

Classification of Astronomical Spectra Using PCA-Compressed Flux and Inverse-Variance Features

Bruno Santos Meneses Barreto and Marcio Eisencraft

Abstract—This paper evaluates a signal-processing and supervised-learning pipeline for classifying SDSS DR17 spectra into stars, galaxies, and quasars. Flux and inverse-variance vectors are standardized, separately compressed by PCA, and used by 12 classifiers. Small MLP and Hist. GBDT tied for the highest F+I accuracy (93.9%), while Small MLP achieved the highest macro-F1 (92.1%) and LightGBM the highest balanced accuracy (91.9%). Ten repeated stratified train/validation splits with a fixed test set showed low variability among the leading models, while LightGBM benefited most clearly from inverse-variance features.

Keywords—Astronomical spectra, signal classification, machine learning, principal component analysis

I. INTRODUCTION

Large-scale astronomical surveys have transformed observational astrophysics into a data-intensive field, producing massive collections of spectra from different types of astronomical objects. In particular, the Sloan Digital Sky Survey (SDSS) has provided a large public database of spectroscopic observations, enabling the development of automatic methods for object identification and physical analysis [1], [2]. Since each spectrum can be interpreted as a one-dimensional sampled signal containing continuum information, absorption features, emission lines, noise, and instrumental effects, spectral classification can be cast as a signal classification and pattern recognition problem.

The automatic classification of astronomical spectra into stars, galaxies, and quasars is an important step in large survey pipelines [3]. Stars are mostly characterized by stellar continua and absorption-line patterns, galaxies often exhibit composite stellar populations, and quasars are typically associated with broad emission lines and non-stellar continua [4], [5]. In this context, machine learning methods provide a practical framework for learning nonlinear decision boundaries from high-dimensional spectral data. Recent end-to-end CNNs have also achieved strong performance on SDSS DR17 spectra [6]; here, we focus instead on compact PCA representations and the added value of inverse-variance information.

In this work, we investigate supervised learning methods for classifying SDSS spectra into three classes: STAR, GALAXY, and QSO (quasi-stellar object). The evaluated pipeline uses both spectral flux and inverse-variance information, allowing the classifiers to exploit the measured signal together with a wavelength-dependent reliability profile. The spectra are resampled onto a common logarithmic wavelength grid, standardized, compressed using principal component analysis (PCA) [7], and then used as input to a set of classification models ranging

from linear methods, such as logistic regression and linear support vector machines (SVMs) [8], [9], to ensemble-based gradient boosting approaches [10], [11]. The main contribution is a controlled comparison of flux-only and flux-plus-inverse-variance representations across 12 classifiers, complemented by repeated-split stability and class-wise error analysis.

II. DATASET

The dataset used in this work was built from spectroscopic data in the SDSS Data Release 17 (SDSS DR17) catalog [12]. The corresponding FITS files [13] were retrieved for the selected objects. We retained only spectra from observations with acceptable quality (`PLATEQUALITY = good` or `marginal`), with no SDSS redshift/classification warning flags (`ZWARNING = 0`), and corresponding to the primary spectroscopic observation of each object (`SPECPRIMARY = 1`). For each spectrum, we extracted the measured flux, the logarithmic wavelength grid, the inverse variance provided by the SDSS pipeline, and the class label.

After data selection, the final dataset contained 32 259 spectra and is publicly available on Hugging Face¹. The dataset was divided into training, validation, and test subsets using a stratified random split with proportions of 70%, 15%, and 15%, respectively. This resulted in 22 581 spectra for training, 4 839 for validation, and 4 839 for testing.

The training set was imbalanced, with GALAXY spectra forming the majority class. It contained 14 243 GALAXY spectra, 4 276 STAR spectra, and 4 062 QSO spectra. Therefore, class weights inversely proportional to class frequencies were computed from the training set and used whenever supported by the classification model. This also motivated the use of balanced accuracy and macro-F1, in addition to overall accuracy, for performance evaluation.

III. DATA PREPROCESSING

The raw SDSS DR17 spectra were preprocessed to obtain a common-grid representation suitable for supervised classification. We first determined the intersection of the log-wavelength ranges covered by the selected spectra, where $\log \lambda = \log_{10}(\lambda)$ and λ is measured in angstroms. A fixed grid with 4 000 uniformly spaced points was then defined over this interval. The resulting grid spans approximately from $\log \lambda = 3.70$ to $\log \lambda = 3.80$, corresponding to about 5 000–6 300 Å.

