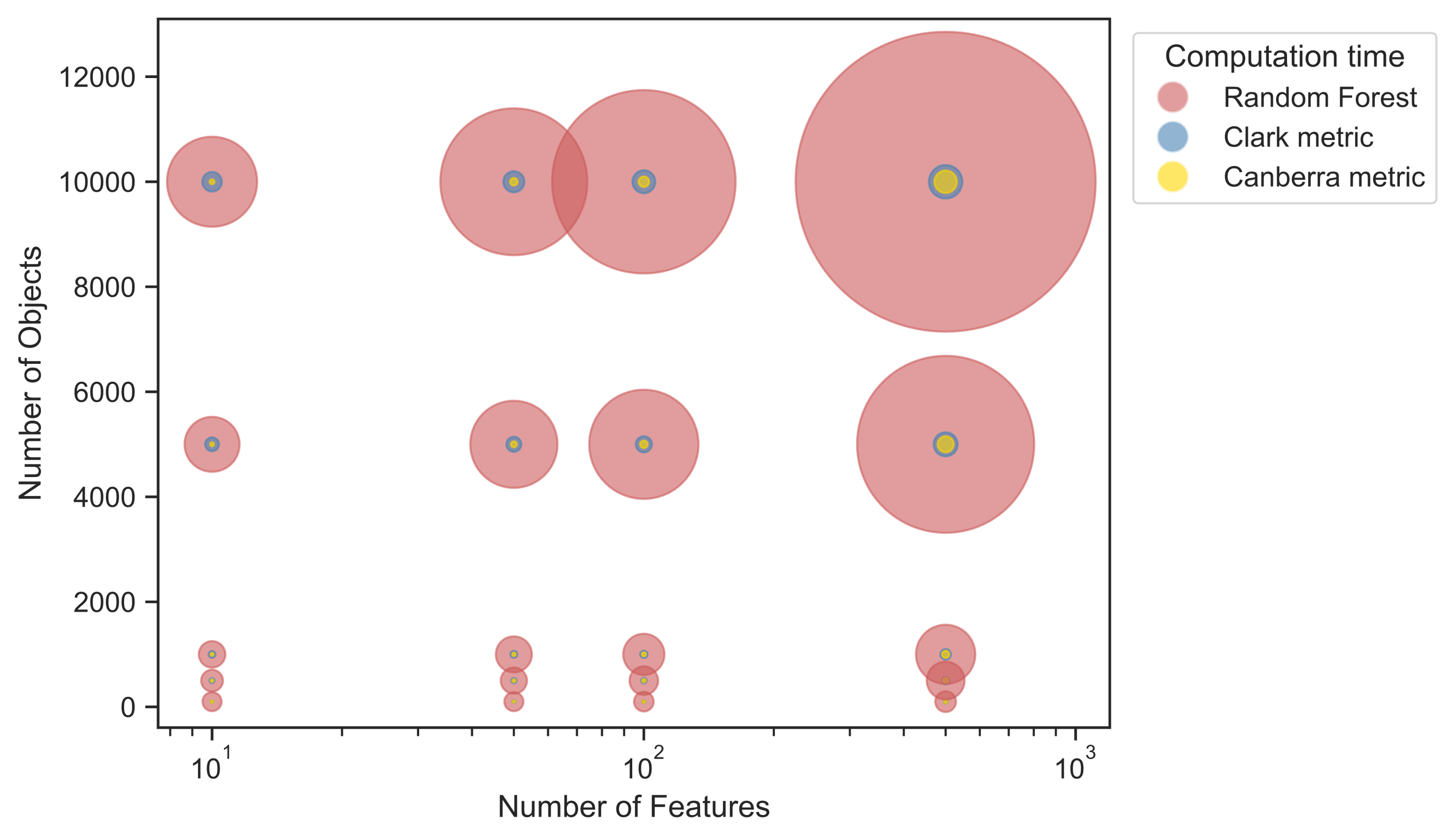
Distance metrics for high-dimensional classifiers:

- As accurate as other ML methods
- Easy to interpret
- Faster than most other ML methods



Light curve classification using distance metrics

Siddharth Chaini[1]
Ashish Mahabal[2]
Federica B. Bianco[1]

