Spectral Feature Extraction for DB White Dwarfs Through Machine Learning Applied to New Discoveries in the SDss DR12 and DR14

Xiao Kong1,2, A-Li Luo1,2*, Xiang-Ru Li3, You-Fen Wang1, Yin-Bi Li1, and Jing-Kun Zhao1

1 Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, People’s Republic of China; lal@nao.cas.cn
2 University of Chinese Academy of Sciences, Beijing 100049, People’s Republic of China
3 School of mathematical Sciences, South China Normal University, Guangzhou, 510631, People’s Republic of China

Received 2018 February 8; accepted 2018 May 23; published 2018 June 26

Abstract

Using a machine learning (ML) method, we mine DB white dwarfs (DBWDs) from the Sloan Digital Sky Survey (SDSS) Data Release (DR) 12 and DR14. The ML method consists of two parts: feature extraction and classification. The least absolute shrinkage and selection operator (LASSO) is used for the spectral feature extraction by comparing high quality data of a positive sample group with negative sample groups. In both the training and testing sets, the positive sample group is composed of a selection of 300 known DBWDs, while the negative sample groups are obtained from all types of SDSS spectra. In the space of the LASSO detected features, a support vector machine is then employed to build classifiers that are used to separate the DBWDs from the non-DBWDs for each individual type. Depending on the classifiers, the DBWD candidates are selected from the entire SDSS data set. After visual inspection, 2808 spectra (2029 objects) are spectroscopically confirmed. By checking the samples with the literature, there are 58 objects with 60 spectra that are newly identified, including a newly discovered AM CVn. Finally, we measure their effective temperatures (Teff), surface gravities (log g), and radial velocities, before compiling them into a catalog.

Key words: catalogs – methods: data analysis – surveys – white dwarfs

Online material: color figures, machine-readable table

1. Introduction

At the final stage of stellar evolution for main sequence stars, white dwarfs (WDs) simply cool off in the absence of nuclear reactions. The energy of most WDs are generated by the radiation of the residual gravitational contraction, instead of nuclear fusion. Generally, the initial masses of the progenitors of WDs are approximately between 0.07 and 8 M_{\odot}, and their radius are often the same order as that of the Earth, implying that they need extremely long cooling times. It is believed that over 97% of the stars in the Galaxy will eventually end up as WDs (Fontaine et al. 2001). The luminosity function of a WD, containing information of the stellar death rate in the local galactic disk, can be used to estimate the density of the matter in the Galaxy. A statistically complete sample is required to measure the luminosity function of WDs (Limoges & Bergeron 2010).

Approximately 80% of all observed WDs belong to DA type with Hydrogen dominated atmospheres, with the remaining 20% falling into the DB (He I) or DO (He II) categories with atmospheres dominated by helium. As these stars are lined up in the WD cooling sequence, they are observed with temperatures of approximately 45,000 K; categorized as hot DO stars with He II rich in spectra, and of effective temperatures (Teff) mostly below 30,000 K; categorized as DB WDs (DBWDs) with only He I lines in the spectra. When the temperature drops to 10,000 K, helium becomes spectroscopically invisible, e.g., featureless smooth DC, carbon present DQ, or metal-rich DZ spectra (Voss et al. 2007).

With nearly pure helium in the neutral form in their atmospheres, the DBWDs represent the best example of hydrogen-deficient stars in the universe. Many hydrogen dominated DA WDs transform into DBWDs with helium atmospheres, and the ratio of DA to non-DA WDs varies as a function of Teff along the cooling sequence (Fontaine et al. 2001). By expanding the DBWD sample size, a better understanding of the evolution of WDs is possible.

The high photospheric purity of DBWDs was first revealed by atmosphere models (Bues 1970). Only about 80 optical spectra and 25 ultraviolet spectrophotometries were investigated in the 1990s (Beauchamp et al. 1996). Later, with the help of the Sloan Digital Sky Survey (SDSS) spectral surveys, systematic searches were completed and a larger sample of DBWDs were obtained, holding great potential for the exploration of the chemical evolution of DB degenerates. Kleinman et al. (2013) provided 922 DBWDs from SDSS Data Release (DR) 7 (Abazajian et al. 2009), Kepler et al. (2015)
added another 450 in DR10 (Eisenstein et al. 2011). DR12 (Alam et al. 2015) increased this number by 121 (Kepler et al. 2016), among which Koester & Kepler (2015) selected 1267 spectra with signal-to-noise ratios (S/N) greater than 10, and of these the atmospheric parameters of 1107 objects were analyzed. Based on the sample of 150 DBWDs and 1733 DA WDs, Kepler et al. (2007) provided the average masses of 0.711 \pm 0.009 M_\odot and 0.593 \pm 0.016 M_\odot for the DB and DA types, respectively. Koester & Kepler (2015) reported that the mass of DB types have a significant increase below T_{\text{eff}} = 16,000 \text{ K}, possibly caused by the imperfect implementation of line broadening of neutral helium atoms, and analyzed the distributions of DBA WDs and DBWDs with the height, z, above the Galactic plane differing toward lower T_{\text{eff}}. Eisenstein et al. (2006b) presented 28 stars as candidate hot DB or cool DO WDs, some of which are the first helium atmosphere WDs found in the range 30,000–45,000 K, in the DB gap.

However, the majority of known DB spectra are obtained through parameter measurement, which may lead to incompleteness in the DBWD findings because of bad data quality or spectral fitting failures. For such a large data set from SDSS DR12, including 4355,200 spectra, some DBWDs will not be identified, especially for data with low S/N. The DB class in the SDSS catalog is lacking as most known DBWDs are misclassified as “O,” “B,” “A,” “QSO” or some other types in the DR12. It should be noted that there is a “WD” class in SDSS DR12, which mainly includes DA WDs, although a few other subtypes of WD are mixed in.

We search for DBWDs in all DR12 and DR14 without color or other limits, which depend on machine learning (ML). After a first manual check, we discard those without obvious He I lines (4471.5 and 5875.6 Å). Then we arrange the remaining spectra in descending order of S/N of the g band (S/N_g), where the majority of He I lines exist, and select the top 300 as our positive samples, which is used to extract the DB features using the LASSO method (Tibshirani 1996). To analyze the features, the mathematical tool Wavelet (Ingrid Daubechies 1992) is used to check the decomposed spectra in different scales. The wavelet transform is able to cut up signals into different frequency components, then each component with a resolution matched to its scale can be studied. We directly employ the two built-in functions of MATLAB, “ wavedec” and “ wrcf,” to aid the wavelet analysis of the spectrum. Li et al. (2015) described the basic properties of wavelet, explained the process of wavelet decomposition, and employed it in combination with LASSO to estimate stellar atmospheric parameters. Luo & Zhao (2001) used it to obtain spectral classification information of galaxies. Afterwards, in the derived feature space rather than the original spectra, SVM is employed to distinguish DB from other types of objects. “Feature” here has the same meaning as flux at some particular wavelength or any specific location of spectra.

This paper is organized as follows. Section 2 describes the spectral data used in this paper. The ML method applied in this paper are explained in detail in Section 3, including the preprocessing of the data set, feature model construction of DBWD through the LASSO algorithm, feature analysis using the wavelet transform, and classifier establishment by the SVM. Then, we apply the method to detect DBWDs from the SDSS data set, as introduced in Section 4. In Section 5, we compare our results with those of the literature and calculate the relative parameters, such as T_{\text{eff}}, surface gravities (\log g), and ultra-violet color (FUV–NUV). Finally, we summarize our work in Section 6.

