Bulletin de la Société Royale des Sciences de Liège, Vol. 88, Actes de Colloques, 2019, p. 174 - 181

Detecting Outliers in SDSS using Convolutional Neural Network

Kaushal Sharma[1*], Ajit Kembhavi[1], Aniruddha Kembhavi[2],
Thirupathi Sivarani[3], Sheelu Abraham[1]

[1] Inter University Centre for Astronomy and Astrophysics (IUCAA), Pune, India
[2] Allen Institute for Artificial Intelligence (AI2), Seattle, USA
[3] Indian Institute of Astrophysics (IIA), Bengaluru, India

Abstract: We propose an automated algorithm based on Convolutional Neural Network (CNN) for the detection of peculiar objects in large databases using their spectral observations. A convolutional neural network is a class of deep-learning algorithms which allows the detection of significant features/patterns in sequential data like images, audio, time-series etc. by applying convolutional neurons (kernels) along the sequence. For detecting unusual spectra, we use eight-layer deep convolutional network with autoencoder architecture on $\sim$ 60,000 spectra collected from the Sloan Digital Sky Survey. The training of the network is done in an unsupervised manner. We show that the trained network is able to retrieve the spectra of rare objects from a large collection of spectra. Such algorithms can easily be rescaled to other surveys and therefore can serve as a potential component of the data reduction pipelines for automatically detecting spectra with unusual features and recovering defective spectra.

Keywords: Astronomical databases – Spectroscopy – Deep learning – Autoencoder – Outlier detection

1 Introduction

Large-scale spectroscopic surveys, such as the Sloan Digital Sky Survey (SDSS, York et al. 2000) and LAMOST (Large Sky Area Multi-Object Fibre Spectroscopic Telescope, Cui et al. 2012), consist of an enormous number of imaging and spectroscopic observations. The latest data release from SDSS, DR 15 (Aguado et al. 2019), contains about four million optical spectra for various astronomical objects. Similarly, the LAMOST survey provides about 700,000 spectral observations in its current data release 6. After obtaining these spectroscopic observations, one of the primary tasks is to classify them. This is performed by the survey specific data-processing pipelines which broadly classify spectra and assign a class ('STAR', 'GALAXY' or 'QSO') depending upon the type of object which they belong to. These classes are referred to as primary classes. A further sub-classification and spectral analysis is also performed in each class by these pipelines. However, the huge amount of data generated by such survey programs prompts the application of machine-learning (ML) and

*kaushals@iucaa.in

deep-learning (DL) algorithms for analysing data more efficiently in lesser time. For example, Ball et al. (2004) demonstrate the use of supervised Artificial Neural Network (ANN) for deriving morphological class, spectral type and redshift for the galaxy photometric data from SDSS DR1. Sánchez Almeida et al. (2010) use unsupervised *k-means* clustering to classify SDSS DR7 galaxy spectra into 17 major classes. Similarly, Parks et al. (2018) have developed and applied a CNN model to detect quasar spectra with strong Lyα absorption and characterize them. They have applied the trained model on quasar spectra from SDSS DR5 and DR7. Busca & Balland (2018) use CNN for the classification and redshift estimation from quasar spectra obtained from the BOSS survey (Dawson et al. 2013) and achieve the human-level accuracy. The application of these algorithms does not only result in a good classification and parameter estimation models but also provides the catalogue of interesting objects from these surveys (Si et al. 2014; Brescia et al. 2015; Li et al. 2018; Barchi et al. 2019).

Since the scope of such surveys is not limited to any specific kind of object, therefore their databases are a rich source for interesting peculiar objects. The observational signatures (e.g. photometric images, spectra, light-curves) of such sources will drastically differ from the rest of the known sources. These are also referred to as 'outliers' or anomalous objects. Discovery of outliers offers the opportunity to explore the unknown phenomena and understand the science behind their characteristic behaviour. Other than the applications mentioned in the previous paragraph, ML/DL techniques prove to be very useful in mining outliers also from large databases.