[1]University of Delaware
[2]Caltech

Download PDF

Introduction

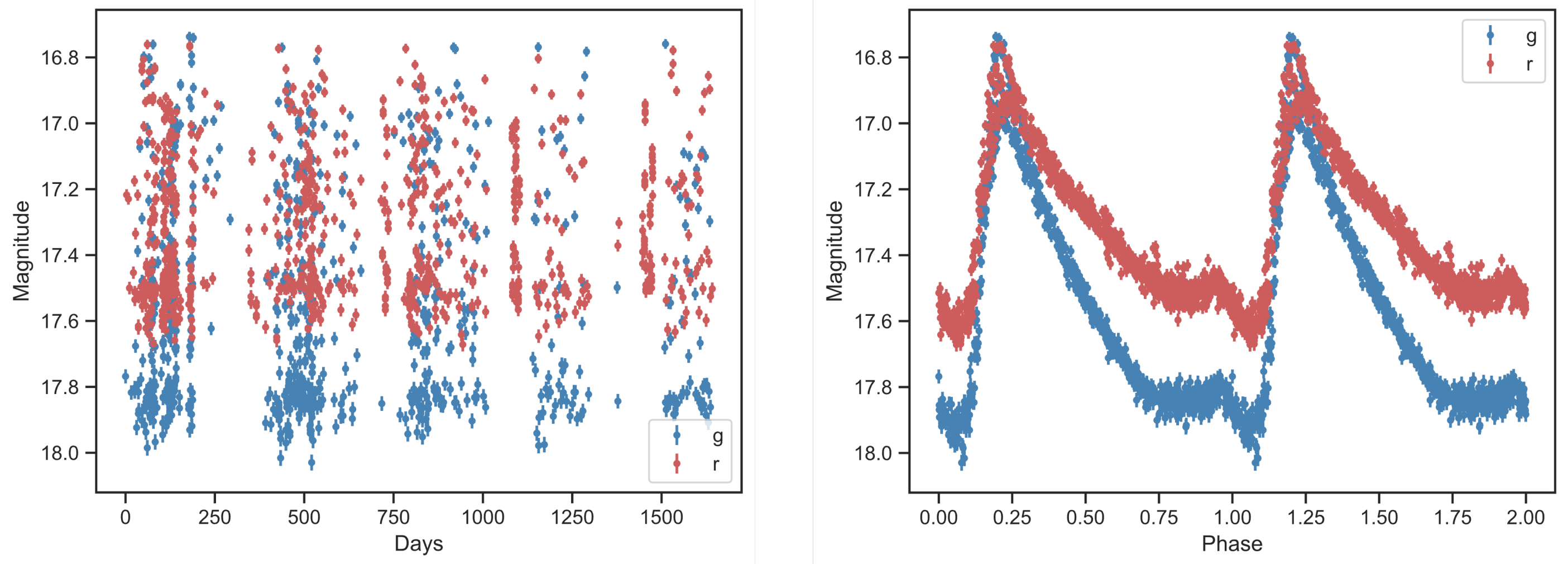
The rise of synoptic sky surveys has ushered in a new era of big data in time-domain astrophysics, making data science essential. While tree-based methods (e.g., Random Forest) are the standard in astrophysical classification, the direct use of distance metrics has not been explored in time-domain astrophysics.

We looked at 18 different distance metrics for classifying astrophysical time series (called light curves) and make recommendations for efficient and physically interpretable classification algorithms.

Data

We use time series data (light curves) from the *Zwicky Transient Facility* (ZTF). We study 4 variable star classes (*CEP, DSCT, RR, RRc*) using light curves that measure brightness in two filters: g (472.2 nm), and r (633.9 nm).

We then extract 118 features based on statistics and model fits from these light curves using a Python package, `lc_classifier`.



