Over 2.7 Million Carbon-Enhanced Metal-Poor stars from BP/RP Spectra in Gaia DR3

Madeline Lucey1, Nariman Al Kharusi1, Keith Hawkins1, Yuan-Sen Ting2,3, Nesar Ramachandra4,5, Timothy C. Beers6,7, Young Sun Lee8, Adrian M. Price-Whelan9 and Jinmi Yoon10

1 Department of Astronomy, The University of Texas at Austin, 2515 Speedway Boulevard, Austin, TX 78712, USA
2 Research School of Astronomy & Astrophysics, Australian National University, Cotter Road, Weston Creek, ACT 2611, Canberra, Australia
3 Research School of Computer Science, Australian National University, Acton ACT 2601, Australia
4 Computational Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Lemont, IL 60439, USA
5 High Energy Physics Division, Argonne National Laboratory, 9700 South Cass Avenue, Lemont, IL 60439, USA
6 Department of Physics and Astronomy, University of Notre Dame, Notre Dame, IN. 46556, USA
7 JINA Center for the Evolution of the Elements (JINA-CEE), USA
8 Department of Astronomy and Space Science, Chungnam National University, Daejeon 34134, Republic of Korea
9 Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA
10 Space Telescope Science Institute, 3700 San Martin Dr., Baltimore, MD 21218, USA

ABSTRACT
Carbon-enhanced metal-poor (CEMP) stars make-up almost a third of stars with [Fe/H]<−2, although their origins are still poorly understood. It is highly likely that one type of CEMP star (CEMP-s stars) is tied to mass-transfer events taking place in binary stars, while another type (CEMP-no stars) has been suggested to be enriched by the nucleosynthetic yields of the first generations of stars. Historically, studies of CEMP stars have been explored in the Galactic halo, but more recently they have also been detected in the thick disk and bulge components of the Milky Way. Gaia DR3 has provided an unprecedented sample of over 200 million low-resolution (R≈50) spectra from the BP and RP photometers. In this work, we use XGBoost to classify these spectra and detect the largest all-sky sample of CEMP stars to date. In total, we find 2,736,922 CEMP stars, with a contamination rate of 5%. This sample spans from the inner to outer Milky Way with distances as close as 0.4 kpc from the Galactic center, and as far as >30 kpc. We also find that 0.28% of these stars are identified as non-single stars in Gaia DR3. By providing the largest, uniformly analyzed sample of CEMP stars we can further investigate the frequency of CEMP-s and CEMP-no stars throughout the Galaxy and constrain the origins of these classes of stars.

Key words:
difficult to compare these fractions across CEMP samples, given the various selection effects and systematic biases in the abundance analysis.

Further analysis of CEMP stars has identified a number of subclasses. A significant fraction of CEMP stars also exhibit enhancement in slow neutron-capture (s-process) elements, and are thus called CEMP-s stars (Beers & Christlieb 2005). There are also CEMP-r stars, which show enhancements in rapid neutron-capture (r-process) elements, CEMP-r/s, which exhibit enhancements in both r- and s-process elements (Gull et al. 2018), and CEMP-i stars, which show enhancement in intermediate neutron capture (i-process) elements (Frebel 2018). Lastly, there are CEMP-no stars which do not exhibit s-, r-, or i-process enhancements (Norris et al. 2013). CEMP-r/(s) and CEMP-i stars are rare, while CEMP-no and CEMP-s stars are the most common.

It is thought that CEMP-s stars, which are more common at [Fe/H] > -3.0, are the result of chemical enrichment by mass-transfer events from (post-)asymptotic giant branch (AGB) stars (Lugaro et al. 2012; Placco et al. 2013). This hypothesis is strongly supported by the high rate of bina-

ary among CEMP-s stars (McClure & Woodsworth 1990; Preston & Sneider 2001; Lucatello et al. 2005; Bisterzo et al. 2010; Abate et al. 2015; Hansen et al. 2016b; Jorissen et al. 2016). In fact, binarity rates as high as 82% have been reported for CEMP-s stars (Hansen et al. 2016b).

On the other hand, CEMP-no stars are not typically in binary systems, and therefore are not likely to have experienced a mass-transfer event (Starkenburg et al. 2014; Hansen et al. 2016a; Arentsen et al. 2019). Hence, CEMP-no stars likely formed from an ISM that was already carbon enriched. Given their low metallicity and increasing frequency at lower metallicities, it is thought that these stars are truly ancient stars, and were primarily enriched by the first generation of stars. It has been suggested that the first stars may have had high rotation rates, which would lead to large carbon production (Chiappini et al. 2006; Meynet et al. 2006). Furthermore, it is possible that the first stars explod as faint supernovae, which also overproduce carbon (Umeda & Nomoto 2003; Nomoto et al. 2013; Tominaga et al. 2014).

Initial studies indicate that the frequency of CEMP-s and CEMP-no stars varies throughout the Galaxy. Specifically, the number of CEMP stars appears to increase with increasing distance from the Sun, although we note that these studies are mostly focused on the Galactic halo (Carillo et al. 2012; Frebel et al. 2006; Lee et al. 2017; Yoon et al. 2018). Furthermore, the relative fraction of CEMP-no stars compared to CEMP-s stars also increases at larger distances (Carollo et al. 2014; Lee et al. 2019). Ultra faint dwarf galaxies have shown similar fractions of CEMP-s stars as the Milky Way, but dwarf spheroidal galaxies have a clear deficit of CEMP-no stars (Norris et al. 2010; Lai et al. 2011; Frebel et al. 2014; Skuladóttr et al. 2015; Salvadori et al. 2015). In a comparative study between Galactic halo and dwarf galaxy CEMP stars, Yoon et al. (2019) suggests that the majority of Galactic halo CEMP-no stars have been accreted from dwarf galaxies. Furthermore, CEMP-no and CEMP-s stars have been discovered in the metal-weak thick disk (MWTD; Beers et al. 2017). Dietz et al. (2021) tenta-

tively tags the retrograde MWTD CEMP-no population to the Gaia-Enceladus system, while suggesting the equivalent prograde population has both in-situ and ex-situ origins.

