# A Comparison of Distance Metrics for Light Curve Classification

Siddharth Chaini[1†], Ashish Mahabal[2,3], Ajit Kembhavi[4], Sukanta Panda[1]

[1] *Department of Physics, Indian Institute of Science Education and Research, Bhopal 462066, India*
[2] *Division of Physics, Mathematics and Astronomy, California Institute of Technology, Pasadena, CA 91125, USA*
[3] *Center for Data Driven Discovery, California Institute of Technology, Pasadena, CA 91125, USA*
[4] *Inter University Centre for Astronomy and Astrophysics (IUCAA), Pune 411007, India*

**Abstract**

The role of machine learning in astronomy has become apparent with the rise of extensive sky surveys like the Zwicky Transient Facility (ZTF). In this work, we present a new machine learning technique for outlier removal and light curve classification, using distance metric analysis. A distance is a quantity that tells us how far away two objects are. Because the distance between two objects from the same class is smaller than two objects from different classes, an analysis of these distances can be used to find outliers, and to separate and classify light curves. We aim to find the most optimum metric for fast and accurate light curve classification by analysing different distance metrics. The use of different distance metrics in this goal is a novel approach that has not been explored in astrophysics.

## 1 Introduction

Astronomy has seen an outburst in the amount of data collected in the last couple of decades. With the advances in computer data storage coupled with the number of large sky surveys conducted, it is now possible to record how millions of astronomical objects change over time. This growth has led to the advent of "time-domain astronomy" and the discovery of an unprecedented number of objects by adopting a data-driven approach. However, this deluge of data has brought new challenges with it. Manual human classification of all these objects has become impossible, with intelligent computer algorithms being the only way out. Machine learning is one such approach to it.

### 1.1 Time-Domain Astronomy

Time-domain astronomy refers to the branch of astronomy and astrophysics focused on studying the lifetime evolution and changes of different cosmic objects. These changes typically occur on short cosmic time scales.

Objects in Time-Domain Astronomy can typically be organised into 3 categories, namely:

1. Moving Objects: These are local objects that are so close to us that their positions change rapidly with time. Some examples are asteroids and comets.

2. Transients: These are events that do not repeat, and fade over time. They are usually the result of some explosion. Some examples are supernovae and gamma ray bursts.

3. Variable Objects: These are objects whose brightness changes on various timescales, either periodically or episodically, but does not fade over time. Some examples are cepheid variables (pulsating) and eclipsing binaries.

[†] siddharthc17@iiserb.ac.in

The Zwicky Transient Facility (ZTF; Bellm et al. (2019)) is a recent time-domain survey that had first light in 2017, at the Palomar Observatory, Caltech. The ZTF is set to advance time-domain astronomy by repeatedly observing the brightness of objects over time. Over the course of the project, the ZTF will produce over 50 terabytes of light curve data.

ZTF is thus in a position to accelerate the field of astroinformatics through the development of new computational techniques for large astronomical datasets. These new algorithms will also provide a "ready-to-use toolkit" (Graham et al., 2019) for data from future telescopes like the Vera C. Rubin Observatory, which will produce even more data, and is expected to help enlighten the unknown areas of dark energy and dark matter by exploring the transient optical sky.

### 1.2 Machine Learning

Machine learning refers to a set of computer algorithms in which the computer automatically learns patterns from data without being programmed. This learning is achieved by applying statistical models to analyse and draw inferences from patterns in data. The field of machine learning has become popular lately, with the availability of large datasets being a driving factor in its success in industrial applications for social media (Backstrom & Leskovec, 2010; Parameswaran et al., 2014; Kalyanam & Lanckriet, 2017), computer vision (Simonyan & Zisserman, 2015; He et al., 2015; Krizhevsky et al., 2017), natural language processing (Liu, 2012; Mikolov et al., 2013; Pennington et al., 2014; Wu et al., 2016), among others.

