilibrium (1D LTE) method, abundances of various elements, including [C/Fe], were determined, and seven of these stars were ultimately identified as CEMP-no stars.

Compared to photometric data, spectroscopic data is like a star's "fingerprint," containing a wealth of diverse spectral line features. By analyzing in detail the position, shape, intensity, and relative relationships of spectral lines, we can gain deeper insights into the chemical composition and physical parameters of stars (Beers et al. 1985, 1992; Frebel et al. 2006; Christlieb et al. 2008). Traditional methods for identifying CEMP stars primarily rely on stellar spectral features, especially the intensity of carbon-related molecular bands. For instance, studies such as Aoki et al. (2007) and Norris et al. (2012a) analyzed the strength of the

CH molecular band in high-resolution stellar spectra to measure [C/Fe], thereby determining whether the star is a CEMP star. However, a notable drawback of these studies is the limited number of stars involved, typically ranging from a few hundred to a few thousand. This limitation arises because traditional identification methods require specialized knowledge and significant manual effort, consequently limiting the utilization of stellar spectra and analytical efficiency.

In contrast, with the advancement of machine learning technologies, researchers have increasingly adopted machine learning methods to process large volumes of spectral data. These methods allow models to automatically extract spectral features without the need for manually defining molecular band characteristics or measuring band intensities, thereby improving spectral utilization and analytical efficiency. For example, Li et al. (2018b) applied machine learning techniques to identify 2651 carbon stars from over 7 million stellar spectra in LAMOST DR4 and reported 17 CEMP candidates. In the search process, adopting a multi-stage strategy can greatly improve search efficiency (Li & Lin 2023; Song et al. 2024; Zhang et al. 2024b), especially given the scarcity of CEMP star samples. High-resolution spectroscopic data, such as those obtained from surveys like APOGEE (Majewski et al. 2017) and the Gaia-ESO Survey (Gilmore et al. 2012), provide detailed insights into stellar properties. However, these data suffer from limited spatial and spectral coverage—meaning that only specific regions of the sky are targeted and a restricted range of wavelengths is observed due to instrumental and operational constraints. In addition, the overall quantity of high-resolution spectra remains relatively small compared to lower-resolution surveys. Consequently, searches based on such data may struggle to achieve efficient large-scale screening for specific targets. Therefore, an effective strategy is to first perform preliminary screening on medium- and low-resolution spectral data. Compared to high-resolution spectral data, medium- and low-resolution spectral data encompass a larger sample of celestial objects, enabling the identification of more potential CEMP candidate stars. After the initial screening, high-resolution spectral data can be used for further analysis, providing more detailed information that helps confirm and distinguish CEMP stars from other types of stars.

2. Carbon-enhanced Metal-poor Stars

Metal-poor stars exhibit significant differences in metallicities and chemical signatures. For instance, while the fraction of CEMP stars increases at lower metallicities, most metal-poor stars with [Fe/H] > -3.0 remain carbon-normal. Beers & Christlieb (2005) defined stars with [C/Fe] $> +1.0$ among metal-poor stars as CEMP stars. Subsequently, Aoki et al. (2007) adjusted this classification criterion to [C/Fe] $\geq +0.7$ (see Equation (1)), a change that reflects the influence of stellar evolution, such as internal mixing and dredge-up processes, on the surface carbon abundances of stars.

where Figure 1 [Figure 1: see original paper]. Elements abundance distribution on the Sun (Lugaro et al. 2023). Se (selenium), Te (tellurium), Xe (xenon), and

Pt (platinum) are characteristic elements formed through the r-process, while Sr (strontium), Y (yttrium), Zr (zirconium), and La (lanthanum) are typical s-process elements. Le, Re, Me, ge, and Teffe represent the luminosity, radius, mass, surface gravity, and effective temperature of the Sun, respectively.

2.1. Classification of CEMP Stars

The neutron-capture process is a nucleosynthesis process in which atomic nuclei inside stars capture neutrons to produce heavier elements. If the timescale of a nucleosynthesis reaction is longer than the timescale of β decay, it is defined as the slow neutron-capture process (s-process), whereas if the timescale is shorter than the β decay timescale, it is defined as the rapid neutron-capture process (r-process) (Wan-qiang et al. 2024). Initially, Beers & Christlieb (2005) classified CEMP stars into four subclasses based on the abundances of two neutron-capture elements—barium (Ba, representing s-process elements) and europium (Eu, representing r-process elements). Subsequently, many studies (Frebel 2018; Hansen et al. 2019; Goswami et al. 2021) have identified many other elements that can serve as representative elements of the s-process and r-process (Figure 1), and proposed different criteria for identifying and classifying CEMP stars.

Based on the relative enrichment of neutron-capture elements, four main subclasses are defined as follows: CEMP-s, CEMP-r, CEMP-r/s, and CEMP-no (see Table 1). These diverse abundance patterns reflect significant differences in the formation processes of various subclasses, highlighting the complexity of chemical evolution in the early universe and providing an important window into understanding chemical enrichment and star formation processes in ancient stellar populations.

