AUTOMATIC SPECTRAL CLASSIFICATION OF STARS USING MACHINE LEARNING: AN APPROACH BASED ON THE USE OF UNBALANCED DATA

Marco Oyarzo Huichaqueo[1] and Renato Muñoz Orrego[2]

[1]School of Engineering, Rovira i Virgili University, 43007 Tarragona, Spain
[2]School of Engineering, Technical University of Madrid, 28006 Madrid, Spain

ABSTRACT

With the increase in astronomical surveys, astronomers are faced with the challenging task of analyzing a large amount of data in order to classify observed objects into hard-to-distinguish classes. This article presents a machine learning-based method for the automatic spectral classification of stars from the latest release of the SDSS database. We propose the combinatorial use of spectral data, derived stellar data, and calculated data to create patterns. Using these patterns as inputs, we develop a Random Forest model that outputs the spectral class of the observed star. Our model is able to classify data into six complex classes: A, F, G, K, M, and Carbon stars. Due to the unbalanced nature of the data, we train our model considering three data use cases: using the original data, using under-sampling, and over-sampling data techniques. We further test our model by using a fixed dataset and a stratified dataset. From this, we analyze the performance of our model through statistical metrics. The experimental results showed that the combinatorial use of data as an input pattern contributes to improve the prediction scores in all data use cases, meanwhile, the model trained with augmented data outperforms the other cases. Our results suggest that machine learning-based spectral classification of stars may be useful for astronomers.

KEYWORDS

Spectral Classification, Machine Learning, Data Analysis, Astronomy

1. INTRODUCTION

Recently, due to the increase in scientific astronomical surveys, there is a need to quickly characterize the data obtained from the observations. Nevertheless, this task is carried out manually, due to the lack of confidence in more sophisticated algorithms, being computationally costly in terms of memory and processing time due to the use of large amounts of data and the long time-consuming analyzes performed by astronomers. For instance, astronomers must perform some challenging tasks such as classifying observed objects into main classes (e.g. galaxy or star) and then their subclasses (e.g. starforming, starburst, B9-type, M0V-type, and others) based on information such as their morphology, recorded spectral data, and calculated stellar parameters.

In the particular case of star classification, astronomers must analyze the spectral data of the candidates. The spectral data is obtained by using instruments and filters to capture specific wavelengths such as ultraviolet (u), green (g), red (r), near-infrared (i), and infrared (z) [1]. These light bands provide useful information that allows a classification based on physical characteristics. In this way, by using stellar classification systems the observed candidates can be classified by human experts. A standard in astrophysics is the well-known Morgan-Keenan (MK)

classification system (Morgan et al., 1943) [2] that is based on both the luminosity and effective temperature of a star. The MK system divides stellar spectra into O, B, A, F, G, K, and M main

classes; from the hottest (O-type: Teff > 30,000 K) to the coolest (M-type: 2,200-3,700 K) and then each letter class can be also subdivided using a numeric digit where 0 correspond to the hottest and 9 to the coolest. Nevertheless, depending on the nature of the candidates, the MK system does not present clear results in some cases. For instance, when the MK system results in irresolvable overlaps in terms of spectral types in both effective temperature and luminosity, some candidates can be classified into L-type [3] and methane dwarfs (T-type), by analyzing the infrared spectra or carbon types (type C) [4] by analyzing the swan bands of their spectra.

On the other hand, from a practical point of view, an important consideration is that with the advancement of technology and computational techniques, an intelligent machine can do this challenging classification task (for a human) more efficiently. Even more when the use of artificial intelligence (AI) and its supervised methods, such as machine learning (ML) and deep learning (DL), have shown their potential in different fields of scientific and industrial application. In this work, a system for the automatic recognition of spectral classes of stars based on ML from spectroscopic and photometric data is proposed. Our proposed system is capable of recognizing six types of stars: A, F, G, K, M, and C-type. For this, we propose the use of the latest data collected up to 2021 from the Sloan Digital Sky Survey (SDSS) [5]. The main contributions of this work are as follows:

- Demonstrating that by using a widely used ML algorithm, it is possible to reach acceptable prediction rates for a challenging task of classification star types under MK spectral class and non-conventional class.
- Introducing an approach to improve the prediction rates from the combinatorial use of astronomical data.
- Evaluating our method based on different data use cases and different types of test sets.

This work is organized into five sections. Section 2 discusses related work on astronomical object classification. In Section 3, we introduce the methodology followed in this work. Section 4 covers the computational experimentation and results obtained. Finally, Section 5 provides research conclusions with some possible directions for future research.

2. RELATED WORK

A review of the literature on astronomical objects classification methods using AI techniques was carried out. This review only includes recent articles (up to 4 years ago) and it is split into star classification and stellar object classification, as follows.

2.1. Star Classification

Several works have proposed the use of AI techniques for the classification of stars based on the MK system. Sharma et al. (2020)[6] proposed an approach for the classification of stellar spectra into O-B-A-F-G-K-M type stars. In their work, they trained some ML and DL algorithms using different spectral libraries and tested them by using the Indo-U.S. Library of Coud´e Feed Stellar Spectra (CFLIB). Using a convolutional neural network (CNN) model, their best-resulting accuracy was 89%. In the work of Lu et al. (2020)[7], a fully connected artificial neural network (ANN) was proposed for binary classification of stars into F-G and G-K types from 2D spectral images. In their development, the Data Release 6 (DR6) from the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) was used for training and testing. Thus, their method

achieved an accuracy of 80%. Similarly, Dafonte et al. (2020)[8] presented a DL method to classify spectral data into A-B-F-G-K-M type stars. To this, the authors proposed the use of an ANN trained and tested using their database. Therefore, their experimental results showed that their method achieved an accuracy of 83%.