Machine learning has seen applications in many scientific disciplines as well, including medicine (Rani, 2011; Korsunsky et al., 2014; Esteban et al., 2017), biology (Tarca et al., 2007; Yuen et al., 2010; Xu & Jackson, 2019), chemistry (Rupp et al., 2011; Smith et al., 2017; Gastegger et al., 2017; Bartók et al., 2018), particle physics (Gligorov & Williams, 2012; Aurisano et al., 2016; de Oliveira et al., 2016; Guest et al., 2016; Long-Gang Pang et al., 2017; Tsaris et al., 2018; Shanahan et al., 2018), and astrophysics.

In astrophysics, it has seen uses in photometric redshift estimation (D'Isanto & Polsterer, 2017; Pasquet et al., 2019b), gravitational waves identification (George & Huerta, 2018), gravitational lensing identification (Cheng et al., 2020), star-galaxy classification (Vasconcellos et al., 2010; Kim & Brunner, 2017) and light curve classification (Lochner et al., 2016; Mahabal et al., 2017; Hinners et al., 2018; Dai et al., 2018; Pasquet et al., 2019a; Mahabal et al., 2019; Boone, 2019; van Roestel et al., 2021).

In this work, we present a new machine learning method for light curves using distance metric analysis. We talk about the data used in Section 2., explain what distance metrics are in Section 3. and talk about how this analysis can be used for outlier removal and classification of light curves in Section 4, and talk about future work to be done in Section 5.

## 2 Data Description

### 2.1 Catalog

For now, we choose to work with a limited dataset of variable stars, to develop techniques properly first. Once the techniques have been polished, we will expand the dataset's size by adding more variable stars as well as transients.

To select variable stars, we use the catalog made available by Chen et al. (2020). It consists of a total of 781,602 variable stars from 11 classes. We choose 600* objects of each of the following 4 classes of variable stars:

1. BY Draconis variables (BYDra) - Quasiperiodic variable stars exhibit brightness variations because of starspots and rotation of the star.

*Arbitrary number chosen based on a small sample of downloaded data

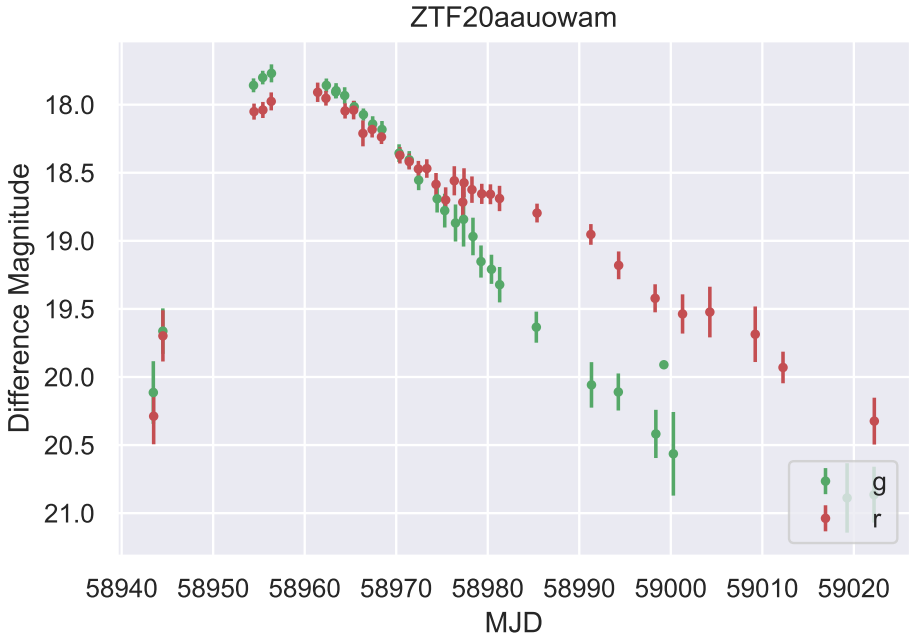


Figure 1: A light curve example of a Supernova Type Ia (ZTF20aauowam) in 2 passbands - g and r.

2. RR Lyrae variables (RR) - Periodic variable stars which undergo pulsation radially. Their periods range between 0.1 and 1 days.

3. Mira variables (Mira) - Another class of periodic variable stars undergoing radial pulsation, but with more stable periods ranging from 100 to 1000 days.