### 2. Data Sets

In the Koester & Kepler (2015) catalog, there are a total of 1107 objects of DBWD from SDSS DR10 and DR12, which are originally classified as O, QSO, B, WD, galaxy, or other types by SDSS 1D pipeline; the quantities of each type are given in Table 1. This table suggests that we can pick out more DB spectra from all types of SDSS spectra, excluding those with a high confidence of classification.

We attempt a ML method to search for more DBWDs, with a focus on the low quality data. The basic idea is to use SVM as a classifier to sort out the DBWDs from the spectral data of SDSS DR12 and DR14, based on features found by LASSO.

We construct two subsets for training and testing from SDSS DR12. The training set is used for learning, that is to fit the

| Class*   | Subclass* | Number | Class*   | Subclass* | Number | Class*   | Subclass* | Number |
|----------|-----------|--------|----------|-----------|--------|----------|-----------|--------|
| QSO      | null      | 416    | galaxy   | null      | 12     | star     | K         | 2      |
| Star     | OB        | 400    | star     | A         | 11     | star     | T         | 2      |
| Star     | O         | 117    | star     | CV        | 9      | star     | carbon    | 2      |
| Star     | B         | 62     | star     | F         | 7      | star     | G         | 1      |
| QSO      | broadline | 30     | star     | L         | 4      | galaxy   | broadline | 1      |
| Star     | WD        | 28     | star     | M         | 3      | ...      | ...       | ...    |

Note. *“Class” and “subclass” are adopted from the data archive of SDSS.
parameters (features and hyper-planes) of the classifier. The testing set is used for the parameter adjustment of the classifier, e.g., to choose the best features and most suitable kernel function of SVM. A validation process of 10-fold cross-validation is built into both the LASSO and SVM packages and conducted automatically using the training set. We select a candidate data set, named experimental data (ED), from the SDSS DR14 and explain the procedure of selection in Section 2.2. Table 2 lists the roles of the training, testing, and ED sets.

In the SDSS data archive, spectra were grouped into “Class” and “Subclass,” which are shown in Table 3. As the basic SVM is a binary-classification algorithm, the classification between DB and each “Subclass” are performed in parallel. For convenience, we abbreviate “Class + Subclass” (CPS) as the ID of each subclass in this experiment, such as “star+O” or “QSO+AGN.”

### Table 2

| Data Set          | Roles                                                                 |
|-------------------|----------------------------------------------------------------------|
| Training Data     | To be used in the training process, i.e.,                           |
|                   | Detecting features by LASSO (Section 3.2);                          |
|                   | Estimating the parameterizing model by LIBSVM (Section 3.4).         |
| Testing Data      | To be used in the training process, i.e.,                           |
|                   | Determining the parameters in LASSO (Section 3.2);                  |
|                   | Determining the hyper-planes in LIBSVM (Section 3.4).               |
| Experimental Data | Application of Section 3, to be used in searching DB spectra from Experimental Data (Section 4). |

### Table 3

| Class\(^a\) | Subclass\(^a\)                           |
|------------|-----------------------------------------|
| Star       | O, B, A, F, G, K, M, L, T, WD, carbon, CV |
| Galaxy     | AGN, broadline, starburst, star-forming, null |
| QSO        | AGN, broadline, starburst, star-forming, null |

Note. \(^a\) “Class” and “Subclass” are adopted from the data archive of SDSS.

Meanwhile, for each CPS, the spectra selected from the total set of spectra, with the ranking of \(S/N_g\) in a descending order, are the negative samples. In the algorithm application, SVM has many limitations when it is applied to the binary-classification procedure from imbalanced data sets, in which the negative instances heavily outnumber the positive ones. Although many applications have been raised to overcome this issue (Akkani et al. 2004), we decided to keep a balance between positive and negative samples to ensure the correction of the classifications, meaning that only 300 spectral data are kept in each group of the negative samples. Furthermore, five groups of each CPS are built as negative samples in order to obtain more comprehensive results; a total of 1500 spectra in every CPS in the Table 3.

### 2.2. Data for Recognition

DR12 and DR14 contain huge amounts of spectral data, most of which have a high quality and correct classification. It would be inefficient if all spectra were included in the searching process, especially for “GALAXY” that corresponds to the largest number in the catalog.

We adopt the full spectral template matching program to classify all spectra from the SDSS data set. DB templates are replaced by all templates of the LAMOST 1D Pipeline (Luo et al. 2015), which differs from the steps introduced in Section 2.1. In order to obtain reliable classification results, the relationship of \(\chi^2\) between the best and second-best fit is also taken into consideration. After this preprocessing, spectral data that are not DBWDs with a high degree of confidence will be excluded, while others remain as ED, from which DBWD is recognized in Section 4. The amount of spectral data from each CPS within the ED and catalog are illustrated in Figure 1, using the red and blue bars, respectively. Compared with the SDSS DR14 data set, the ED is eventually built by reducing the quantity by an average of 75 percent, after the reduction process.

### 3. Method in the Training Process

The flowchart of the training process is given in Figure 2. All of the prepared data begin with the preprocessing procedure, including the normalization and redshift measurement. The
spectral features are then extracted through LASSO for each group, followed by a bi-classify modular of SVM to separate the DBWDs from all of the types. If the accuracy (see Section 3.4.2 in detail) is not high enough, an optimization process needs to be performed, i.e., remove some contamination of DB from the negative sample sets and restart this loop. Once the training process is completed, the unique features of the DBWDs and hyper-planes of each group can be derived by LASSO and SVM, respectively. In addition, we analyze the features extracted by LASSO using Wavelet and provide a multi-scale explanation.

3.1. Data Preprocessing

3.1.1. Normalization for Positive and Negative Samples

To ensure the consistency of different spectral data, a simple normalization is required before feature extraction. Let a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ represent a spectrum, where $n (n > 0)$ is
the number of points. The component \( x_i \) represents the flux of the spectrum \( x, i \in \{1, 2, \ldots, n\} \). We simply put all of the flux between \(-1\) and 1,

\[
\hat{x} = \frac{x - \bar{x}}{\sigma_x}
\]

where \( \bar{x} \) and \( \sigma_x \) are the mean value and variance of \( x \), respectively.

3.1.2. Redshift Measurement for Positive Sample Groups

In our investigation, we find that most DB spectra from the SDSS data set have incorrect redshifts, especially for those with large redshifts that are classified as “GALAXY” and “QSO” in the catalog. To ensure consistency of the feature extraction, all samples in the training set should be in the rest frame.

3.2. Feature Extraction

DBWDs account for about 20% of all WDs and have atmospheres dominated by neutral helium, represented as the He I line in the spectrum. Compared with all other spectra, the most significant spectral line in a DBWDs spectrum is the He I at 4471 Å (Kepler et al. 2016). The spectral lines shown in Figure 3 were checked using the atomic line table from the National Institute of Standards and Technology (NIST 2017) to aid the analysis of the feature extraction results. In this plot, we show a typical DB spectrum with all of the He I lines ranging from 3800 to 7400 Å.

Instead of full spectral template matching, ML can be applied to imitate visual recognition to detect the features in noisy spectral data. The line wings, rather than the line center, may be more sensitive to the distinction between DB and non-DB. The LASSO algorithm has the ability to obtain such positions at some particular wavelength as features, which satisfy the demand. In other words, some dominant spectral lines do not work when spectra are similar, such as DB and B that both have a He I line at 5015.7 Å, which cannot be used as a component of the classifier.

3.2.1. LASSO

LASSO is an interpretable model that minimizes the residual sum of matrixes subject to the sum of the absolute value of the coefficients smaller than a constant (Tibshirani 1996). This model is successfully applied to extract linearly supporting...
features from stellar spectra, and the atmospheric parameters are automatically estimated including \( T_{\text{eff}} \), \( \log g \), and \([\text{Fe/H}]\) (Li et al. 2015).