We further identify some works in which the Random Forest and XGBoost algorithms have recently been used for variable star classification. Zhang et al. (2020) [9] proposed a method to classify RR Lyrae candidates into A-type and F-type stars. They combined photometric and spectroscopic data obtained from SDSS achieving a completeness of ~97%. Adassuriya et al. (2021) [10] used stellar lights curves observed by the Kepler mission to train a classifier that achieved a sensitivity of 87% during the identification β-Cephei, δ-Scuti, γ-Doradus, Red Giants, RR Lyrae, and RV Tauri star types. In the work of Naydenkin et al. (2020) [11] a ML model to classify data of the Zwicky Transient Facility (ZTF) into β-Cephei, δ-Scuti, and RR Lyrae variable stars is proposed. Their method achieved a 95% of the area under the curve (AUC) metric. Similar works have been performed by Hosenie et al. (2019 & 2020) [12][13] where they proposed a system for classification into β-Cephei, δ-Scuti, and RR Lyrae variable stars with ~98% of accuracy using Catalina Real-time Transient Survey (CRTS) database.

2.2. Stellar Object Classification

On the other hand, some authors proposed methods to identify stellar objects from astronomical data. Acharya et al. (2018) [14] presented ML algorithms to classify into star, quasar, and galaxy class from SDSS photometric data. Their best-resulting accuracy was 94.10% achieved using Random Forest. Wierzbinski et al. (2021) [15] used SDSS photometric data and principal component analysis (PCA) to develop some ML models to classify into star, quasar, and galaxy. Thus, Wierzbinski et al. reached an accuracy of 99.16% by using a voting classifier that contains estimators such as quadratic discriminant analysis (QDA), support vector machine (SVM), Decision Tree, Random Forest, XGBoost, Bagging classifier, multilayer perceptron (MLP), Extra Trees, and Naive Bayes classifier. Similarly, Martinazzo et al. (2020) [16] proposed an approach to classify data from Southern Photometric Local Universe Survey (S-PLUS) into star and galaxy using DL models. They demonstrated that using DenseNet-121 was possible to reach an accuracy of 99.2%. A comparison between classical ML models and DL models was proposed by Ethiraj & Kumar (2022) [17] using SDSS data to classify into star, quasar, and galaxy. In their work, the Extra Trees classifier reached an accuracy of 96%, meanwhile, DL models such as EfficientNetB2 and Xception achieved 91% of accuracy.

This paper uses an approach similar to that proposed by Sharma et al., Lu et al., Dafonte et al., and Zhang et al. Particularly, in our case, we propose a system that considers a greater classification capacity, also including types of stars not considered in the MK system, such as C-type stars.

3. MATERIALS AND METHOD

3.1. Proposed Architecture

The proposed method in this work consists of the complementary use of spectroscopic and photometric data of observed stars to estimate their spectral class by using supervised learning. As shown in Fig. 1, our architecture includes a processing module and a supervised learning module. The function of the processing module is to prepare the data to be used in the next module, while the function of the supervised learning module is to analyze and classify the data. A summary of the methodology followed in this work is as follows.

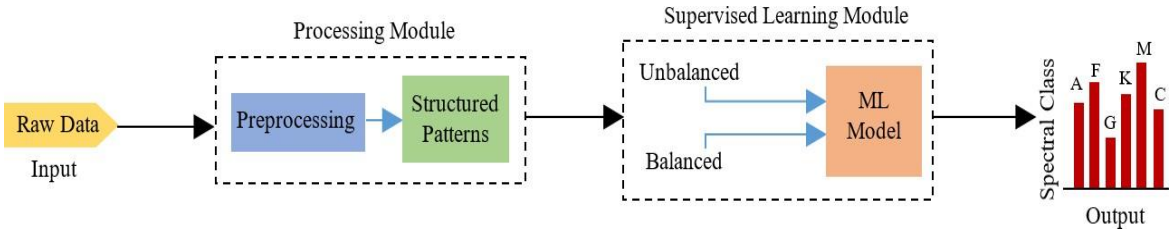


Figure 1. Overview of the proposed spectral classification architecture

In the processing module, the raw data from an astronomical survey is processed in order to obtain labeled data samples. To this end, we use data from SDSS, which contain astronomical data for stars previously manually classified by astronomers. Within the information provided by SDSS is the spectral subclass (e.g. A0, F2, and others) of the observed stars, but in our approach, we propose a system designed to estimate the main spectral class (e.g. A, F, and others). Therefore, we perform data processing to obtain the spectroscopic and photometric data of each observed star, thus obtaining labeled data that are denoted as samples for our purpose. We also propose to use the samples as patterns to allow the combinatorial use of the data obtained. The details of both data processing and pattern usage will be explained later in this article.

In the supervised learning module, we propose the use of a widely used ML algorithm. In this way, using the patterns as inputs we train our ML model to classify the observed stars into A, F, G, K, M, and C-type. Thus, our system is able to identify which spectral class belongs to each input pattern. Furthermore, due to the unbalanced nature of the data from astronomical surveys, which significantly influences the predictive behaviour of our ML model, the training of the ML model is proposed using the real unbalanced data and balanced data through techniques that will be explained later in this article.

Finally, in particular, the spectroscopic data correspond to measured u, g, r, i, and z light bands in a range of 3543-9134 Å. In the photometric data case, this information stands for derived stellar parameters such as redshift, effective temperature (Teff), surface gravity (log(g)), metallicity [Fe/H] of ELODIE star (Prugniel & Soubiran, 2001) [18]. In addition, all implementation of the proposed architecture was carried out using Python codes on a laptop workstation with a 1.6 GHz Intel® Core™ i5-10210U CPU and 8 GB of memory. The detail of each module is described below.

3.2. Experimental Data

SDSS is a scientific project where the main aim is to map the universe and identify astronomical objects such as galaxies, quasars, and stars. Since 1998, SDSS has progressed through several phases that involve multiple surveys with interlocking science goals.

In this work, we use the Data Release 17 (DR17) (Abdurro'uf et al., 2022) [19] of the fourth phase of the project (SDSS-IV) which contains observations through January 2021 considering a dual hemisphere view of the sky, observing from both Las Campanas Observatory, using the du Pont Telescope (Bowen & Vaughan, 1973) [20] and the Sloan Foundation 2.5m Telescope (Gunn et al. 2006) [21] at Apache Point Observatory. DR17 includes different open access types of data such as images, optical spectra, infrared spectra, integral field unit spectra, stellar library spectra, and catalog data. From our approach, we obtained both spectroscopic and photometric data through the science archive server (SAS) catalogues of the SDSS. In this way, all data is provided

by using the standard flexible image transport system (FITS) file format widely used in astronomy.

