Classifying Blazars and Cataclysmic Variables from the Catalina Real-Time Transient Survey

Adrian Markelov
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
amarkelo@andrew.cmu.edu

Kai Wen Wang
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
kaiwenw1@andrew.cmu.edu

Yizhou Xu
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
yizhoux@andrew.cmu.edu

Abstract

To properly delegate follow-up facilities with the growing number of recorded events from synoptic sky surveys, an automated and reliable classification method is necessary. In previous works, the problem was tackled as a whole and a multi-class classification problem was researched, but with limited success. In this paper, we focus on the binary classification of two important cosmological events, Blazars and Cataclysmic Variables (CVs), using data from the Catalina Real-Time Transient Survey. This task is very challenging due to the limited amount of reliable data, the high dimensionality of the problem, the irregularity in sample times and the high bias in the dataset towards different classes. We analyze the dataset and explore methods of feature extraction with a wide range of classification models, all with their pros and cons. Our best method yields an accuracy of around 90%.

1 Introduction

In the cosmological sciences there is a growing abundance of recorded transient and variable events from large, digital, synoptic sky surveys. Many follow-up facilities are specialized to analyze specific types of astrophysical phenomenon for various characteristics in terms of distance, cadence, wavelengths, light intensity, etc. [6]. To delegate the follow-up facilities without wasting and duplicating their efforts, an automated, probabilistic classification of the detected variables and transients becomes necessary. These decisions must be made fast, as many of the events need rapid follow-up for proper analysis. One of these important classifications to make is Blazars vs. Cataclysmic Variables (CVs). [1]

As of now, approaches for classifying significant celestial phenomena such as Blazars and CVs require very long time series of a single object to make a reliable inference. In many cases, a full time series is collected in periods of around 8 years. As a result, any available data for this classification problem is highly limited and likely to be incomplete. Even with long streams of data, on the order of

[1] Blazars are a very especially large type of Quasar. They contain a super-massive black hole at the center of a spiraling galaxy shooting massive amounts of matter out of the central axis. CVs are a binary system with a neutron star at the center being orbited by a second smaller 'donor' star that is being absorbed by the neutron star.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

hundreds of data points, the classification accuracy can be improved. Although Blazars and CVs are very different in their physical appearance and their behaviors, the observed light magnitudes are very similar in nature, and almost seemingly random. Another challenge is the imbalance in data, as CVs are almost 3 times more likely to occur then Blazars. This inherent bias towards classifying CVs along with the shortage of data makes this problem very challenging.

The ability to make very accurate inferences, especially on shorter streams of data, can significantly boost the productivity of facilities to capture and discover useful astronomical information. Success with the binary classification of Blazars and CVs can lead to promising approaches to be extended to other astronomical phenomenon, such as supernovae, asteroids/flares, etc.

2 Dataset

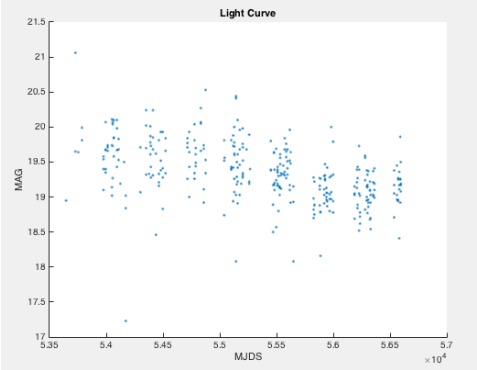

Figure 1: Example plot of raw Blazar data.

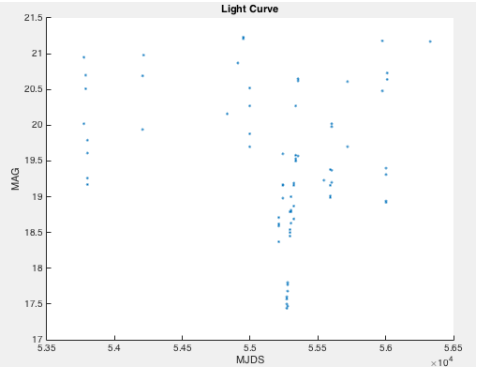

Figure 2: Example plot of raw CV data.