In simple terms, consider a sample consisting of \( N \) spectra, each of which includes \( n \) points. Let \( y_i \) be the outcome and \( x_i \) be the covariate vector for the \( i \)th case. Then, the objective of LASSO is to solve

\[
\hat{w} = \arg \min_w \sum_{i=1}^{N} (y_i - f(x^i; w))^2 + \lambda \|w\|_1,
\]

where

\[
\|w\|_1 = \sum_{i=1}^{n} |w_i|,
\]

\[
f(x^i; w) = \sum_{i=1}^{n} w_i x_i
\]

and \( \lambda > 0 \) is a tuning parameter that controls the value of non-zero parameters, \( w \), and the complexity of the model.

James et al. (2017) has proved that LASSO can effectively filter out most of the irrelevant or redundant variables by reducing the amount of non-zero parameters of \( w_i \). We use the LASSO program based on MATLAB (Efron et al. 2004), in which the parameter \( \lambda \) can be equivalently replaced with the number \( m \) of non-zero parameters, \( w_i \).

### 3.2.2. Feature Selection

Here, we adopt LASSO to extract features between the DBWD and other types of spectra. At first, we build five groups of negative samples for each CPS with the highest \( S/N_g \) (see Table 3 for details), as the features can be affected by data quality or parameters. Each group has 300 negative samples, and we combine all features within a CPS into one as the final output. Features from different wavelength ranges may vary, and most DB spectral lines are, in theory, mainly on the blue band of a spectrum. Thus, positions of the features within \( 3900 \)–\( 5900 \) Å and \( 3900 \)–\( 8900 \) Å need to be analyzed.

1. **Feature with different CPSs.** Features from different CPSs are not similar, which represent the difference between spectra of this type and DB spectra. Figure 4 is one example that illustrates distinct features extracted from WD and QSO groups, which are shown by the short solid blue and red lines, respectively. From the experiment, we note that the features extracted from QSO almost cover all wavelengths, indicating that the difference between QSO and DB is large. This is also the main reason why the majority of QSO spectra with high confidence are excluded from the ED in advance. Conversely, features from the CPSs of A star groups are mainly near \( 4026.2 \), \( 4471.5 \), \( 4713.1 \), \( 4921.9 \), \( 5875.6 \), and \( 4341.7 \) Å, which is the position of the \( \text{HeI} \) and \( \text{H}_\gamma \) lines. Face that none of the \( \text{HeI} \) line exist in normal early A stars, as panel (a) of Figure 2 in Takeda et al. (2007) shows, these locations represent the major differences between A and DB stars.

There are five groups in CPS in which negative samples are all classified as O or OB in the DR14 catalog. Figure 5 shows the features extracted from two groups. It can be shown that the main features of the same classification are similar, but the details are slightly different. All these spectra were observed by the telescope on Earth, which may lead to uncertainties in the data due to noise from sky light or instrument efficiencies. Besides, ML is a data-based approach, providing practical results that differ from the results obtained from the theory of spectral analysis. Therefore, a few more wavelengths, that are not characteristic spectral
lines, are recognized as features between O and DB. We show these features in Table 10 in Section 6.

2. **Feature within different wavelength range.** The characteristics in the blue bands (3900–5900 Å) are not exactly the same as those in full spectrum (3900–8900 Å). The possible cause is changes in the original points, while the wavelength increases. Figure 6, for instance, compares features within 3900–5900 Å (blue, above) and 3900–8900 Å (red, below).

In Figure 6, all features within the blue band of the two groups are almost the same, except for some minor differences. For example, He I 5015.7 Å is an important feature within the wavelength range 3900–5900 Å, but it is not so significant within 3900–8900 Å. In addition, many data points, such as 7010 Å, are also significant characteristics, that cannot be ignored. Despite the importance of the blue band in early type stars and DB WDs, there obviously exists features for wavelengths redder than 5900 Å, such as Hα (6564.6 Å), He I (6678.2, 7065.2, and 7281.4 Å), and other positions that theoretically have no spectral lines in low resolution spectra.

In conclusion, we use the full spectrum to extract features when performing the following experiment.

### 3.3. Features at Multi-scale

Wavelet decomposition is adopted to analyze features of DB WD at multi-scales.

Li et al. (2015) evaluated the performance of various wavelet basis functions and decomposition levels for the estimation of stellar atmospheric parameters, using several evaluation methods. In some situations, the most essential difference between wavelet applications is the selection of the basis function and decomposition level. The efficiency can differ when some variables change, such as the basis function and decomposition level. For example, Meyer and Biorthogonal wavelets can sometimes lead to a large distinction when estimating $T_{\text{eff}}$ of some specific stellar spectra (Panel (e) of Figure 7 in Li et al. 2015).

When it comes to analyzing the distribution of DB WD features extracted by LASSO, the wavelet basis function becomes less important since the classification algorithm is only concerned with the wavelength location of the features. We compare different basis functions and obtain almost similar locations of the features, which is relatively fixed at some wavelengths in the fourth or fifth wavelet coefficients. As a result, we simply employ the simplest basis form—Haar wavelet—to conduct the decompose procedure.

Some of the most important features stay at the same position of line wings on the same scale for DB WDs with various temperatures and gravities. Features extracted from one group are served as an example shown in Figure 5, in which line wings of He I (4026.2, 4471.5, 4921.9, and 5875.6 Å), and He II, Ar II, and Ne II, are recognized as features.

We decompose a spectrum into a low-frequency approximation signal and high-frequency details by wavelet transform, and discover that most features fall on the crest of the detail coefficients at the fourth layer of the wavelet domain. For example, 4471.5, 5875.6, and 6678.2 Å in Figure 7 are the three main typical features in DB.

The features located at the 16th point from the line center or nearby, coincide with the coefficients of the fourth layer of the
wavelet decomposition ($2^4$). This part of the spectral line should become a major character of WDs when distinguishing from other types of spectra. We believe that spectral line broadening is a significant characteristic of WDs and that the positions are the most dramatic changes in the spectral data.

### 3.4. SVM

#### 3.4.1. Hyper-plane

As a supervised ML method, SVM is used in classification and regression analysis. In a dual clustering system, given a set of training examples that are each marked by positive or negative categories, an SVM training algorithm builds a robust binary linear classifier model that sorts new data to one type or the other.

We apply the LIBSVM (Chang & Lin 2011) software to pick DB spectra from all of the data set with features obtained in Section 3.2. LIBSVM is an integrated software for support vector classification, regression, and distribution estimation. It supports multi-type classifications.

Let $(x_i, y_i), i = 1, 2, ..., n$ represent a training set, where $x_i \in \mathbb{R}^n$ and $y \in \{-1, 1\}^n$ are the spectral data at the feature points and label, respectively. LIBSVM tries to seek a linear separating hyper-plane with the maximal margin in this higher dimensional space by solving the following optimization problem:

\[
\begin{align*}
\min_{w,b,\xi} & \quad \frac{1}{2}w^Tw + C\sum_{i=1}^{n}\xi_i \\
\text{subject to} & \quad y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i, \\
& \quad \xi_i \geq 0.
\end{align*}
\]

Here, training vectors $x_i$ are mapped into a higher dimensional space by the function $\phi$ and $C > 0$ is the penalty parameter of the error term. Furthermore, $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ is the kernel function. LIBSVM provides four basic kernels below:

1. Linear: $K(x_i, x_j) = x_i^T x_j$.
2. Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$.
3. Radial basis function (RBF): $K(x_i, x_j) = \exp(-\gamma|x_i - x_j|^2), \gamma > 0$.

