Classification of Stars using Stellar Spectra collected by the Sloan Digital Sky Survey

Michael Brice
Computer Science Department
Central Washington University
Ellensburg, WA, USA
michael.brice@cwu.edu

Răzvan Andonie
Computer Science Department
Central Washington University
Ellensburg, WA, USA
andonie@cwu.edu

***Abstract*—The classification of stellar spectra is a fundamental task in stellar astrophysics. There have been many explorations into the automated classification of stellar spectra but few that involve the Sloan Digital Sky Survey (SDSS). We use the SDSS dataset since it is the most important stellar spectra database available today. In our approach, we apply redshift corrections to the spectra and reduce the number of flux measurements by feature selection. Then we apply standard classifier methods: Random Forest and Support Vector Machine. We compare the accuracy of feature selection and classifier combinations for redshifted stellar spectra and rest stellar spectra. Even though redshifted stellar spectra create feature matrix discrepancies, classifiers utilizing redshifted stellar spectra perform with high accuracy. This creates a viable option for automated classification of stellar spectra without having to identify the redshift value.**

***Index Terms*—Stellar Spectra Classification, Redshift, Feature Selection, Random Forest, Support Vector Machine, Chi Squared, Fisher, Imbalanced Data, Sloan Digital Sky Survey**

I. Introduction

Stellar classification is a fundamental task in stellar astrophysics. Traditionally, stellar spectra are classified using statistical tests. Stellar spectra are either classified by determining the wavelengths of emission and absorption lines using wavelet transformations, statistical analysis, and using references to the Harvard spectral classification system or they are classified by comparing the approximate curve of the spectra to that of templates using statistical tests. The traditional classification schemes require complex data transformations and analysis to identify the class of a star based on its spectrum.

Presently, the Sloan Digital Sky Survey (SDSS) is creating the most detailed three-dimensional maps of the Universe ever made, with deep multi-color images of one-third of the sky, and spectra for more than three million astronomical objects [1]. The SDSS provides stellar spectra with redshift wavelengths. In our work, we will use the SDSS data run 14 optical spectra dataset for stars, the most important stellar spectra database available today. So far, there are only a few classification results reported for this database.

In previous work related to machine learning classification of stellar spectra, Xing and Guo [2], Zhang *et al.* [3], and Yi and Pan [4] used stellar spectra sources such as Pickles [5] and Jacoby [6]. Bazarghan and Gupta [7] also used the stellar spectra sources Jacoby and SDSS [8].

Xing and Guo [2] used Principle Component Analysis (PCA) and wavelet reduction to reduce the large number of flux measurements, which for SDSS range from 3,800 to 4,000 measurements. We have to note that they reduced the number of flux measurements without using machine learning feature selection or extraction methods. Zhang *et al.* [3] pre-processed the data by filtering out the high frequency components of stellar spectra using continuum normalization and did not reduce the number of features.

Recently, Pasquet-Itam *et al.* applied several methods to classify quasars in the SDSS Stripe 82 and also to predict the photometric redshifts of quasars [9]. SDSS Stripe 82 is a different SDSS dataset than data run 14 optical spectra for stars. It is important to note that a quasar is an extremely luminous galactic nuclei, fundamentally different from a star.

Pasquet-Itam *et al.* used Convolutional Neural Network (CNN), K-nearest neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and a Gaussian Process classifier. The input for the CNN were Light Curve Images which are built from light curves of each object in the five ugriz filters, so as to include both the crucial information of the variability and the colors in the learning of the network. The *ugriz filters* are u (Ultraviolet), g (Green), r (Red), i (Near Infrared), z (Infrared). Theses filters are like a filter on a camera and it only lets certain wavelengths through, which is not data filtering. A *light curve* is the apparent, or differential, brightness or magnitude taken over multiple nights [10]. For SDSS Stripe 82, the maximum number of nights of observation is 3,340 nights [9]. Our results do not compare to the results found in [9] because our data is significantly different: Pasquet-Itam *et al.* utilize light curve images where we utilize spectra.

The amount of astronomical data increases exponentially, and the complexity and dimension of astronomical data are also growing rapidly. The SDSS is a representative example. Extracting information from such data becomes a challenging problem. For example, some classification algorithms can only be employed in the low-dimensional spaces, so feature selection and feature extraction become important topics. A review of feature selection methods for high dimensional data in astronomy can be found in [11].

Our contribution is a novel approach to stellar classification.

In contrast to previous approaches, our method is characterized by the following:

- We avoid complex transformation and statistical analysis of the spectra space by using machine learning.
- We use spectra without redshift corrections.
- We use feature selection to reduce the number of flux measurements. Feature selection may destroy the shape and the structure of each stellar spectrum by only using the most relevant flux measurements. This is in contrast to [2], where PCA and wavelet reduction are used to reduce the number of flux measurements while maintaining the shape and structure of each spectra.

For the SDSS spectra, we obtained the best results using the Fisher feature selection and the RF classifier. Our method can be used to classify stars using redshifted stellar spectra (e.g. stellar spectra that has not been redshift corrected). Our experimental results outperform the ones reported in [2], where wavelet reduction and SVMs were used for SDSS spectra.

The structure of the paper is as follows. Section II provides relevant astronomy background. Section III introduces our approach to stellar classification. In Section IV the experimental setup and results are explained. Section V presents our conclusions.

II. Astronomy Background

We start by introducing some basic definitions related to astronomy. Details may be found in standard sources like [12] and [10].