We used the Blazars and CVs data from the Catalina Real-time Transient Survey [5]. The Survey is a synoptic exploration of thirty three thousand square miles of the sky to observe transient events. In general, there are times when data is not collected at all when the observing facilities are not looking in this area of the sky. Each data point contains relevant information about the time of the observation in Modified Julian Date Time (MJDT), the observed light magnitude (Unfiltered CSS Magnitude), and error bars for the observed light magnitude.

Our dataset has four main challenges that make this classification problem incredibly challenging:

1. Irregular sampling: The observations made have highly irregular spacings in time, which is unavoidable as the rotation of the Earth and weather conditions makes it nearly impossible to have reliably timed and consistent observations at a specific area of the sky. Also, this leads to an inconsistent number of data points per sample. The histogram of the number of data points is plotted in Figure 3 and the histogram of the lengths of observations is plotted in Figure 4.
2. High dimensionality: Each time-series data sample may contain up to hundreds of points.
3. Small sample size: While it is the most comprehensive dataset for this problem, there are only around 938 reliable observations to date, 704 CVs and 234 Blazars. After basic preprocessing of the data, around 100 of these observations are highly incomplete and comprises of less than 50 data points.
4. High bias towards CVs: As mentioned in the previous point, there are almost 3x more CVs than Blazars.

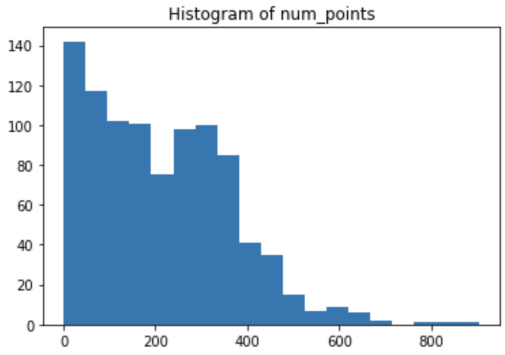

Figure 3: Number of observations per each sample.

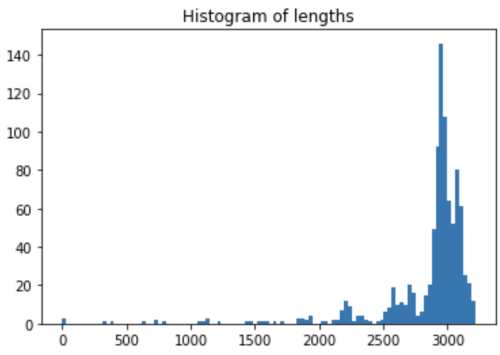

Figure 4: Length of each observation.

3 Related Works

To our knowledge, the state-of-the-art on this type of work is the seminal paper by the team in charge of the Catalina Real-Time Survey who posed the problem in the first place [1]. In their paper, they had much more extensive data and achieved a completion rate of 81% on Blazars and 96% on CVs, and had no reports on accuracy.

We also have an accuracy baseline, as we have been given this data set by the CMU Statistics Professor Chad Schafer [14]. He has done some previous work with this data achieving around 90% accuracy on the full length data. His models also used various feature transformations including the structure function to filter out unknown human made time irregularities during the data collection process and a quantile regression transformation to reduce the dimensionality of the data.

We have found other research into variable stars classification, not Blazars vs. CVs specifically [12]. The paper investigated into classification of variable stars using recurrent neural network using unevenly sampled data. Their accuracy is above 90% on several datasets, including ASAS, LINEAR, and MACHO. This is not a good baseline since our datasets are different, but it serves as a reference for us to see how well our model performance should be to be considered useful.