---

**Figure 8.** Features of DB vs. O, B, A, F, WD(DA), and QSO, from top to bottom in order. To make it more explicit, the wavelength of the features are marked with red lines below each DB spectrum (black); the other six types of spectra are plotted in blue. (A color version of this figure is available in the online journal.)
4. Sigmoid: \( K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) \).

Here, \( \gamma \), \( r \), and \( d \) are the kernel parameters.

The kernel function and parameters are important for the SVM algorithm to be adjusted. Experiments show that the linear and RBF kernels should provide better discrimination results for spectral data. 10-fold cross-validation is utilized to automatically determine all parameters using LIBSVM software.

### 3.4.2. Verification

In this section, we verify the reliability of the algorithm by labeling all of the testing set. There are some measures for information retrieval and statistical classification to evaluate the quality of the algorithm: accuracy, precision, and recall. The accuracy is based on our prediction, which shows how many of the positive predictions are true positives (TPs). The recall rate shows how many positive examples in the sample were predicted correctly. During the training process, almost all of the positive samples, except a few, have been recognized correctly. Therefore, the recall TP/(TP+FN) can reach an approximate percentage of 100%. Generally, the mean accuracy (TP+TN)/(TP+FN+FP+TN), and precision TP/(TP+FP), of all of the CPS can reach 99.9% and 99.7%, respectively, indicating a very high stability and reliability for this algorithm.

### 4. Recognition

#### 4.1. Input of the SVM

We collect all of the features derived from Section 3.2, and demonstrate some of them obtained from the CPSs of O, B, A, F, WD, and QSO (from top to bottom) in Figure 8.

The features are marked with short red lines between the DB spectra and that of the other CPSs. In general, features on either side of the spectral lines are not perfectly symmetrical. Distinctions near He I 4026.2 Å, for instance, between DB...
and O, only appeared on the right side of this HeI line. It can be intuitively derived that a large number of features are either just a single data points or narrow ranges of wavelengths. When they exist in a relatively long range of wavelength, this part should display the most dramatic changes in the spectral lines, as discussed below. The content of metal elements increases with changes in the stellar types O, B, A, and F, corresponding to a raise in the number of features within the red band.

Afterwards, the DB candidates will be identified from all of the spectra during the searching process. The models of the feature derived from the training process, Section 3.2, are employed as the input of SVM. A flowchart is presented in Figure 9 to demonstrate the start-to-end flow of this procedure.

### 4.2. Recognition and Results

After the reduction procedure described in Section 2.2, the ED is generated as the original data set to conduct recognition. Similar to the preprocessing in Section 3.1, we normalize the ED and then move all of them to the rest frame using cross-correlation.

Then a hyper-plane in feature space is applied by SVM to distinguish the DBWDs from the ED. We inspect the output and find a sample of 2808 spectra of 2029 different objects from SDSS DR12 and DR14, in which 58 are newly identified objects. In Table 4, we illustrate those spectra classified as DBWDs to evaluate the performance of the algorithm model.

![Figure 10. Three typical spectra.](image)

Table 5

| Source          | QSO* | Galaxy* | Star* | Total |
|-----------------|------|---------|-------|-------|
| AMC             | 1    | 1       |       | 2     |
| ELG             | ...  | ...     | 1     | 1     |
| HOT_STD         | 22   | 3       | 514   | 539   |
| LRG             | 2    | ...     | 1     | 3     |
| NONLEGACY       | 24   | 3       | 413   | 440   |
| Null            | 4    | ...     | 3     | 7     |
| QA              | ...  | ...     | 2     | 2     |
| QSO             | 1    | ...     | 20    | 21    |
| QSO_EBOSS_W3ADM | 1    | ...     | 2     | 3     |
| QSO_VAR         | ...  | ...     | 1     | 1     |
| QSO_VAR_SDSS    | 1    | ...     | 1     | 2     |
| ROSAT_D         | ...  | ...     | 2     | 2     |
| SEGUEI          | 5    | ...     | 1     | 6     |
| SEQUELS_TARGET  | 1    | 1       | 3     | 5     |
| SERENDIPITY_BLUE | 29   | 5       | 180   | 214   |
| SERENDIPITY_DISTANT | 23   | 3       | 207   | 233   |
| STAR            | ...  | ...     | 2     | 2     |
| STAR_CATY_VAR   | 1    | ...     | 11    | 12    |
| STAR_WHITE_DWARF | 6    | 3       | 84    | 93    |
| WHITEDWARGE_NEW | 462  | 3       | 236   | 701   |
| WHITEDWARGE_SDSS | 295  | 2       | 188   | 485   |

Total           | 878  | 24      | 1872  | 2774  |

**Notes.**

* Target selection of SDSS DR12 and DR14.