The carbon enhancement phenomenon observed in CEMP stars likely originates from one of the following sources: (1) a primordial mechanism from massive stellar progenitors, (2) intrinsic internal production by low-mass stars of extremely low [Fe/H], or (3) extrinsic production of carbon by stars of intermediate mass, which can be prodigious manufacturers of carbon during their Asymptotic Giant Branch (AGB) stages, followed by mass transfer to a surviving lower-mass companion. It remains possible that all three sources have played a role (Beers & Christlieb 2005). The first source suggests that the carbon in some CEMP stars is primordial or close to primordial, possibly produced by the first generation of stars in the early universe. These stars undergo nucleosynthesis in a zero-metallicity environment and release large amounts of carbon during their supernova explosions, which then become the carbon source for subsequent CEMP stars (Woosley & Weaver 1995). The second source proposes that, in the early universe, when heavy elements were scarce, low-mass stars experienced unusually effective mixing processes during the helium core flash phase. This process dredges up the internally produced carbon and deposits it on the surface of the star (Fujimoto et al. 1999; Schlattl et al. 2002; Picardi et al. 2004; Weiss et al. 2004). The third source suggests that intermediate-mass stars in the early Galaxy produce large amounts of carbon during their AGB evolution. If these

stars are part of binary systems, carbon-rich material could be transferred to the lower-mass companion via Roche-lobe overflow or stellar winds, causing the companion star to become a CEMP star (Lucatello et al. 2005).

Goswami et al. (2024) pointed out that most CEMP-s stars are binaries (Lucatello et al. 2005; Starkenburg et al. 2014; Hansen et al. 2016; Jorissen et al. 2016). By comparing the observed abundances of these stars with theoretical model predictions, it is possible to confirm binary mass transfer from the AGB companion (Bisterzo et al. 2011; Placco et al. 2013; Hollek et al. 2015). Cowan & Rose (1976) proposed the intermediate neutron-capture process (i-process), in which the neutron density is intermediate between the s-process and the r-process. This process can simultaneously produce both s-process and r-process elements within a single star. Therefore, many studies have used the yield of the i-process model for low-mass, low-metallicity AGB stars to explain the observed abundance patterns in CEMP-r/s stars (Hampel et al. 2016; Shejeelammal & Goswami 2021, 2022).

However, CEMP-r stars are extremely rare among all the CEMP subclasses. Hansen et al. (2011, 2015) found that the abundance anomalies of CEMP-r stars are not caused by binary mass transfer but are instead due to the enrichment of their birth clouds by r-process elements from external sources. These external sources may include core-collapse supernovae (Argast et al. 2004; Arcones & Thielemann 2012), fallback supernovae (Fryer et al. 2006), neutron-star mergers (Drout et al. 2017; Lippuner et al. 2017), or neutron star-black hole mergers (Surman et al. 2008).

As for CEMP-no stars, they are believed to be the most chemically primitive stars known, directly reflecting the chemical composition of the early universe (Norris & Yong 2019; Yoon et al. 2020). The exact origin of CEMP-no stars remains unclear, but several hypotheses have been proposed to explain the source of their carbon overabundance, including rotating massive stars (Maeder et al. 2015), faint supernovae (Umeda & Nomoto 2005), and inhomogeneous metal mixing (Hartwig & Yoshida 2019). These hypotheses provide valuable clues for studying the chemical evolution of the early universe and the formation of the first generation of stars.

2.2. Spectral Characteristics of CEMP Stars

Spectroscopy is an important tool in astronomy and stellar physics for studying the properties of celestial bodies. By analyzing the absorption lines of different elements and molecules in a spectrum, we can directly reveal the abundance characteristics of elements in stars. CEMP stars are a special class of metal-poor stars, characterized by low metal abundances and high carbon abundances. As a result, the spectra of these objects exhibit distinctive features. For example, two early survey projects—the HK survey (Beers et al. 1985, 1992) and the Hamburg/ESO Survey (Frebel et al. 2006; Christlieb et al. 2008)—used the strength of the Ca II H & K lines (CaHK lines) to identify metal-poor candidate stars.

The basic idea of using the CaHK lines is to check their relative strength, as these two lines are significantly stronger than Fe lines and can still be detected in stars with the lowest metallicity, even when Fe lines are completely absent. By measuring the equivalent width of this spectral line and comparing it with theoretical spectra or template spectra with known metallicity, potential metal-poor stars can be selected, which in turn helps in identifying CEMP stars. Christlieb et al. (2001) published a list of HES stars with strong carbon molecular lines (such as CN, C_2, and CH), based on the total strength of these lines, measured using line indices defined as the ratios of the mean photographic densities in the carbon molecular absorption features to those in the continuum bandpasses. However, both of these search methods have limitations: the CaHK lines-based method misses many CEMP stars with [Fe/H] > -2.5, while the method based on strong carbon molecular lines tends to identify cool stars (Placco et al. 2010). Therefore, these methods are likely to miss a large number of CEMP stars.