Previous studies in astronomy have introduced structure functions to analyze irregularly sampled data [4]. In this paper, we use a modified version, which is elaborated in Section 4.2 [14].

4 Method

4.1 Summary of Methods and Motivation

One of the biggest pitfalls of our data set comes from the complex nature of the sampling and generation of the time processes. The combination of uniquely dimensioned data samples, irregular time sampling and aperiodic signals makes it very difficult to create a basis of comparison between each of our object observations that is necessary for all machine learning models. This problem calls for the structure function transformation, but at the cost of squaring the dimensionality of our data set.

Both our raw data and our transformed data are very high dimensional, which is susceptible to the curse of dimensionality. To deal with the high dimensionality of the data, extraction of good features is arguably the most important step of our process. Quantiles provide a highly accurate characterization of the data distribution. By picking quantiles at many regular points, we could compactly represent the high dimensional data distribution as a much shorter vector that could be passed into classification models.

Although we have manually crafted some specialized methods to extract features, such as the structure function and PCA, we sought to do this automatically by using a convolutional neural network (CNN). Since our dataset is very small, we designed our CNN to be very simple to decrease the number of parameters to learn and highly regularized to prevent overfitting.

4.2 Feature Transformations:

Structure Function Transformation

The structure function is a feature transformation of the data that takes in a single light curve, which is a function of light magnitude with respect to MJDT and returns a function of the log of the absolute differential magnitudes with respect to the corresponding absolute time differentials. From now on all of the transformations and models we work with are done on the basis of the structure function space and one structure function corresponds to a single cosmological object. This is very important as this puts our irregularly sampled data into a domain which is comparable.

Light Curve Function: $\phi : \mathbb{T} \to \mathbb{M} \quad s.t.$

$\mathbb{T} = \{\text{MJDT}\}$

$\mathbb{M} = \{\text{Light Magnitudes}\}$

Structure Function: $\psi : \mathbb{T}' \to \mathbb{M}' \quad s.t.$

$\mathbb{T}' = \{t' : t' = |t_i - t_j| \quad s.t. \quad t \in \mathbb{T}, i, j \in \mathbb{N}\}$

$\mathbb{M}' = \{m' : m' = \log_{10} |m_i - m_j| \quad s.t. \quad m \in \mathbb{M}, i, j \in \mathbb{N}\}$

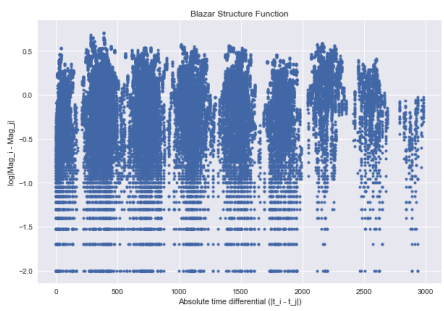

Figure 5: Structure function for Blazar 77.

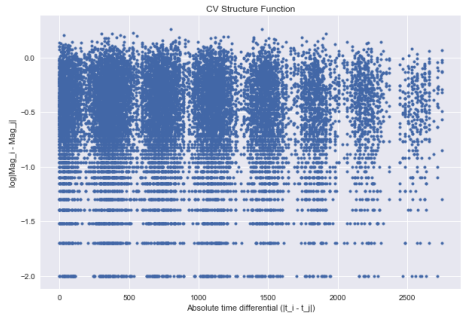

Figure 6: Structure function for CV 251.

The gaps we see in the data are superficial and are artifacts of the irregularity in sample times. The gaps are also visible in the structure function, but this domain is one where the data becomes directly comparable.

Quantile Regression Transformation

The quantile regression technique essentially tries to estimate the quantile function of the y-axis. It is analogous to linear regression, which tries to learn $y = f(x) + \epsilon$. Here, f is the quantile function defined by

$$Q : [0, 1] \to \mathbb{R} \qquad Q(p) = \inf\{x \in \mathbb{R} : p \leq F(x)\}$$

where F is the cumulative distribution function.