* "Class" and "subclass" are adopted from the data archive of SDSS.
| Designation | P–M–F | Type | $R_{\text{DB}}$ (km s$^{-1}$) | $R_{\text{M}}$ (km s$^{-1}$) | $T_{\text{eff}}$ (K) | log $g$ (cgs) | FUV (mag) | NUV (mag) | S/N | Mass $M_\odot$ | Age Myr | Ref |
|------------|-------|-----|-----------------|-----------------|--------|---------|--------|--------|------|-------------|--------|------|
| J094038.80+364645.6 | 1275–52996–0037 | DB | 449 ± 0 | ... | 18200 ± 2166 | 8.09 ± 0.223 | 19.8 ± 0.2 | 19.1 ± 0.1 | 2.5 | 0.60 | 106.0 | 0 |
| J115601.31+293115.4 | 2224–53815–0171 | DB | −41 ± 31 | ... | 42897 ± 1577 | 7.51 ± 0.088 | 19.5 ± 0.1 | 19.7 ± 0.1 | 3.2 | −9999 | −9999 | 0 |
| J008010.20+272906.1 | 2824–54452–0037 | DBA | 149 ± 32 | ... | 16673 ± 757 | 7.92 ± 0.225 | ... | ... | 4.7 | 0.59 | 130.7 | 0 |
| J222646.14+061921.3 | 4410–56187–0506 | DB | 70 ± 37 | ... | 20120 ± 1567 | 8.75 ± 0.159 | 20.0 ± 0.1 | 19.8 ± 0.1 | 10.8 | 0.91 | 204.3 | 0 |
| J012752.18+140622.9 | 4665–56209–0726 | DBA | −28 ± 16 | ... | 32695 ± 441 | 8.87 ± 0.023 | ... | ... | 22.1 | 1.20 | 148.2 | 0 |
| J222711.11+073510.7 | 5057–56209–0276 | DB | 50 ± 12 | ... | 15391 ± 82 | 8.88 ± 0.029 | 19.7 ± 0.1 | 18.9 ± 0.1 | 22.7 | 1.19 | 1100.0 | 0 |
| J095403.47+223919.2 | 5787–56254–0254 | DB | 118 ± 59 | ... | 13661 ± 339 | 8.17 ± 0.111 | 20.5 ± 0.2 | 19.5 ± 0.1 | 13.2 | 0.59 | 245.9 | 0 |
| J094852.66+233004.1 | 5787–56254–0500 | DB | 58 ± 14 | ... | 17018 ± 157 | 8.75 ± 0.044 | 18.8 ± 0.1 | 18.4 ± 0.1 | 11.1 | 1.19 | 822.0 | 0 |
| J094852.66+233004.1 | 5788–56255–0028 | DB | 68 ± 12 | ... | 17615 ± 129 | 8.77 ± 0.030 | 18.8 ± 0.1 | 18.4 ± 0.1 | 16.7 | 1.19 | 705.9 | 0 |
| J090730.35+270413.6 | 5780–56274–0018 | DB | 113 ± 22 | ... | 18200 ± 275 | 8.88 ± 0.048 | 19.8 ± 0.1 | 19.3 ± 0.0 | 10.1 | 1.19 | 705.9 | 0 |
| J100104.94+302543.5 | 5800–56279–0870 | DB | 27 ± 33 | ... | 13939 ± 110 | 8.92 ± 0.053 | 20.2 ± 0.2 | 19.3 ± 0.1 | 12.2 | 1.19 | 1268.0 | 0 |
| J135815.93+290525.5 | 6009–56313–0624 | DB | 46 ± 3 | ... | 22148 ± 208 | 8.76 ± 0.013 | 17.1 ± 0.1 | 16.9 ± 0.0 | 30.1 | 1.19 | 372.9 | 11 |
| J091256.90+430023.0 | 4687–56338–0324 | DB | 4 ± 24 | ... | 13893 ± 184 | 9.03 ± 0.073 | ... | ... | 11.1 | 1.19 | 1268.0 | 0 |
| J091256.90+430023.0 | 4687–56369–0326 | DB+M1 | 181 ± 2893 | 1147 ± 52 | 17954 ± 409 | 9.32 ± 0.087 | ... | ... | 11.8 | −9999 | −9999 | 0 |
| J091638.24+475253.6 | 5813–56363–0640 | DB | −7 ± 7 | ... | 17900 ± 99 | 8.67 ± 0.024 | 18.4 ± 0.1 | 18.0 ± 0.0 | 17.5 | 0.91 | 278.0 | 0 |
| J142046.13+554201.4 | 6803–56402–0201 | DB | 217 ± 60 | ... | 17200 ± 778 | 8.34 ± 0.179 | ... | ... | 3.0 | 0.91 | 326.3 | 0 |
| J091534.70+513610.3 | 5729–56958–0121 | DB | 37 ± 17 | ... | 16071 ± 115 | 8.77 ± 0.041 | 19.4 ± 0.2 | 19.0 ± 0.1 | 12.1 | 1.19 | 951.5 | 0 |
| J092540.56+511229.6 | 5730–56607–0940 | DB | 45 ± 30 | ... | 17247 ± 293 | 8.78 ± 0.049 | ... | ... | 9.6 | 1.19 | 822.0 | 0 |
| J012644.96–025633.9 | 7877–56898–0048 | DB | 172 ± 2893 | ... | 13162 ± 949 | 8.89 ± 0.460 | ... | ... | 2.6 | 1.19 | 1458.0 | 0 |
| J080710.33+485259.6 | 7324–56935–0828 | AMCVa | 462 ± 56 | ... | 8260 ± 23 | 8.18 ± 0.032 | ... | ... | 3.5 | −9999 | −9999 | 0 |
| J231213.74+185713.8 | 7611–56496–0897 | DBO | 1260 ± 593 | ... | 26200 ± 18432 | 6.33 ± 1.958 | ... | ... | 1.9 | −9999 | −9999 | 0 |
| J022756.30+445404.2 | 8127–56957–0899 | DB | 1245 ± 107 | ... | 7653 ± 7143 | 7.27 ± 38.056 | ... | ... | 0.7 | −9999 | −9999 | 0 |
| J012920.84+191241.5 | 7628–56978–0465 | DB | 877 ± 48 | ... | 9228 ± 221 | 7.23 ± 0.141 | ... | ... | 0.8 | 0.19 | 351.9 | 0 |
| J235607.30+025254.8 | DBZ | 99 ± 38 | ... | 16037 ± 236 | 8.86 ± 0.087 | 20.5 ± 0.2 | 20.0 ± 0.1 | 6.8 | 1.19 | 951.5 | 0 |
| Designation       | P–M–F          | Type | RV$_{DB}$ (km s$^{-1}$) | RV$_M$ (km s$^{-1}$) | $T_{\text{eff}}$ (K) | log g (cgs) | FUV (mag) | NUV (mag) | S/N | Mass $M_\odot$ | Age Myr | Ref |
|------------------|----------------|------|-------------------------|----------------------|---------------------|------------|---------|----------|-----|----------------|---------|-----|
| J020022.48+242343.3 | 7849–56980–0914 | DB   | 1245 ± 1076            | ...               | 18200 ± 6743       | 6.08 ± 3.551 | ...      | ...       | -0.0 | -9999         | -9999   | 0   |
| J074325.35+432027.7 | 7692–57064–0409 | DB   | 149 ± 44               | ...               | 13841 ± 1083       | 8.88 ± 0.500   | ...      | ...       | 2.7  | 1.19          | 1268.0  | 0   |
| J005436.04–041940.6 | 8276–57067–0470 | DBA  | 268 ± 28               | ...               | 16939 ± 508        | 6.58 ± 0.162   | 18.9 ± 0.1 | 19.4 ± 0.1 | 5.5  | -9999         | -9999   | 0   |
| J013634.37–001109.9 | 7912–57310–0460 | DBA  | 27 ± 58                | ...               | 16415 ± 200        | 8.88 ± 0.056   | 20.3 ± 0.2 | 20.0 ± 0.1 | 9.4  | 1.19          | 951.5   | 0   |
| J234924.30–025209.5 | 8792–57364–0358 | DB   | 1152 ± 70              | ...               | 17200 ± 4598       | 6.19 ± 1.638   | ...      | ...       | 1.6  | -9999         | -9999   | 0   |
| J232933.23+212015.6 | 7851–56932–0403 | DB   | 1127 ± 77              | ...               | 18273 ± 1310       | 8.95 ± 0.137   | ...      | ...       | 2.2  | 1.19          | 705.9   | 0   |
| J082623.07+555006.2 | 7375–56981–0144 | DB   | -299 ± 87              | ...               | 15200 ± 1122       | 9.46 ± 0.366   | ...      | ...       | 6.7  | -9999         | -9999   | 0   |
| J081453.04+555033.1 | 7375–56981–0487 | DB   | -149 ± 74              | ...               | 17200 ± 738        | 8.95 ± 0.178   | ...      | ...       | 3.2  | 1.19          | 822.0   | 0   |
| J234527.43+215712.3 | 7600–56984–0082 | DB   | 1315 ± 696             | ...               | 28200 ± 3881       | 8.77 ± 0.184   | ...      | ...       | 5.2  | 1.19          | 184.3   | 0   |
| J234131.56+224240.6 | 7600–56984–0313 | DB+M9 | 67 ± 35                | 1174 ± 30         | 21428 ± 791        | 8.84 ± 0.049   | ...      | ...       | 9.5  | 1.19          | 436.0   | 0   |
| J025818.61+004131.3 | 7820–56984–0106 | DB   | 308 ± 134              | ...               | 20158 ± 1915       | 7.90 ± 0.203   | ...      | ...       | 3.2  | 0.60          | 68.8    | 0   |
| J025720.06–003812.0 | 7820–56984–0182 | DB   | 884 ± 59               | ...               | 15400 ± 992        | 7.84 ± 0.282   | ...      | ...       | 2.1  | 0.59          | 198.7   | 0   |
| J225345.44+223258.6 | 7613–56988–0380 | DB+M9 | 29 ± 37                | 1164 ± 41         | 12473 ± 486        | 9.35 ± 0.190   | 22.0 ± 0.3 | 21.1 ± 0.1 | 7.8  | -9999         | -9999   | 0   |
| J093021.79+544359.5 | 7285–56991–1000 | DB   | -2098 ± 81             | ...               | 17194 ± 761        | 8.95 ± 0.222   | ...      | ...       | 3.3  | 1.19          | 822.0   | 0   |
| J080128.06+554004.7 | 7281–57007–0548 | DB   | 29 ± 795               | ...               | 17707 ± 1093       | 7.66 ± 0.262   | 21.4 ± 0.4 | 21.2 ± 0.3 | 2.0  | 0.36          | 56.2    | 0   |
| J023303.84–022104.8 | 7829–57011–0397 | DB   | 1245 ± 1578            | ...               | 16429 ± 1429       | 8.35 ± 0.493   | ...      | ...       | 1.8  | 0.91          | 385.3   | 0   |
| J084704.97+511056.0 | 7303–57013–0675 | DB   | 1245 ± 2586            | ...               | 15650 ± 1067       | 8.58 ± 0.502   | ...      | ...       | 1.5  | 0.91          | 385.3   | 0   |
| J104624.26+490908.9 | 7387–57038–1000 | DB   | 173 ± 41               | ...               | 36204 ± 1786       | 8.60 ± 0.099   | 20.7 ± 0.3 | 20.5 ± 0.2 | 3.2  | 0.93          | 16.1    | 0   |
| J091854.84+515603.2 | 7289–57039–0049 | DB   | -1498 ± 76             | ...               | 20200 ± 3648       | 7.45 ± 0.617   | ...      | ...       | 2.8  | 0.37          | 39.1    | 0   |
| J010633.14+203043.1 | 7624–57039–0927 | DB   | 1191 ± 23              | ...               | 21763 ± 1440       | 8.59 ± 0.111   | ...      | ...       | 3.0  | 0.91          | 152.4   | 0   |
| J090702.69+430612.5 | 8282–57041–0092 | DB   | 1140 ± 97              | ...               | 15436 ± 1381       | 7.21 ± 0.535   | ...      | ...       | 3.5  | 0.23          | 77.4    | 0   |
| J085309.15+584336.3 | 8197–57064–0537 | DBAZ  | 29 ± 62                | ...               | 16715 ± 466        | 9.23 ± 0.112   | ...      | ...       | 9.2  | 1.19          | 822.0   | 0   |
| J103107.17+520854.7 | 7849–56980–0914 | DB   | 216 ± 24               | ...               | 16535 ± 531        | 8.76 ± 0.146   | 21.3 ± 0.5 | 20.7 ± 0.2 | 3.1  | 1.19          | 822.0   | 0   |
### Table 6
(Continued)