3.3. Data Processing

From the data obtained, first we selected the best observations following some recommendations from the SDSS catalogue. These observations correspond to those that satisfy some conditions that are described below.

The particular data was selected based on our ML approach. In this way, the data classified as stars by astronomers were selected using the "Type" condition that distinguishes stars based on their morphology. In addition, when the bitmask called "ZWARNING" is equal to zero indicates no problems were identified during the redshift determination. Furthermore, as criteria, only those primary observations of the objects were selected through both "SpecPrimary" and "Mode"

conditions set to true for spectroscopic and photometric data respectively. The resulting processed data contains different spectral subclasses previously determined by astronomers that are also unevenly distributed. The data correspond to both u, g, r, i, and z light bands values and some stellar parameters such as Teff, log(g), [Fe/H], and redshift for each sample. For future ML classification, we cluster the data based on its main spectral class as shown in Table 1.

Table 1. Distributions of the samples obtained from the SDSS database

Original subclasses from SDSS	**Samples**	**Class**
O, OB	1,957	O
B6, B9	6,719	B
A0, A0p	86,623	A
F2, F5, F9	337,929	F
G0, G2, G5	72,025	G
K1, K3, K5, K7	150,543	K
M0V, M2V, M1, M2, M3, M4, M5, M6, M7, M8	159,625	M
L0, L1, L2, L3, L4, L5, L5.5, L9	4,643	L
T2	290	T
Carbon, Carbon_lines, CarbonWD, CV	37,315	C

3.4. Machine Learning Approach

In our supervised learning approach, we propose the use of the spectral data combined with stellar parameters as input patterns to a meta-estimator based on the ensemble learning method. Therefore, we use the widely used Random Forest algorithm (Breiman, 2001) [22] set as 100-trees in the forest and entropy function to measure the information gain of a split.

In order to get a robust classifier, we also propose an interesting experimental focus based on both the complementary use of the fairly reddening-insensitive pseudocolor parameter, proposed by Cáceres & Catelan (2008) [23], and the statistical analysis of spectral data and stellar parameters in each sample. In this way, we trained the Random Forest classifier using an input pattern based on the combinatorial use of the data as described below and denoted in Table 2:

- Pattern 1: Structured by the spectral data.
- Pattern 2 to 5: Structured by adding the stellar parameters.

- Pattern 6: Structured by adding the fairly reddening-insensitive pseudocolor calculated value, denoted as: $C0 = (u - g)0 - (g - r)0$.
- Pattern 7: Structured by adding the provided variance calculated from spectral data, denoted as: $Var1 = Var(u, g, r, i, z)$.
- Pattern 8: Structured by adding the provided variance calculated from derived stellar parameters data, denoted as: $Var2 = Var(Teff, \log(g), [Fe/H], redshift)$.

Table 2. The proposed input pattern for supervised learning

Pattern	**Structure of proposed input pattern**
1	u, g, r, i, z
2	u, g, r, i, z, T_{eff}
3	u, g, r, i, z, T_{eff}, $\log(g)$
4	u, g, r, i, z, T_{eff}, $\log(g)$, [Fe/H]
5	u, g, r, i, z, Teff, log(g), [Fe/H], redshift
6	u, g, r, i, z, Teff, log(g), [Fe/H], redshift, C0
7	u, g, r, i, z, Teff, log(g), [Fe/H], redshift, C0, Var1
8	u, g, r, i, z, Teff, log(g), [Fe/H], redshift, C0, Var2

In this work, we are concerned that we have a severe class imbalance distribution that will make it difficult to obtain good prediction rates. To face this challenging issue, first, from the processed data shown in Table 1, we do not consider the O, B, L, and T-type classes due to their poor amount of samples, and then we propose the following experimental data use case: the use of the unbalanced original data distribution of the classes, and the use of the balanced data through under-sampling, and over-sampling data techniques. These three proposed cases are denoted as 1, 2, and 3, respectively in Fig. 2. In order to reduce the data, we consider the smallest class as the target. Thus, it is possible to undersample all classes greater than the smallest class. On the other hand, to augment the data the largest class was considered as the target. For this, we propose the use of the Synthetic Minority Over-sampling Technique (SMOTE, Chawla et al. 2002) [24] in order to simplify the issue and oversample the minority classes.