We performed a linear quantile regression for the quantiles 0.05, 0.15, ..., 0.95 on the output of the structure function. Each regressor produces a slope and an intercept for the regression, for a total of 20 points.

Image Transformation

While the quantile regression model is great at simplifying the problem it does have a massive loss of information that is encoded in the structure function. The quantile regression solution was also great for placing all of our observation data into the same dimension. Now, to solve the problem of loss of information while still placing each of our data observations in the same dimension we turn our structure function into an image. To do this we use the vector of tuples from the structure function and performed a distribution estimation with 2D histograms of evenly-spaced bins. Using

this method we create a 100x100 image where each pixel corresponds to a bin and the intensity of the pixel corresponds to the bin size.

PCA Transformation

After transforming our structure function into a 100x100 image, all of our data samples are guaranteed to be 10,000 dimensions. This is obviously very problematic because we have less than 1000 data samples. To solve this problem we use Principle Component Analysis (PCA) [9] to transform our features into a much smaller dimensional space that preserves a majority of the information in the original data set. First, though we must use spectral decomposition to find what proportion of the features are actually contributing a significant amount of information. Spectral decomposition allows us to quickly find all of the eigenvalues which correspond to the max variance of the corresponding principle component. Using each of these eigenvalues normalized by the sum of all eigenvalues or variances we can create a mapping between each principle component feature and its contribution to the explained variance of the data. This is defined as follows:

Let $x_i' =$ i'th principle component and $\lambda_i =$ i'th eigenvalue

$$\text{Explained Variance} = f(x_i') = \left(\frac{\lambda_i}{\sum_j^N \lambda_j} \right) \cdot 100$$

This explained variance is proportional to the amount of information that the feature provides with respect to all of the other features. Using the cumulative distribution function of this mapping we can decide on the number of dimensions we want to keep in our new space after the PCA transformation. We define this CDF as follows:

$$\text{CDF of Explained Variance} = F(x_k') = \sum_{i=1}^{k} \left(\frac{\lambda_i}{\sum_j^N \lambda_j} \right) \cdot 100$$

The realization of these equations can be seen in figure 7. The ideal is to achieve a balance between removing enough redundant features and preserving most of the information. Therefore, we will pick a dimension size that will provide us with at least 90% of the information that all of the dimensions could provide us. We also superficially evaluate PCA by looking at the reconstruction of our original data set from the best newly generated features (Seen in Figure 9).

GMM Transformations

Another model that may have a lot of potential in representing our data are the Gaussian Mixture Models (GMM). The new representation of the data under the GMM Transformation would be the means and covariances of K Gaussians that we fit to each structure function. To do this we use clustering algorithms (such as Expectation Maximization) developed in many standard machine learning packages like SK-Learn. While quantile regression just creates K bins along the $\log|m_i - m_j|$ axis that contains no information of the instantaneous density in each bin, GMM's can stack Gaussians on top of each other along the $\log|m_i - m_j|$ axis and make stronger implications of the distribution of the structure function through the continuous curvature of the data in just about the same number of parameters as quantile regression. Thus we can now view our new space for classification in the following mathematical definitions:

Structure Function Distribution Estimation

$$\text{Structure Function} \approx \psi \sim \sum_{i}^{K} \mathcal{N}(\vec{\mu}_i, \Sigma_i) \quad \text{s.t. } K = \text{ number of estimated cluster}$$

$$\text{New Space} = \theta = \{(\vec{\mu}_i, \Sigma_i) \quad \forall i \in [K]\}$$

4.3 Classification Models:

Basic Models on Quantile Regression Transformation

Our first tests used simple models, including SVMs, Random Forests, Multi-layer Perceptrons and Adaboost with Decision Trees. Since our dataset is very small, we chose to use these simple methods to reduce overfitting. This especially motivates the use of SVMs, as they are usually robust to overfitting. The input was a 20-dimensional vector comprising of reduced quantile information, specified in Section 4.2. These models serve as a baseline on the accuracy that we aimed to get from more sophisticated models, which we will explain below.