Detection of outliers in astronomy has been done in various ways, e.g. probabilistic and statistical approaches (Henrion et al. 2013), distance and dis-similarity in the feature space (Agnello 2017), PCA based techniques (Dutta et al. 2007), and classification based (Nun et al. 2014) to name few. In this work, we use the deep CNN (DCNN) with the autoencoder architecture for detecting anomalous objects using optical stellar spectroscopic data from SDSS. DCNN is a type of unsupervised approach which is trained to learn how to encode the input features into a smaller dimensional space and reconstruct (decode) the input features from the data in the reduced dimensional space. Discrepancy in the reconstruction is considered as a measure of *outlierness* and used to identify the outliers.

The paper is organized as follows: In Sect. 2, we describe the process of obtaining the data from SDSS. In the next section, Sect. 3, we provide a brief overview of the technique used for detecting the anomalous spectra. Finally, we present our results in Sect. 4 followed by a conclusion and discussion in the last section, Sect. 5.

2 Data

For this work, the spectral observations are acquired from SDSS DR 13 (Albareti et al. 2017) database. All the spectroscopic observations under the SDSS program are passed through a data reduction (`spec2d`, Stoughton et al. 2002) and processing pipeline (`spec1d`, Bolton et al. 2012) which associates each observation to the primary spectroscopic class, either 'STAR', 'GALAXY' or 'QSO' for star, galaxy and quasar respectively using template matching and χ^2-minimization. The pipeline also computes instrumental and astrophysical parameters like signal-to-noise ratio (SNR), flux in different photometric bands, redshift, error and flag on redshift estimate etc. Each observation labelled as 'STAR' is also supplemented with stellar sub-classification, effective temperature, surface gravity, metallicity.

We query the SDSS DR13 database using Structured Query Language (SQL) for all spectra classified as 'STAR' by setting initial constraints on SNR (median SNR > 20) and redshift warning flag (`zWarning` $= 0$ indicating a good confidence in the pipeline classification) to ensure that we get spectra with good quality and reliable classification. This returns 61,627 individual stellar spectra with SNR in the range 20-165. The majority of these spectra ($\sim$99%) is classified as one of the seven MK spectral classes (O, B, A, F, G, K, M) by the SDSS pipeline out of which more than half ($\sim$34000)

belong to the F-type class only. Assuming that the SDSS classification is done properly, we can safely assume that our sample is dominated by MK type stellar spectra. A very small fraction of the downloaded spectra do not belong to the canonical MK classification system like 'CV' (cataclysmic variables) or 'Carbon' (carbon stars) etc.

3 Outlier Detection Methodology

We use convolutional neural network (CNN; LeCun & Bengio 1995), also referred to as ConvNet, for outlier detection. CNNs is a class of deep-learning algorithms inspired by the structure of the visual system (Hubel & Wiesel 1962; Fukushima 1980, Kuzovkin et al. 2018). The architecture of a deep convolutional neural network includes multiple layers containing convolutional and max-pooling filters which are applied on the input one/two/three-dimensional data to extract local features and map it to the output feature space. LeNet (LeCun et al. 1989), AlexNet (Krizhevsky 2014), GoogLeNet (Szegedy et al. 2014) are few famous DCNN architectures. Primarily developed for the computer vision (LeCun et al. 1989), CNNs have also been implemented for various other applications like document recognition (LeCun et al. 1998, Ranzato & Tapporo 2007), audio/video classification (Amodei et al. 2015; Zha et al. 2015), and speech recognition (Graves et al. 2013). In astronomy, CNNs have been applied for star-galaxy classification (Kim & Brunner 2017), galaxy morphology studies (Barchi et al. 2019), detecting transits in the light curves of exoplanet host stars (Pearson et al. 2018), detecting bar-like structures in galaxies (Abraham et al. 2018), to name a few.

Another DCNN architecture which has been extensively applied for the task of anomaly detection is autoencoder (Rumelhart et al. 1986; Japkowicz et al. 1995; Chong & Tay 2017; Luo & Nagarajan 2018). Deep autoencoder architecture (also called Stacked Autoencoder) is a combination of two networks, each having typically four or five layers of neurons. The first set of layers, called encoding layers, projects the high-dimensional input data to a lower-dimensional latent space (also referred to as *bottleneck*). The second set of decoding layers maps the latent space representation back to the original higher-dimensional space. Since the training of autoencoder does not require a labelled dataset, therefore it comes under the category of unsupervised method of anomaly detection. In our case, the input sample is dominated by 'normal' spectra, therefore while training, network weights and biases are fine-tuned to learn the reconstruction only for 'normal' spectra. A well-trained DCNN autoencoder will give higher reconstruction errors for any spectra with abnormal features like emission lines, observations with defects like improper cosmic-ray correction, poor flux-calibration etc.