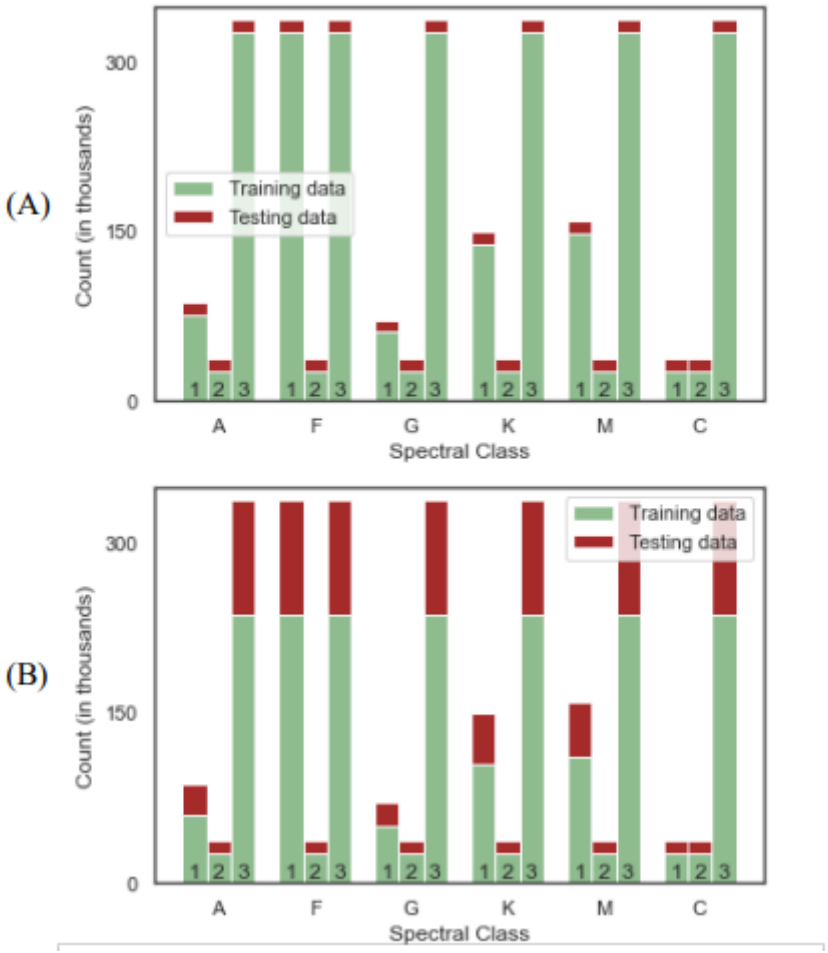


Figure 2. Data distribution for (A) fixed testing dataset and (B) stratified testing dataset

Furthermore, to correctly evaluate our ML model we propose the use of two datasets for testing. Our testing approach is useful for testing our method with different datasets that have been created from randomly shuffled data. In both cases, a seed value was used to initialize the random number generator to ensure that the experiments are reproducible and the same in any case. First, we consider a fixed testing dataset containing the same number of samples per class. This amount is distributed in a balanced way for all classes and corresponds to 30% of the smallest class. In this way, it is possible to guarantee a static and formed testing dataset with original data.

Secondly, we consider a stratified testing dataset, which contains the preserved proportion of samples per class. This proportion corresponds to 30% of the samples for each class. In this way, it is possible to guarantee a dynamic and formed testing dataset based on data that can be real or synthetic (e.g. oversampled data). For this purpose, the model_selection.train_test_split method of the scikit- learn library (available on the official website https://scikit-learn.org/stable/) was used. Fig. 2 shows in a simple way the data distribution for each spectral class and each data use case.

3.5. Statistic Metrics and Performance Evaluation

In order to evaluate our ML model, we consider some statistic metrics by using some categorical labels obtained from a prediction task such as true positive (TP) and false positive (FP), true

negative (TN) and false negative (FN). Thus, we propose some widely used statistical evaluation metrics such as accuracy (acc.), precision (prec.), recall, F-score (F), and AUC for multiclass classification task (Sokolova & Lapalme, 2009) [25]. It is important to clarify that these metrics will be calculated using the mean value obtained for each class. As shown in Table 3, the focus of these metrics is to represent the classifier's ability to both correctly predict and avoid false classification.

We further consider some non-conventional statistical metrics in order to get a better performance evaluation. These metrics are described in the following equations. In our approach, to represent the level of agreement between the predictions obtained for our ML models, we propose the Cohen's Kappa (K) score [26] as shown in Eq. (1), where Pa standing for the observed agreement ratio and Pe is the expected agreement. On the other hand, to indicate how close the prediction probability (p) is to the corresponding outcome (y) for N observations, we propose the use of the Log-loss (L-L) function [27] as denoted in Eq. (2).

Table 3. Description of the statistic evaluation metrics

Measure	Formula	Evaluation focus
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	How much the predictions are correctly classified.
Precision	$\frac{TP}{TP+FP}$	How often the positive predictions are actually positive.
Recall	$\frac{TP}{TP+FN}$	How much the predictions are correctly classified as positive.
Fscore	$\frac{TP}{TP+\frac{1}{2}(FP+FN)}$	Harmonic mean of the precision and recall.
AUC	$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)$	Classifier's ability to avoid false classification.

$$Kappa = \frac{P_a - P_e}{1 - P_e} \quad (1)$$

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N} y_i \ln p_i + (1 - y_i)\ln(1 - p_i) \quad (2)$$

4. RESULTS AND DISCUSSION

4.1. Evaluation of Data use Cases

As illustrated in Fig. 2, for the proposed unbalanced data use case, 844,060 original samples were used. On the other hand, for the balanced data use case, we used 223,890 samples obtained by using the under-sampling technique; meanwhile, 2,027,574 samples obtained by using the over-sampling technique were used. The data distribution is as follows.

For the unbalanced data case, when the fixed testing dataset was used the distribution was: 776,896 samples for training (A: 75,429 - F: 326,735 - G: 60,831 - K: 139,349 - M: 148,431 - C:26,121), and 67,164 samples for testing (11,194 samples per class). Likewise, when the

stratified testing dataset was used the distribution per class was: 590,842 samples for training (A: 60,636 - F: 236,550 - G: 50,418 - K: 105,380 - M: 111,738 - C: 26,120), and 253,218 samples for testing (A: 25,987 - F: 101,379 - G: 21,607 - K: 45,163 - M: 47,887 - C: 11,195).

On the other hand, for the balanced data case from the use of the under-sampling technique, when the fixed testing dataset has used the distribution was: 156,726 samples for training (26,121 samples per class), and 67,164 samples for testing (11,194 samples per class). Likewise, when the stratified testing dataset has used the distribution per class was: 156,726 samples for training (26,121 samples per class), and 67,164 samples for testing (11,194 samples per class).

For the balanced data case from the use of the over-sampling technique, when the fixed testing dataset has used the distribution was: 1,960,410 samples for training (326,735 samples per class), and 67,164 samples for testing (11,194 samples per class). Likewise, when the stratified testing dataset has used the distribution per class was: 1,419,300 samples for training (236,550 samples per class), and 608,273 samples for testing (101,379 samples per class).