A. Stellar Spectra

The *spectra* are defined as the way in which light is distributed with wavelength [10]. Incandescent light bulbs, when viewed with the unaided eye, emit white light, but when viewed through a prism or a diffraction grating, the bulb appears to be emitting a rainbow or spectrum of light. Stellar spectra work the same way. Stellar spectra are collected by a spectrograph, which utilizes a fine tuned diffraction grating to collect the spectrum of a star. Unlike the example with the incandescent light bulb, stellar spectra have interesting properties.

The stellar spectra from the SDSS dataset contain wavelength measurements and flux measurements. The *wavelengths* are discrete values that represent a range of wavelengths of light. For example, a wavelength value can be 6,000$\pm$5 $\mathring{A}ngstr\ddot{o}ms$ ($\mathring{A}$). *Flux measurements* are the number of photons that pass through an area per second per measured wavelength. The shape and features of stellar spectra are defined by the flux measurements at particular wavelength values. Fig. 1 shows an example spectrum of a flux scaled G2 (Sun like) star. The downward spikes seen in Fig. 1 are absorption lines.

Absorption lines are one of the many interesting properties found in stellar spectra. Absorption lines indicate the types of elements present in a star such as hydrogen, helium, and heavy metals and are key to classifying stars.

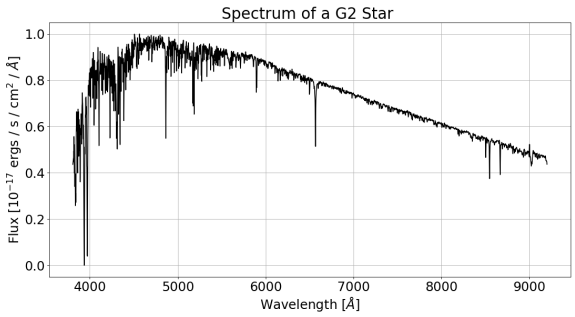

Fig. 1. Example: Stellar Spectrum of a flux scaled G2 star collected by the SDSS.

B. Stellar Classification Types

There are two types of stellar classification schemes: The Harvard spectral classification and the Morgan - Keenan Luminosity Classes (M-K) [12]. The Harvard spectral classification is a surface temperature classification that uses the spectra from stars to categorize into groups labeled O, B, A, F, G, K, and M. These groups are then divided into 0 - 9. For example, A0 - A9 where A0 is an early A type star and A9 as a late A star [12]. The Harvard spectral classification is a one dimensional system that utilizes absorption lines to characterize stellar types. The M-K Classes extends on the Harvard spectral classification by adding the luminosity of the star. The M-K adds the identifiers of I to VI for Super Giants, Bright Giants, Giants, Sub-Giants, Main Sequence (Dwarfs), and Sub-Dwarfs [12]. For example, Betelgeuse is a red super giant and a stellar class of M1 (Harvard) Ia (M-K) star and Proxima Centauri is a main sequence red dwarf and a stellar class of M6 V star.

The data provided by the SDSS has the Harvard spectral classes of the spectra and not the M-K luminosity classes. This paper is only concerned with the Harvard spectral classification.

C. Importance of Stellar Classes

Astronomers use stellar classes, in conjunction with other data, to learn many details about stars. The Hertzsprung-Russell (H-R) diagram, shown in Fig. 2, is a plot of stars where the horizontal axis is the spectral class or surface temperature and the vertical axis is the luminosity or absolute magnitude.

The location of a star on the H-R diagram tells astronomers information about that star. For example, stellar evolution can be mapped out on the H-R diagram [12]. The location of a star on the H-R diagram defines whether it is in the main sequence, a giant, a super giant, or whether the star has stopped fusion of Hydrogen and is fusing Helium, or if it has stopped fusing Helium and so on [12].

D. Redshift

Redshift is caused by the relative motion of a star with respect to an observer through the Doppler effect on light. When a star moves away from an observer, the wavelengths of

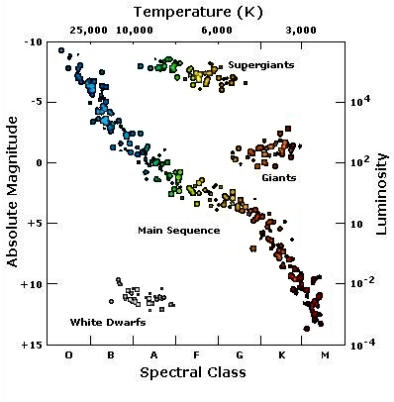

Fig. 2. Hertzsprung Russell Diagram [13]

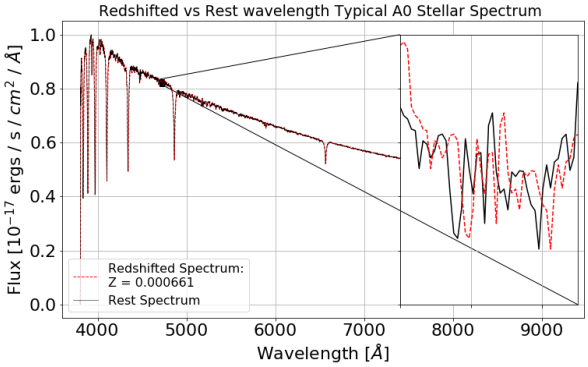

Fig. 3. Example of redshift

light appear to be longer, or, in terms of the visible spectrum, redder. Redshift causes the flux measurements to be shifted compared to what would have been observed at rest [12], [10].