To address the limitations of the above methods, Placco et al. (2010) designed a line index GPE (GPHES Extended) for the CH G-band. This index has a wide wavelength coverage, effectively avoiding the decrease in accuracy caused by the influence of carbon features on the band edges. As a result, it can precisely identify CEMP stars that may have been overlooked in previous screening processes due to their unique temperature and carbon characteristics. The CEMP star search technique based on the CH G-band is an important direction for exploration. In addition to the GPE index, the G-band index (GP; Beers et al. 1999) and the GPHES (GP index, measured in HES spectrum) index (Christlieb et al. 2008) existed earlier. However, the wavelength coverage of these two indices is relatively narrow, which prevents them from fully capturing all the details of the CH G-band.

Subsequently, Placco et al. (2011) optimized the GPE and proposed a line index EGP. With its unique calculation method, the EGP index can avoid interference from strong Hγ lines and CN bands on the blue side of the line in cool stars, greatly enhancing its resistance to contamination. Using this method, they yielded a list of 5288 new CEMP candidates. Li et al. (2018a) used a combination of the line index G1 and EGP to select 636 CEMP candidate stars from the LAMOST DR3 VMP (Very Metal-Poor) sample set, within the effective temperature range of 4000 K $<$ Teff $<$ 7000 K.

In addition, the Swan bands, as characteristic absorption features of the C_2 molecule, are one of the spectral signatures of stars with excessively high carbon abundance. They can also be used to estimate carbon abundance and identify CEMP stars. Mikolaitis et al. (2011) used the Swan bands to derive the carbon abundances of two first-ascent giants and two core-helium-burning "clump" stars in the NGC 2506 cluster. Čotar et al. (2018) used the Swan band features and employed both supervised and unsupervised classification algorithms to search for carbon-enhanced stars in the GALAH data set, thereby identifying CEMP candidate stars.

Figure 2 [Figure 2: see original paper]. A comparison of the spectral features

between CEMP stars and Carbon-Normal Metal-Poor (CnMP) stars in the 4000–4800 Å range. The solid lines at the top represent the normalized spectra of CEMP stars, the dashed lines in the middle represent the spectra of CnMP stars, and the dashed–dotted lines below show the difference between the spectra of some CEMP stars and their corresponding CnMP stars. The shaded region in the figure represents the wavelength range involved in the EGP (see Equation (2) for details).

3. Current Status of CEMP Star Searches

In-depth study of the properties and origin of the oldest CEMP stars is crucial for revealing the characteristics of the first generation of stars and the chemical evolution of the early universe. Currently, CEMP star searches based on spectral data can be roughly divided into three categories: line index methods, template matching methods, and machine learning methods.

3.1. Line Index Methods

The line index method is a commonly used spectral analysis technique that can be employed to quantify and extract the characteristics of specific spectral lines. This method quantifies the intensity, width, shape, or other related characteristics of spectral lines by integrating, taking the weighted average, or calculating ratios over specific wavelength ranges, and defining an "index" to reflect the intensity or characteristics of the line. For example, one might calculate the full width at half maximum (FWHM) or equivalent width (EW) for a specific wavelength range. In the search for CEMP stars, since these stars typically exhibit low metallicity and high carbon abundance, the line index method is used to quantify spectral line features related to elements such as carbon. In particular, certain spectral lines (such as the CH G-band, CN band, C_2 band, etc.) often exhibit abnormal intensities in CEMP stars. These features can effectively help distinguish CEMP stars from normal stars or other types of celestial objects.

For example, the GP index and GPHES index mentioned in Section 2.2 are line indices designed based on the CH G-band. These line indices, by quantifying the intensity of the CH G-band, assist in the identification and classification of CEMP stars, playing a crucial role especially in the study of their relationships with other chemical compositions such as metallicity and carbon abundance. The advantage of the line index method lies in its ability to extract spectral information even at low resolution by measuring the ratio of mean spectral densities within specific wavelength bands, making it commonly used in low-resolution or slitless spectroscopy. However, this method may be limited by wavelength selection, signal-to-noise ratio, and its dependence on empirical calibrations or theoretical templates.

3.2. Template Matching Methods

The template matching method is a widely used spectral analysis technique in astronomy, where theoretical or empirically generated standard spectra (i.e., templates) are compared with observed spectra to derive the physical parameters of celestial objects, such as effective temperature, surface gravity, and metallicity. For example, Li et al. (2018a) aimed to identify extremely metal-poor stars and CEMP stars in the LAMOST DR3 data set by using SPECTRUM (Gray & Corbally 1994) to generate a set of synthetic spectra, and employed two methods to derive the physical parameters of the stars. One method involves calculating the line indices of the synthetic and observed spectra, and then matching these indices with a grid of synthetic spectra to derive the physical parameters of the star. Another method is to compare the observed spectra with the synthetic spectra in the wavelength range from 4360 Å to 5500 Å in order to derive the physical parameters of the star.