In this way, the Random Forest algorithm was trained using the patterns proposed in Table 2, and its resulting statistic metrics are presented in Table 4. The results obtained from our ML approach are as follows. We can see that the best performance (0.94 of specificity and 0.94 of sensitivity for top prediction) was reached for the model developed from the use of balanced data through over-sampling technique. This is to be expected as the stratified testing dataset inevitably contains synthetic data that is likely to be easier to predict than the real data. On the other hand, we can note that there is no significant improvement between the use of original data (0.88 of specificity and 0.73 of sensitivity for top prediction) and the use of balanced data through the under-sampling technique (0.86 of specificity and 0.80 of sensitivity for top prediction).

4.2. The Effect of the Combinatorial use of Astronomical Data

The input pattern number seven outperforms the other input cases in all data use cases. Therefore, by analyzing the resulting metrics presented in Table 4 we can see how much each stellar parameter contributes to the predictive results of the trained model. In this way, we note that the complementary use of the derived stellar parameters helps to strengthen the information provided as input data to the Random Forest algorithm.

On the other hand, the contribution of the combinatorial use of astronomical data is interesting. This is because a contribution from the stellar data was demonstrated, contrary to the Pearson correlation [28] analysis between spectral and stellar data presented in Fig. 3. For instance, we can see this in the decrease of the Log-loss rate when the model is trained using patterns that contain more astronomical information. We further can see that both the fairly reddening-insensitive pseudocolor and the calculated variance from spectroscopic data contribute to slightly improving the prediction capacity of the model. In the same way, when using the calculated variance from stellar parameters the model did not show improvements.

Table 4. The resulting evaluation metrics from the proposed ML approach

Data use case	Testing dataset	Input pattern	Acc.	Prec.	F	AUC	K	L-L
Unbalanced (original)	Fixed	1	0.54	0.59	0.52	0.87	0.45	2.74
		2	0.53	0.58	0.51	0.86	0.44	2.96
		3	0.52	0.53	0.49	0.86	0.43	2.47
		4	0.53	0.58	0.50	0.88	0.44	1.73
		5	0.53	0.57	0.50	0.88	0.44	1.79
		6	0.55	0.59	0.51	0.89	0.46	1.74
		7	0.56	0.60	0.51	0.90	0.47	1.57
		8	0.54	0.57	0.50	0.88	0.45	1.89
	Stratified	1	0.79	0.78	0.78	0.94	0.72	0.76
		2	0.86	0.85	0.85	0.97	0.81	0.48
		3	0.86	0.86	0.86	0.97	0.81	0.48
		4	0.87	0.87	0.87	0.98	0.82	0.44
		5	0.87	0.87	0.87	0.98	0.83	0.41
		6	0.87	0.87	0.87	0.98	0.83	0.38
		7	0.87	0.88	0.87	0.98	0.83	0.35
		8	0.87	0.87	0.87	0.98	0.83	0.37
Balanced (undersampled)	Fixed	1	0.38	0.55	0.38	0.72	0.26	9.15
		2	0.46	0.70	0.48	0.79	0.35	6.01
		3	0.47	0.71	0.49	0.80	0.37	5.04
		4	0.52	0.68	0.53	0.82	0.42	4.44
		5	0.50	0.74	0.52	0.82	0.40	4.53
		6	0.51	0.70	0.53	0.86	0.41	3.18
		7	0.50	0.73	0.52	0.84	0.39	3.54
		8	0.51	0.73	0.53	0.86	0.41	2.92
	Stratified	1	0.75	0.75	0.75	0.95	0.70	0.91
		2	0.75	0.75	0.75	0.95	0.70	0.91
		3	0.83	0.83	0.83	0.97	0.80	0.56
		4	0.85	0.85	0.85	0.98	0.82	0.53
		5	0.85	0.85	0.85	0.98	0.82	0.49
		6	0.86	0.86	0.86	0.98	0.83	0.44
		7	0.86	0.86	0.86	0.98	0.83	0.41
		8	0.86	0.86	0.86	0.98	0.83	0.44
Balanced (oversampled)	Fixed	1	0.46	0.59	0.47	0.78	0.35	5.82
		2	0.53	0.67	0.55	0.84	0.44	3.48
		3	0.54	0.63	0.55	0.85	0.45	2.56
		4	0.55	0.63	0.55	0.86	0.45	2.17
		5	0.55	0.64	0.56	0.86	0.46	2.21
		6	0.60	0.68	0.60	0.90	0.52	1.68
		7	0.59	0.67	0.59	0.90	0.51	1.50
		8	0.58	0.65	0.58	0.90	0.49	1.59
	Stratified	1	0.86	0.86	0.86	0.98	0.83	0.45
		2	0.92	0.92	0.92	0.99	0.91	0.24
		3	0.93	0.93	0.93	0.99	0.92	0.23
		4	0.94	0.94	0.94	0.99	0.92	0.22
		5	0.94	0.94	0.94	0.99	0.93	0.20
		6	0.94	0.94	0.94	0.99	0.93	0.19
		7	0.94	0.94	0.94	0.99	0.93	0.18
		8	0.94	0.94	0.94	0.99	0.93	0.19

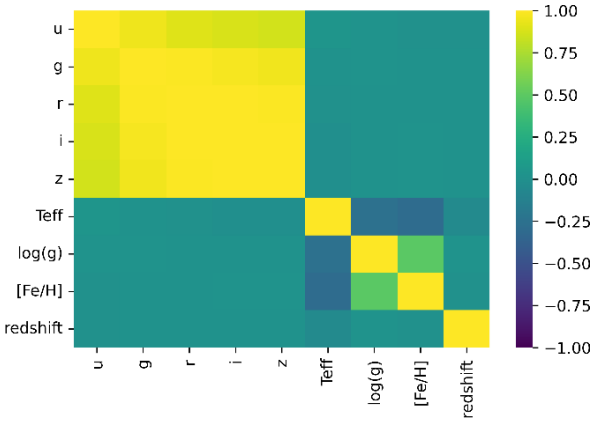


Figure 3. Pearson correlation between spectral data and derived stellar parameters

4.3. Evaluation using Different Testing Datasets

From Table 4, it is possible to notice that when evaluating the model using a stratified and a fixed testing dataset, the results are different. The stratified dataset achieves significantly better results than the fixed testing dataset. For instance, we can notice an average increase of 35% and 41% in the F-score and Cohen's Kappa score, respectively, in those models that achieved the best predictive rates (trained with input pattern number seven).