There are fewer studies of metal-poor stars towards the center of the Galaxy than compared to the Galactic halo. This is partly due to the difficulty of targeting metal-poor stars in a region of the Galaxy that is dense with metal-rich stars. Furthermore, high levels of extinction demand long exposure times and large-aperture telescopes in order to achieve sufficient signal-to-noise spectroscopic observations for metallicity measurements. Fortunately, the advent of metallicity-sensitive photometric surveys (e.g., Skymapper and Pristine; Bessell et al. 2011; Wolf et al. 2018; Starkenburg et al. 2017) have led to studies of thousands of metal-poor inner Galaxy stars. Studies using SkyMapper photometry have found a much lower fraction of CEMP stars in the inner Galaxy compared to the Galactic halo (Howes et al. 2014, 2015, 2016; Lucey et al. 2022). However, metallicity estimates from Skymapper photometry have proven to be biased against CEMP stars (Da Costa et al. 2019; Chiti et al. 2020). Targeting with Pristine photometry, Arentsen et al. (2021) found a CEMP frequency that is consistent with the Galactic halo for stars with [Fe/H] < -3, but also found that it is much lower than the halo at higher metallicities.

Measuring and understanding the frequency and relative rates of CEMP-no/CEMP-s stars throughout the Galaxy will be crucial for shedding new light on the origins and formation mechanisms of these stars, including whether or not CEMP-no stars are true inheritors of the elements created by the first stars. Given that the measured properties of CEMP samples have been shown to vary across different samples (Arentsen et al. 2022), creating a uniformly analyzed sample with limited selection effects across the Milky Way will be essential for achieving this goal. The release of the Gaia BP/RP spectra in DR3 presents a unique opportunity to discover the largest, all-sky sample of CEMP stars to date. However, the BP/RP spectra have quite low resolution (R = λ/Δλ ≈ 50), and require unconventional methods for analysis.

In this work, we present a novel method for detecting CEMP stars in the Gaia BP/RP spectra with machine learning, specifically the XGBoost classification algorithm, and apply it to the spectra released in DR3. In Section 2 we describe the BP/RP spectra along with other data used in our analysis. We introduce XGBoost, our chosen classification algorithm in Section 3. We evaluate the accuracy and completeness of our classification in Section 4. In Section 5, we interpret the XGBoost model. Finally, in Section 6 we present the sample along with an investigation into the binarity rate and Galactic distribution of these stars.
Figure 1. Examples of synthetic mock BP and RP spectra (top) and the impact of carbon enhancement on the spectra (bottom). Specifically, in the top panel we compare stars of the same stellar parameters and reddening, except with different carbon abundances. The BP and RP spectra are plotted separately, with the BP spectra at lower wavelengths. The BP and RP spectra overlap at ≈650 nm. The dotted lines have [C/Fe]=+0.5 while the solid lines have [C/Fe]=+1.0. Starting from the dark blue spectrum, which has typical stellar parameters for a metal-poor giant (\(T_{\text{eff}}=4500\) K, \(\log g=2.5\) and \([\text{Fe/H}]=-2.0\)), we change one stellar parameter or extinction value at a time to observe how the spectra are impacted. The green spectra have increased extinction with \(A_V=9.0\) mag. The light blue spectra are hotter, with \(T_{\text{eff}}=5500\) K. The red spectra are more metal-rich, with \([\text{Fe/H}]=0.0\). In the bottom panel, we have subtracted the dotted spectrum (\([C/Fe]=+0.5\)) from the solid spectrum (\([C/Fe]=+1.0\)) of the same stellar parameters and reddening. The colors of the lines correspond to the spectra that were subtracted. The impact of carbon on the spectra changes drastically with the stellar parameters; we therefore require a flexible classification model in order to achieve low contamination of our detected carbon-enhanced stars.
Given the wavelength coverage of 3300–10500 Å (Carrasco et al. 2021), we expect to be able to detect carbon-enhanced stars from the plethora of carbon molecular bands in that range (e.g., C2 Swan bands at \( \approx 4500–6000 \) Å, and CN bands at \( \approx 7000–10500 \) Å).

To explore the impact of carbon on the BP/RP spectra, we create mock synthetic spectra. We employ the MARCS carbon-enhanced model atmosphere grids (Gustafsson et al. 2008) and the TURBOSPECTRUM radiative transfer code (Alvarez & Plez 1998; Plez 2012) to construct these spectra. We also use the fifth version of the Gaia-ESO atomic linelist (Heiter et al. 2020) with the addition of molecular lines for CH (Masseron et al. 2014), CN, NH, OH, MgH, C2 (T. Masseron, private communication), SiH (Kurucz linelists1), and TiO, ZrO, FeH, CaH (B. Plez, private communication). To apply the instrumental profile of the BP/RP photometers, we use the DR3 calibrated passbands (Riello et al. 2021). The true spectral resolution is a function of wavelength for both BP and RP spectra, with the BP spectra ranging from \( R \approx 100 \) at the blue edge to \( R \approx 30 \) at the red edge, and the RP spectra ranging from \( R \approx 100 \) at the blue edge to \( R \approx 70 \) at the red edge. To simplify the calculation, we assume a resolution of 50 for the BP spectra and 70 for the RP spectra. We also model the impact of dust extinction using the DUST\_EXTINCTION2 package with the extinction curve from Fitzpatrick (2004), assuming \( R_V = 3.1 \).