For implementing deep convolutional autoencoder, we use KERAS API in Python with TensorFlow (Abadi et al. 2015) at the backend. We build an autoencoder architecture with 4-layer deep encoding and 4-layer deep decoding network (Fig. 1).

4 Results: Application on SDSS data

Before supplying SDSS spectra (Refer to Sect. 2) to the autoencoder model described in the previous section, we normalize each spectrum to 1 at 5550 Å and resample the data every 5 Å wavelength step to reduce the dimensionality. We use only the 4000 -7200 Å region of the spectrum to avoid the noisy features at the blue and red ends of the wavelength range. This returns 640 wavelength points for each spectrum.

We apply the autoencoder architecture shown in Fig. 1 on normalized and pre-processed SDSS data. For unsupervised training of the network, we split the whole sample in three sets: training, validation and test set. Out of 61,627 spectra, we keep 55000 for the training plus validation, leaving

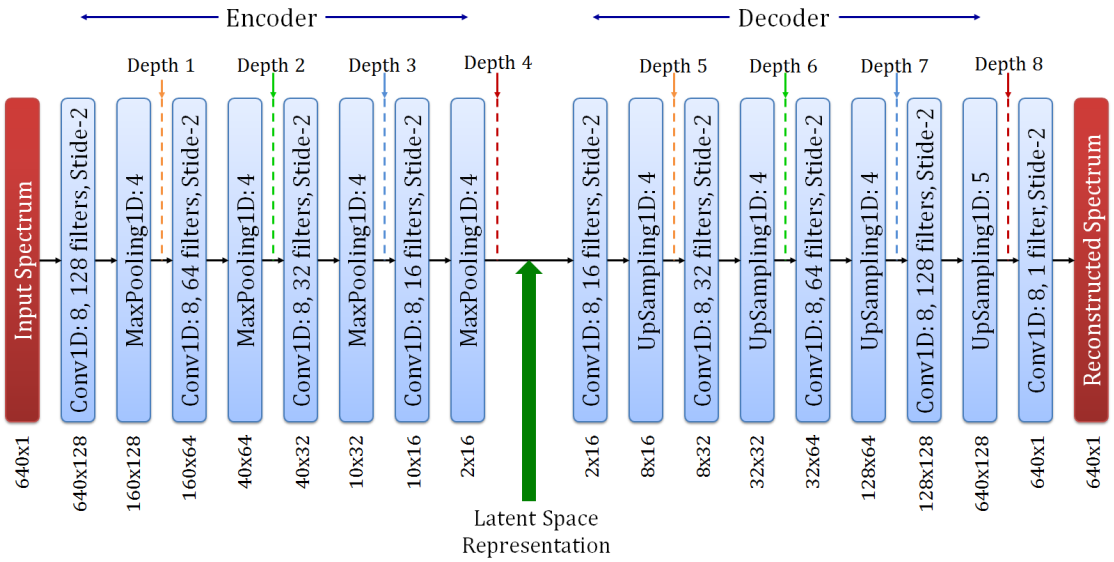

Figure 1: Schematic diagram representing the Deep autoencoder architecture used in this work. Each operation (convolution/maxpooling/upsampling) is presented as a box. Convolutional layers in the network uses filters of size 8 (or 8×1) with the Rectified Linear Unit (ReLU) activation (Nair & Hinton 2010). The last convolutional layer after the decoder part uses sigmoid activation (Han & Moraga 1995) and gives the reconstructed spectrum as the output. The output dimension from each layer is indicated below the box representing the layer.