On the other hand, an interesting case to evaluate is the case of using balanced data through the under-sampling technique, since an adequate representation of the samples was ensured, for both testing datasets, using the same amount of data per class. Therefore, it was possible to demonstrate the usefulness of our approach by comparing both testing datasets. This is because, although they appear similar, stratified sampling differs from simple random sampling by dividing the samples into strata, based on shared characteristics.

4.4. Performance Evaluation

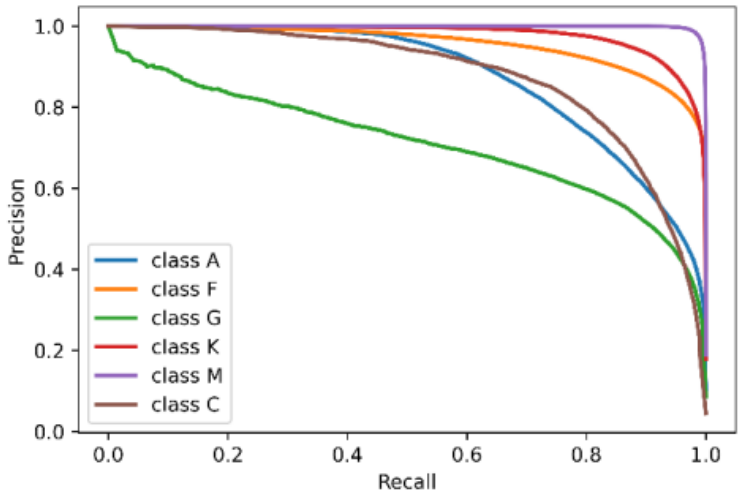
In order to analyze the class imbalance, we plot the Precision-Recall curves of the models trained using the proposed input pattern number seven. The resulting scores and curves for the three data use cases are shown in Table 5 and Fig. 4, respectively. Thus, we can note that A, F, and G-type star classes are more difficult to classify for our models; meanwhile, a better performance level is always reached for the M-type stars. Furthermore, it is also interesting that by using the under-sampling technique, our predictive model improves the precision and recall rates for some spectral classes such as G and C-type, but at the same time reduces its performance for the F-type star class. On the other hand, we noted that by using synthetic data it is possible to improve precision and recall rates, but the F-type star class does not achieve a significant improvement.

Table 5. The resulting mean precision and recall scores per class

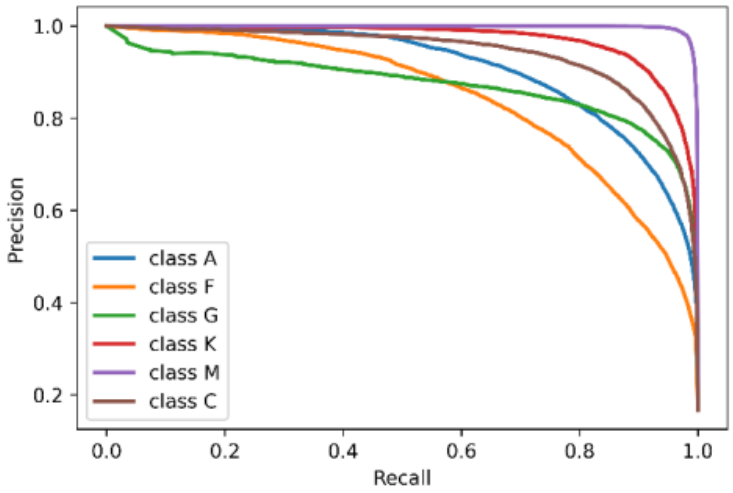
Class	Data use case	Prec.	Recall
A	Unbalanced (original)	0.81	0.73
	Balanced (undersampled)	0.82	0.80
	Balanced (oversampled)	0.93	0.94
F	Unbalanced (original)	0.86	0.91
	Balanced (undersampled)	0.78	0.73

	Balanced (oversampled)	0.90	0.83
G	Unbalanced (original)	0.68	0.63
	Balanced (undersampled)	0.80	0.86
	Balanced (oversampled)	0.89	0.95
K	Unbalanced (original)	0.92	0.92
	Balanced (undersampled)	0.91	0.90
	Balanced (oversampled)	0.96	0.95
M	Unbalanced (original)	0.99	0.98
	Balanced (undersampled)	0.98	0.98
	Balanced (oversampled)	0.99	0.99
C	Unbalanced (original)	0.82	0.77
	Balanced (undersampled)	0.86	0.88
	Balanced (oversampled)	0.96	0.97

(A)



(B)



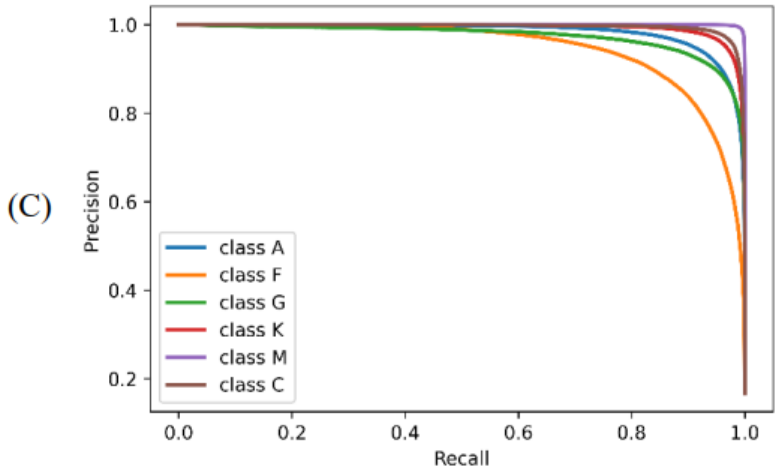


Figure 4. The resulting Precision-Recall curve from the use of (A) unbalanced original data, and balanced data by using (B) under-sampling and (C) over-sampling techniques

4.5. Benchmark Evaluation

Finally, we further analyze our resulting scores with some benchmark results obtained from similar works (Sharma et al. [6], Lu et al. [7], and Dafonte et al. [8], Zhang et al. [9]). These state-of-the-art methods were briefly detailed in Section 2. Thus, based on the comparative results shown in Table 6, we can highlight the following relevant aspects:

- Our proposed method was developed using the most recent astronomical data.
- Our results are likely to be more realistic because our model was developed with a larger amount of data. This means that more relevant and reliable data were considered.
- Our method proposes a higher classification capacity by considering not only classes ofthe MK system, but also having the capacity to recognize C-type stars.
- The method presented by Zhang et al. reached the best accuracy, however, it wasproposed for the binary classification task.