Hansen et al. (2016) used the spectral synthesis code MOOG (Sneden 1973) to derive the [C/Fe] abundance ratio of stars by fitting synthetic spectra of the CH G-band and Swan band to the observed data. They employed the ^{2} minimization method to match the synthetic spectra with the observed spectra in the selected wavelength range. Through this process, they successfully derived the carbon abundances of 27 metal-poor stars and further classified these stars into CEMP-no stars and CEMP-s stars in subsequent analysis. Molaro et al. (2023) aimed to investigate whether four known metal-poor stars belong to the CEMP-no class. They used a synthetic spectrum grid generated with the SYNTHE code (Kurucz 2005) and the ATLAS model, and combined it with the Markov chain Monte Carlo method to derive the $^{12}C/^{13}C$ isotope ratios of these objects.

To improve the accuracy of parameter determination, many studies that use template matching have also incorporated other star parameter measurement techniques. For example, the star parameter measurement program SSPP from the Sloan Extension for Galactic Understanding and Exploration (SEGUE) of the SDSS survey (Lee et al. 2008) combines various measurement methods, including the ^{2} minimization spectral fitting method based on theoretical spectral libraries and multiple line index methods. The template matching method compares the observed spectrum with a set of known template spectra to find the template that best matches the observed spectrum. As a result, the template matching method relies on the quality of the templates, and the choice and quality of the templates determine the effectiveness of the matching. In addition, the effectiveness of template matching is also constrained by the signal-to-noise ratio and resolution of the observed spectrum. With low-quality data, significant measurement errors are more likely to be introduced.

3.3. Machine Learning Investigations for CEMP Candidate Searching

With the rapid increase in astronomical observation data, machine learning (ML) techniques have been applied in astronomy. Due to its ability to handle

large-scale, high-dimensional data and uncover underlying patterns, machine learning has become an important tool for spectral data analysis. Depending on the task, machine learning methods in the search for CEMP star candidates can be divided into unsupervised learning, supervised classification, and regression methods.

The unsupervised learning method is used to automatically discover the structure, patterns, or regularities in unlabeled data by analyzing the intrinsic structure of the data to automatically cluster it into different categories or uncover hidden patterns within it. Common methods include dimensionality reduction and clustering. Typically, astrophysical features are interrelated, so we can create a mapping between the sparse high-dimensional space and a lower-dimensional space that captures most of the information in the data. Čotar et al. (2018) used the t-SNE method to visualize spectral data containing CEMP stars and other types of stars. By projecting these high-dimensional data onto a two-dimensional plane, the researchers were able to more intuitively identify the distribution patterns of different types of stars. Clustering methods automatically assign stars to different groups or categories, thereby further revealing the similarities and differences between stars. Zhang et al. (2024) and Shank et al. (2022) performed clustering analysis on several metal-poor stars using the Friends-of-Friends (FoF) algorithm and the HDBSCAN algorithm, respectively. By comparing the clustering results with the distribution of known CEMP stars, they revealed the distribution characteristics of CEMP stars in different Galactic substructures. The contribution of these studies lies in identifying potential aggregation regions for CEMP stars, further uncovering possible associations between CEMP stars and specific substructures.

The advantage of the unsupervised learning method lies in its ability to uncover potential patterns or structures within unlabeled data, without relying on manually labeled data sets. This is particularly valuable in astronomy, where data volumes are immense and labeling costs are high. Notably, unsupervised learning methods demonstrate unique flexibility when exploring unknown sample categories, making them an excellent foundation for subsequent supervised learning tasks. However, these methods have limitations, including their sensitivity to the quality of input features and parameter settings. Poor feature selection or improper parameter configurations can result in outcomes that deviate from actual physical meanings. Additionally, unsupervised methods cannot provide explicit classification labels, meaning the patterns or structures they identify require further validation through other approaches, which adds to the complexity and uncertainty of the analysis.

The supervised classification method is a machine learning technique that learns from labeled data. It aims to classify new data by learning the relationship between the features and labels of existing samples. In supervised classification, the model is “trained” to predict the label of each data point. The principle is to construct a function or model that can map input features to corresponding class labels. For example, Lucey et al. (2023) trained an XGBoost classifica-

tion model using the stellar spectrum catalog provided by SDSS/SEGUE, and after training, applied the model to the BP/RP spectral data from Gaia DR3 to identify CEMP stars. In addition, the study also validated the feasibility of using machine learning methods to process spectral data with a resolution as low as R = 50. The advantage of the supervised classification method lies in its ability to leverage labeled data for learning the mapping between input features and class labels, thereby achieving accurate classification of new samples. This approach is particularly well-suited for tasks with abundant labeled data sets. Compared to unsupervised learning, supervised classification provides results with explicit class labels, making them easier to interpret and validate. However, its limitations include a reliance on large labeled data sets. For rare targets such as CEMP stars, the lack of labeled data can limit the model's ability to recognize small-sample classes. More importantly, supervised classification methods typically output only class information and are unable to provide detailed parameters of samples (e.g., stellar atmospheric parameters and elemental abundances), which hinders further research and scientific interpretation.