6627 spectra for testing the trained network. To minimize the presence of anomalous spectra in the training, we visually inspect the complete training sample of 55000 spectra and identify 1044 cases with anomalous features. After removing these cases, we are left with 53,956 spectra in the training sample out of which 15% (8094) are used for the validation, leaving 45,862 spectra for the training. During each training epoch, we use a batch size of 256 spectra and set maximum number of training epochs as 1000. For minimizing the loss function (mean squared error) between the input and reconstructed spectra, we use the Adam optimizer (Kingma & Ba 2014). We also keep track of the loss function for the validation set after each training epoch and set an early stopping condition to prevent over-fitting of the model. For the early stopping, we put a condition on the validation loss function such that if it does not improve beyond 0.0001 for 20 consecutive epochs, the training should stop. To track the performance of the model, we also check the R^2-Score (Glantz 1990; Draper 1998) after each training epoch. The R^2-Score can take values in the range 0-1 and is defined as:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}, \tag{1}$$

where SS_{res} (sum of square of residuals) and SS_{tot} (total sum of squares) are defined as:

$$SS_{\text{res}} = \sum_i (x_i - d_i)^2,$$
$$SS_{\text{tot}} = \sum_i (x_i - \bar{x})^2.$$

Here x_i and d_i denote the original and reconstructed flux values. $\bar{x}$ denotes the mean of the original flux values at a given wavelength point. For a perfect classification model, the value of R^2-Score should be equal to 1.

We apply the trained model on 6627 spectra from the test sample. The original and reconstructed spectra for the three test cases are shown in Fig. 2 with their respective classification provided by the SDSS pipeline. Residuals between the two series of spectra in the lower panel of the figure show that the reconstruction matches well the original spectra.

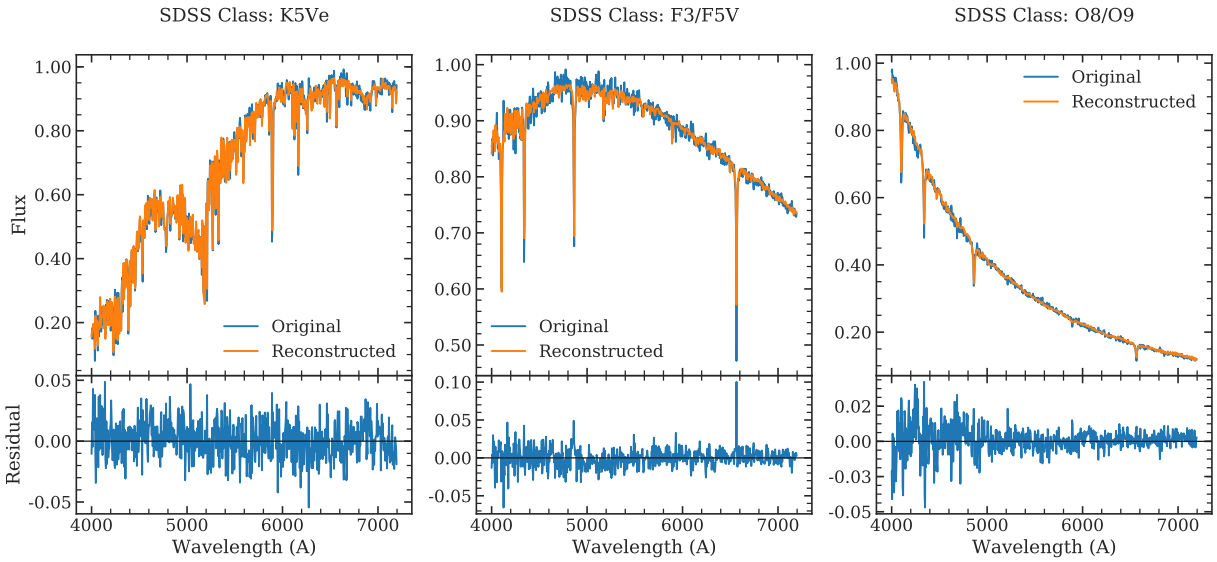

Figure 2: Comparison between original and reconstructed spectra for three samples in the test set. Top panel shows original and reconstructed spectra plotted on top of each other whereas the bottom panel shows the residual (reconstructed−original) between the two. The title of each panel indicates the classification provided in the SDSS database.

For the test set, we compute the statistics for three types of error, namely mean squared error (MSE), root mean squared error (RMSE), and mean difference (Original-reconstructed). The statistics is presented in Table 1. Since the errors are very small, we present the numbers up to 4th decimal point. The very small average RMSE of 0.0222 ± 0.0079 is indicative of extremely good quality of reconstruction. As indicated in the table, 75% of the spectra (75th percentile) have RMSE less than 0.0268. We also evaluate the influence of SNR on the RMSE by dividing the test sample in three SNR bins such that equal number of test spectra (2209) lie in each bin. This exercise results in three SNR-bins: 20-27, 27-38, and >38 with mean RMSE of 0.0299, 0.0239, and 0.0198 respectively in each bin. This implies that the lower SNR spectra have larger reconstruction error, but the difference is not significant as compared to the higher SNR spectra.