Table 6. Comparative results between some state-of-the-art methods and our proposed method

Method	**Classification capacity**	**Database used**	**Test set**	**Acc.**
Sharma et al.	O-B-A-F-G-K-M types	CFLIB	850	0.89
Lu et al.	F-G and G-K types	LAMOST (DR6)	273 and 260	0.80
Dafonte et al.	A-B-F-G-K-M types	Built by themselves	100	0.83
Zhang et al.	A-F types	SDSS (DR15)	9,500	0.97
Ours (unbalanced (original))	A-F-G-K-M-C types	SDSS (DR17)	253,218	0.87
Ours (balanced (undersampled))	A-F-G-K-M-C types	SDSS (DR17)	67,164	0.86
Ours (balanced (oversampled))	A-F-G-K-M-C types	SDSS (DR17)	608,273	0.94

5. CONCLUSION

In this work, we introduced a ML approach to classify stars according to the spectral main classes: A, F, G, K, M, and C-type. In this way, we demonstrate that using the widely used Random Forest algorithm it is possible to analyze spectral data and stellar parameters as input patterns and then classify them. In particular, the proposed approach is also interesting for its simplicity of implementation, thus being able to contribute to the analysis carried out by astronomers. Despite the unbalanced nature of the data collected from astronomical surveys, by using both balanced data technique and supervised learning we reached acceptable prediction rates for this challenging multiclassification task.

In future works, from the supervised learning point of view, we intend to study the use of DL models to carry out a comparative analysis between ML and DL applied to the task of spectral classification of stars.

REFERENCES

[1] M. Fukugita, T. Ichikawa, J. E. Gunn, M. Doi, K. Shimasaku, D. P. Schneider, “The Sloan Digital Sky Survey Photometric System”, Astronomical Journal, vol. 111, no. 1, pp. 1748, Apr. 1996, doi: 10.1086/117915.

[2] W. W. Morgan, P. C. Keenan, E. Kellman, An Atlas of Stellar Spectra with an Outline of Spectral Classification in Astrophysical Monographs sponsored by The Astrophysical Journal, P. W. Merril, J. H. Moore, H. Shapley and O. Struve, Eds., Chicago, IL, USA: University of Chicago Press, 1943.

[3] J. D. Kirkpatrick, I. N. Reid, J. Liebert, R. M. Cutri , B. Nelson, C. A. Beichman, C. C. Dahn, D. G. Monet, J. E. Gizis, M. F. Skrutskie, “Dwarfs Cooler than "M": The Definition of Spectral Type "L" Using Discoveries from the 2-Micron All-Sky Survey (2MASS)”, The Astrophysical Journal, vol. 519, no. 1, pp. 802-833, Jul. 1999, doi: 10.1086/307414.

[4] A. Gonneau, A. Lançon, S. C. Trager, B. Aringer, M. Lyubenova, W. Nowotny, R. F. Peletier, P. Prugniel, Y. P. Chen, M. Dries, O. S. Choudhury, J. Falcón-Barroso, M. Koleva, S. Meneses- Goytia, P. Sánchez-Blázquez, A. Vazdekis, “Carbon stars in the X-Shooter Spectral Library”, Astronomy and Astrophysics, vol. 589, no. 1, pp. 1-26, Apr. 2016, doi: 10.1051/0004- 6361/201526292.

[5] M. R. Blanton, M. A. Bershady, B. Abolfathi, F. D. Albareti, C. A. Prieto, A. Almeida, J. A. García, F. Anders, S. F. Anderson, B. Andrews, et al., “Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe”, The Astronomical Journal, vol. 154, no. 1, pp. 28-62, Jul. 2017. doi:10.3847/1538-3881/aa7567.

[6] K. Sharma, A. Kembhavi, A. Kembhavi, T. Sivarani, S. Abraham, K. Vaghmare, “Application of convolutional neural networks for stellar spectral classification”, Monthly Notices of the Royal Astronomical Society, vol. 491, no. 2, pp. 2280-2300, Jan. 2020. doi:10.1093/mnras/stz3100.

[7] Y. Lu, B. Qiu, G. Xiang, M. Li, Z. He, “Stellar Spectral Classification with 2D Spectrum and Fully Connected Neural Network”, Journal of Physics: Conference Series, vol. 1626, no. 1, pp. 12-16, Oct. 2020. doi:10.1088/1742-6596/1626/1/012016.

[8] C. Dafonte, A. Rodriguez, M. Manteiga, A. Gomez, B. Arcay, “A Blended Artificial Intelligence Approach for Spectral Classification of Stars in Massive Astronomical Surveys”, Entropy, vol. 22, no. 5, pp. 1-26, May. 2020. https://doi.org/10.3390/e22050518.

[9] J. Zhang, Y. Zhang, Y. Zhao, "RR Lyrae Star Candidates from SDSS Databases by Cost-sensitive Random Forests", The Astrophysical Journal Supplement Series, vol. 246, no. 8, pp. 1-8, Jan. 2020, doi: 10.3847/1538-4365/ab5a7c.

[10] J. Adassuriya, J. A. N. S. S. Jayasinghe, K. P. S. C. Jayaratne, "Identifying Variable Stars from Kepler Data Using Machine Learning", European Journal of Applied Physics, vol. 3, no. 4, pp. 32-37, Jul. 2021, doi: 10.24018/ejphysics.2021.3.4.93.