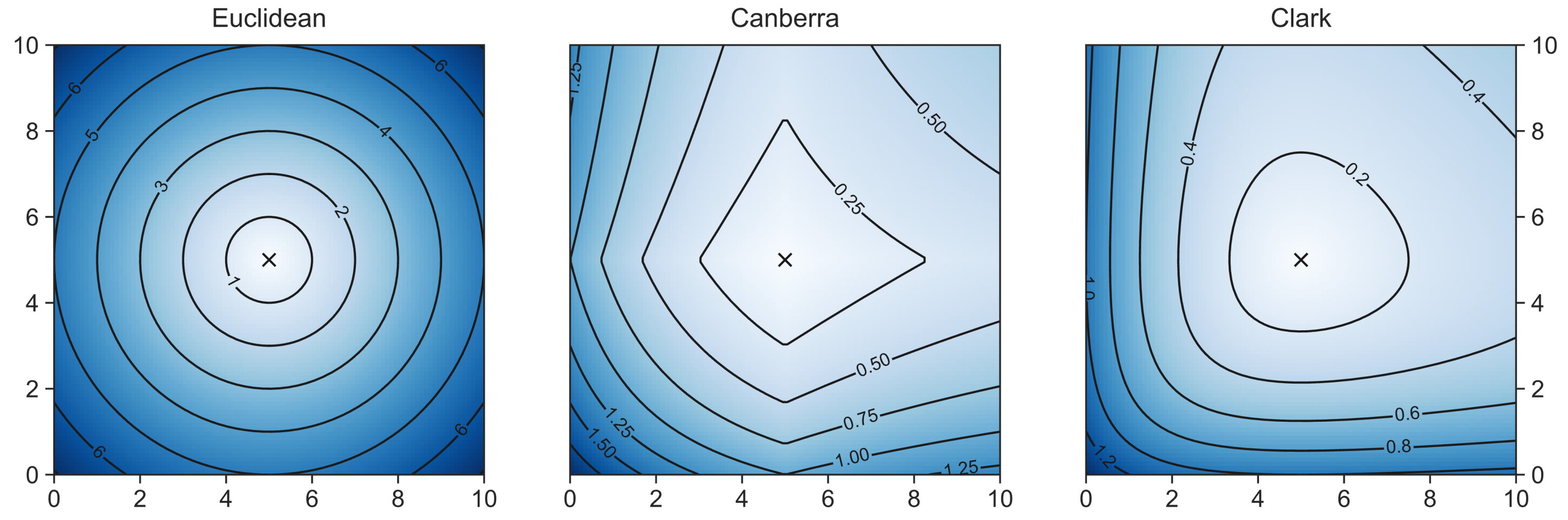
Example light curve of an RR Lyrae star.
(Left: Raw light curve, Right: Period folded light curve)

Distance Metrics

A distance metric is any mathematical function that is:

- Symmetric
- Follows triangle inequality
- Zero only when points coincide

We can thus use different distance metrics for machine learning! For e.g., Euclidean, Canberra, Clark, etc.



Algorithm

Training

For each class and each feature,

1. Compute the median
2. Computer the std dev.

The median set of features for a class is the representative set from which distances are computed in the prediction step.

Predicting

1. Choose a metric.
2. For each class,

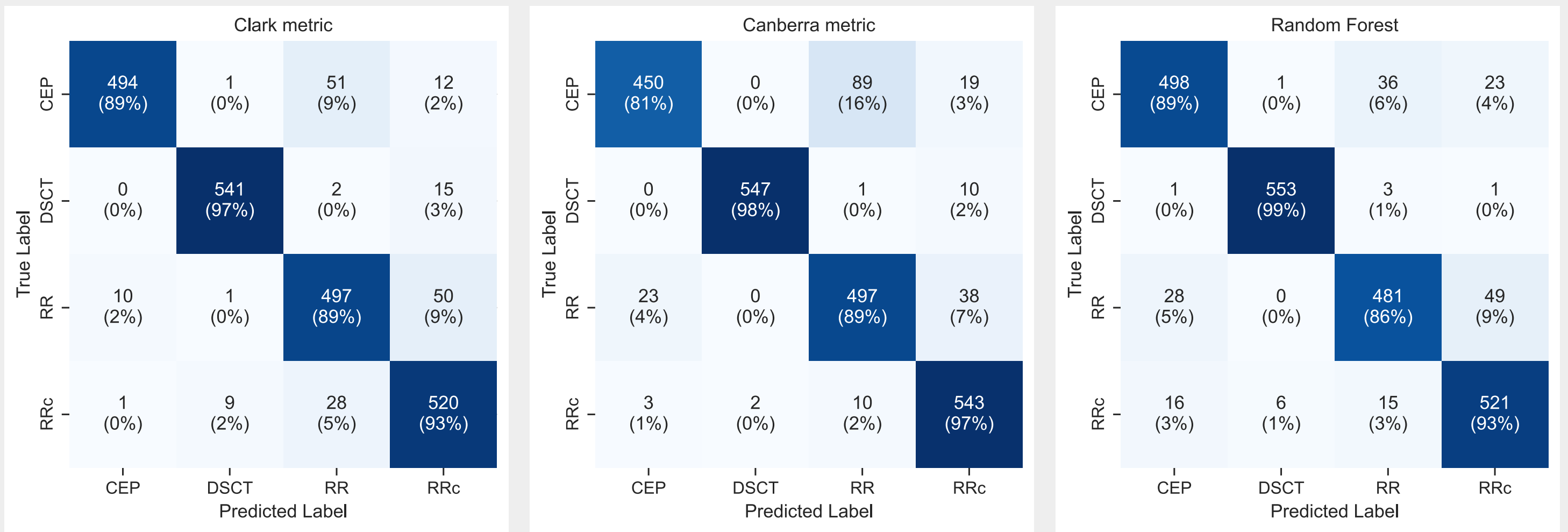
a) Scale both the median set and test object by std dev for that class.