| Designation   | P–M–F   | Type | RV\(_{DB}\) (km s\(^{-1}\)) | RV\(_{M}\) (km s\(^{-1}\)) | \(T_{\text{eff}}\) (K) | log g (cgs) | FUV (mag) | NUV (mag) | S/N | Mass \(M_\odot\) | Age Myr | Ref |
|---------------|---------|------|-----------------------------|-----------------------------|------------------------|-------------|-----------|-----------|-----|----------------|---------|-----|
| J112752.26+565539.5 | 8167–57071–0548 | DB | 449 ± 98 | ... | 17200 ± 878 | 7.94 ± 0.206 | ... | ... | 2.2 | 0.59 | 130.7 | 0 |
| J120156.39+493707.4 | 7423–57135–0578 | DB+M2 | 29 ± 37 | 1146 ± 8 | 29415 ± 3758 | 8.80 ± 0.168 | 21.8 ± 0.5 | 20.9 ± 0.1 | 3.0 | 1.19 | 166.8 | 0 |
| J121027.14+502735.7 | 7423–57135–0833 | DBZ | 172 ± 37 | ... | 15530 ± 307 | 9.07 ± 0.096 | 20.4 ± 0.3 | 19.7 ± 0.1 | 13.5 | 1.19 | 951.5 | 0 |
| J214544.31+270923.4 | 7641–57307–0622 | DB | 1245 ± 364 | ... | 21503 ± 2935 | 7.90 ± 0.210 | ... | ... | 2.5 | 0.60 | 43.8 | 0 |
| J223318.14+244812.3 | 7654–57330–0204 | DB | 618 ± 43 | ... | 36971 ± 2568 | 8.74 ± 0.135 | ... | ... | 2.0 | 0.93 | 13.6 | 0 |
| J231304.25+265057.9 | 7703–57333–0554 | DB | 659 ± 43 | ... | 14268 ± 1085 | 6.83 ± 0.488 | ... | ... | 1.7 | 0.22 | 96.8 | 0 |
| J005242.64+285411.0 | 7674–57359–0834 | DB | 570 ± 45 | ... | 18200 ± 840 | 9.19 ± 0.180 | 21.6 ± 0.4 | 21.1 ± 0.3 | 3.1 | 1.19 | 705.9 | 0 |
| J001334.89+262424.5 | 7694–57359–0405 | DB | −999 ± 23 | ... | 16025 ± 343 | 9.44 ± 0.161 | ... | ... | 4.3 | −9999 | −9999 | 0 |
| J001627.53+281843.5 | 7694–57359–0737 | DB | 703 ± 66 | ... | 16495 ± 606 | 7.90 ± 0.165 | ... | ... | 2.6 | 0.59 | 161.0 | 0 |
| J075925.84+414454.4 | 8291–57391–0933 | DB+M9 | 29 ± 68 | 1169 ± 61 | 16303 ± 473 | 8.23 ± 0.108 | 18.8 ± 0.1 | 19.5 ± 0.1 | 8.4 | 0.59 | 161.0 | 0 |
| J014803.85+005317.7 | 8793–57391–0826 | DB | 29 ± 55 | ... | 16485 ± 510 | 8.85 ± 0.143 | 22.1 ± 0.4 | 21.7 ± 0.2 | 2.3 | 1.19 | 951.5 | 0 |
| J131316.23+51428.8 | 8210–57426–0532 | DBA | −2 ± 41 | ... | 14982 ± 554 | 8.61 ± 0.261 | ... | ... | 2.0 | 0.91 | 459.1 | 0 |

(This table is available in machine-readable form.)
In this part, we could not provide the precise ratio mentioned above because we cannot confirm how many real DBWDs reside in the predication negative category. Assuming all labeled negative samples are non-DB, then the mean percent of correctly identified samples using the algorithm can reach 99.5%.

Clearly, most DBWDs are identified from O, B, WD, and QSO in the DR14 data set. In a different way, this also indicates that the qualities, such as spectral lines, of DBWDs are more likely with these types; or perhaps they are usually mixed together when matched with the full spectrum instead of some particular wavelengths (or features).

The target selections of all DBWDs are given in Table 5.

Table 7
Columns Provided in Table 6 and Online Table

| Column No. | Heading        | Description                                                                 |
|------------|----------------|-----------------------------------------------------------------------------|
| 1          | Designation    | SDSS object name (SDSS 2000J+)                                             |
| 2          | P–M–F          | SDSS Plate number–Modified Julian date–Fiber                               |
| 3          | Type           | Classification of objects derived from ML method                             |
| 4          | RVDB           | Radial velocity and uncertainty of each spectrum (km s\(^{-1}\))            |
| 5          | RV\(_M\)       | Radial velocity and uncertainty of M companions (km s\(^{-1}\))             |
| 6          | \(T_{\text{eff}}\) | Effective temperature (K)                                                   |
| 7          | \(\log g\)     | Surface gravity (cgs)                                                       |
| 8          | FUV            | Magnitude of FUV from GALEX, \(-9999\): there is no corresponding value (mag) |
| 9          | NUV            | Magnitude of NUV from GALEX, \(-9999\): there is no corresponding value (mag) |
| 10         | S/N            | Median S/N from catalog of SDSS DR14                                        |
| 11         | Mass           | Obtained from Bergeron & Gilles Fontaine (2016) \((M_\odot)\)               |
| 12         | Age            | Obtained from Bergeron & Gilles Fontaine (2016) \((\text{Myr})\)            |
| 13         | Ref            | ID of the literature, 0: newly identified in this paper, see Section 5.1.1 for detail |

Table 8
Numbers of Identified DBWD Types

| Type       | No. of Objects | No. of Spectra |
|------------|----------------|----------------|
| DB         | 1895           | 1395           |
| DB+Ma      | 89             | 79             |
| DB+DC      | 23             | 21             |
| DBA\(^b\)  | 627            | 465            |
| DBO        | 23             | 18             |
| DBQ        | 5              | 4              |
| DBZ        | 112            | 81             |

Notes.