The grid was restricted to the wavelength interval shared by all spectra to avoid extrapolation. Using the union of the observed ranges would require padding or extrapolating

Departamento de Engenharia de Telecomunicações e Controle, Universidade de São Paulo, São Paulo-SP, e-mail: bruno.smbarroto@usp.br; marcioft@usp.br. This work was partially supported by CNPq (404081/2023-1).

¹https://huggingface.co/datasets/BrunoBarreto/sdss_dr17_spectra_for_classification

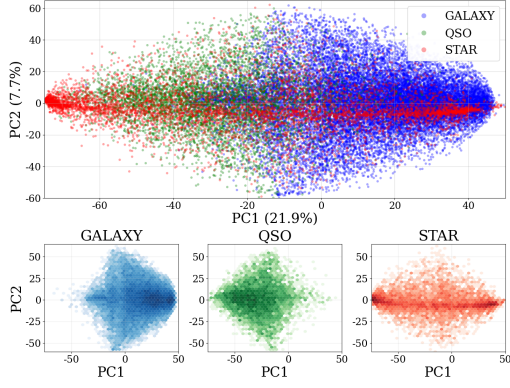


Fig. 1

FLUX PROJECTION ONTO PC1 AND PC2, EXPLAINING 21.9% AND 7.7% OF THE VARIANCE.

many STAR and GALAXY spectra, potentially introducing nonphysical features.

To reduce the effect of absolute flux scaling, each spectrum was standardized independently. Given a flux vector $\mathbf{x}_i \in \mathbb{R}^p$, where $p = 4000$, the standardized spectrum $\tilde{\mathbf{x}}_i$ is defined by

$$\bar{x}_i = \frac{1}{p} \sum_{j=1}^p x_{ij}, \quad s_i = \sqrt{\frac{1}{p} \sum_{j=1}^p (x_{ij} - \bar{x}_i)^2}, \quad (1)$$

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}.$$

The same per-spectrum standardization was applied to the inverse-variance vectors before dimensionality reduction. For the flux vectors, this standardization makes the model focus on relative spectral shape and line patterns rather than on absolute flux scale. For the inverse-variance vectors, the standardized representation is interpreted as a relative wavelength-dependent reliability profile within each spectrum.

PCA was applied separately to the standardized flux and inverse-variance matrices. In both cases, the PCA transformation was fitted only on the training data and then applied to the validation and test sets, avoiding information leakage. The resulting flux and inverse-variance components were concatenated only after projection, preserving modality-specific bases while giving both modalities the same number of coefficients.

Figure 1 shows the projection of the standardized flux spectra onto the first two principal components, which explain 21.9% and 7.7% of the variance, respectively. The projection shows substantial class overlap in the first two components, motivating the use of a higher-dimensional PCA representation followed by nonlinear classifiers.

IV. CLASSIFICATION MODELS

For each spectrum, let $\mathbf{c}_i^{(\text{flux})}, \mathbf{c}_i^{(\text{ivar})} \in \mathbb{R}^d$ denote the PCA representations. They are concatenated as

$$\mathbf{z}_i = \begin{bmatrix} \mathbf{c}_i^{(\text{flux})} \\ \mathbf{c}_i^{(\text{ivar})} \end{bmatrix} \in \mathbb{R}^{2d}. \quad (2)$$

We used $d = 24$, beyond which validation performance showed no substantial improvement. The retained components explain

55.39% and 85.53% of the variance in the flux and inverse-variance representations, respectively; thus, F and F+I used 24 and 48 features.

The benchmark covered complementary classifier families. Logistic regression and linear SVM [8], [9] served as linear baselines, testing whether the compressed representations could be separated by simple decision boundaries. Gaussian NB, LDA, QDA, and k-NN [14] provided probabilistic, covariance-based, and local-neighborhood alternatives. Random forest [15], Extra Trees [16], gradient boosting [10], XGBoost [17], and LightGBM [11] were included to capture nonlinear interactions among the PCA coefficients.

The Small MLP [18] used two hidden layers with 64 and 32 units, ReLU activations, and early stopping, providing a compact nonlinear neural baseline rather than a large end-to-end spectral model. All classifiers used the same partitions and the same F or F+I representations. Class weights were applied when supported, and hyperparameters were selected through small model-specific grid searches maximizing validation macro-F1 (e.g., for LightGBM, $n_{\text{estimators}} \in \{120, 180\}$ and $\text{num_leaves} \in \{16, 32\}$, resulting in four candidate configurations). We report accuracy, balanced accuracy, macro-F1, and weighted F1 using scikit-learn [19].