Figure 1 shows examples of a number of synthetic mock spectra, along with the impact of carbon on the calculated flux. The top panel shows eight spectra, in sets of two with the same stellar parameters, except one spectrum has \([C/\text{Fe}]=+0.5\) (dotted line) and the other has \([C/\text{Fe}]=+1.0\) (solid line). We start with the stellar parameters of a typical metal-poor giant star \((T_{\text{eff}}=4500 \text{ K}, \log g=2.5\text{ and }[\text{Fe}/\text{H}]=2.0)\), which is shown as the dark blue spectra. We then change one parameter at a time to test how the carbon enhancement impacts spectra of different stellar parameters and reddening. The green spectra have \(A_V=9.0\) mag, the blue spectra have \(T_{\text{eff}}=5500 \text{ K}\) and the red spectra have \([\text{Fe}/\text{H}]=0.0\). This is a worst case example, given that we expect very few spectra released in Gaia DR3 to have \(A_V \geq 9.0\) mag, given that most stars have \(G<17.6\) mag. In the bottom panel, we subtract the spectrum with \([C/\text{Fe}]=+1.0\) from the spectrum with the same stellar parameters, but with \([C/\text{Fe}]=+0.5\).

The impact of carbon enhancement on the BP/RP spectra (Fig. 1) is very dependent on the stellar parameters and reddening. The metal-rich spectra (with \([\text{Fe}/\text{H}]=0.0\) have the largest difference between the carbon-enhanced and carbon-normal stars. This is likely because they have the largest \([\text{C}/\text{H}]=+1.0\) and \([\text{C}/\text{H}]=+0.5\) compared to the metal-poor spectra \(([\text{Fe}/\text{H}]=2.0)\) with \([\text{C}/\text{H}]=−1.0\) and \([\text{C}/\text{H}]=−1.5\). Generally, when \([\text{C}/\text{H}]=+1.0\) compared to the metal-poor giant \((T_{\text{eff}}=4500 \text{ K})\), the hotter star \((\text{blue} \lesssim T_{\text{eff}} \lesssim \text{ red})\) has the largest \([\text{C}/\text{H}]=+0.5\). We find that \(36,169\) of the total \(58,015\) stars have \([C/\text{Fe}]>0.7\). In preliminary tests, we found that our algorithm is incapable of detecting carbon enhancement in hot stars so we apply a cut in the \(G_{\text{BP}} - G_{\text{RP}}\) color. Specifically, we only include stars with \(G_{\text{BP}} - G_{\text{RP}} > 0.8\) mag. This leaves us 1,279 carbon-enhanced stars and 1,316 carbon-normal stars.

From the AEGIS Survey, we start with a total of 58,015 stars that have measured carbon abundances, of which 1,476 have \([C/\text{Fe}]>0.7\). We find that 36,169 of the total 58,015 stars have BP/RP spectra released in Gaia DR3, of which \(8,885\) \((\approx 57\%)\) have \([C/\text{Fe}]>0.7\). In preliminary tests, we found that our algorithm is incapable of detecting carbon enhancement in hot stars so we apply a cut in the \(G_{\text{BP}} - G_{\text{RP}}\) color. Specifically, we only include stars with \(G_{\text{BP}} - G_{\text{RP}} > 0.8\) mag. This leaves us 1,279 carbon-enhanced stars and 1,316 carbon-normal stars.

In total, we employ 32,589 spectra for the training and testing, of which 1,740 have \([C/\text{Fe}]>0.7\). We randomly se-

---

1 http://kurucz.harvard.edu
2 https://github.com/karllark/dust_extinction

2.1 Training and Testing Sample

In order to teach our model how to accurately detect carbon-enhanced stars, we require a sample of stars that are already classified. To acquire this, we use the spectroscopic catalogs from the Sloan Extension for Galactic Understanding and Exploration (SEGUE; Yanny et al. 2009; Lee et al. 2013; Rockosi et al. 2022) and AAOmega Evolution of Galactic Structure (AEGIS; Yoon et al. 2018) surveys. The SEGUE survey was a sub-survey of the Sloan Digital Sky Survey (SDSS-II and SDSS-III; Fukugita et al. 1996; Gunn et al. 1998, 2006; York et al. 2000; Stoughton et al. 2002; Abazajian et al. 2003, 2004, 2005, 2009; Pier et al. 2003; Adelman-McCarthy et al. 2006, 2007, 2008; Aihara et al. 2011; Ahn et al. 2012), and obtained \(\approx 500,000\) low-resolution \((R=2000)\) optical spectra. However, these spectra are primarily obtained for stars in the Northern Hemisphere, so we supplement this sample with stars from the Southern AEGIS survey. The AEGIS survey observed \(\approx 70,000\) low-resolution \((R=1,300)\) optical spectra, from which both metallicity and carbon abundance could be measured. For both of these surveys, we use the measured carbon abundances that are not corrected for evolutionary effects (e.g., the first and second dredge ups).