b) Compute distance between test object and median set.

c) Choose the class for which distance is minimum.

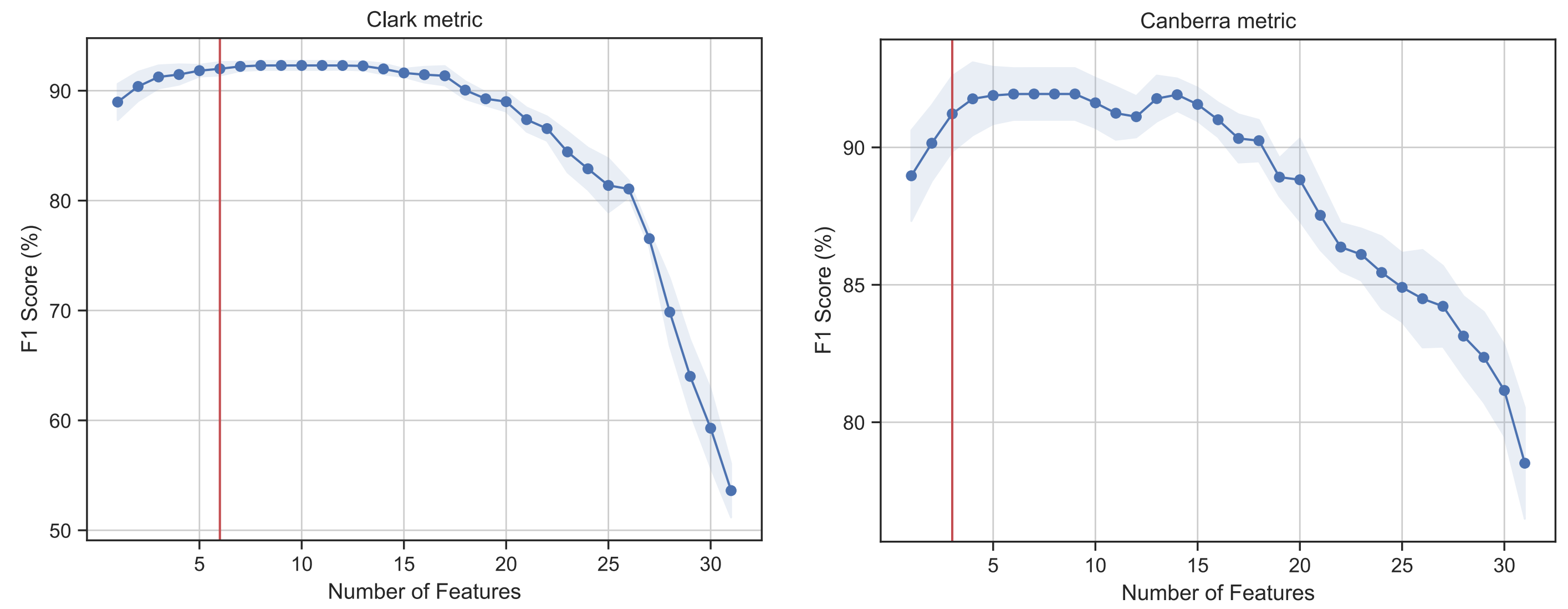
Results

Classification Performance: As accurate as a random forest