The regression method transforms the search for CEMP stars into a parameter estimation problem, where the goal is to analyze the stellar spectrum to estimate atmospheric parameters and elemental abundances. Based on these estimated parameters, CEMP stars can then be identified. In this process, the regression model learns the relationship between the spectral data of known stars and their corresponding parameters in order to predict the parameters of new stars. For example, Ardern-Arentsen et al. (2025) used artificial neural networks (ANNs) to predict metallicity and carbon content from low-resolution stellar spectra. The work first constructed a training set based on LAMOST spectra and high-resolution samples. Then, the trained network was applied to the identification of CEMP candidate stars and giant star samples. Finally, around 2000 high-confidence CEMP stars ([Fe/H] < -2.0 and [C/Fe] $> +0.7$) were successfully identified. To quantitatively assess the impact of UV band spectra on EMP ([Fe/H] < -3.0) and Carbon-Enhanced EMP (CE-EMP, [Fe/H] < -3.0 and [C/Fe] $> +1.0$) stars, and to provide theoretical basis and methodological support for the effective identification of EMP and CE-EMP stars in future China Space Station Telescope (CSST) observations, Zhang et al. (2024b) developed a dual-branch model based on spectral transformer (SPT) (Zhang et al. 2024a) to predict the stars' [Fe/H] and [C/Fe], while also exploring the stellar classification problem. Xie et al. (2021) used the S-shaped folding technique to convert the observed spectra into a 64×64 matrix form, and employed a convolutional neural network (CNN) to estimate the stellar spectral parameters [Fe/H] and [C/Fe]. The S-shaped folding technique converts one-dimensional data into a two-dimensional matrix by arranging the data in a serpentine pattern—alternating between left-to-right and right-to-left across rows. This allows the data to be processed by a CNN as if it were an image. In the subsequent analysis, the model successfully identified 260 out of 414 known CEMP stars based on the criteria of [Fe/H] < -2 and [C/Fe] > -1, achieving a recall rate of 62.80%.

Regression methods play a significant role in the study of identifying CEMP star candidates, with their primary advantage being the ability to directly predict stellar atmospheric parameters and elemental abundances. This provides fine-grained information crucial for subsequent scientific research. Such methods are particularly suited for scenarios requiring parameterized analysis of stellar spectra, enabling the rapid identification of potential CEMP star candidates. The use of ANNs or CNNs to process low-resolution spectral data for parameter prediction has proven effective in numerous studies. However, regression methods also have certain limitations. First, these models require large, high-quality training data sets that include both spectral information and precise parameter annotations. The availability of high-resolution, well-annotated samples is often limited, which may constrain model performance. Second, regression methods are sensitive to the quality and distribution of input data. Low signal-to-noise ratios or noisy spectra can result in prediction biases. Additionally, the complexity of regression models increases the risk of overfitting, particularly when training data is insufficient or parameters are not appropriately configured. Finally, regression methods tend to have lower interpretability, especially for deep learning models, where it is often unclear how specific spectral features influence the model's outputs. This lack of transparency can limit their broader application in scientific research.

In conclusion, the application of machine learning methods in stellar spectral data analysis has shown great potential. From unsupervised learning methods to supervised classification methods and regression methods, each approach has provided new ideas and methods for the identification and analysis of CEMP stars. In particular, regression methods, through neural networks and other deep learning models, can accurately estimate stellar parameters such as metallicity and carbon abundance, helping us effectively distinguish between different types of stars.

4. Challenges in Searching for CEMP Stars

The line index method and template matching method are commonly applied to low-resolution or slitless spectroscopy, where individual spectral lines are difficult to resolve, making direct metallicity and carbon abundance determinations challenging. In high-resolution spectroscopy, individual spectral lines can be directly measured, reducing the need for such methods. However, low-resolution spectra often suffer from broadened or even indistinguishable spectral features, which can impact the accuracy of parameter extraction. The line index method, in particular, is highly dependent on the definition of spectral bandpasses, making it sensitive to noise and spectral resolution. Furthermore, in large-scale low-resolution spectral surveys, significant variations in signal-to-noise ratio (SNR) introduce additional uncertainties. The template matching method, widely used in such surveys, is especially vulnerable to noise-induced mismatches between the observed spectra and the template library, ultimately affecting parameter estimation reliability. For instance, the official stellar spectral processing pipeline

of LAMOST, LASP, employs template matching. However, in the LAMOST DR9 release, out of 11.22 million low-resolution spectra, stellar parameters were published for only 6.18 million spectra.

Unsupervised learning and supervised classification methods, while effective for categorization, do not directly yield detailed stellar parameters. These methods are thus more suitable for preliminary selection of potential CEMP candidates, serving as a reference for subsequent precise measurements. Although these methods perform well in classification tasks, they have inherent limitations in deriving comprehensive stellar physical properties. Moreover, the scientific potential of low-resolution spectral data remains underexplored. In contrast, regression models can predict key parameters, facilitating precise identification of CEMP stars. When applying regression methods to CEMP star searches, several challenges must be considered:

1) **Sparsity of CEMP Star Samples**. Although CEMP stars make up a significant proportion of metal-poor stars, metal-poor stars are relatively rare in the Milky Way, accounting for less than 0.1% of the total stellar population in the Galaxy (Zhang et al. 2024b). The sparsity of such samples can lead to poor performance of the model in identifying CEMP stars during the training process.