We further examine the distribution of these errors using histograms. Fig. 3 shows the histograms for the three types of errors. The average value of difference, -0.0011, between the original and reconstructed spectra is very close to zero and most of the spectra lie within $\pm 3\sigma$ about the mean. There are only 86 spectra ($\sim 1\%$) which lie beyond -0.0011 ± 0.0050. Similarly in the middle panel, we see that a large number of spectra (more than 5000) fall in the first histogram bin close to 0.00. In the bottom panel of the figure, we present the RMSE distribution, where there are only nine spectra which have RMSE > 0.075.

We check nine spectra which are extreme RMSE outliers and plot them in Fig. 4. According to their SDSS classification, we notice that these spectra are actually peculiar and have been classified as either cataclysmic variables with emission lines or carbon stars as indicated in the plots. We notice that out of nine spectra, three show an unusually large spike which points at a possible glitch while processing these spectra.

Table 1: Statistical summary of different error estimates.

Statistics	Mean squared error	Root mean squared error	Mean difference
Count	6627	6627	6627
Mean	0.0006	0.0222	-0.0011
Standard dev.	0.0007	0.0079	0.0017
Minimum	0.0000	0.0068	-0.0276
25th percentile	0.0003	0.0168	-0.0019
50th percentile	0.0005	0.0214	-0.0011
75th percentile	0.0007	0.0268	-0.0002
Maximum	0.0372	0.1928	0.0112

5 Conclusion and Discussion

In this work, we introduce a deep convolutional stacked autoencoder network which is capable of retrieving spectra with peculiarity from SDSS without any supervised training. We show that the spectra with the maximum reconstruction error are actually not 'normal'. Network training takes only 314 seconds for 57 epochs using a single GPU node equipped with three Tesla M 2050 GPU cards and 24 GB RAM which proves to be very time-efficient. Due to the unsupervised nature of the outlier detection algorithm, the algorithm is easily generalizable for other spectroscopic surveys, e.g. LAMOST, and can be used for discovering objects with unusual spectra, for re-calibrating the flux continuum, and recovering defective spectra. By choosing different thresholds on the reconstruction error, it can also be used to pick spectra of specific type of objects (say quasar) as well.

Autoencoders are considered to be representation learning algorithms which can extract the important representative features from the input data and therefore are useful for various applications other than detection of anomalies. One such important application of stacked autoencoder is in those situations where the size of the labelled training data set is small and not sufficient enough to train a supervised deep neural network. For such problems, autoencoder can be used as a semi-supervised classifier in two phases. The first pre-training phase uses unlabelled large dataset for unsupervised learning of the representation and adjusting the network weights and biases. In the second-phase, the reduced representation in the latent space obtained from the encoder network is used for the supervised classification by adding a new fully connected *softmax* layer (Goodfellow et al. 2016). In another work which is under preparation, we demonstrate the potential of autoencoder as semi-supervised classifier for the stellar spectral classification. Reduced representation can also be used for regression and clustering analysis. Another variant of autoencoder, called denoising autoencoder can be useful in cleaning the observed spectra and increasing the SNR.

Acknowledgements

In this work, we have extensively used the SDSS database. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is `www.sdss.org`.