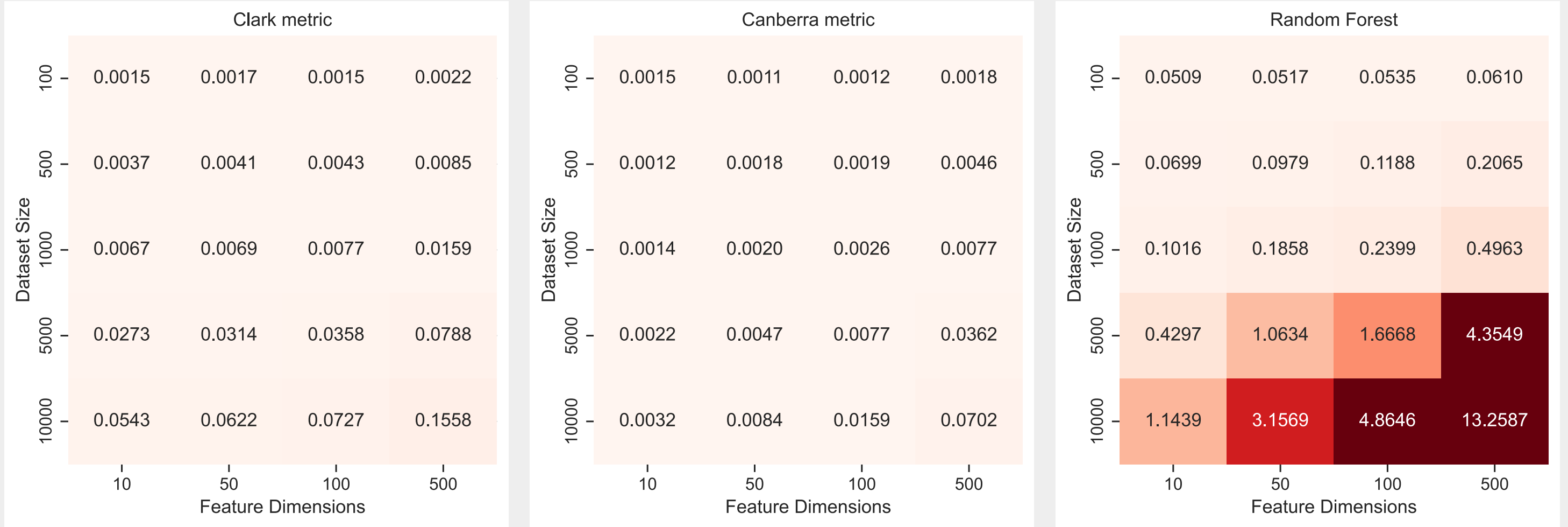
Confusion matrices for the multi-class classification problem with the Clark Distance Classifier (F_1= 92.0%) and Canberra Distance Classifier (F_1= 91.2%), compared to a Random Forest Classifier (F_1= 92.0%).