All stars have different redshift values. Redshift is defined by (1), where Z is the redshift value and λ is the wavelength [12], [10]. Equation (2) defines how the flux density changes due to redshift, where f is the flux density at wavelength λ (see [10]). However, for the purpose of this paper, since Z for stars from the SDSS database have $Z \ll 1$, (3) is used:

$$Z = \frac{\lambda_{observed} - \lambda_{rest}}{\lambda_{rest}} \tag{1}$$

$$f(\lambda_{rest}) = (1+Z)^2 f(\lambda_{observed}) \tag{2}$$

$$f(\lambda_{rest}) \approx f(\lambda_{observed}) \tag{3}$$

Fig. 3 shows an example of redshift. Notice that the two spectra are identical, but shifted. The dashed line represents spectra with redshift and the solid line represents the same spectra but with redshift corrections applied. Redshift correction shifts the wavelengths from their observed values to the values that would have been observed if the star was at rest relative to the observer, which is referred to as *rest wavelengths* in this paper.

III. Our Approach

Our goal is to predict star classes from the SDSS database based on stellar spectra. In this section, we describe the data, the pre-processing steps, and the feature selection and classification techniques for stellar classification.

A. Data

The stellar spectra collected by the SDSS are pre-processed by SDSS scientists through the methods presented by Dawson *et al.* [14] and Stoughton *et al.* [15]. There are two spectrographs used by the SDSS to collect the stellar spectra: the Baryon Oscillation Spectroscopic Survey (BOSS) spectrograph and the SDSS spectrograph. The data used in this paper was collected using the SDSS spectrograph. The SDSS spectrograph is identical to the BOSS spectrograph except the BOSS spectrograph has upgraded Charged Coupled Device (CCD) cameras, a larger wavelength sensitivity, and a larger number of astronomical objects that can be simultaneously observed [16]. Therefore, due to the limited literature of the SDSS spectrograph, the more in depth descriptions of the BOSS spectrograph will suffice and are presented by Smee *et al.* [17].

The dataset used in this paper comes from SDSS data run 14, which collected 600,967 stellar spectra. Some of the stellar spectra were rejected from this study because they were not able to be classified (e.g. the documented stellar class is not O, B, A, F, G, K, and M with sub-classes of 0 - 9). Other spectra were rejected because large portions of flux measurements were missing due to CCD or other instrument failures. Stellar spectra with more than half of their spectrum missing were also rejected. The usable dataset contains 578,346 stellar spectra and 22 of the 70 classes. These spectra are used as the input vectors, with each spectrum having 3,800 to 4,000 features. These features are the flux measurements. The distribution of the classes from this dataset is highly imbalanced, as seen in Fig. 4. We balanced the dataset using: *i)* undersampling (removing samples); and *ii)* hybrid sampling, consisting in both undersampling and oversampling (duplicating samples) [18]. Undersampling results in each class having 572 samples for a total of 12,584 samples. Hybrid sampling results in each class having 16,682 samples for a total of 367,004 samples. Data balancing and data pre-processing (Section III-B) do not change the 22 output classes found in Fig. 4.

B. Data pre-processing

The flux intensity for a given class of star varies for a variety of reasons, including proximity to the spectrograph and the star's luminosity. To resolve this issue, the flux is scaled from 0 to 1 using (4), where f_i is the ith flux measurement, f_{max} and f_{min} are the maximum and minimum flux measurements respectively, and $f_{i,scaled}$ is the resulting scaled flux:

$$f_{i,scaled} = \frac{f_i - f_{min}}{f_{max} - f_{min}} \tag{4}$$

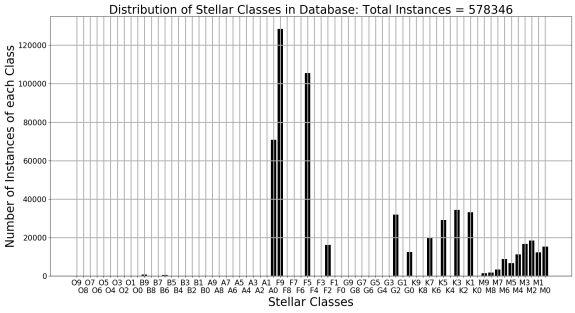

Fig. 4. Distribution of classes in the dataset

Each stellar spectrum in the dataset collected different amounts of flux measurements. This creates a problem when building the feature matrix, which will be described in Section III-C. To overcome this problem, an average number of flux measurements is computed using the first 5,000 spectra. This average is then used to fit each spectrum to a standardized number of flux measurements. The resulting average number of flux measurements is 3,834.

The next pre-processing phase deals with redshift. The data presented by the SDSS is redshifted. Due to the dataset being redshifted, the dataset is redshift corrected using (1) and the provided redshift for each spectrum. This way we create artificial rest spectra.

When the wavelengths are represented in logarithmic space, the difference between two adjacent wavelength values is 0.0001 (see [19]). Therefore, the margin for each wavelength value is $\lambda \pm 0.00005$. Using the wavelengths in logarithmic space, the artificial rest wavelengths are fitted to the nearest λ, within the margin of 0.00005.

Building the feature matrix requires a template for wavelength values. The purpose of the template is to fit each stellar spectrum to identical wavelength values, because each stellar spectrum starts collecting flux measurements at different wavelength values [19]. The first spectrum in the database with 3,834 flux measurements and wavelength values was chosen as the template for all spectra wavelength fitting. The wavelengths of this spectrum are then extracted into the template and used to fit each spectrum to a uniform set of wavelength values.

The process of fitting takes the spectrum's first wavelength value and locates it in the template's wavelength array (Fig. 5). For example, if the first wavelength value (index 0) in the spectrum's wavelength array is located at index 1 in the template's wavelength array, then the spectrum's wavelength array is "shifted" to the right. This causes index 0 of the spectrum's wavelength array to be shifted to index 1. If index 0 of the spectrum's wavelength array does not appear in the template's wavelength array, then the fitting process locates the template's index 0 in the spectrum's wavelength array (Fig. 6).