V. EXPERIMENTAL RESULTS

Table I summarizes the test-set performance. The table compares classifiers trained using both flux and inverse-variance information (F+I) against models trained only with flux features (F). The highest F+I point estimates were distributed among the leading models: Small MLP and Hist. GBDT tied in accuracy (0.939), Small MLP achieved the highest macro-F1 and weighted-F1 scores (0.921 and 0.939), and LightGBM achieved the highest balanced accuracy (0.919). These point estimates are close among the leading models, and the repeated-split results indicate that their ranking may vary across development partitions.

TABLE I

TEST PERFORMANCE WITH $d = 24$ COMPONENTS PER MODALITY. F+I USES FLUX AND INVERSE VARIANCE; F USES FLUX ONLY.

Model	Acc.		Bal. Acc.		Macro-F1		W-F1	
	F+I	F	F+I	F	F+I	F	F+I	F
LightGBM	0.934	0.921	0.919	0.913	0.916	0.902	0.935	0.922
Hist. GBDT	0.939	0.933	0.905	0.898	0.919	0.912	0.938	0.932
XGBoost	0.931	0.927	0.889	0.885	0.907	0.905	0.929	0.926
Small MLP	0.939	0.935	0.914	0.904	0.921	0.916	0.939	0.934
Random Forest	0.916	0.915	0.861	0.856	0.891	0.888	0.914	0.912
QDA	0.902	0.883	0.849	0.833	0.871	0.850	0.900	0.880
Logistic Reg.	0.839	0.815	0.828	0.803	0.801	0.776	0.847	0.828
Linear SVM	0.860	0.843	0.816	0.792	0.810	0.787	0.863	0.847
Extra Trees	0.926	0.924	0.881	0.878	0.902	0.899	0.925	0.922
LDA	0.851	0.832	0.775	0.739	0.792	0.760	0.850	0.830
Gaussian NB	0.837	0.865	0.766	0.802	0.789	0.822	0.833	0.862
k-NN	0.852	0.868	0.750	0.770	0.794	0.814	0.841	0.858

The four strongest classifiers were also assessed over ten repeated stratified train/validation splits while keeping the original test set fixed. At each repetition, train/validation data were stratified and reshuffled, PCA was refitted, and models were retrained with fixed hyperparameters. Table II shows low variability: Hist. GBDT led mean accuracy, macro-F1, and weighted F1, while LightGBM led balanced accuracy. The standard deviations remained between 0.003 and 0.007 across

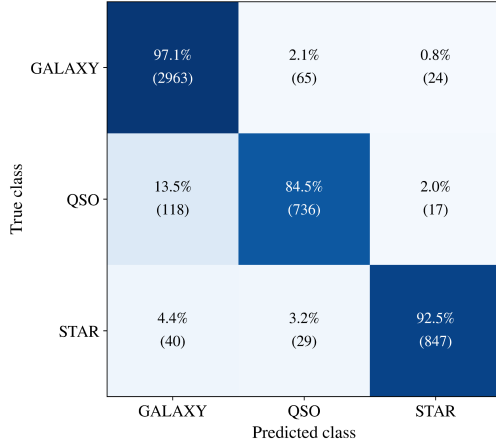


Fig. 2

ROW-NORMALIZED SMALL MLP CONFUSION MATRIX WITH F+I FEATURES.

the reported metrics, showing that the leading methods were not strongly dependent on a single development split. At the same time, changes in ranking caution against treating the fixed-test ordering as definitive.

TABLE II

 STABILITY OVER TEN REPEATED STRATIFIED TRAIN/VALIDATION SPLITS WITH A FIXED TEST SET AND F+I FEATURES (MEAN $\pm$ STANDARD DEVIATION).

Model	Acc.	Bal. Acc.	Macro-F1	W-F1
LightGBM	0.936 ± 0.003	0.917 ± 0.005	0.917 ± 0.004	0.936 ± 0.003
Hist. GBDT	0.939 ± 0.004	0.904 ± 0.007	0.920 ± 0.006	0.938 ± 0.004
XGBoost	0.932 ± 0.005	0.892 ± 0.007	0.910 ± 0.006	0.931 ± 0.005
Small MLP	0.935 ± 0.004	0.905 ± 0.006	0.914 ± 0.005	0.934 ± 0.004

PCA compression preserved strong discrimination, but the value of inverse variance was model dependent. For LightGBM, adding inverse-variance features increased test accuracy from 0.921 to 0.934 and macro-F1 from 0.902 to 0.916. This was the clearest F+I improvement among the leading models. Differences between the strongest F+I classifiers were comparatively small, so the results do not indicate a definitive overall winner.