Balanced Bagging Model

A previous issue that we have yet to address is that the dataset is imbalanced, with only around 25% of the training data consisting of Blazars. This raises an important issue for many machine learning classifiers: they tend to perform worse in predicting the minority class because traditional machine learning classifiers are often biased towards the majority class. They usually use error rates such as cross entropy loss, while ignoring the data distribution. Although our dataset is not considered highly imbalanced (i.e. 90% majority class), our classification model still runs the high risk of bias towards the majority class of CVs.

Another problem that arises from our imbalanced dataset is that it is not clear whether the ratio of Blazars and CVs accurately represents the likelihood of the two cosmological events. For that reason, it would make more sense for our model to focus on the time-magnitude distribution of our data instead of the prior. To address this particular problem, we investigated into classifiers specifically suited for imbalance datasets. Here, we include several common techniques for dealing with imbalanced datasets, which are over-sampling, under-sampling, and increments on minority sample weights [2].

One particular classifier that we have found to be useful is the balanced bagging classifier [8]. The purpose is to balance the dataset prior to training, in order to improve the classification performance. It is an ensemble algorithm that combines multiple decision trees classifiers. Within each classifier, it randomly samples training data so that the two classes are balanced. This is also known as under-sampling.

Convolutional Neural Networks

We used a simple 3-layer CNN [11] implemented with Keras [3] on both the 100^2 and 20^2 image to perform feature extraction. We also included a 0.5 Dropout [15] for regularization. We used strided convolution since it allows the network to learn to downsample the data and extract the most useful features. A detailed diagram of the architecture is shown in Figure 7.
We trained our model using Adam [10] with learning rate 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and decay rate 0.001.

Distribution Estimation Models (GMM)

After performing the GMM Transformation, each Structure Function was mapped to a Kx2 vector of means and variances. We flattened and concatenated them into a $4K$-dimensional vector and trained a neural network and an SVM based on this data. This method was not as successful as our other methods but we believe this is a promising area of continued research since it has the potential to completely learn the distributions of the structure function data. In a sense, it is a more sophisticated distribution estimation approach than quantile regression, because it does not discretized the data but rather learns continuous distributions.

5 Experiments and Results

We conducted all of our experiments in the same order that we explained our methods, starting with the simple methods and increasing with sophistication. In all of these results, we commit to a train-test split of $80 : 20$ ratio. The reported accuracies are all accuracies on the test set.

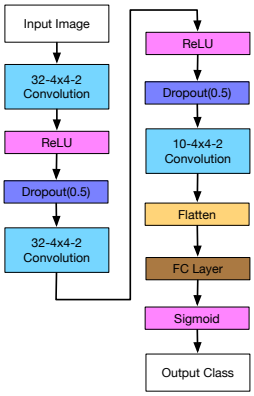

Figure 7: CNN Architecture. The Convolutional layers have the specification $N - K\text{x}K - S$ which stands for N filters, kernel size K^2, and stride S.

Model Results on Quantile Regression Transformation:

model	kNN	SVM	Adaboost	Random Forest	Neural net
accuracy (%)	86.1	85.8	84.5	84.6	85.0

The table shows the testing accuracies for simple models trained and tested on the quantiles extracted from quantile regression. Most of these models produce consistent accuracies, hovering around the 85% range. In particular, none of these methods performs significantly better than 85%. This could be bottlenecked by our method of feature extraction. The distribution estimates could have been too crude and were not very representative of the distinguishing features between Blazars and CVs. For our purposes, these results served as a baseline for our later, more sophisticated methods.