\(^a\) The subtype and RV of the M companion can be found in the online table.

\(^b\) Some of the DBA are actually DBAZ or DBAQ; they are all counted in “DBA.”

In this part, we could not provide the precise ratio mentioned above because we cannot confirm how many real DBWDs reside in the predication negative category. Assuming all labeled negative samples are non-DB, then the mean percent of correctly identified samples using the algorithm can reach 99.5%.

Clearly, most DBWDs are identified from O, B, WD, and QSO in the DR14 data set. In a different way, this also indicates that the qualities, such as spectral lines, of DBWDs are more likely with these types; or perhaps they are usually mixed together when matched with the full spectrum instead of some particular wavelengths (or features).

The target selections of all DBWDs are given in Table 5.

Many spectra, with sources that are stars, are mis-classified into QSO, such as “WHITEDWARF_NEW” listed in Table 5. We believe this is due to a weakness in the algorithm of pipeline. Without an efficient feature wavelength, many spectra may not be correctly classified by full spectral template matching. There are also some quite broad spectral lines in both the QSO and WD (DB) spectra that may mislead the classification results.

5. Analysis

5.1. Comparison with the Literature

Altogether, there are 1309 pure DB objects, including double stars, in the literature (see Section 5.1.1). In this paper, we present 1999 objects of DBWDs (including, but not limited to, DB, DBA, or DBZ) with 2774 spectra in SDSS DR12 and DR14, among which 58 objects are newly spectroscopically confirmed.

A total of 176 pure DB spectra from the literature are omitted in our catalog. Most of them (96 spectra) have no apparent He I lines or only one possible He I line so that our method could not be used to recognize them. We present one example in Figure 10; SDSS Plate–MJD–Fiber (p–m–f), 0804–52286–0262. Another
38 spectra are generally of very poor quality, and the hyperplanes in the feature spaces can be ineffective. Besides, an incorrect radial velocity (RV) may lead to a failure of recognition, which includes 33 of these spectra. As for measuring RV of A type star using DB templates, there are no He I lines in A-type star spectra and only He I lines in DBWD spectra, the errors of RV arise because Balmer lines in the A-type star spectrum are being identified with He I lines in the DB spectrum.

Nine spectra are supposed to be DAB instead of DB. It is worth mentioning that some spectra, for example SDSS J092604.91+264225.0, which is mis-classified as DA in Kleinman et al. (2013), are typical DBWDs. More details can be found in Table 6 and the online table. Table 7 lists the columns of data provided in our online catalog, Table 6.

Furthermore, we add 15 pure DB spectra of 14 objects in the SDSS DR14. Consider the number of DB objects presented in the literature and our catalog in SDSS DR12; the completeness of our ML method should be about 96.0%. Strong noises at the wave range of features may cause mis-classifications in this paper, which is the main disadvantage of SVM that is in need of improvement.

In general, Table 8 lists the numbers of each types of DBWD identified in this study.

5.1.1. ID of Literature

In the last column of Table 6, the numbers represent the IDs of specific literatures, which are listed as follows.

0: first reported in this paper; 1: Kleinman et al. (2013); 2: Koester & Kepler (2015); 3: Kepler et al. (2015); 4: Kepler et al. (2016); 5: Atlee & Gould (2007); 6: Eisenstein et al. (2006a); 7: Croom et al. (2004); 8: Croom et al. (2001); 9: Rebassa-Mansergas et al. (2010); 10: West et al. (2008); 11: Girven et al. (2011); 12: Stepanian (2005); 13: Bicay et al. (2000); 14: Gentile Fusillo et al. (2015); 15: Levitan et al. (2015); 16: Drake et al. (2014); 17: Vennes et al. (2011); 18: Rau et al. (2010); 19: Carter et al. (2013); 20: Jura & Xu (2012); 21: Girven et al. (2012); 22: Bergeron et al. (2011); 23: Zuckerman et al. (2010); 24: Voss et al. (2007); 25: Koester et al. (2005); 26: McCook & Sion (1999); 27: Bradley (2000); 28: Bradley (1998); 29: Lépine & Shara (2005); 30: Stepanian et al. (1999); 31: Stark & Wade (2003); 32: Calcaferro et al. (2017); 33: Kleinman et al. (2004).

5.2. Noteworthy Individual Objects

Panel “b” in Figure 10 shows the AM Canum Venaticorum (AM CVn) type spectrum (p–m–f 7324–56935–0828) that we spectroscopically identified for the ML method that only requires the intensity of the change. The AM CVn binaries are a rare ultra-compact double degenerate system and only 43 such objects are known (Campbell et al. 2015; Levitan et al. 2015).

We provide 66 DB spectra with M type stars as companions. However, there are more than 30 DB M double stars in the literature (Kleinman et al. 2013; Kepler et al. 2015, 2016). The reason why these double stars cannot be discovered by our method is that the flux of M exceeds that of DB, which could lead to much weaker features of the DB in a spectrum. After a visual inspection, we select 23 “DB+M” double stars with relative good qualities from the literature. For these 89 double stars, we provide subtype and RV of the M companion in Table 6. One DB+M spectrum (p–m–f 1057–52522–0613) together with templates of DB and M are illustrated in the panel “c” of Figure 10.
Koester & Kepler (2015). With the method of full spectral template matching mentioned in Section 2.1, we measure $T_{\text{eff}}$ and $\log g$ on several He I lines, and presented the results in Table 6. The average errors of $T_{\text{eff}}$ and $\log g$ are 30.1% and 10.6%, respectively.

5.4. $T_{\text{eff}}$ and Ultra-violet Color

WDs are a type of stars that have strong intensities in the ultraviolet waveband. The FUV–NUV color from the Galaxy Evolution Explorer (GALEX) is almost reddening-free (Bianchi et al. 2017). All DB objects with photometric data of GALEX are cross-matched and those objects with errors in both FUV and NUV less than 0.3 mag are selected. From Figure 11, we conclude that $T_{\text{eff}}$ and FUV–NUV color are roughly linear using Equation (1). The fitting variance is $\sigma \approx 0.19$:

$$y = -0.90 \times 10^{-4}x + 2.05,$$

where $x$ is $T_{\text{eff}}$ and $y$ is the FUV–NUV color. The majority of sources fall within the $\pm 3\sigma$ region of Equation (1), which is illustrated by the red dashed line in Figure 11.

6. Conclusion and Discussion

We have spectroscopically identified 1999 DBWDs in the SDSS DR12 and DR14, including 58 newly identified objects, using ML, i.e., LASSO and SVM. A total of 176 DB objects from the literature are not included in our catalog, $T_{\text{eff}}$ mostly varies around 11,000 K, and $\log g$ is fixed at 8.0. The DB spectra in this parameter range have almost no He I lines; hence, our method failed to identify these spectra.