For the SEGUE survey data, we start with 52,273 stars from T. Beers (2022, personal communication), of which 33,582 have \([C/\text{Fe}]>0.7\). We find that 15,508 of the 52,273 stars have BP/RP spectra released in Gaia DR3, of which 8,885 \((\approx 57\%)\) have \([C/\text{Fe}]>0.7\). In preliminary tests, we found that our algorithm is incapable of detecting carbon enhancement in hot stars so we apply a cut in the \(G_{\text{BP}} - G_{\text{RP}}\) color. Specifically, we only include stars with \(G_{\text{BP}} - G_{\text{RP}} > 0.8\) mag. This leaves us 1,279 carbon-enhanced stars and 1,316 carbon-normal stars.
\section*{2.2 Gaia BP/RP Spectra}

The Gaia BP/RP spectra are a unique data set, not only due to their wide wavelength coverage, low-resolution (R\approx50), and unprecedented number of stars, but also because the spectra have been released as a linear combination of basis functions, specifically as Hermite function coefficients (Carrasco et al. 2021). This was done because of the complexity of the Gaia instrument, which has two wide fields of view and 14 detectors. To create the calibrated mean spectra, multiple epochs of observations with different instrumental conditions need to be combined. In this work, we use the coefficients as the input data for our model rather than convert them to sampled spectra, which results in some information loss (Carrasco et al. 2021).

In total, the BP and RP spectra comprise 55 coefficients each, but also come with a recommended truncation. This is possible because the coefficients have been rotated to an optimized basis so that the bulk of the spectral information is contained in the first few coefficients. A truncation is then recommended based on the magnitude of the coefficients compared to their corresponding uncertainties. For more details see Carrasco et al. (2021). As XGBoost requires the input data to be vectors of the same length, we apply the largest recommended truncation to avoid losing potentially useful information. The largest recommended truncation is 55 for both BP and RP spectra (i.e., all coefficients are relevant). Therefore we do not truncate the coefficients. Because we do not want to include apparent magnitude information,
we normalize the coefficients by the maximum of the combined BP and RP coefficients.

The XGBoost algorithm cannot reliably extrapolate. Therefore, we ensure to only classify stars that are similar to our training sample. Specifically, we constrain our sample using the absolute G magnitude and $G_{BP} - G_{RP}$ color. As shown in rightmost panel of Figure 2 in the black dotted lines, we only classify stars that satisfy the following criteria:

(i) $0.8 < G_{BP} - G_{RP} < 2.0$
(ii) $G + 5\log_{10}(\pi) + 5 < 7.0$.

Although our training sample includes stars outside of this range, we chose to restrict to these values where most of the sample resides. Note that we do not place a lower limit on the absolute G magnitude, since stars with very low absolute G magnitudes likely have small, uncertain parallaxes that cause an overestimation in the brightness, and therefore underestimate in the absolute G magnitude. As we do not want to introduce a selection bias by removing stars with uncertain parallaxes, we chose to include them. In Section 2.1, we investigate how the precision and completeness of our classification behaves at the edges of this region where the training sample is less dense. In total, we find 160,422,878 BP/RP spectra in Gaia DR3 that are within our color and magnitude cuts.

### 3 XGBOOST

In order to detect carbon-enhanced stars across a wide range of stellar parameters and reddenings, we require a flexible model. We chose to use XGBoost, which is powerful but still easy to interpret. Furthermore, XGBoost is optimized for efficiency, allowing fast training and inference. XGBoost is quickly becoming a popular machine learning algorithm in astronomy, with applications in a large variety of sub-fields (e.g., Hayden et al. 2020; Machado Poletti Valle et al. 2021; Li et al. 2021; He et al. 2022; Pham & Kaltenegger 2022).

Figure 3 shows the general architecture of XGBoost for classification. For a detailed description of the algorithm see Chen & Guestrin (2016). In short, XGBoost sequentially builds decision trees to fit the residuals from the previous tree. XGBoost continues to train trees until it reaches the maximum number of trees set by the user or the residuals stop consistently shrinking. The results from each tree are then summed together, weighted by the learning rate, $\eta$. This value is then plugged into the sigmoid function, $\sigma(x) = 1/(1 + e^{-x})$, to calculate the probability that the star is carbon enhanced ($\hat{y}$). If $\hat{y} > 0.5$, then the star is classified as carbon enhanced.

To train the XGBoost algorithm, a number of hyperparameters need to be set. Namely, we can set the maximum number of trees, the learning rate ($\eta$), the percentage of the training sample and the percentage of input coefficients to use for each tree, as well as the maximum depth of each tree. We can also set limits on the purity of a sample for a given leaf to prevent overfitting. To explore the parameter space and find the optimal set of hyperparameters, we use RandomSearchCV from scikit-learn (Pedregosa et al. 2011).

### 4 PRECISION AND COMPLETENESS

In order to estimate the true positive rate (i.e., precision) and completeness of our sample of newly identified CEMP stars, we use our testing sample (described in Section 2.1) where we already know the observed carbon abundance from the catalogs of Lee et al. (2013) and Yoon et al. (2018). As we constrain the full set of spectra that we classify by the absolute G magnitude and $G_{BP} - G_{RP}$ color of the training/testing sample (see Section 2.2), we can be confident that the precision and completeness estimates are representative of the full sample. In addition, we have stellar parameters and carbon abundances for these stars, which allows us to study how the precision and completeness behaves as a function of various parameters, including observational effects.