Model Results on PCA Transformation:

Dimension Analysis Results:

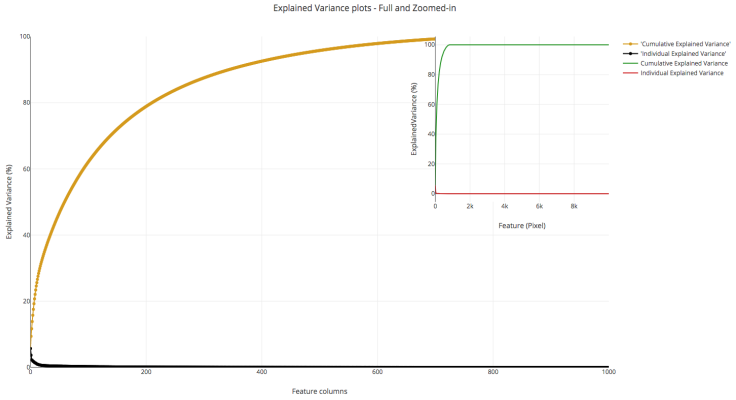

Figure 8: The orange line is the zoomed in cumulative explained variance curve of the first 1000 principle components. The black line models the individual explained variance of the same zoomed in corresponding principle components. The other 2 lines in the top right corner contain the fully zoomed out cumulative and individual explained variance of all 10,000 principle components.

Using spectral decomposition on the covariance matrix of each objects structure function we can see how much information each of our features is providing compared to all of the other features. The graph in Figure 8 shows the cumulative explained variance of each of the features starting from the feature that provides the most explained variance to the least. This graph tells us that 92.6% of

the explained variance comes from 400 of the 10,000 features. Thus, we will use PCA to reduce our 10,000 dimensional feature space to 400 dimensions. We can see an example of this new low dimensional space/ image in Figure 9 and the reconstruction of the original space using only the first principle component can be seen in Figure 10.

Model Results on PCA Transformation:

With our new PCA transformed data, we tested our CNN model with a 20x20 image and achieved approximately 90% accuracy. This is much better than simply running the CNN on the 100x100 image, which gave us an accuracy of around 87%. This shows that the PCA reduction was very useful in making the data more interpretable for the CNN to learn, and in removing redundant information that made our data very high dimensional. This result specifically shows that PCA is the best feature extraction method for the dataset.

Model Results on Balanced Bagging classifier:

We trained this model on 80% of the data, and achieved an accuracy of 86.8%. There are more useful metrics to evaluate performance of imbalanced datasets than accuracy, such as precision-recall and ROC (Receiver Operating Characteristics) [7]. For precision vs. recall, our model has the following confusion matrix:

	Predicted CVs	Predicted Blazars
Actual CVs	101	14
Actual Blazars	7	37

As we can see from the matrix above, we have relatively accurate predictions. More importantly, we notice that our model is not highly biased towards the majority class CVs, as we are able to predict most of the Blazars correctly as well. ROC is a graphical representation of the balance between True Positive Rate and False Positive Rate at every possible decision boundary. Our model has an ROC AUC score of 86.0%, which demonstrates good performance from our classifier.

6 Conclusion

The dataset is highly problematic due to the four reasons outlined in Section 2, and the extraction of good features is paramount. From our results, PCA showed the most promise in extracting features and CNNs performed the best with classification. Also, we developed methods to account for the bias in the data, and not purely optimize for accuracy. This work is an important step in understanding the complexity of the data and possible ways of handling this complexity. Although we did not significantly improve the accuracy achieved by Professor Schafer, we were able to attain similar results with different approaches.

For future research, we can try to make use of other cosmological events, such as supernovae and flares, and see how the results compare to the state-of-the-art [1]. Our methods of feature extraction can potentially lead to better results in multiclass classification. Although we did not achieve any good results with GMM, this method of learning the distribution of the visible clusters is a promising direction for future works, such as Adaptive GMMs [16]. Another route that has a lot of potential is using the Cross-Correlation Kernel on the original time series data to create a new basis of regularized comparison [13]. This technology comes from the merger of signal processing domain science and machine learning and has the potential for providing classification models with simple features that preserve almost all of the original information. In the future, RNN networks can also be attempted to learn the time-series data [12].