2) **Limitations of Low-Resolution Spectra**. Low-resolution spectra may obscure the subtle differences between CEMP stars and other metal-poor stars, making feature extraction more challenging (Figure 3(a)). Compared to low-resolution spectra, high-resolution spectra can clearly resolve individual spectral lines, which are associated with accurate atomic physics data (such as energy levels, transition probabilities, etc.), allowing for more precise extraction of stellar atmospheric properties. At low resolution, each "feature" is actually a mixture of multiple spectral lines, making it difficult to accurately distinguish between meaningful information and noise in the model's spectrum. As a result, the robust information content in low-resolution spectra may be affected by the imperfections of the model (Ting et al. 2017).

3) **Noise Issues**. In real spectral data, noise is inevitable, especially for spectra with low SNR. A decrease in SNR means that useful information in the signal is drowned out by background noise (Figure 3(b)), significantly reducing the detectability and accuracy of spectral lines. This makes it difficult for the model to learn accurate patterns during training and increases the likelihood of overfitting (Wu et al. 2020b).

4) **Model Interpretability Issues**. Machine learning methods have demonstrated strong potential in the search for CEMP stars. However, these models are often "black-box" models, lacking interpretability. For small-sample targets like CEMP stars, astronomers want to understand which spectral features the model is using to make predictions and which features are most important for the model's decision-making. However, the

interpretability of learning models has not yet been well addressed.

Looking ahead, addressing the challenges in CEMP star searches will require multifaceted efforts. First, to address the issue of sample sparsity, researchers can increase the training sample size by using data augmentation and synthetic data generation techniques. For example, synthetic spectra can be generated using existing CEMP star data, and recent CEMP star samples can be compiled and aggregated to expand the sample size. A feasible solution to the limitation of low-resolution spectra is to use the attention mechanism. The attention mechanism is a computational method in deep learning that mimics human cognitive attention, allowing models to dynamically focus on the most relevant parts of input data (e.g., specific spectral features in stellar spectra) while suppressing less important information. In recent years, the attention mechanism has been widely applied in deep learning tasks due to its outstanding performance and flexible applicability. For example, models like SE-Net (Hu et al. 2018), ECA-Net (Wang et al. 2020), and CBAM (Woo et al. 2018) have achieved automatic focus and enhancement of important features in the input data through different methods. In low-resolution spectral analysis, the attention mechanism can help the model effectively focus on key wavelength regions within the limited resolution information, enhancing the ability to recognize important features and reducing the impact of feature blurring caused by low resolution. In addition, the emergence of Spectra-GANs (Wu et al. 2020a) has provided a new breakthrough for denoising using deep learning models.

To address the issue of model interpretability, interpretable machine learning techniques such as LIME (Ribeiro et al. 2016) and SHAP (Lundberg & Lee 2017) can be introduced. These methods help us understand how the model makes decisions and identify which spectral features are most crucial for the predictions. Using attention mechanisms in deep learning, such as Grad-CAM (Selvaraju et al. 2019), can visualize the spectral regions that the model focuses on when making predictions. This enhances the model's transparency and interpretability, assisting astronomers in understanding the features of CEMP stars. These improvements not only help enhance the detection accuracy of CEMP stars but also expand the application of machine learning methods in other astronomical data analyses, driving the entire field of astronomy toward greater automation and efficiency.

Figure 3 [Figure 3: see original paper]. Spectrum comparisons. (a) Comparison of spectra of the same star at different resolutions; the spectrum with R = 1800 is provided by LAMOST DR5, while the spectrum with R = 200 is obtained by degrading the resolution using iSpec (Blanco-Cuaresma et al. 2014; Blanco-Cuaresma 2019). (b) Comparison of high and low signal-to-noise ratio (SNR) spectra of the same star. Both spectra are from LAMOST DR5, with a wavelength range selected from 4000 to 8096 Å. The spectral data have been normalized; the target star's celestial coordinates are R.A. 57.°033750 and decl. 49.°865000. The high SNR spectrum has an SNRg value of 101.72, while the low SNR spectrum has an SNRg value of 3.67.

5. Conclusion

In summary, this paper summarizes and analyzes the research progress on CEMP stars. We review the importance of CEMP stars in the study of the origin and evolution of the Milky Way, as well as their classification and unique spectral features, emphasizing their distinctive role in understanding the chemical evolution of the early universe and stellar formation. We highlight the search techniques for CEMP stars, with a particular focus on the significant development of machine learning methods in CEMP star identification, driven by the rapid increase in astronomical data. Despite the significant advantages of modern machine learning methods in detecting CEMP stars, challenges remain, including sample scarcity, limitations of low-resolution spectra, noise interference, and insufficient model interpretability. This paper proposes potential solutions to these issues, including increasing training samples through data augmentation, incorporating advanced denoising techniques, enhancing model performance by integrating attention mechanisms, and utilizing explainable machine learning techniques to improve model transparency, helping astronomers better understand the characteristics of CEMP stars.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (grant No. 12373108).