Notice in Figs. 5 and 6 that the shifted spectrum's wavelengths appear in the same column as the template's wavelengths. The shifted spectrum's wavelength array is now fitted

Template's Wavelengths [$\AA$]						
3,806.27	3,807.15	3,808.03	3,808.90	3,809.78	3,810.66	3,811.54
Spectrum's Wavelengths [$\AA$]						
3,807.15	3,808.03	3,808.90	3,809.78	3,810.66	3,811.54	3,812.41
$\lambda_{missing}$	3,807.15	3,808.03	3,808.90	3,809.78	3,810.66	3,811.54
Shifted Spectrum's Wavelengths [$\AA$]						

Fig. 5. Example of fitting a spectrum's wavelengths to the template wavelengths: Right Shifting.

Template's Wavelengths [$\AA$]						
3,806.27	3,807.15	3,808.03	3,808.90	3,809.78	3,810.66	3,811.54
Spectrum's Wavelengths [$\AA$]						
3,805.40	3,806.27	3,807.15	3,808.03	3,808.90	3,809.78	3,810.66
3,806.27	3,807.15	3,808.03	3,808.90	3,809.75	3,810.66	$\lambda_{missing}$
Shifted Spectrum's Wavelengths [$\AA$]						

Fig. 6. Example of fitting a spectrum's wavelengths to the template wavelengths: Left Shifting.

to the template. Since each flux measurement is directly associated with a wavelength value, the flux array is fitted in parallel to the wavelength arrays. When the wavelength array is shifted to the left or right, the flux array is shifted to the left or right by the same amount, which is apparent in Fig. 7.

C. Feature Matrix

Each stellar spectra has two arrays, a flux array that stores flux measurements and a wavelength array that stores the wavelengths that correspond to the flux measurements. The feature matrix is constructed with the rows as stellar spectra and the columns as the flux measurements that are measured at a particular wavelength (see Table I). Specifically, the wavelength array is used as column headers in the feature matrix.

As stated in Section III-B, each stellar spectra starts at different wavelength values. To ensure that each column of the feature matrix represents flux measured at the same wavelength, each wavelength array is fitted to the wavelength array template.

$w_{missing}$	3,807.15	3,808.03	3,808.90	3,809.78	3,810.66	3,811.54
Shifted Spectrum's Wavelengths [$\AA$]						
$f_{missing}$	0.09193	0.73104	0.80613	0.73711	0.80371	0.77967
Shifted Spectrum's Flux [$ergs\ s^{-1}cm^{-2}$]						

Fig. 7. Example of a spectrum's wavelengths and flux arrays after fitting to a template.

TABLE I
EXAMPLE OF THE FEATURE MATRIX

	Wavelength: 3,807.15 $\AA$	Wavelength: 3,808.03 $\AA$	Wavelength: 3,808.90 $\AA$	Wavelength: 3,810.66 $\AA$
Spectrum 1	Flux	Flux	Flux	Flux
Spectrum 2	Flux	Flux	Flux	Flux

When the wavelength array is being fitted to the wavelength template array, it can create missing values (represented by $\lambda_{missing}$), as seen in Figs. 5 and 6. When a wavelength array is right shifted, missing values are guaranteed to form at the beginning. This is due to the first wavelength value being shifted to the right leaving no wavelength values. When a wavelength array is left shifted, a missing value forms if the last wavelength value in the spectrum is smaller than the last wavelength value in the template. After shifting the spectrum's wavelength and flux arrays, if the array length is larger than 3,834, then the spectrum's wavelength and flux arrays are cut off at index 3,833 to ensure all spectra have the same number of flux measurements.

Missing wavelength values are replaced with the values from the template's wavelength array. Missing flux values (represented by $f_{missing}$) are replaced using the average of the next or last *K* flux measurements. A moving average is used to ensure the continuation of the trend and that no artificial absorption lines are created. Artificial absorption lines can lead to misclassification.

When the missing flux measurement is at the beginning of the flux array, then the next *K* flux measurements are averaged using (5), where *j* is the index of $f_{missing}$ being replaced. The sequence of missing values at the beginning of the flux array is as follows: [$\bar{f}_0$, $\bar{f}_1$, ..., $\bar{f}_w$], where *w* is the index of the last missing value. The missing values are replaced from index *w* to index 0. The computed moving average includes any estimated missing values.

$$\bar{f}_j = \frac{1}{K} \sum_{i=j+1}^{K+j+1} flux_i \tag{5}$$

When the missing flux measurement is at the end of the flux array then the last *K* flux measurements are averaged using (6), where *j* is the index of $f_{missing}$ being replaced and $j > K$. The sequence of missing values at the end of the flux array is as follows: [$\bar{f}_v$, $\bar{f}_{v+1}$, ..., $\bar{f}_{3,833}$], where *v* is the index of the first missing value. The missing values are replaced starting at index *v* to index 3,833. The moving average when computed includes any estimated missing values.

$$\bar{f}_j = \frac{1}{K} \sum_{i=K-j-1}^{j-1} flux_i \tag{6}$$

A problem arises with the feature matrix because redshift causes the flux measurements to be shifted in wavelength. This causes the columns of the feature matrix to have the same flux measurement at different wavelengths for the same class. Certain flux measurements have real physical meaning such as the absorption lines. For example, the Hydrogen Alpha (Hα) absorption line is observed at rest with wavelength at 6,562.8 $\mathring{A}$ in air, but due to redshift, the Hα line can be observed at other wavelengths such as 6,563.8 $\mathring{A}$ or 6,561.8 $\mathring{A}$. An example of how redshift changes the meaning of each column is seen in Table II. Notice how Hα appears in different columns of the feature matrix due to redshift for the same stellar class. This is comparable to taking the pedal width of the iris dataset's [20] feature matrix and putting it in the pedal length column.