4. Eclipsing Algol variables (EA) - Periodic variable stars comprising of binary stars, whose brightness dips when one star passes in front of the other, forming an eclipse.

We choose an equal number of objects from each class, to form what is called a balanced dataset. The advantage of doing this is that there is no hidden bias due to the relative frequencies when the machine learns from the data.

### 2.2 Light Curves

After choosing the subset of objects from the catalog, we download the light curves for each from the catalog website[†]. The light curves are essentially multivariate time series, and consist of photometric readings taken by the ZTF in 2 filters - g (centred around 550 nm) and r (centred around 650 nm). An example of a light curve is given in Fig. 1. These light curves have been obtained from the ZTF Data Release 2 (DR2) by Chen et al. (2020).

### 2.3 Preprocessing and Feature Extraction

Since the ZTF is a ground-based telescope, the light curves obtained sparse and unevenly sampled, noisy and heteroskedastic. Because of this, the direct comparison of light curves is difficult. However, we can instead extract different features from these light curves - attributes that give us some information about the light curve. These features allow us to make a better comparison between different light curves.

[†]http://variables.cn:88

To calculate these features, we use the `lc_classifier`[‡] package released by the ALeRCE Broker[§] (Sánchez-Sáez et al., 2021).

Some of the features we use are: Amplitude, Beyond1Std, MaxSlope, Mean, Meanvariance, Multiband_period, Q31, Skew, Std, among others. A full list of all 108 features used, along with explanation of these features is available here and in Sánchez-Sáez et al. (2021).

The final dataset used for the distance metric analysis thus consists of 108 total features (i.e., $n = 108$), for a total of 2400 objects (600 from 4 classes each).

# 3 Methods

## 3.1 Distance Metrics

The notion of distance is intuitive to most of us. It tells us about the degree of closeness of two physical objects or ideas. The shorter the distance, the closer the objects or ideas are.

However, the existence of such a quantity does not tell us how to calculate it. Because of this, we can have different types of distances, each calculated differently. These distances can have different physical meanings, but they can also be entirely abstract.

### 3.1.1 Mathematical Formulation

**Definition 1.** *The distance $d$ between two points, in a set $X$, is a function $d : X \times X \to [0, \infty)$ that gives a distance between each pair of points in that set such that, for all $x, y, z \in X$, the following properties hold:*

1. $d(x, y) = 0 \iff x = y$ *(identity of indiscernibles)*
2. $d(x, y) = d(y, x)$ *(symmetry)*
3. $d(x, y) \leq d(x, z) + d(z, y)$ *(triangle inequality)*

The above three axioms also imply the following condition:

$$d(x, y) \geq 0, \text{for all } x, y \in X \tag{1}$$

Meanwhile, the term metric refers to the way of computing the distance between any two elements of the set. In this text, we will use the terms *metric*, *distance* and *distance metric* interchangeably[¶], all to refer to the distance as defined in Definition 1.

**Definition 2.** *A metric space $(X, d)$ is a set $X$ equipped with a metric $d$.*

A simple example of a metric space is the 2-dimensional $\mathbb{R}^2$ plane with the Euclidean distance. Similarly, we can also define metrics on matrices, functions, sets of points or any other mathematical object.

### 3.1.2 Distance in Machine Learning

In machine learning, the features (columns of the data) for any given problem form an abstract space known as the feature space. The points in this space are given by the individual data points (rows of the data). The feature space is similar to the n-dimensional real coordinate space, $\mathbb{R}^n$.

[‡] https://github.com/alercebroker/lc_classifier

[§] https://alerce.online

[¶] In some texts, the metric and the distance have slightly different definitions, but here we will use them interchangeably.

Different distance metrics can be defined in this higher dimensional feature space, as per definition 1. Because the distance is always a positive real number, it can be compared easily for different points even if the feature space is highly dimensional.