Features of DB versus several other types of spectra were also extracted by LASSO using this procedure. Although we cannot guarantee the completeness of our samples, we have proposed a significant scheme to extract linearly supporting features from spectra to identify DBWDs.
Furthermore, we define all of the features illustrated in Figure 12, and Tables 8 and 9. A DB spectrum with high S/N is plotted in blue, and the features in red with a light gray background. The features within the area of most of the HeI lines are asymmetric about the two sides, illustrating the characteristic of HeI lines in a DB spectrum. These extracted features are demonstrated in Table 9, and do not differ from one or two atomic lines; the spectral line’s name and position are also given in this table. We consider the flux in the range 4693.5–4699.0 Å as a characteristic of HeI 4713.1 Å although it appears to go beyond the range of this spectral line because the flux of a DB spectrum begins to decrease slightly in this part. This kind of tiny variation is usually overlooked by the human eye, but is observable when using programs to carry out the classifications. More positions of other features are also defined in a similar fashion. The width of the feature to the left of some spectral lines, such as HeI 5875.6 Å, is smaller than that on the right. On the other side, many features that are not linked to any specific spectral lines are present in Table 10. These features are purely data based or due to the residual sky background.

Finally, we measure the parameters, \( T_{\text{eff}} \) and \( \log g \), of DBWDs using DB templates from Koester & Kepler (2015). The consistency of \( T_{\text{eff}} \) of DB objects between Koester & Kepler (2015) and our catalog is demonstrated through the FUV–NUV colors from GALEXY. The distribution of \( \log g \) also indicates capability of our method of searching for DBWDs in fainter objects.

The authors would like to thank Drs. Hai-Feng Yang, Peng Wei and Zhen-Ping Yi for valuable discussion, and also appreciate Dr. Anthony E. Lynas-Gray for the helpful suggestions. This work was funded by the National Basic Research Program of China (973 program, 2014CB845700) and National Natural Science Foundation of China (grant No. 11390371/4). This work is supported by the Astronomical Big Data Joint Research Center, co-founded by the National Astronomical Observatories, Chinese Academy of Sciences and the Alibaba Cloud. 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 Astrofísica 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.

ORCID iDs
A-Li Luo, https://orcid.org/0000-0001-7865-2648

References
Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543
Akbani, R., Kwek, S., & Japkowicz, N. 2004, in Machine Learning: ECML 2004, ed. J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Berlin: Springer Berlin Heidelberg), 39
Alam, S., Albareti, F. D., Allende Prieto, C., et al. 2015, ApJS, 219, 12
Atlee, D. W., & Gould. A. 2007, ApJ, 664, 53
Beauvachamp, A., Wesemael, F., Bergeron, P., Liebert, J., & Saffer, R. A. 1996, in ASP Conf. Ser. 96, Hydrogen Deficient Stars, ed. C. S. Jeffery & U. Heber (San Francisco, CA: ASP), 295
Bergeron, P., & Gilles Fontaine, P. 2016, Synthetic Colors and Evolutionary Sequences of Hydrogen- and Helium-Atmosphere White Dwarfs, http://www.astro.umontreal.ca/~bergeron/CoolingModels/
Bergeron, P., Wesemael, F., Dufour, P., et al. 2011, ApJ, 737, 28
Bianchi, L., Shiao, B., & Thilker, D. 2017, ApJS, 230, 24
Bicak, M. D., Stepian, J. A., Chavushyan, V. H., et al. 2000, A&AS, 147, 169
Bradley, P. A. 1998, BalTa, 7, 355
Bradley, P. A. 2000, BalTa, 9, 485
Bues, I. 1970, A&A, 7, 91
Calcaferro, L. M., Córreo, A. H., & Althaus, L. G. 2017, A&A, 600, A73
Campbell, H. C., Marsh, T. R., Fraser, M., et al. 2015, MNRAS, 452, 1060
Carter, P. J., Marsh, T. R., Steeghs, D., et al. 2013, MNRAS, 429, 2143
Chang, C.-C., & Lin, C.-J. 2011, ACM Transactions on Intelligent Systems and Technology, 2, 27
Croom, S. M., Smith, R. J., Boyle, B. J., et al. 2001, MNRAS, 322, L29
Croom, S. M., Smith, R. J., Boyle, B. J., et al. 2004, MNRAS, 349, 1397
Drake, A. J., Graham, M. J., Djorgovski, S. G., et al. 2014, ApJS, 213, 9
Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. 2004, arXiv:math:0406456
Eisenstein, D. J., Liebert, J., Harris, H. C., et al. 2006a, ApJ, 667, 40
Eisenstein, D. J., Liebert, J., Koester, D., et al. 2006b, ApJ, 132, 676
Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, ApJ, 142, 72
Fontaine, G., Brassard, P., & Bergeron, P. 2001, PASP, 113, 409
Gentile Fusillo, N. P., Gänsicke, B. T., & Greiss, S. 2015, MNRAS, 448, 2260
Girven, J., Brinkworth, C. S., Farhi, J., et al. 2012, ApJ, 749, 154
Girven, J., Gänsicke, B. T., Steeghs, D., & Koester, D. 2011, MNRAS, 417, 1210
Ingrid D. 1992, Ten Lectures on Wavelets
James, G., Witten, D., Hastie, T., & Tibshirani, R. 2017, An Introduction to Statistical Learning (1st ed.; Berlin: Springer)
Jura, M., & Xu, S. 2012, AJ, 143, 6
Kepler, S. O., Kleinman, S. J., Nitta, A., et al. 2007, MNRAS, 375, 1315
Kepler, S. O., Pelisoli, I., Koester, D., et al. 2015, MNRAS, 446, 4078
Kepler, S. O., Pelisoli, I., Koester, D., et al. 2016, MNRAS, 455, 3413
Kleinman, S. J., Harris, H. C., Eisenstein, D. J., et al. 2004, ApJ, 607, 426
Kleinman, S. J., Kepler, S. O., Koester, D., et al. 2013, ApJS, 204, 5
Koester, D., & Kepler, S. O. 2015, A&A, 583, A86
Koester, D., Rollehnagen, K., Napierwotzki, R., et al. 2005, A&A, 432, 1025
Lee, Y. S., Beers, T. C., Sivaraman, T., et al. 2008, AJ, 136, 2022
Lépine, S., & Shara, M. M. 2005, AJ, 129, 1463
Levitan, D., Groot, P. J., Prince, T. A., et al. 2015, MNRAS, 446, 391
Li, X., Lu, Y., Conte, G., et al. 2015, ApJS, 218, 3
Limoges, M.-M., & Bergeron, P. 2010, ApJ, 714, 1037
Stepanian, J. A., Chavushyan, V. H., Carrasco, L., Tovmassian, H. M., & Erastova, L. K. 1999, PASP, 111, 1099
Takeda, Y., Kawanomoto, S., & Ohishi, N. 2007, PASJ, 59, 245
Tibshirani, R. 1996, Journal of the Royal Statistical Society, 58, 267
Vennes, S., Kawka, A., & Németh, P. 2011, MNRAS, 410, 2095
Voss, B., Koester, D., Napiwotzki, R., Christlieb, N., & Reimers, D. 2007, A&A, 470, 1079
West, A. A., Hawley, S. L., Bochanski, J. J., et al. 2008, AJ, 135, 785
Zuckerman, B., Melis, C., Klein, B., Koester, D., & Jura, M. 2010, ApJ, 722, 725

Luo, A., & Zhao, Y. 2001, NCimB, 116, 879
Luo, A.-L., Zhao, Y.-H., Zhao, G., et al. 2015, RAA, 15, 1095
McCook, G. P., & Sion, E. M. 1999, ApJS, 121, 1
NIST 2017, Physical Measurement Laboratory, National Institute of Standards and Technology, https://www.nist.gov/pml
Rau, A., Roelofs, G. H. A., Groot, P. J., et al. 2010, ApJ, 708, 456
Rebassa-Mansergas, A., Gänsicke, B. T., Schreiber, M. R., Koester, D., & Rodríguez-Gil, P. 2010, MNRAS, 402, 620
Stark, M. A., & Wade, R. A. 2003, AJ, 126, 1455
Stepanian, J. A. 2005, RMxAA, 41, 155

18