TABLE II
EXAMPLE OF THE FEATURE MATRIX THAT HAS REDSHIFTED DATA

Wavelength: 6,560.8 $\mathring{A}$	Wavelength: 6,561.8 $\mathring{A}$	Wavelength: 6,562.8 $\mathring{A}$	Wavelength: 6,563.8 $\mathring{A}$	Wavelength: 6,564.8 $\mathring{A}$	Star Class
		Hα			A0
	Hα				A0
			Hα		A0
				Hα	A0
Hα					A0

As described in Sections II-A and II-B, absorption lines are a key parameter in classifying stars with the Harvard spectral classification system. The classification of a star defines the expected wavelength at which certain absorption lines occur. However, to redshift correct a spectrum, the redshift is computed using (1), where the wavelengths of known absorption lines are used. By definition, knowing the absorption lines to perform redshift corrections also means that the class of the star is known. The other approach to determine the redshift of a spectrum utilizes statistical analysis to compare the spectra to templates, which as a by product the class of the star is known.

Machine learning approaches to classifying stars with redshift corrected data are redundant because the by product of redshift correcting a spectrum is the star's class.

The redshift of stars in the SDSS dataset is small. *Z* is on the order of magnitude of 0.0001. As such, absorption lines, key features of the dataset, are still grouped closely together. Classification with a large number of stellar spectra of each class with a large range of redshift values overcome the redshift problem.

D. Feature selection and stellar classification

We used the Chi Squared [21] and Fisher [22] filter feature selection algorithms. For a large dataset as SDSS, the Chi Squared and Fisher methods performed equally.

We used the RF and SVM classifiers, chosen because other works using these two classifiers on similar data reported good results. For instance, Xing and Guo [2] used a SVM to classify stellar spectra. Yi and Pan [4] applied RF to the classification of stellar spectra.

IV. EXPERIMENTS

We performed experiments on stellar spectra from the SDSS database. In the following, we present the results of the Chi Squared and Fisher feature selection algorithm in conjunction with the RF, and SVM classifiers. The resulting best feature selection and classifier algorithm pair is then expanded on with its precision, recall, and F1 score statistics.

The results are then compared for redshifted spectra and artificial rest spectra. Artificial rest spectra are used as a baseline for comparison for the redshifted spectra.

We did not consider any misclassification costs.

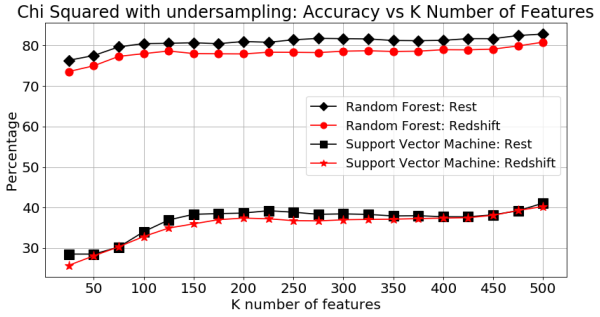

Fig. 8. 10-Fold cross validation results for Chi Squared feature selection with undersampling for redshifted spectra and artificial rest spectra

TABLE III
10-FOLD CROSS VALIDATION RESULTS FOR CHI SQUARED FEATURE SELECTION WITH UNDERSAMPLING (12,584 SAMPLES) FOR REDSHIFTED SPECTRA AND ARTIFICIAL REST SPECTRA

Classifier	Accuracy (%) for Chi Squared for K number of features									
	50.0	100.0	150.0	200.0	250.0	300.0	350.0	400.0	450.0	500.0
redshifted spectra										
RF	76.60	80.81	80.66	80.88	80.92	81.64	81.99	81.79	82.13	83.59
SVM	28.33	33.60	37.64	37.95	38.73	39.01	39.02	38.97	39.13	40.41
artificial rest spectra										
RF	81.81	86.17	86.03	86.13	86.72	86.85	86.82	86.89	86.92	88.24
SVM	29.13	35.69	38.63	39.13	39.59	39.82	40.27	39.87	39.68	41.27

A. Experimental Framework and Results

The first step of the experiments is to perform data pre-processing, as described in Section III-B. The flux is scaled to ensure that similar classes have similar flux measurements using (4). There are two datasets used in this paper, redshifted spectra and artificial rest spectra. Both datasets are created from the SDSS dataset. For artificial rest spectra, the wavelengths are adjusted using (1) through the process described in Section III-B. For redshifted spectra, this process is forgone because all of the spectra in the SDSS database are redshifted. The remaining parts of the data pre-processing are applied to both datasets.

The wavelengths are fitted to a template to ensure that each spectrum's wavelength coverage is identical, which is supported by the work presented by Bazarghan and Gupta [7]. This process is described in Section III-B. Since each flux measurement is associated with a wavelength value when the wavelengths are fitted to the template, the flux array gets transformed identically. The process of fitting the wavelengths to the template and the flux array transformation causes missing values to appear. These missing values are approximated using a moving average, as described in Section III-C, and by (5) and (6).