To calculate these distances, we use the following distance metrics:

1. **Euclidean distance**: The most widely used metric, this is the length of the shortest straight line joining 2 points, given by:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} \left(y_i - x_i\right)^2} \tag{2}$$

2. **Cityblock distance**: This gives the shortest path between 2 points if we restrict movement to a square grid. Named after the grid layout of the streets in Manhattan, New York, this is the shortest distance for a taxi to traverse between two places. It is given by:

$$d(x,y) = \sum_{i=1}^{n} \left|y_i - x_i\right| \tag{3}$$

3. **Canberra distance**: The Canberra distance is a modified, weighted version of the Cityblock distance, and is given by:

$$d(x,y) = \sum_{i=1}^{n} \frac{|y_i - x_i|}{|x_i| + |y_i|} \tag{4}$$

4. **Braycurtis distance**: Another modified version and weighted version of the Manhattan distance, the Braycurtis distance is different from how it is weighted, and is given by:

$$d(x,y) = \frac{\sum_{i=1}^{n} |y_i - x_i|}{\sum_{i=1}^{n} |x_i + y_i|} \tag{5}$$

A detailed description of each of these distances is available in (Deza & Deza, 2013). Furthermore, this list will be expanded further, in the second half of the thesis.

# 4 Results

Using the four distance definitions defined in Section 3.1.2, we calculate the four distances between all pairs of 2400 objects.

Ideally, we expect the distance between two objects from the same class (the intraclass distance) to be very low and almost equal to zero. This is because, variable stars of the same class have similar features and properties, and thus we expect them to have less distance between them. Using this, we demonstrate two applications of distance analysis - outlier removal and classification.

## 4.1 Outlier Removal

To visualise the distribution of the distances calculated between objects from the same class, we use a density plot, a continuous form of the histogram. For outlier removal, we use the Cityblock distance.

Initially, we find a large peak centred near 0, accompanied by a small peak at a vast distance (left subfigure; Figure 2). Because the standard deviation of this distribution is a couple of magnitudes higher than the median, we note that these objects are probably misclassifications

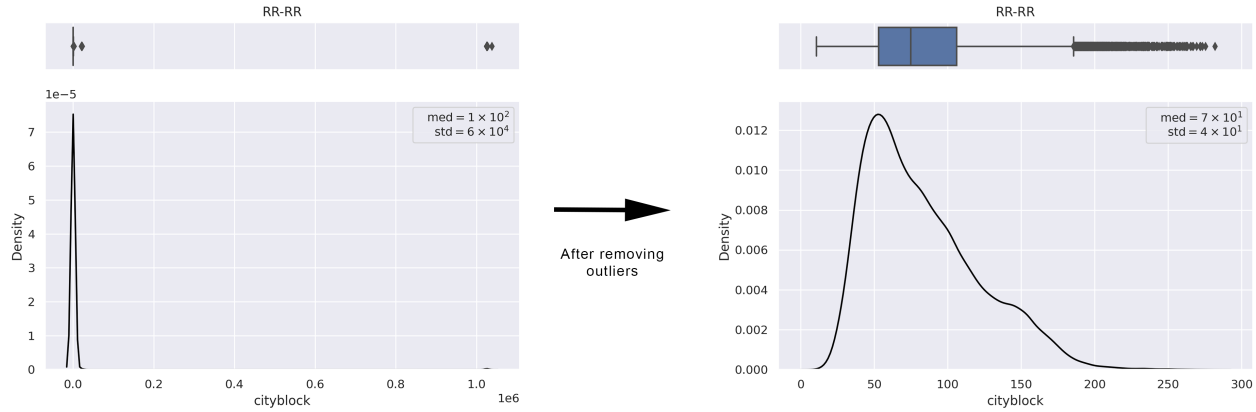


Figure 2: Density Plot for Cityblock distance distribution between RR Lyrae objects with RR Lyrae objects. (Left): Before outlier removal. (Right): After outlier removal.

in Chen et al. (2020)'s catalog, or, are objects for which feature extraction did not work. We thus claim that the objects responsible for this second peak are outliers, and remove them from our dataset.

To remove these outliers, we first calculate the pairwise distance between all objects from the same class. We then find the median distance for a particular object, and sort it based on this value. Finally, we drop 20 % of the objects with the largest median intraclass distances.

After removing the outliers, we obtain a much more constrained plot (right subfigure; Figure 2), with significantly smaller intraclass distance values as compared to before.