The same benefit did not hold for every classifier. Gaussian NB and k-NN performed better with flux alone, while the F+I improvement of Random Forest was small. The representation should therefore be evaluated jointly with the classifier: additional reliability coefficients are useful only when the model can exploit them without being adversely affected by the larger feature space.

Figure 2 shows the Small MLP confusion matrix. GALAXY and STAR recalls were 97.1% and 92.5%, while QSO recall was 84.5%. The main error was QSO classified as GALAXY (118 of 871, or 13.5%); class-wise F1 scores were 0.960, 0.865, and 0.939 for GALAXY, QSO, and STAR.

A limitation of the current representation is its restriction to the observed-frame wavelength interval shared by all selected spectra. This avoids extrapolation, but discards class-specific coverage and potentially informative features outside 5,000–6,300 Å. The larger QSO error rate motivates future evaluation of broader grids with explicit coverage masks or missing-data treatment, as well as validation on spectra from other surveys.

VI. CONCLUSION

This work evaluated PCA-compressed flux and inverse-variance features for classifying SDSS DR17 spectra into

STAR, GALAXY, and QSO. The Small MLP tied with Hist. GBDT for the highest F+I accuracy and achieved the highest macro-F1, whereas LightGBM led balanced accuracy. Repeated splits indicated stable performance among the leading models, and LightGBM showed the clearest test-set improvement from the addition of inverse-variance features. QSO remained the most difficult class, mainly through confusion with galaxies. This behavior may partly reflect the restricted observed-frame interval: depending on redshift, discriminative QSO emission features may fall outside the common grid, potentially reducing class separability.

The results show that a compact 48-feature F+I representation can support strong nonlinear classification performance without operating directly on all 8,000 wavelength-grid values. Inverse-variance information can therefore provide a complementary reliability signal, although its effect depends on how the classifier combines it with flux information. Future work should evaluate coverage-aware representations over wider wavelength ranges and test transfer to spectra from other surveys.

REFERENCES

- [1] D. G. York *et al.*, “The Sloan Digital Sky Survey: Technical summary,” *The Astronomical Journal*, vol. 120, p. 1579–1587, Sept. 2000.
- [2] Ž. Ivezić, A. J. Connolly, J. T. VanderPlas, and A. Gray, *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*. Princeton University Press, 2014.
- [3] A. S. Bolton *et al.*, “Spectral classification and redshift measurement for the SDSS-III baryon oscillation spectroscopic survey,” *The Astronomical Journal*, vol. 144, no. 5, p. 144, 2012.
- [4] C.-W. Yip *et al.*, “Distributions of galaxy spectral types in the Sloan Digital Sky Survey,” *The Astronomical Journal*, vol. 128, no. 2, pp. 585–609, 2004.
- [5] D. E. Vanden Berk *et al.*, “Composite quasar spectra from the Sloan Digital Sky Survey,” *The Astronomical Journal*, vol. 122, no. 2, pp. 549–564, 2001.
- [6] R. Moradi, F. Rastegarnia, Y. Wang, and M. T. Mirtorabi, “FNet II: Spectral classification of quasars, galaxies, stars, and broad absorption line (BAL) quasars,” *Monthly Notices of the Royal Astronomical Society*, vol. 533, no. 2, pp. 1976–1985, 2024.
- [7] I. T. Jolliffe, *Principal Component Analysis*. Springer Series in Statistics, New York, NY: Springer, 2 ed., Oct. 2002.
- [8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 ed., 2009.
- [9] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, p. 273–297, Sept. 1995.
- [10] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189 – 1232, 2001.
- [11] G. Ke *et al.*, “LightGBM: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.
- [12] Abdurro’uf *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, p. 35, mar 2022.
- [13] D. C. Wells, E. W. Greisen, and R. H. Harten, “FITS: a flexible image transport system,” *Astronomy and Astrophysics Supplement Series*, vol. 44, pp. 363–370, 1981.
- [14] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [15] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.
- [16] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Mach. Learn.*, vol. 63, pp. 3–42, Apr. 2006.
- [17] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, (New York, NY, USA), p. 785–794, Association for Computing Machinery, 2016.
- [18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [19] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.