Due to the number of samples obtained after hybrid sampling (367,004 samples), a random balanced subset of 110,000 samples is used to compute the feature rankings for Chi Squared and Fisher feature selection methods. After the feature selection phase, the RF and SVM classifiers are applied to the resulted reduced feature matrix. Due to the large execution time for SVM (approximately 21 hours for the first experiment) and the higher accuracy of RF, SVM is not computed for hybrid sampling. The accuracy, precision, recall, and F1 score are computed. In each experiment, we use 10-fold cross validation to split our dataset into test and training sets.

B. Implementation

The experiments presented in this paper are implemented using Python and scikit-learn [21]. Due to the size of the imbalanced raw dataset (35.4 GB for artificial rest spectra and 35.4 GB for redshifted spectra), the Python NumPy memmap [23], [24] module was utilized to read very large arrays from

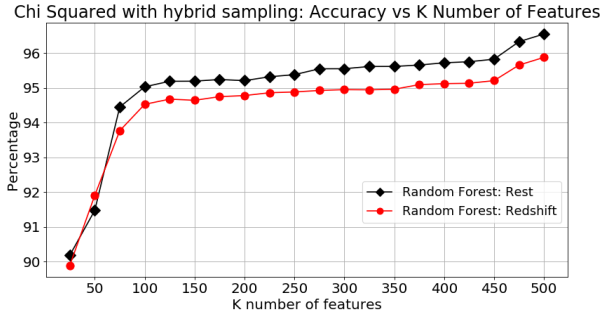

Fig. 9. 10-Fold cross validation results for Chi Squared feature selection with hybrid sampling for redshifted spectra and artificial rest spectra spectra

TABLE IV
10-FOLD CROSS VALIDATION RESULTS FOR CHI SQUARED FEATURE SELECTION WITH HYBRID SAMPLING (367,004 SAMPLES) FOR REDSHIFTED SPECTRA AND ARTIFICIAL ARTIFICIAL REST SPECTRA USING RANDOM FOREST

Accuracy (%) for Chi Squared for K number of features									
50.0	100.0	150.0	200.0	250.0	300.0	350.0	400.0	450.0	500.0
redshifted spectra									
91.24	95.11	95.25	95.24	95.46	95.62	95.65	95.76	95.90	96.55
artificial rest spectra									
94.34	97.02	97.03	97.14	97.31	97.45	97.43	97.54	97.65	98.21

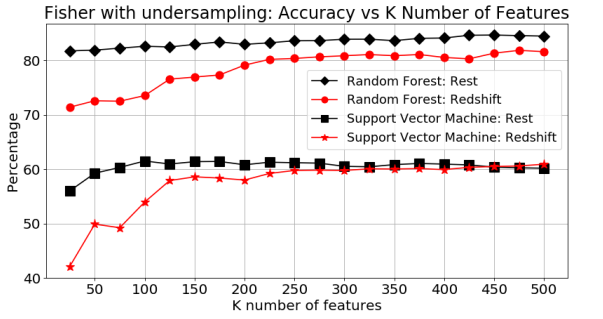

Fig. 10. 10-Fold cross validation results for Fisher feature selection with undersampling for redshifted spectra and artificial artificial rest spectra

TABLE V
10-FOLD CROSS VALIDATION RESULTS FOR FISHER FEATURE SELECTION WITH UNDERSAMPLING (12,584 SAMPLES) FOR REDSHIFTED SPECTRA AND ARTIFICIAL ARTIFICIAL REST SPECTRA

Classifier	Accuracy (%) for Fisher for K number of features									
	50.0	100.0	150.0	200.0	250.0	300.0	350.0	400.0	450.0	500.0
redshifted spectra										
RF	81.177	82.00	82.79	83.51	83.45	83.73	83.26	83.90	84.52	84.95
SVM	60.43	66.51	65.62	64.83	64.44	64.48	65.21	65.33	64.94	64.66
artificial rest spectra										
RF	83.64	83.46	85.29	86.29	86.27	86.19	86.43	87.16	87.19	87.40
SVM	55.63	56.06	57.29	58.73	61.02	62.23	62.59	63.21	63.63	63.81

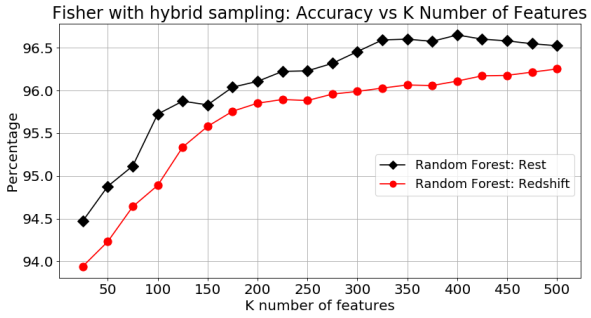

Fig. 11. 10-Fold cross validation results for Fisher feature selection with hybrid sampling for redshifted spectra and artificial artificial rest spectra

TABLE VI
10-FOLD CROSS VALIDATION RESULTS FOR FISHER FEATURE SELECTION WITH HYBRID SAMPLING (367,004 SAMPLES) FOR REDSHIFTED SPECTRA AND ARTIFICIAL ARTIFICIAL REST SPECTRA USING RANDOM FOREST

Accuracy (%) for Fisher for K number of features									
50.0	100.0	150.0	200.0	250.0	300.0	350.0	400.0	450.0	500.0
redshifted spectra									
94.65	95.89	96.22	96.38	96.49	96.59	96.62	96.74	96.79	96.87
artificial rest spectra									
96.30	97.22	97.51	97.59	97.58	97.44	97.37	97.36	97.32	97.32