Figure 4 shows the confusion matrix which shows the number of true and false positives along with the number of true and false negatives for our entire testing sample. The boxes are colored by the number of stars that fall in that part of the table. Stars which are known to be carbon enhanced from the SEGUE and AEGIS data are in the top row while the known carbon normal stars are in the bottom row. If the XGBoost algorithm classifies the star as carbon enhanced it goes in the first column, while stars classified as carbon-normal go in the second column. Therefore, the top-left box and the bottom-right box contain stars that are correctly classified, while the bottom-left box and top-right box contain stars that were falsely classified. From these numbers, we can calculate the precision and the completeness percentage. To calculate the precision, we divide the number of true positives (488) by the summation of the false positives (24) and true positives. We measure a precision of 95%. We also calculate our completeness percentage by dividing the number of true positives by the summation of the true positives and the false negatives (181). We achieve a completeness of 73%.

Figure 5 shows 2D maps of the false positive (left) and true positive rates (right) as a function of absolute G magnitude, and $G_{BP} - G_{RP}$ color (top), as well as $[C/Fe]$ and $[Fe/H]$ (bottom). In these plots, we also show the underlying density distribution of stars as gray hexagonal bins. The left panels show the false positive rate, which is calculated by taking the number of false positives divided by the number of true carbon-normal stars in that bin. The right panels show the true positive rate, which is calculated by dividing the number of true positives (i.e., correctly identified CEMP stars) by the total number of true CEMP stars in that bin.

In the bottom row of Figure 5, we do not see a trend in the false positive and true positive rate with the $G_{BP} - G_{RP}$ color or absolute G magnitude. In the top right panel, we see that at the edges of our distribution the true positive percentage highly varies. These is merely because of the few true positives in these bins.

In the bottom row of Figure 5, we see a slight trend in the false positive and true positive rates with the $[C/Fe]$ abundance. Specifically, we see the false positive rate is higher for stars with $+0.5 < [C/Fe] < +0.7$, and the true positive rate is lowest for stars with $+0.7 < [C/Fe] < +1.0$. This indicates the our classification is most inaccurate for stars with $[C/Fe] \approx +0.7$, which is to be expected given that is is unlikely we could measure a $[C/Fe]$ abundance from these...
Figure 3. General architecture of XGBoost for classification. In short, XGBoost iteratively creates trees to fit the residuals from the prediction of the previous tree. The first tree made gives a prediction of zero for each spectrum. For the final output, the predictions of each tree are summed after being multiplied by the learning rate, \( \eta \). This value is then input into the sigmoid function, \( \sigma(x) = \frac{1}{1+e^{-x}} \), to calculate the final probability, \( \hat{y} \). If \( \hat{y} > 0.5 \), then the star is classified as carbon enhanced.

extremely low-resolution (R≈50) spectra that is more precise than ≈0.2 dex, at best.

Figure 6 shows the true positive percentage and contamination rate, as a function of the absolute Galactic latitude (top panel) and G magnitude (bottom panel). Here, we use the absolute Galactic latitude as a proxy for the extinction, given that stars with low absolute Galactic latitudes will be in the disk, and be more impacted by extinction than stars with high absolute Galactic latitudes. The gray histogram shows the arbitrarily scaled underlying distribution of the testing sample for these parameters. Here the contamination rate is calculated by taking the number of false positives divided by the total number of stars classified as CEMP stars in each parameter bin. The green line shows the contamination rate with the scaling provided by the right y-axis, while the dark blue lines shows the true positive percentage (calculated in the same way as Figure 5), with corresponding scaling on the left y-axis.

As expected, we find that the classification performs better at high absolute Galactic latitudes where extinction is low. Specifically, we see that the true positive percentage increases from ≈70% at Galactic latitudes <\( |10|° \) to ≈79% at Galactic latitudes >\( |80|° \). We also see that the contamination rate drops from ≈10% to ≈0% over the same range. Furthermore, we find that the classification also improves with brighter G magnitude, as expected, since the G magnitude is directly related to the signal-to-noise for these spectra. We find that the true positive percentage increases from ≈70% at G of ≈17 mag to ≈100% at G =12.5 mag. Over the same range, we find that the contamination rate drops from ≈5% to ≈0%. Therefore, we recommend using bright, high absolute Galactic latitude stars to retrieve a higher fidelity sub-sample for our provided catalog of carbon-enhanced candidate stars.

5 MODEL INTERPRETATION

As we have used a data-driven method to classify stars in this work, we want to ensure that the final model matches our physical intuition. It is also important to ensure that
the model is not using any confounding variables (e.g., extinction or metallicity) that could be correlated with carbon enhancement but are not a direct measurement of the true carbon abundance. Typically, model interpretation is done by determining the importance of each input feature for the final inference and comparing to expectations from physical models. Here, we investigate the importance of two key Hermite coefficients and how they relate to the true carbon abundance.

We measure the impact of each Hermite basis coefficient on the classification using SHAP (Shapley Additive exPlanations) values (Lundberg & Lee 2017). SHAP values allow us to explore the importance of individual features, as a function of various parameters, since each star’s coefficients are assigned individual values. The SHAP values are defined so that their summation plus the average predicted value \( \bar{y} \) is equal to the predicted value \( \hat{y} \). Explicitly,

\[
\hat{y} = \phi_0 + \sum_{i=1}^{M} \phi_i
\]

where \( \phi_0 \) is the average value of \( \hat{y} \) for all of the spectra, \( \phi_i \) is the SHAP value for coefficient \( i \), and \( M \) is the total number of coefficients per spectrum. Therefore, each SHAP value directly measures the impact of the selected coefficient on the inference of \( \hat{y} \) for a given star. We refer the interested reader to Lundberg & Lee (2017) for further details on the calculation of SHAP values.

