

Boosting the Classification of Complex Large Synoptic Survey Telescope (LSST) Data

Rao Farhat Masood¹, Imtiaz Ahmad Taj¹

¹Department of Electrical Engineering, Capital University of Science and Technology (CUST), Islamabad, Pakistan

Corresponding author: Rao Farhat Masood (e-mail: farhatmasood.fm@gmail.com).

ABSTRACT Analysis of light curves emanating from various celestial bodies is of paramount importance in order to enable ourselves with the ability to quantify the variability in sky and discover time-varying objects. Large Synoptic Survey Telescope (LSST) [1] gather voluminous time-series data, however, classifying these events from large-scale surveys is a challenging task that requires efficient and robust machine learning methods. In this paper, we present a novel approach for astronomical time series classification using gradient boost, a powerful ensemble technique that combines weak learners into a strong classifier. We apply our method to two datasets from the Catalina [2] and Zwicky Transient [3] Facility surveys, which contain lightcurves of various types of transients and variables. We compare our results with state-of-the-art methods that use different features and models. We show that our method achieves superior performance in terms of accuracy with comparable computational complexity. We also discuss the advantages and limitations of our method and suggest possible directions for future work.

INDEX TERMS LSST Dataset, Light Curves, Random Forest, Gradient Boosting

I. INTRODUCTION

The exploration of our cosmic surroundings has transitioned into a revolutionary phase with the implementation of advanced astronomical surveys, prominently featuring the Large Synoptic Survey Telescope (LSST) [1]. These surveys meticulously curate large amounts of time-series data [4], capturing the ever-evolving dynamism of celestial entities. In order to reveal the concealed intricacies within this data mandates an adept classification methodology; necessitating application of sophisticated machine learning [5] methodologies.

The Large Synoptic Survey Telescope (LSST), being positioned at the Cerro Pachón mountain in Chile, is designed to execute and collect time-series data of the night sky. It has the ability to capture the wide-field images covering the entire Southern hemisphere. Equipped with a 3.2-gigapixel camera, LSST promises to capture dynamic cosmic events, contributing immensely to our understanding of the universe's evolution and structure.

Unarguably, the utilization of machine learning algorithms [5] to classify time-series data from the Large Synoptic Survey Telescope (LSST) marks a significant leap forward in the field of observational astronomy. Although, originally based on the Photometric LSST Astronomical Time-Series Classification Challenge (PLAsTiCC), Kaggle Challenge [6] (2018), researchers have contributed ably to unveil the hidden aspects of night sky by proposing novel methodologies. A possibility still exists in order to substantiate the usefulness and relevance of modern machine learning sophisticated algorithms to solve a complex problem without having the expert-domain knowledge.

In this paper, we present a Gradient Boost [7] based classifier to classify the objects observed by LSST, advocating the ease and accuracy to implement a complex problem without being hindered by absence of expert-domain knowledge. The data used for training our model is acquired from Kaggle Dataset [6]. The trained model is able to classify with promising accuracy and reasonable computational complexity.

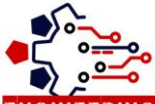
The paper is organized as follows: Section II covers the literature review, Section III highlights the aspects related to dataset, Section IV showcases the proposed methodology. Results are presented in Section V and finally conclusion with future direction for related work is mentioned.

II. LITERATURE REVIEW

Keeping the scope of this work restricted to application of Machine Learning models to classify the light curves, the related material was evaluated. The Kaggle Challenge (PLAsTiCC) [8] attracted entries from over 1,000 teams, showcasing diverse approaches to tackle the classification problem. The challenge enables bench-marking of various proposed solutions in the form of algorithms.

The winner of Competition Kyle Boone [9], utilized boosted Decision Trees (BDTs) with Light GBM achieving exceptional performance in Supernova and Kilonova classification. It was revealed that BDTs offered higher interpretability and robustness to overfitting. Various other participants [10], [11] in their post-challenge analysis, mentioned exploring KNN and K-means clustering but ultimately found them less effective than their final approach.

Many teams [10] combined individual weak



learners, leveraging their strengths and have established that stacking and blending yielded significant performance. It was also revealed that ensembles often provided better generalization and robustness than single models. In another work, Khan et al. [12] used Random Forests [13] and achieved a reasonable accuracy. As per the researchers they have achieved a score of 63% on the same dataset. It is revealed that traditional machine learning models struggle with high-volume datasets and have shown to exhibit poor generalization.

Based on the reviewed literature, ensemble methods have the potential to produce promising results for addressing the classification problems, without having the constraint of limited expert-domain knowledge. The research void is further elaborated in the sections where we best present the gradient boost method to produce the best results.