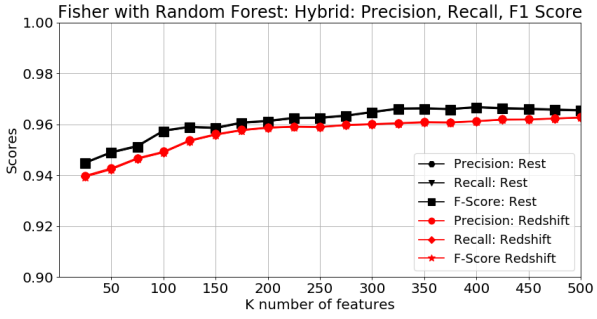

Fig. 12. Average results of five experiments for Fisher feature selection for random forest

TABLE VII
PRECISION, RECALL, AND F1 SCORE FOR FISHER FEATURE SELECTION WITH HYBRID FOR RANDOM FOREST

	Results for Fisher + Random Forest for K number of features									
	50.0	100.0	150.0	200.0	250.0	300.0	350.0	400.0	450.0	500.0
redshifted spectra										
Precision	0.9474	0.9593	0.9626	0.9642	0.9652	0.9662	0.9665	0.9677	0.9682	0.9690
Recall	0.9465	0.9589	0.9622	0.9638	0.9649	0.9659	0.9662	0.9674	0.9679	0.9687
F1 Score	0.9467	0.9589	0.9622	0.9639	0.9649	0.9659	0.9662	0.9674	0.9680	0.9687
artificial rest spectra										
Precision	0.9634	0.9724	0.9753	0.9761	0.9760	0.9746	0.9740	0.9739	0.9735	0.9735
Recall	0.9630	0.9722	0.9751	0.9759	0.9758	0.9744	0.9737	0.9736	0.9732	0.9732
F1 Score	0.9630	0.9722	0.9751	0.9759	0.9758	0.9744	0.9738	0.9736	0.9732	0.9732

TABLE VIII
EXECUTION TIME FOR HYBRID SAMPLING FOR RANDOM FOREST WITH ARTIFICIAL REST SPECTRA

	Execution times (Seconds) for K number of features									
	50.0	100.0	150.0	200.0	250.0	300.0	350.0	400.0	450.0	500.0
Fisher + artificial rest spectra										
Feature Selection	7.94 $\times 10^5$	7.94 $\times 10^5$	7.94 $\times 10^5$	7.94 $\times 10^5$	7.94 $\times 10^5$	7.94 $\times 10^5$	7.94 $\times 10^5$	7.94 $\times 10^5$	7.94 $\times 10^5$	7.94 $\times 10^5$
Train	39.83	55.49	66.23	76.77	82.10	93.07	99.46	110.21	115.85	121.24
Test	0.65	0.58	0.61	0.62	0.63	0.68	0.69	0.70	0.75	0.79
Chi Squared + artificial rest spectra										
Feature Selection	6.84	7.98	9.12	10.41	11.21	12.40	13.24	14.72	16.13	17.48
Train	37.11	50.08	60.03	69.79	75.07	84.80	90.28	100.28	105.45	109.06
Test	0.57	0.55	0.57	0.60	0.62	0.64	0.68	0.70	0.73	0.74

storage rather than RAM. We run our test on an IBM S822LC cluster with IBM POWER 8 nodes, NVLink and NVidia Tesla P100 GPUs [25].

RF and SVM are implemented with scikit-learn with default parameters [21]. SVM is implemented using a Gaussian kernel. Precision, recall, and F1 score are computed using functions implemented by the scikit-learn sklearn.metrics package [21]. The Chi Squared feature selection algorithm is implemented using scikit-learn [21]. The Fisher feature selection algorithm is implemented using scikit-feature [22].

C. Discussion

Figs. 8, 9, 10 and 11 show that classification using the redshift spectra has essentially the same accuracy as the artificial rest spectra, despite the fact that redshift has introduced problematic data into the feature matrix. Figs. 8 and 10 also demonstrate that RF is the best classifier for both redshifted spectra and artificial rest spectra.

Tables III and V show that the Fisher method performs better than Chi Squared method for both redshifted spectra and artificial rest spectra when using the RF classifier. From Tables IV and VI we observe that with the Fisher feature selection method we obtain roughly the same accuracy as with the Chi Squared method. However, Table VIII demonstrates that Chi Squared has a significantly shorter execution times than Fisher.

Tables III, IV, V, and VI demonstrate that hybrid sampling outperforms undersampling. Tables III, IV, V, and VI show that accuracy increases with more samples. Supplementing

SDSS data run 14 with samples from SDSS data run 12, 13, and 15 may improve accuracy.

Fig. 12 shows that the precision, recall, and F1 scores for redshifted spectra are lower than for artificial rest spectra. However, Table VII show that the precision for redshifted spectra is only lower by 0.0045, recall is only lower by 0.0045 and the F1 score is lower by 0.0045 (for 500 features).

For redshifted spectra, we achieved an accuracy of 96.87% using the Fisher feature selection with hybrid sampling using the RF classifier applied to 500 selected flux measurements. Using the same work flow, we obtained 97.32% accuracy for artificial rest spectra. These results are superior by more than three percent than the ones achieved by Xing *et al.* [2], since they reported 93.26% accuracy using wavelet reduction and SVMs for SDSS spectra.

Bolton *et al.* used Chi Squared minimization to determine the redshift and classification of SDSS spectra of galaxies, quasars, and stars [26]. They applied the Least Squares minimization to compare each spectrum to a full range of templates that span through galaxies, quasars, and stars. Our approach takes considerably fewer steps than the one in [26] and produces excellent results.