Sequential Feature Selection: Makes interpretation easy



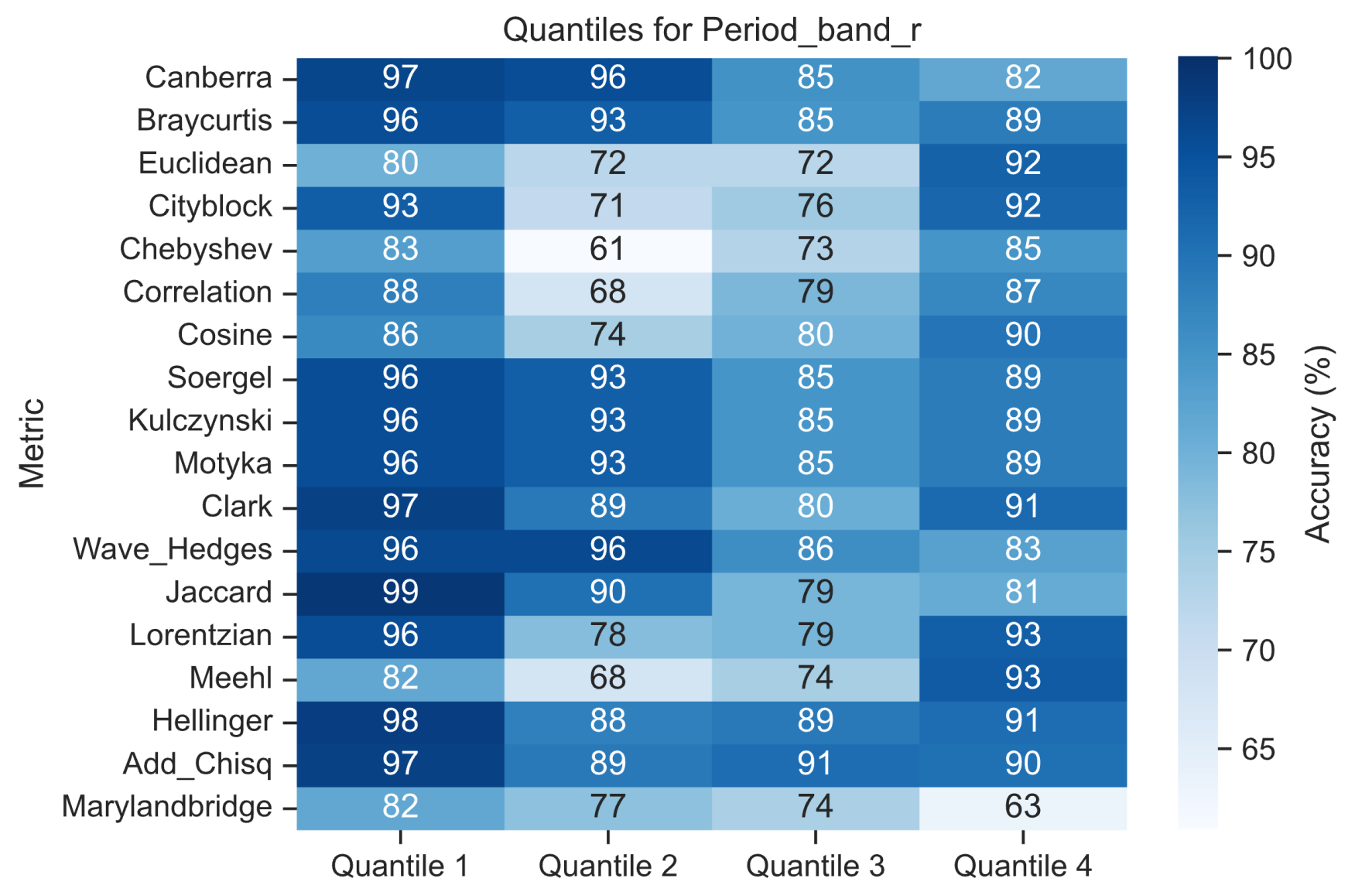
Features are chosen iteratively, one at a time, to maximize overall performance. The final set of n features is chosen to be the smallest number of features for which the F_1 is within 1 std dev of the maximum F_1 score. The final set of features effectively reduces dimensionality and varies for different distance metrics.

Computation Time: Faster than a random forest!



Computation time (s) for different methods

Robustness:



A comparison of performance when test objects are taken from different quantiles based on the feature r -band period.

We see that, for different quantiles, different metrics perform better. This can serve as a powerful way to choose the metric for a given problem!

Conclusion

The use of distance metrics is a very promising approach which may have important benefits for "dynamic" classification, dimensionality reduction and anomaly detection in time-domain astrophysics, astroinformatics and beyond.

References

[1] Bellm et al. (2019), PASP 131, 018002.
[2] van Roestel, J. et al. (2021), AJ 161, 267.
[3] Cha, S.-H. (2007), International Journal of Mathematical Models and Methods in Applied Sciences 1, 300–307.
[4] Chaini, S. N., Mahabal, A., Kembhavi, A. K. & Panda, Indian Institute of Science Education and Research Bhopal, 2022.