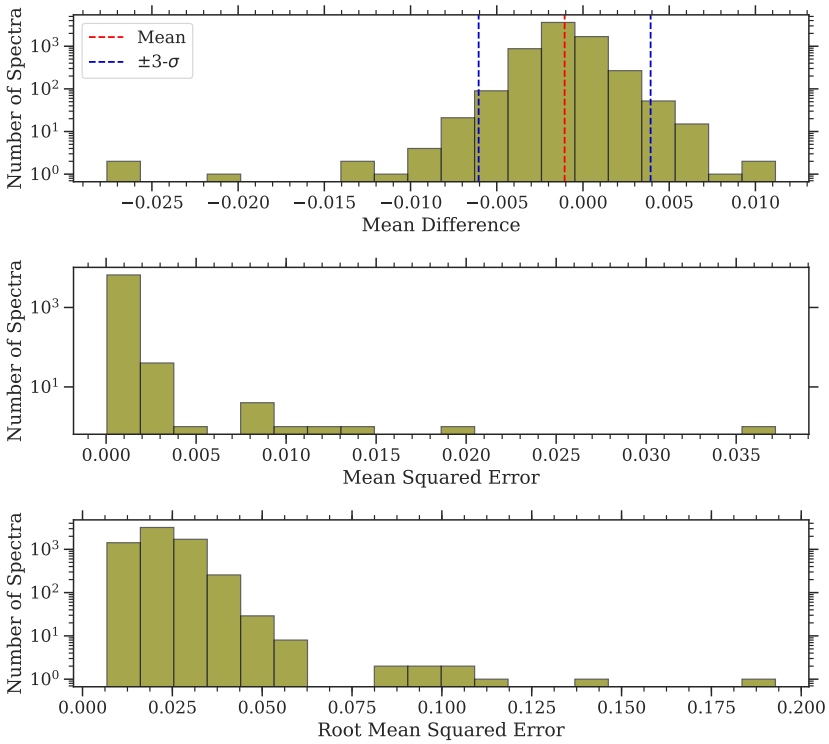

Figure 3: Histogram for three error estimates, MSE, RMSE and mean difference in top, middle and bottom panels respectively.

References

Abadi M. et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems
Abraham S., Aniyan A. K., Kembhavi A. K. et al. 2018, MNRAS, 477, 894
Agnello A., 2017, MNRAS, 471, 2013
Aguado D. S., Ahumada R., Almeida A. et al. 2019, ApJS, 240, 23
Albareti F. D., Allende Prieto C., Almeida A. et al. 2017, ApJS, 233, 25
Amodei D., Anubhai R., Battenberg E. et al. 2015, arXiv:1512.02595
Ball N. M., Loveday J., Fukugita M. et al. 2004, MNRAS, 348, 1038
Barchi P. H., de Carvalho R. R., Rosa R. R. et al. 2019, submitted to MNRAS, arXiv:1901.07047
Bolton A. S., Schlegel D. J., Aubourg E. et al. 2012, AJ, 144, 144
Brescia M., Cavuoti S., Longo G. 2015, MNRAS, 450, 3893
Busca N., Balland C. 2018, arXiv:1808.09955
Chong Y. S., Tay Y. H. 2017, in Cong F., Leung A., Wei Q., eds, Advances in Neural Networks - ISNN 2017. Springer International Publishing, Cham, pp 189196
Cui X.-Q., Zhao Y.-H., Chu Y.-Q. et al. 2012, Research in Astronomy and Astrophysics, 12, 1197
Dawson K. S., Schlegel D. J., Ahn C. P. et al. 2013, AJ, 145, 10
Draper N. 1998, Applied regression analysis. Wiley, New York
Dutta H., Giannella C., Borne K. et al. 2007, Distributed Top-K Outlier Detection from Astronomy Catalogs using the DEMAC System. pp 473478
Fukushima K. 1980, Biological Cybernetics, 36, 193
Glantz S. 1990, Primer of applied regression and analysis of variance. McGraw-Hill, Health Professions Division, New York
Goodfellow I., Bengio Y., Courville A. 2016, Deep Learning. MIT Press
Graves A., Mohamed A.-R., Hinton G. 2013, International Conference on Acoustics, Speech and Signal Processing (ICASSP), arXiv:1303.5778
Han J., Moraga C. 1995, in Mira J., Sandoval F., eds, From Natural to Artificial Neural Computation. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 195201