[11] K. Naydenkin, K. Malanchev, M. Pruzhinskaya, "Variable Stars Classification with the Help of Machine Learning", in CEUR Workshop Proceedings, 22nd International Conference on Data Analytics and Management in Data Intensive Domains, DAMDID/RCDL 2020, Voronezh, Oct. 2020, pp. 296-303.

[12] Z. Hosenie, R. Lyon, B. Stappers, A. Mootoovaloo, "Comparing Multi-class, Binary and Hierarchical

Machine Learning Classification schemes for variable stars", Monthly Notices of the Royal Astronomical Society, vol.1 , no. 1, pp. 1-16, Jul. 2019, doi: 10.1093/mnras/stz1999.

[13] Z. Hosenie, R. Lyon, B. Stappers, A. Mootoovaloo, V. McBride "Imbalance Learning for Variable Star Classification", Monthly Notices of the Royal Astronomical Society, vol.1 , no. 1, pp. 1-11, Feb. 2020, doi: 10.1093/mnras/staa642.

[14] V. Acharya, P. S. Bora, K. Navin, A. Nazareth, P. S. Anusha, S. Rao, " Classification of SDSS photometric data using machine learning on a cloud", Current Science, vol. 115, no. 2, pp. 249- 257, Jul. 2018, doi: 10.18520/cs/v115/i2/249-257.

[15] M. Wierzbinski, P. Plawiak, M. Hammad, U. R. Acharya, Development of accurate classification of heavenly bodies using novel machine learning techniques"", Soft Computing, vol. 25, no. 1, pp. 7213-7228, Feb 2021, doi: 10.1007/s00500-021-05687-4.

[16] A. Martinazzo, M. Espadoto, N. S. T. Hirata, " Deep Learning for Astronomical Object Classification: A Case Study", in Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020), Valleta, Feb. 2020, pp. 87-95.

[17] S. Ethiraj, B. K. Bolla, "Classification of Quasars, Galaxies, and Stars in the Mapping of the Universe Multi-modal Deep Learning", ArXiv preprint, vol. 1, no.1, pp. 1-7, May. 2022, doi: 10.48550/arXiv.2205.10745 .

[18] P. Prugniel, C. Soubiran, "A database of high and medium-resolution stellar spectra", Astronomy and Astrophysics, vol. 369, no. 3, pp. 1048-1057, Jan. 2001, doi: 10.1051/0004-6361:20010163.

[19] Abdurro'uf , K. Accetta, C. Aerts, V. S. Aguirre, R. Ahumada, N. Ajgaonkar, N. F. Ak, S. Alam, C. A. Prieto, A. Almeida, et al., "The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar, and APOGEE-2 Data", The Astrophysical Journal Supplement Series, vol. 259, no. 35, pp. 1-39, Apr. 2022, doi: 10.3847/1538-4365/ac4414.

[20] I. S. Bowen, A. H. Vaughan, "The Optical Design of the 40-in. Telescope and of the Irenee DuPont Telescope at Las Campanas Observatory, Chile," Appl. Opt., vol. 12, no. 7, pp. 1430-1435, Jul. 1973, doi: 10.1364/AO.12.001430.

[21] J. E. Gunn, W. A. Siegmund, E. J. Mannery, R. E. Owen, C. L. Hull, R. F. Leger, L. N. Carey, G. R. Knapp, D. G. York, W. N. Boroski, et al., "The 2.5 m Telescope of the Sloan Digital Sky Survey", The Astronomical Journal, vol. 131, no. 4, pp. 2332-2359, Apr. 2006, doi: 10.1086/500975.

[22] L. Breiman, "Random Forests", Machine Learning, vol. 45, no. 1, pp. 5-32, Oct. 2001, doi: 10.1023/A:1010933404324.

[23] C. Cáceres, M. Catelan, "The Period-Luminosity Relation of RR Lyrae Stars in the SDSS Photometric System", The Astrophysical Journal Supplement Series, vol. 179, no. 1, pp. 242-248, Nov. 2008, doi: 10.1086/591231 .

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, " SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, vol. 16, no. 1, pp. 321-357, Jun. 2002, doi: 10.1613/jair.953

[25] M. Sokolova, G. Lapalme, " A systematic analysis of performance measures for classification tasks ", Information Processing and Management, vol. 45, no. 1, pp. 427-437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[26] M. L. McHugh, "Interrater reliability: the kappa statistic", Biochemia Medica, vol. 22, no. 3, pp. 276–282, Oct. 2012, PMCID: PMC3900052.

[27] V. Vovk, The Fundamental Nature of the Log Loss Function in Fields of Logic and Computation II, L. Beklemishev, A. Blass, N. Dershowitz, B. Finkbeiner, W. Schulte, eds., Springer, Cham, 2015, pp. 307-318.

[28] W. Kirch, Pearson’s Correlation Coefficient in Encyclopedia of Public Health, W. Kirch, eds., Springer Dordrecht, 2008, pp. 1090-1091.

AUTHORS

Marco Oyarzo Huichaqueo holds a B.S. degree in Automation Engineering and Industrial Control from the Technological University of Chile Inacap, since 2015. In 2021, he received his M.Sc. degree in Computational Engineering and Mathematics from the Rovira i Virgili University of Catalonia. Currently he is a member of the Institute of Electrical and Electronics Engineers (IEEE) and affiliated to the International Federation of Automatic Control (IFAC). He has five years of industrial research and development experience and has also published some papers in international journals and conferences. His research interests include robotics and artificial intelligence applications, machine learning and deep learning methods, computer vision, and industrial automation. He can be contacted at email: markooyarzoh@gmail.com.

Renato Muñoz Orrego holds a B.S. degree in Industrial Control and Instrumentation Engineering from the Federico Santa María Technical University, since 2016. He is currently a student of the M.Sc. degree in Robotics and Automation at the Technical University of Madrid, since 2021. He is currently carrying out his research thesis at the European Organization for Nuclear Research (CERN). His research interests include intelligent robotics, human-robot interaction, and environment perception using machine learning and computer vision. He can be contacted at email: rena.munozo@gmail.com.