To decide which spectral coefficients contain the majority of information about the carbon abundance, we calculate the Pearson Correlation coefficient for each coefficient and the carbon abundance for the entire testing sample. We find that the second, third, fifth and sixth BP coefficients have the highest absolute Pearson Correlation coefficients. Therefore we check to ensure that our XGBoost model gives high SHAP values for these coefficients for carbon-enhanced stars.

We calculate the SHAP values for all of the spectra in our testing sample. Figure 7 shows how the SHAP values relate to the carbon abundance for the four selected coefficients that have the largest Pearson Correlation coefficients. Specifically, we have the \([C/Fe]\) on the x-axis and the SHAP values on the y-axis. We show the points colored by the coefficient’s value. As large positive SHAP values indicate that the given coefficient increased the probability that the star is carbon enhanced, it is expected that stars with \([C/Fe]>+0.7\) should have higher SHAP values than carbon-normal stars. We find that the XGBoost model is able to pick up this sensitivity, in that the SHAP values for these coefficients are large and positive for carbon-enhanced stars. Although there are some coefficients were the SHAP value is negative for a carbon-enhanced star, these stars may still be classified as carbon-enhanced based on the value of other coefficients. Similarly, these are many carbon-normal stars that have positive SHAP values, but the combined effect of all of the other coefficients effectively decrease the probability of the star being carbon enhanced, so that the false positive rate is not exceedingly high. Given the positive correlation between \([C/Fe]\) and the SHAP value, it is likely that XGBoost model is using the carbon information in the spectra to determine whether a given star is carbon enhanced rather than using another confounding variable.

### 6 PROPERTIES OF THE CEMP SAMPLE

Out of the \(\approx 160\) million stars that we classify, we find 2,736,922 CEMP stars. This is the largest, homogeneously identified sample of CEMP stars to date. In this section, we briefly investigate a few properties of this sample, including the binarity rate and the Galactic distribution.

For each star in Gaia DR3, a flag indicating whether the star is an astrometric, spectroscopic, or eclipsing binary is provided (Gaia Collaboration et al. 2022b). We use these flags and the reported metallicities from the Gaia GSPPHOT pipeline (Andrae et al. 2022), along with our classification, to explore the relationship between metallicity and binarity for carbon-enhanced stars. In Figure 8, we plot the occurrence and binarity rate of CE stars as a function of metallicity. In addition, we show the metallicity distribution of our sample of CEMP stars in a gray histogram, arbitrarily scaled. Given that carbon-enhanced stars were not accounted for, it is likely that the \([M/H]\) from the GSPPHOT pipeline and the iron abundance of these stars are actually much lower.

Specifically, in dark blue, we show the ratio of the number of stars classified as CE to the total number of BP/RP spectra we analyzed for \([M/H]<x\). In other words, the dark blue line gives the fraction of CEMP stars for stars with \([M/H]<x\). Compared to results from the literature, we find a fraction that is roughly an order of magnitude lower (Lucatello et al. 2006; Lee et al. 2013; Placco et al. 2014; Yoon et al. 2018; Arentsen et al. 2022). This may be a result of low completeness in that we are able to classifying many CEMP stars as carbon-normal. From our testing sample we estimated our completeness as 73%, but these results may indicate it it closer to 10%. Another possibility is that there
Figure 5. The false positive and true positive rates, as functions of absolute G magnitude, $G_{BP} - G_{RP}$ color, $[C/Fe]$, and $[Fe/H]$. We show the distribution of stellar parameters for the testing sample in a grayscale, with darker areas corresponding to more stars. In general, the model tends to struggle most with redder stars, which are likely cooler and more extincted. Furthermore, the false positives tend to have $+0.5 \leq [C/Fe] \leq +0.7$, while the false negatives tend to have $+0.7 \leq [C/Fe] \leq +1.0$. Therefore, our model likely only can interpret the carbon abundance to $\approx +0.5$ for some stars.

may be inaccuracies with the $[M/H]$ measurements from the Gaia DR3 GSPPHOT pipeline. It is quite unexpected that the CEMP fraction for stars with $[M/H]<0$ would be higher than for $[M/H]<-2.5$. This is likely a result of the $[M/H]$ for CEMP stars being overestimated since carbon-enhanced stars were not included in the models used for the analysis in the GSPPHOT pipeline. Furthermore, the GSPPHot pipeline may be over estimating the number of metal-poor stars observed which would cause our CEMP fraction to be underestimated.

In addition, we also show the fraction of non-single stars among the CE (green) and non-CE (light blue) populations. As expected from the literature (Hwang et al. 2021), the binary fractions generally increase with increasing $[M/H]$. It is interesting that the fraction of CEMP in binaries is largely than the fraction of carbon-normal stars in bina-
likely that their iron abundance is actually much lower. This

... and normal stars, we want to ensure we do not introduce such a bias. We show the distribution of our CEMP sample in Galactic X and Y coordinates in the right panel, while the left panel shows the distribution in Galactic X and Z coordinates. The Galactic center is located at (0,0,0) kpc with the Sun located at (8.3,0,0) kpc (Reid et al. 2014).