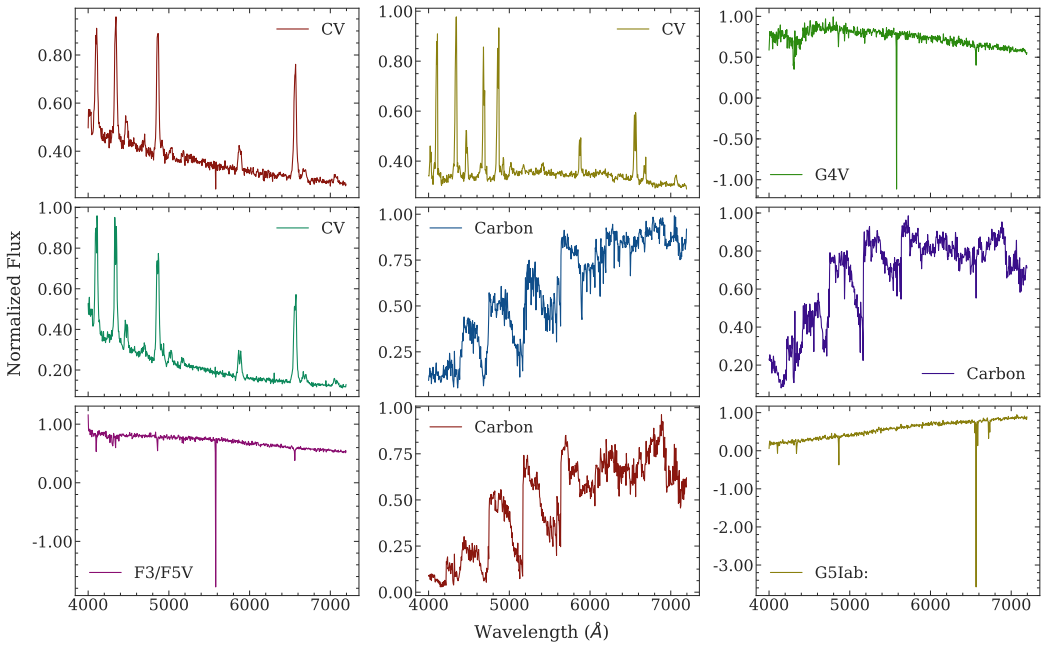

Figure 4: SDSS spectra for nine extreme RMSE outliers. The SDSS classification for these spectra is indicated in the respective panel as legend.

Henrion M., Mortlock D. J., Hand D. J. et al. 2013, Classification and Anomaly Detection for Astronomical Survey Data. Springer New York, New York, NY, pp 149184
Hubel, D., Wiesel, T. 1962, Receptive fields, binocular interaction, and functional architecture in the cats visual cortex. Journal of Physiology (London), 160, 106154
Japkowicz N., Myers C., Gluck M. 1995, in Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1. IJCAI95. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 518523
Kim E. J., Brunner R. J. 2017, MNRAS, 464, 4463
Kingma D. P., Ba J. 2014, 3rd International Conference for Learning Representations, San Diego, arXiv:1412.6980
Krizhevsky A. 2014, arXiv:1404:5997
Kuzovkin I., Vicente R., Petton M. et al. 2018, Communications Biology, 1, 107
LeCun Y., Bengio Y. 1995, The handbook of brain theory and neural networks, 3361, 1995
LeCun Y., Boser B., Denker J. S. et al. 1989, Neural Comput., 1, 541
Li Y.-B., Luo A.-L., Du C.-D. et al. 2018, ApJS, 234, 31
Luo T., Nagarajan S. G. 2018, in 2018 IEEE International Conference on Communications (ICC). pp 16
Nair V., Hinton G. E. 2010, Proc. 27th Int. Conf. Int. Conf. Machine Learning. I Omnipress, p. 807.
Nun I., Pichara K., Protopapas P., Kim D.-W. 2014, ApJ, 793, 23
Parks D., Prochaska J. X., Dong S., Cai Z. 2018, MNRAS, 476, 1151
Pearson K. A., Palafox L., Griffith C. A. 2018, MNRAS, 474, 478
Ranzato F., Tapporo F. 2007, arXiv:0709:4118
Rumelhart D. E., Hinton G. E., Williams R. J. 1986, MIT Press, Cambridge, MA, USA, Chapt. Learning Internal Representations by Error Propagation, pp 318362
Sánchez Almeida J., Aguerri J. A. L., Muñoz- Tuñón C., de Vicente A. 2010, ApJ, 714, 487
Si J., Luo A., Li Y. et al. 2014, Science China Physics, Mechanics, and Astronomy, 57, 176
Stoughton C., Lupton R. H., Bernardi M. et al., 2002, AJ, 123, 485
Szegedy C., Liu W., Jia Y. et al. 2014, arXiv:1409.4842
York D. G., Adelman J., Anderson J. E. et al. 2000, AJ, 120, 1579
Zha S., Luisier F., Andrews W. et al. 2015, arXiv:1503.04144