### 4.2 Classification

Using the intraclass distances, we can also classify these objects. Firstly, we drop the outliers using the method mentioned in Section 4.1, leaving a total of 480 objects.

Then, we randomly shuffle and split the remaining dataset into a training set (85% - roughly 400 objects each) and a test set (15% - roughly 80 objects each). The machine learns the classification using the training set. Meanwhile, the true classification of the test set is hidden from the machine, and is used only for final performance evaluation.

Following this, we create a "canonical" feature set for each class in the training set. The canonical feature set tells us what features an "average" light curve of that class has. To create this canonical feature set, we calculate the median value for each feature for that class. It is at this point we say that the training is complete, and the machine now has learnt what the features for each class' light curve look like.

Now, to predict and classify the objects from the test set, we take the unknown test object's features and calculate the distance between it and the canonical set features for all four classes. Finally, the class corresponding to which this distance is minimum is the prediction for the test object.

Once we have the predictions for all objects in the test set, we can evaluate the classifier's performance through its accuracy, which is the ratio of the number of correct predictions to the total number of predictions.

To compare the performance of different distance metrics, we calculate the accuracies for all four of them in Table 1.

| Distance Metric | Accuracy |
|---|---|
| Euclidean | 73.61% |
| Cityblock | 82.29% |
| Canberra | 94.10% |
| Braycurtis | 84.03% |

Table 1: Comparison of performance of the 4 distance metrics for light curve classification.

We thus see that the Canberra distance gives us the best result for classification, with an accuracy of 94.10%. An open question as of now is why the Canberra distance performs better than the other metrics.

# 5 Future Work

In the second half of the thesis, we wish to cover the following parts:

- Explain why the Canberra distance performs better than the others.
- Create a classifier that combines the results of multiple distance metrics
- Increase the number of distance metrics used.
- Increase the size of the dataset.
- Decrease the dimensionality of the feature space by eliminating redundant features based on distance metric analysis, leading to faster performance.
- Come up with custom distance metrics to calculate the distances directly from light curves, without the need for feature extraction, which is computationally expensive and slow.
- Finding the least amount of data (number of samples, number of points, number of bands) for successful classification.
- Going beyond machine learning algorithms and exploring the use of different advanced genetic algorithms for the classification.

## References


Aurisano A., et al., 2016, Journal of Instrumentation, 11, P09001

Backstrom L., Leskovec J., 2010, arXiv:1011.4071 [physics, stat]

Bartók A. P., Kermode J., Bernstein N., Csányi G., 2018, Physical Review X, 8, 041048

Bellm E. C., et al., 2019, Publications of the Astronomical Society of the Pacific, 131, 018002

Boone K., 2019, arXiv:1907.04690 [astro-ph 10.3847/1538-3881/ab5182

Chen X., Wang S., Deng L., de Grijs R., Yang M., Tian H., 2020, The Astrophysical Journal Supplement Series, 249, 18

Cheng T.-Y., Li N., Conselice C. J., Aragón-Salamanca A., Dye S., Metcalf R. B., 2020, Monthly Notices of the Royal Astronomical Society, 494, 3750

D'Isanto A., Polsterer K. L., 2017, arXiv:1706.02467 [astro-ph 10.1051/0004-6361/201731326

Dai M., Kuhlmann S., Wang Y., Kovacs E., 2018, arXiv:1701.05689 [astro-ph 10.1093/mnras/sty965

Deza M. M., Deza E., 2013, Encyclopedia of Distances. Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-642-30958-8

Esteban C., Hyland S. L., Rätsch G., 2017, arXiv:1706.02633 [cs, stat]

Gastegger M., Behler J., Marquetand P., 2017, Chemical Science, 8, 6924

George D., Huerta E. A., 2018, Physics Letters B, 778, 64

Gligorov V. V., Williams M., 2012, arXiv:1210.6861 [hep-ex, physics:physics 10.1088/1748-0221/8/02/P02013

Graham M. J., et al., 2019, Publications of the Astronomical Society of the Pacific, 131, 078001

Guest D., Collado J., Baldi P., Hsu S.-C., Urban G., Whiteson D., 2016, arXiv:1607.08633 [hep-ex, physics:physics 10.1103/PhysRevD.94.112002