Consistent with a Galactic halo population, our CEMP sample is extended to far distances from the Galactic center. It is important to note that some of the largest distances may be unreliable given that the fractional parallax uncertainty for faint stars can be quite large. Given that the majority of BP/RP spectra released in DR3 have $G < 17.6$ mag and assuming an absolute $G$ magnitude of $-2.5$ mag (roughly the tip of the red giant branch), we expect to have detected CEMP stars at distances of up to $\approx 30$ kpc. We find that 45,423 stars in our CEMP sample that have parallaxes corresponding to distances from the Galactic center $>30$ kpc. Therefore, it is likely that we have detected CEMP stars in other galaxies.

In addition, we find there is a dearth of stars at low distances from the plane ($Z$), especially towards the Galactic center. Although there are fewer than expected for a disk or bulge population, we do still find a number of stars in the inner Galaxy, with distances as low as $\approx 0.4$ kpc from the Galactic Center. However, is is difficult to determine how our observed distribution is impacted by the selection function of our algorithm and Gaia DR3. Further work is required to fully understand and model the selection function of our sample which will enable us to draw strong conclusions from population statistics.

We also investigate the relative Galactic distribution of CEMP stars compared to carbon-normal stars in Figure 10. Specifically, in the left panel, we show the reverse cumulative density distribution of the distance from the Galactic center for CEMP stars (green) and for carbon-normal stars. We find that the CEMP stars tend to be farther from the Galactic plane than the carbon-normal stars. This is consistent with previous work, which found the frequency of CEMP stars to increase with increasing distance from the Sun (Carollo et al. 2012; Frebel et al. 2006; Lee et al. 2013, 2017; Yoon et al. 2018). In addition, the right panel shows the density distribution of the distance from the Galactic center for the CEMP stars (green) and the carbon-normal stars (blue). We also mark the distance of the Sun from the Galactic center as a black dashed line at 8.3 kpc. In general, we find that the CEMP stars are more distant from the Sun than the carbon-normal stars. Principally, we find a higher density of CEMP stars in the outer regions of the Galaxy, with distances from the Galactic center $>10$ kpc. This indicates these stars may originate in the outer Galactic halo, as suggested by Yoon et al. (2018, 2019).

Figure 9 shows the Galactic distribution of the CEMP sample. To calculate the Galactic positions of each CEMP star we detect, we simply invert the reported Gaia eDR3 parallax. We choose not to use distances that are calculated using photometry since it is unknown how the carbon enhancement will bias these results. As we want to compare to the Galactic distribution of carbon-normal stars, we want to ensure we do not introduce such a bias. We show the distribution of our CEMP sample in Figure 9 (top panel) and G magnitude (bottom panel). For each panel, the scale for the true positive percentage (dark blue) is on the left y-axis; the right y-axis shows the scale for the contamination rate (green). As expected, we find that our classification improves for bright stars, which likely have high signal-to-noise spectra, and stars at high absolute Galactic latitude, which is consistent with low extinction.

|Galactic Latitude| (degrees) | G (mag) |
|------------------|-----------|---------|
|0.72              | 12.5      |
|0.74              | 15.0      |
|0.76              | 17.5      |
|0.78              | 0.00      |

Figure 6. The contamination rate and true positive percentage, as a function of G magnitude and absolute Galactic latitude, as a proxy for extinction. The gray bins show the arbitrarily scaled number density of stars as a function of absolute Galactic latitude (top panel) and G magnitude (bottom panel). For each panel, the scale for the true positive percentage (dark blue) is on the left y-axis; the right y-axis shows the scale for the contamination rate (green). As expected, we find that our classification improves for bright stars, which likely have high signal-to-noise spectra, and stars at high absolute Galactic latitude, which is consistent with low extinction.

Effect is likely increased for CEMP-s stars because of the high abundance of s-process elements.

|Galactic Latitude| (degrees) |
|------------------|-----------|
|0.00              | 0.02      |
|0.01              | 0.04      |
|0.02              | 0.06      |
|0.03              | 0.08      |

|G (mag) | Contamination Rate |
|--------|---------------------|
|12.5    | 0.00                |
|15.0    | 0.02                |
|17.5    | 0.04                |
|20.0    | 0.06                |

This indicates these stars may originate in the outer Galactic halo, as suggested by Yoon et al. (2018, 2019).
Figure 7. The SHAP values, as a function of [C/Fe] (x-axis), for the four spectral coefficients that are most correlated with the carbon abundance. Each point corresponds to an individual star in our testing sample and are colored by the given coefficient’s value. We show a horizontal black dashed line at SHAP value=0 and a vertical black dashed line at [C/Fe]=+0.7. Large positive SHAP values indicate that the measured coefficient value greatly increased the inferred probability of the star being carbon enhanced. Therefore, if the model uses the carbon information in the coefficients, we expect that stars with [C/Fe] >+0.7 to have high positive SHAP values.