References

Aoki, W., Beers, T. C., Christlieb, N., et al. 2007, ApJ, 655, 492 Arcones, A., & Thielemann, F.-K. 2012, JPhG, 40, 013201 Ardern-Arentsen, A., Kane, S. G., Belokurov, V., et al. 2025, MNRAS, 537, 1984 Argast, D., Samland, M., Thielemann, F. K., & Qian, Y.-Z. 2004, A&A, 416, 997 Banerjee, P., Qian, Y.-Z., & Heger, A. 2018, ApJ, 865, 120 Beers, T. C., & Christlieb, N. 2005, ARA&A, 43, 531 Beers, T. C., Preston, G. W., & Shectman, S. A. 1985, AJ, 90, 2089 Beers, T. C., Preston, G. W., & Shectman, S. A. 1992, AJ, 103, 1987 Beers, T. C., Rossi, S., Norris, J. E., Ryan, S. G., & Shefler, T. 1999, AJ, 117, 981 Bisterzo, S., Gallino, R., Straniero, O., Cristallo, S., & Käppeler, F. 2011, MNRAS, 418, 284 Blanco-Cuaresma, S., Soubiran, C., Heiter, U., & Jofré, P. 2014, A&A, 569, A111 Blanco-Cuaresma, S. 2019, MNRAS, 486, 2075 Bonifacio, P., Sbordone, L., Caffau, E., et al. 2012, A&A, 542, A87 Christlieb, N., Green, P. J., Wisotzki, L., & Reimers, D. 2001, A&A, 375, 366 Christlieb, N., Schörck, T., Frebel, A., et al. 2008, A&A, 484, 721 Čotar, K., Zwitter, T., Kos, J., et al. 2018, MNRAS, 483, 3196 Cowan, J., & Rose, W. 1976, BAAS, 8, 320 Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, RAA, 12, 1197 Drout, M. R., Piro, A. L., Shappee, B. J., et al. 2017, Sci, 358, 1570 Frebel, A. 2018, ARNPS, 68, 237 Frebel, A., Christlieb, N., Norris, J. E., et al. 2006, ApJ, 652, 1585 Frebel, A., & Norris, J. E. 2015, ARA&A, 53, 631 Fryer, C. L., Herwig, F., Hungerford, A., & Timmes, F. X. 2006, ApJL, 646, L131 Fujimoto, M. Y.,