He K., Zhang X., Ren S., Sun J., 2015, arXiv:1512.03385 [cs]

Hinners T., Tat K., Thorp R., 2018, The Astronomical Journal, 156, 7

Kalyanam J., Lanckriet G., 2017, UC San Diego, Retrieved from https://escholarship.org/uc/item/6545w71z

Kim E. J., Brunner R. J., 2017, Monthly Notices of the Royal Astronomical Society, 464, 4463

Korsunsky I., Ramazzotti D., Caravagna G., Mishra B., 2014, arXiv:1408.6032 [cs, q-bio, stat]

Krizhevsky A., Sutskever I., Hinton G. E., 2017, Communications of the ACM, 60, 84

Liu B., 2012, Synthesis Lectures on Human Language Technologies, 5, 1

Lochner M., McEwen J. D., Peiris H. V., Lahav O., Winter M. K., 2016, arXiv:1603.00882 [astro-ph 10.3847/0067-0049/225/2/31

Long-Gang Pang Zhou K., Su N., Petersen H., Stocker H., Xin-Nian Wang 2017, Training and Testing Data Used in the Paper "An Equation-of-State-Meter of QCD Transition from Deep Learning", doi:10.6084/M9.FIGSHARE.5457220.V1

Mahabal A., Sheth K., Gieseke F., Pai A., Djorgovski S. G., Drake A., Graham M., Collaboration t. C., 2017, 2017 IEEE Symposium Series on Computational Intelligence (SSCI), pp 1–8

Mahabal A., et al., 2019, Publications of the Astronomical Society of the Pacific, 131, 038002

Mikolov T., Sutskever I., Chen K., Corrado G., Dean J., 2013, arXiv:1310.4546 [cs, stat]

Parameswaran A., Boyd S., Garcia-Molina H., Gupta A., Polyzotis N., Widom J., 2014, Proceedings of the VLDB Endowment, 7, 685

Pasquet J., Pasquet J., Chaumont M., Fouchez D., 2019a, arXiv:1901.01298 [astro-ph 10.1051/0004-6361/201834473

Pasquet J., Bertin E., Treyer M., Arnouts S., Fouchez D., 2019b, Astronomy & Astrophysics, 621, A26

Pennington J., Socher R., Manning C., 2014, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp 1532–1543, doi:10.3115/v1/D14-1162

Rani K. U., 2011, International Journal of Data Mining & Knowledge Management Process, 1, 1

Rupp M., Tkatchenko A., Müller K.-R., von Lilienfeld O. A., 2011, arXiv:1109.2618 [cond-mat, physics:physics, stat 10.1103/PhysRevLett.108.058301

Sánchez-Sáez P., et al., 2021, The Astronomical Journal, 161, 141

Shanahan P. E., Trewartha D., Detmold W., 2018, Physical Review D, 97, 094506

Simonyan K., Zisserman A., 2015, arXiv:1409.1556 [cs]

Smith J. S., Isayev O., Roitberg A. E., 2017, Chemical Science, 8, 3192

Tarca A. L., Carey V. J., Chen X.-w., Romero R., Drăghici S., 2007, PLoS Computational Biology, 3, e116

Tsaris A., et al., 2018, Journal of Physics: Conference Series, 1085, 042023

Vasconcellos E. C., de Carvalho R. R., Gal R. R., LaBarbera F. L., Capelato H. V., Velho H. F. C., Trevisan M., Ruiz R. S. R., 2010, arXiv:1011.1951 [astro-ph 10.1088/0004-6256/141/6/189

Wu Y., et al., 2016, arXiv:1609.08144 [cs]

Xu C., Jackson S. A., 2019, Genome Biology, 20, 76, s13059

Yuen H., Shimojo F., Zhang K. J., Nomura K.-i., Kalia R. K., Nakano A., Vashishta P., 2010, arXiv:1012.0900 [physics, q-bio]

de Oliveira L., Kagan M., Mackey L., Nachman B., Schwartzman A., 2016, Journal of High Energy Physics, 2016, 69

van Roestel J., et al., 2021, arXiv:2102.11304 [astro-ph]