7 SUMMARY

The origins of CEMP stars are poorly understood, even though they comprise ≈30% of stars with [Fe/H]<−2 (Lucatello et al. 2006; Lee et al. 2013; Placco et al. 2014; Yoon et al. 2018; Arentsen et al. 2022). A significant fraction of CEMP stars have enhancements in s-process elements and are called CEMP-s stars (Beers & Christlieb 2005). These stars are thought to receive their overabundant carbon and s-process material from a mass-transfer event with their binary companion, which has evolved to or past the AGB (Lugaro et al. 2012; Placco et al. 2013). On the other hand, CEMP stars without neutron-capture enhancements are thought to have been primarily enriched by material from the first generation of stars (Umeda & Nomoto 2003; Chiappini et al. 2006; Meynet et al. 2006; Nomoto et al. 2013; Tominaga et al. 2014). However, there are many remaining questions about these unique stars, including why they seem to be less frequent in the central regions of our Galaxy, where we expect the highest concentration of ancient stars to reside (Howes et al. 2014, 2015, 2016; Arentsen et al. 2021; Lucey et al. 2022). As suggested by Yoon et al. (2019), this dearth of CEMP stars may be caused by the dilution of CEMP stars in more massive subsystems (e.g., dwarf galaxies) with prolonged star formation histories, which could be the origin of metal-poor stars in the inner Galaxy.

In this work, we leverage the data from the all-sky Gaia survey to increase the number of known CEMP stars from ≈50,000 to over 2.7 million. We accomplish this by leveraging the ∼200 million BP/RP spectra made available in Gaia DR3. Using the XGBoost algorithm for classification, we achieve a contamination rate of 5% and completeness of 73%. We ensure that the XGBoost algorithm matches our physical intuition and primarily performs the classification using spectral features that are correlated with the carbon abundance.

We briefly investigate a few of the properties of our sample including the binarity rate and the Galactic distribution. Specifically, we use the non-single stars flag provided in Gaia DR3 and compare the binarity rate of CEMP as a function of metallicity to carbon-normal stars. We also look at the distributions of the distances from the Galactic plane and the Galactic center of the CEMP stars compared to the carbon-normal stars. In general, we find that the the CEMP
stars are farther from the Galactic plane. In addition, we find they are generally farther from the Sun with a stronger tail towards larger radii.

In future work, we plan to look at the orbital properties of these stars as well as the rate of s-process enhancement. We plan to perform high-resolution follow-up of many of these targets in order to confirm our contamination rate, measure radial velocities, and determine s-process abundances. We plan to specifically follow-up targets towards the central region of the Galaxy where the number of known CEMP stars is much lower. Following future Gaia releases, which will include many millions more BP/RP spectra, we expect to continue this work and again increase the known number of CEMP stars by orders of magnitude.

ACKNOWLEDGEMENTS

This project was developed in part at the Gaia Fête, held at the Flatiron institute Centre for Computational Astrophysics in 2022 June. We’d like to thank David Hogg and Jason Hunt for helpful discussions and computational resources. We also thank Gary Da Costa for observational data from the AEGIS survey which made this work possible.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 000392968. T.C.B. acknowledges partial support for this work from grant PHY 14-30152; Physics Frontier Center/JINA Center for the Evolution of the Elements (JINa-CEE), awarded by the US National Science Foundation. Y.S.T. acknowledges financial support from the Australian Research Council through DECRA Fellowship DE220101520. Work at Argonne National Laboratory was supported by the U.S. Department of Energy, Office of High Energy Physics. Argonne, a U.S. Department of Energy Office of Science laboratory, is operated by UChicago Argonne LLC under contract no. DE-AC02-06CH11357. N.R. acknowledges the Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-06CH11357. Y.S.L. acknowledges support from the National Research Foundation (NRF) of Korea grant funded by the Ministry of Science and ICT (NRF-2021R1A2C1008679).

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.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

This research made use of Astropy, a community-developed core Python package for Astronomy (Astropy Collaboration et al. 2013, 2018). Other software used includes IPython (Pérez & Granger 2007), matplotlib (Hunter 2007), numpy (Harris et al. 2020), scikit-learn (Pedregosa et al. 2011) and scipy (Virtanen et al. 2020).

DATA AVAILABILITY

The data underlying this article are publicly available through the Gaia Archive. The generated catalog of CEMP stars is available in the online supplementary material and CDS upon acceptance.

REFERENCES

Abate C., Pols O. R., Karakas A. I., Izzard R. G., 2015, A&A, 576, A118

Abazajian K., et al., 2003, AJ, 126, 2081

http://www.astropy.org

https://gea.esac.esa.int/archive/
Figure 9. The Galactic distribution of our carbon-enhanced (CE) sample. Specifically, in the left panel, we show the distribution of stars in the Galactic coordinates $X$ and $Y$. The right panel shows the distribution in the Galactic $X$ and $Z$ coordinates. The Galactic center is located at $(0,0,0)$ kpc while the Sun is at $(8.3,0,0)$ kpc. The distribution of our CEMP sample is very halo-like and is quite extended rather than staying confined to low $Z$ as one would expect for a disk population.

Figure 10. The Galactic distribution of our carbon-enhanced (CE) sample compared to the non-carbon-enhanced (non-CE) sample. Specifically, in the left panel, we show the reverse cumulative density distributions of the distance from the Galactic plane ($|Z|$) for both samples, while the right panel shows the density distributions of the distance ($R_{GC}$) from the Galactic center. We also show the Sun’s position at 8.3 kpc as a black dashed line. The carbon-enhanced stars are generally farther from the Galactic center and the Galactic plane than the non-carbon-enhanced stars.
Yanny B., et al., 2009, AJ, 137, 4377
Yoon J., et al., 2018, ApJ, 861, 146
Yoon J., Beers T. C., Tian D., Whitten D. D., 2019, ApJ, 878, 97
York D. G., et al., 2000, AJ, 120, 1579

This paper has been typeset from a TeX/LaTeX file prepared by the author.