Ikeda, Y., & Iben, I. 1999, ApJL, 529, L25 Gilmore, G., Randich, S., Asplund, M., et al. 2012, Msngr, 147, 25 Goswami, A., Shejeelammal, J., Goswami, P. P., & Purandardas, M. 2024, BSRSL, 93, 406 Goswami, P. P., Rathour, R. S., & Goswami, A. 2021, A&A, 649, A49 Gray, R. O., & Corbally, C. J. 1994, AJ, 107, 742 Hampel, M., Stancliffe, R. J., Lugaro, M., & Meyer, B. S. 2016, ApJ, 831, 171 Hansen, C. J., Hansen, T. T., Koch, A., et al. 2019, A&A, 623, A128 Hansen, C. J., Nordström, B., Hansen, T. T., et al. 2016, A&A, 588, A37 Hansen, T., Andersen, J., Nordström, B., Buchhave, L. A., & Beers, T. C. 2011, ApJL, 743, L1 Hansen, T., Andersen, J., Nordström, B., et al. 2016, A&A, 586, A160 Hansen, T., Hansen, C. J., Christlieb, N., et al. 2015, ApJ, 807, 173 Hartwig, T., & Yoshida, N. 2019, ApJL, 870, L3 Hollek, J. K., Frebel, A., Placco, V. M., et al. 2015, ApJ, 814, 121 Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. 2020, TPAMI, 42, 2011 Huang, Y., Beers, T. C., Xiao, K., et al. 2024, ApJ, 974, 192 Jacobson, H. R., Keller, S., Frebel, A., et al. 2015, ApJ, 807, 171 Jorissen, A., Van Eck, S., Van Winckel, H., et al. 2016, A&A, 586, A158 Kurucz, R. L. 2005, MSAIS, 8, 14 Lee, Y. S., Beers, T. C., Sivarani, T., et al. 2008, AJ, 136, 2022 Li, H., Tan, K., & Zhao, G. 2018a, ApJS, 238, 16 Luo, A., Zhao, Y., Zhao, G., et al. 2015, RAA, 15, 1095 Li, X., & Lin, B. 2023, MNRAS, 521, 6354 Li, Y.-B., Luo, A.-L., Du, C.-D., et al. 2018b, ApJS, 234, 31 Lippuner, J., Fernández, R., Roberts, L. F., et al. 2017, MNRAS, 472, 904 Lucatello, S., Tsangarides, S., Beers, T. C., et al. 2005, ApJ, 625, 825 Lucey, M., Al Kharusi, N., Hawkins, K., et al. 2023, MNRAS, 523, 4049 Lugaro, M., Pignatari, M., Reifarth, R., & Wiescher, M. 2023, ARNPS, 73, 315 Lundberg, S. M., & Lee, S.-I. 2017, A unified approach to interpreting model predictions, in Proc. of the 31st Int. Conf. on Neural Information Proc. Systems (Red Hook, NY: Curran Associates Inc.), 4768 Maeder, André, Meynet, G., & Chiappini, C. 2015, A&A, 576, A56 Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94 Mikolaitis, Š., Tautvaišiene, G., Gratton, R., Bragaglia, A., & Carretta, E. 2011, MNRAS, 416, 1092 Molaro, P., Aguado, D. S., Caffau, E., et al. 2023, A&A, 679, A72 Norris, J. E., Bessell, M. S., Yong, D., et al. 2012a, ApJ, 762, 25 Norris, J. E., & Yong, D. 2019, ApJ, 879, 37 Norris, J. E., Yong, D., Bessell, M. S., et al. 2012b, ApJ, 762, 28 Perottoni Hélio, D., Placco, V. M., Almeida-Fernandes, F., et al. 2024, A&A, 691, A138 Picardi, I., Chieffi, A., Limongi, M., et al. 2004, ApJ, 609, 1035 Placco, V. M., Frebel, A., Beers, T. C., & Stancliffe, R. J. 2014, ApJ, 797, 21 Placco, V. M., Frebel, A., Beers, T. C., et al. 2013, ApJ, 770, 104 Placco, V. M., Kennedy, C. R., Beers, T. C., et al. 2011, AJ, 142, 188 Placco, V. M., Kennedy, C. R., Rossi, S., et al. 2010, AJ, 139, 1051 Ribeiro, M. T., Singh, S., & Guestrin, C. 2016, in Proc. of the 22nd ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining (New York, NY: Association for Computing Machinery), 1135 Schlattl, H., Salaris, M., Cassisi, S., & Weiss, A. 2002, A&A, 395, 77 Selvaraju, R. R., Michael, C., Abhishek, D., et al. 2019, Int. J. Comput. Vis., 128, 336 Shank, D., Komater, D., Beers, T. C., Placco, V. M., & Huang, Y. 2022, ApJS, 261, 19 Shejeelammal, J., & Goswami, A. 2021, ApJ, 921, 77 Shejeelammal, J., & Goswami, A. 2022, ApJ, 934, 110 Shejeelammal, J., & Goswami, A. 2023, MNRAS, 527, 2323 Sneden, C. 1973, PhD thesis, The Univ. of Texas at Austin Song, S., Kong, X., Bu,

Y., Yi, Z., & Liu, M. 2024, ApJ, 974, 78 Starkenburg, E., Shetrone, M. D., McConnachie, A. W., & Venn, K. A. 2014, MNRAS, 441, 1217 Surman, R., McLaughlin, G. C., Ruffert, M., Janka, H.-T., & Hix, W. R. 2008, ApJL, 679, L117 Ting, Y.-S., Conroy, C., Rix, H.-W., & Cargile, P. 2017, ApJ, 843, 32 Umeda, H., & Nomoto, K. 2005, ApJ, 619, 427 Wan-qiang, H., Guo-chao, Y., Ping, N., & Bo, Z. 2024, ChA&A, 48, 73 Wang, Q., Wu, B., Zhu, P., et al. 2020, in Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (Piscataway, NJ: IEEE), 11531 Weiss, A., Schlattl, H., Salaris, M., & Cassisi, S. 2004, A&A, 422, 217 Whitten, D. D., Placco, V. M., Beers, T. C., et al. 2021, ApJ, 912, 147 Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. 2018, in Proc. of the European Conf. on Computer Vision (ECCV), ed. V. Ferrari et al. (Cham: Springer) Woosley, S. E., & Weaver, T. A. 1995, ApJS, 101, 181 Wu, M., Bu, Y., Pan, J., Yi, Z., & Kong, X. 2020a, IEEEA, 8, 107912 Wu, M., Pan, J., Yi, Z., & Wei, P. 2020b, IEEEA, 8, 66475 Xie, J., Bu, Y., Liang, J., et al. 2021, AJ, 162, 155 Yoon, J., Beers, T. C., Dietz, S., et al. 2018, ApJ, 861, 146 Yoon, J., Whitten, D. D., Beers, T. C., et al. 2020, ApJ, 894, 7 Zhang, M., Bu, Y., Wu, F., et al. 2024b, A&A, 691, A21 Zhang, M., Wu, F., Bu, Y., et al. 2024a, A&A, 683, A163 Zhang, R., Matsuno, T., Li, H., et al. 2024, ApJ, 966, 174

Note: Figure translations are in progress. See original paper for figures.