7 Appendix

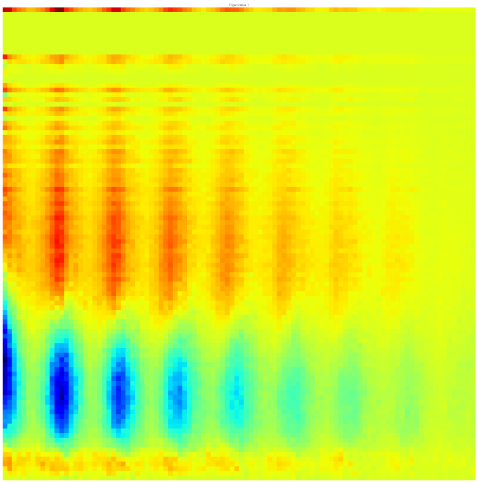

Figure 9: Plot of the first eigenspace of a structure function image.

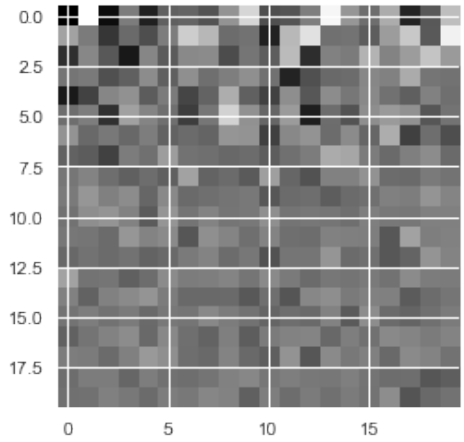

Figure 10: Plot of a low dimensional structure function transformed by PCA (100x100) -> (20,20).

References

[1] Gabriele Campanella, Arjun R. Rajanna, Lorraine Corsale, Peter J. Schüffler, Yukako Yagi, and Thomas J. Fuchs. Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology. *Comp. Med. Imag. and Graph.*, 65:142–151, 2018.

[2] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[3] François Chollet et al. Keras, 2015.

[4] Wim H de Vries, RH Becker, RL White, and C Loomis. Structure function analysis of long-term quasar variability. *The Astronomical Journal*, 129(2):615, 2005.

[5] SG Djorgovski, AJ Drake, AA Mahabal, MJ Graham, C Donalek, R Williams, EC Beshore, SM Larson, J Prieto, M Catelan, et al. The catalina real-time transient survey (crts). *arXiv preprint arXiv:1102.5004*, 2011.

[6] AJ Drake, SG Djorgovski, A Mahabal, E Beshore, S Larson, MJ Graham, R Williams, E Christensen, M Catelan, A Boattini, et al. First results from the catalina real-time transient survey. *The Astrophysical Journal*, 696(1):870, 2009.

[7] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[8] Shohei Hido, Hisashi Kashima, and Yutaka Takahashi. Roughly balanced bagging for imbalanced data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2(5-6):412–426, 2009.

[9] Ian T Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pages 115–128. Springer, 1986.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[12] Brett Naul, Joshua S Bloom, Fernando Pérez, and Stéfan van der Walt. A recurrent neural network for classification of unevenly sampled variable stars. *Nature Astronomy*, 2(2):151, 2018.

[13] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000.

[14] Chad Schafer. Statistical challenges in the search for dark matter, 2 2018. Remarks by Professor Chad Schafer at the Banff International Research Station for Mathematical Research and Discovery, Banff, Alberta Canada [Accessed: March-May 2018].

[15] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[16] Gabriel Wachman, Roni Khardon, Pavlos Protopapas, and Charles R Alcock. Kernels for periodic time series arising in astronomy. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 489–505. Springer, 2009.