The execution times and the obtained accuracy demonstrate that, for a real application of this work, the automated classification of redshifted stellar spectra not only achieves a high accuracy but is also fast.

V. Conclusions

According to our results, accurate stellar classifications can be obtained using a combination of domain specific data pre-processing, feature selection and classification techniques. Compared to previous work of other authors, we have two interesting conclusions:

1) Redshift corrections are not necessary to obtain this high level of accuracy.
2) Aside from wavelength fitting and flux scaling, any additional spectrum pre-processing after the processes presented by Dawson *et al.* [14] and Stoughton *et al.* [15] is unnecessary.

Therefore, when a star can be accurately classified prior to redshift correction, many stellar properties can be easily attained without the complex data transformations and statistical analyzes used by other authors.

Even though the feature matrix for redshifted stellar spectra produces an undesired feature matrix because of flux discrepancies, high accuracy was still achieved. This can be accounted for by redshift values for stars being small and a large sample of redshift values to train the classifier.

References

[1] Sloan Digital Sky Survey, "Sloan Digital Sky Survey." [Online]. Available: https://www.sdss.org/

[2] F. Xing and P. Guo, "Classification of stellar spectral data using svm," in *Advances in Neural Networks (ISNN 2004)*, F.-L. Yin, J. Wang, and C. Guo, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 616–621.

[3] J. N. Zhang, A. L. Luo, and L. P. Tu, "A stratified approach for automatic stellar spectra classification," in *2008 International Congress on Image and Signal Processing (CISP 2008)*, 2008, pp. 249–252.

[4] Z. Yi and J. Pan, "Application of random forest to stellar spectra classification," *978-1-4244-6516-3/10/26.00 2010 IEEE*, pp. 3129–3132, 2010.

[5] A. J. Pickles, "A Stellar Spectral Flux Library: 1150-25000 Å," *pasp*, vol. 110, pp. 863–878, jul 1998.

[6] G. H. Jacoby, D. A. Hunter, and C. A. Christian, "A library of stellar spectra," *apjs*, vol. 56, pp. 257–281, oct 1984.

[7] M. Bazarghan and R. Gupta, "Automated classification of sloan digital sky survey (SDSS) stellar spectra using artificial neural networks," *Astrophysics and Space Science*, vol. 315, no. 1, p. 201, May 2008. [Online]. Available: https://doi.org/10.1007/s10509-008-9816-5

[8] D. G. York *et al.*, "The Sloan Digital Sky Survey: Technical Summary," *The Astronimical Journal*, vol. 120, pp. 1579–1587, Sep. 2000.

[9] Pasquet-Itam, J. and Pasquet, J., "Deep learning approach for classifying, detecting and predicting photometric redshifts of quasars in the sloan digital sky survey stripe 82," *Astronomy & Astrophysics*, vol. 611, p. A97, 2018. [Online]. Available: https://doi.org/10.1051/0004-6361/201731106

[10] F. R. Chromey, *To Measure The Sky: An Introduction to Observational Astronomy*. Cambridge University Press, 2010.

[11] H. Zheng and Y. Zhang, "Feature selection for high-dimensional data in astronomy," *Advances in Space Research*, vol. 41, pp. 1960–1964, 09 2007.

[12] B. W. Carroll and D. A. Ostlie, *An Introduction to Modern Astrophysics*, 2nd ed. Cambridge University Press, 2017.

[13] University of Iowa Department of Physics and Astronomy, "Imaging the universe." [Online]. Available: http://astro.physics.uiowa.edu/ITU/labs/foundational-labs/exploring-hertzsprung-russe/part-1-the-hr-diagram.html

[14] K. S. Dawson *et al.*, "The baryon oscillation spectroscopic survey of SDSS-III," *The Astronomical Journal*, vol. 145, no. 10, pp. 1–41, 2013.

[15] C. Stoughton *et al.*, "Sloan digital sky survey: Early data release," *The Astronomical Journal*, vol. 123, pp. 485–548, 2002.

[16] Sloan Digital Sky Survey, "BOSS spectrograph." [Online]. Available: https://www.sdss.org/instruments/boss_spectrograph/

[17] S. A. Smee *et al.*, "The multi-object, fiber-fed spectrographs for SDSS and the baryon oscillation spectroscopic survey," *The Astronomical Journal*, vol. 146, no. 32, pp. 1–40, 2013.

[18] N. Japkowicz, "Learning from imbalanced data sets: A comparison of various strategies." AAAI Press, 2000, pp. 10–15.

[19] Sloan Digital Sky Survey, "Understanding the optical data." [Online]. Available: https://www.sdss.org/dr14/spectro/spectro_basics/

[20] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[22] J. Li, K. Cheng, S. Wang, F. Morstatter, T. Robert, J. Tang, and H. Liu, "Feature selection: A data perspective," *arXiv:1601.07996*, 2016.

[23] T. E. Oliphant, *A guide to NumPy*. Published by Continuum Press, a division of Continuum Analytics, Inc., 2015.

[24] Numpy, "Memmap." [Online]. Available: https://docs.scipy.org/doc/numpy-1.14.0/reference/generated/numpy.memmap.html

[25] "Turing: The cwu supercomputer." [Online]. Available: http://www.cwu.edu/faculty/turing-cwu-supercomputer

[26] A. S. Bolton *et al.*, "Spectral classification and redshift measurement for the SDSS-III baryon oscillation spectroscopic survey," *The Astronomical Journal*, vol. 144, no. 144, pp. 1–20, 2012.