III. DATASET

A. ORIGINAL DATASET WITH LIGHT-CURVES DATA

The PLAsTiCC dataset consists of training and test sets provided in multiple CSV files. There are two types of files: header files containing summary information about astronomical objects and light-curve data containing time series of fluxes in six filters, including flux uncertainties.

1) Header File Information

- **Object ID:** A unique identifier, typically an integer (int32), assigned to each astronomical object for identification and tracking purposes.
- **ra:** Right Ascension, a sky coordinate representing the east-west position of an object on the celestial sphere (float32).
- **decl:** Declination, a sky coordinate representing the north-south position of an object on the celestial sphere (float32).
- **gal l:** Galactic longitude, a coordinate in the galactic coordinate system indicating the angular distance of an object from the Galactic Center (float32).
- **gal b:** Galactic latitude, a coordinate in the galactic coordinate system indicating the angular distance of an object from the Galactic plane (float32).
- **ddf:** A Boolean flag indicating whether the object is within the Deep Drilling Field (DDF) survey area (1 for DDF).
- **hostgal specz:** The spectroscopic redshift of the source, providing a precise measure of the source's distance by examining its spectral lines (float32).
- **hostgal photoz:** The photometric redshift of the host galaxy, estimating the source's distance based on its color (float32).
- **hostgal photoz err:** The uncertainty associated with the photometric redshift of the host galaxy (float32).
- **distmod:** The distance modulus calculated from the photometric redshift of the host galaxy, providing a measure of the object's luminosity distance (float32).

- **MWEBV:** The extinction of light due to Milky Way dust, accounting for the dimming of light from astronomical objects as it passes through interstellar dust in the Milky Way (float32).
- **target:** The class of the astronomical source, represented as an integer (int8). This could refer to various types of astronomical objects such as stars, galaxies, or supernovae.

2) Light-Curve Information

- **Object ID:** Unique identifier, typically an integer (int32), for reference and tracking.
- **MJD:** Modified Julian Date, a continuous count of days since November 17, 1858, used in astronomy.
- **Passband:** LSST-specific range of wavelengths, usually represented as an integer (int8).
- **Flux:** Measured brightness of an object in a specific passband, corrected for Milky Way extinction (MWEBV). Typically represented as a floating-point number (float32).
- **Flux Err:** Uncertainty or error in the measured flux, also a floating-point number (float32).
- **Detected:** Boolean flag (True/False) indicating a significant brightness difference at the 3σ level, considered statistically significant in statistics.

The light curves in this study span six distinct bands identified as *u*, *g*, *r*, *i*, *z* and *y*, each associated with specific light wavelengths. The wavelength ranges for these bands are showcased in Table 1

TABLE 1. Wavelength Range of Light Bands

Band	Wavelength Range (nm)
u	300 to 400
g	400 to 600
r	500 to 700
i	650 to 850
z	800 to 950
y	950 to 1050

B. FEATURE EXTRACTION

Feature extraction played a crucial role in this work, as highlighted by Khan et al. [12] for effective classification of light sources into multiple categories. To capture diverse characteristics of the light bands, we extracted both statistical, shape-based and higher-order features from the training dataset.

1) Statistical Features

These features quantify properties such as the average brightness (μ), weighted average (μ_w), standard deviation (σ), skewness (γ), kurtosis (κ), maximum (X_{max}) and minimum (X_{min}) values, median (\tilde{X}), and median absolute deviation (MAD). Additionally, the percentage of values beyond one standard deviation from the weighted average is also calculated. Comprehensive list of statistical features is

shown in table 2.

TABLE 2. Statistical Features

Statistical Feature	Equation
Mean (μ)	$\mu = \frac{1}{n} \sum_{i=1}^n x_i$
Weighted Average (μ_w)	$\mu_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$
Standard Deviation (σ)	$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$
Percent Beyond 1 STD	Number of x_i beyond $\mu + \sigma$
Skewness (γ)	$\gamma = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^3}{\sigma^3}$
Kurtosis (κ)	$\kappa = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^4}{\sigma^4}$
Maximum Value (X_{\max})	$X_{\max} = \max(x_1, x_2, \dots, x_n)$
Minimum Value (X_{\min})	$X_{\min} = \min(x_1, x_2, \dots, x_n)$
Median (\tilde{X})	$\tilde{X} = \text{Median}(x_1, x_2, \dots, x_n)$
Median Absolute Deviation (MAD)	$\text{MAD} = \text{Median}(x_i - \text{Median})$

2) Shape-based Features

Eight features capture visual distinctions in light curve shapes, essential for the classification of astronomical entities. These features encompass:

Maximum Slope: The maximum slope represents the steepest rate of change in the signal and is calculated as:

$$\text{MaxSlope} = \max \left(\frac{x_i - x_{i-1}}{t_i - t_{i-1}} \right)$$

- Amplitude:** The amplitude measures the range between the maximum and minimum values of the signal and is computed as:

$$\text{Amp} = \frac{2}{\text{Max} - \text{Min}}$$

- Peaks Above the Weighted Average:** This feature counts the number of data points in the signal that are above the weighted average (μ_w) and is expressed as:

$$\text{Peaks}^{above} = \sum_{i=1}^n 1(x_i > \mu_w)$$

- Maximum Peak Prominence (MaxProm) and Minimum Peak Prominence (MinProm):** These features represent the highest and lowest values of peak prominence in the signal, respectively:

$$\text{MaxProm} = \max(\text{PeakProm}_i)$$

$$\text{MinProm} = \min(\text{PeakProm}_i)$$

- Average Time to Brighten (AvgBrightTime) and Average Time to Fade (AvgFadeTime):** These features denote the average time taken for the signal to brighten and fade, respectively. They are computed as the average of the time differences between consecutive data points where the signal experiences significant changes.

$$\text{AvgBrightTime} = \frac{\text{Peaks}^{above}}{n} \sum_{i=1}^n \Delta t_i$$

$$\text{AvgFadeTime} = \frac{\text{Peaks}^{below}}{n} \sum_{i=1}^n \Delta t_i$$

3) Wavelet Decomposition

Using the DB1 wavelet, light curves are decomposed into three levels, resulting in decomposed components. The energy of these components, comprising four features, provides insights into object activity across different passbands. The features include the energy of wavelet decomposition at detail levels 1, 2, and 3, as well as the approximation level 3. The energy of the decomposed component at level k , denoted by E_k , is calculated using the equation 1

$$E_k = \sum_{i=1}^n |d_k(i)|^2 \quad (1)$$

where n represents the number of coefficients in the decomposed component, and $d_k(i)$ denotes the i -th coefficient of the decomposed component at level k . The features are concatenated to form a feature vector of length 161. For each of the six channels, features are computed both by concatenating sequentially and adding them point-by-point, offering a comprehensive view of the object's profile. Our analysis utilized synthetic data provided for training, containing 7,848 samples distributed among 14 classes. For evaluation, a separate set of approximately 3.5 million light curves was reserved. We identified a significant class imbalance within the training data, with only three categories comprising over half of the samples. Figure 1 provides a detailed breakdown of class distribution showcasing the class imbalance challenge within the training samples.

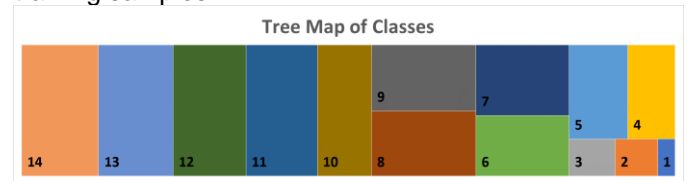


FIGURE 1. Distribution of 14 x Classes.

IV. PROPOSED METHODOLOGY

As already well-established by various researchers [10]–[12], [14]–[16] regarding suitability of ensemble methods for improvement in classification results; we also exhibited the same through extensive experimentation. Initially, we used Random Forest [13] Ensemble method as used by Khan et al. [12] and achieved the claimed accuracy score of 63%. In order to improve the claimed score various other methods were extensively experimented. Class imbalance issue was resolved by splitting the training and testing set in the ratio of 60:40 for every class. We didn't perform data augmentation, as it improved the classification accuracy but compromised the integrity of original data. Summarized overview of each model with the methods is presented in the subsequent paragraphs.

- Using the Random Forest [13] ensemble technique, the overall training time of the model was 345 seconds, while achieved the claimed accuracy of 63 %.
- kNN was only able to achieve accuracy of 43 % however, the training time was less than 10 seconds.
- The value of k was used keeping in view the Rule-of-Thumb defined as $k = \sqrt{n}$, where k is the number of neighbors in k-Nearest Neighbors (k-NN) and n is the number of samples in the dataset.
- Making use of CART Model, we achieved 56% accuracy score with the training time of 6 seconds.
- SVM [17] gave the accuracy of 57% with the training time of 19.5 seconds.
- We also used MLP, with Sigmoid as activation function and found out that accuracy score was improved to 60.9%. We also found out that the time required to train the MLP was increased to 698 seconds.
- As a first to establish the ensemble methods, we used Majority Voting Ensemble with kNN, Decision Tree as base classifier and achieved the overall accuracy of 61%.
- We used other ensemble methods including Adaboost, gradient boost and bagging to test the respective model performance and found that Gradient Boost gave the most promising accuracy, surpassing the claimed accuracy of Random Forest [12] model by 5 percent. Adaboost and bagging classifiers achieved the accuracy scores of 50% and 65% respectively.

V. PERFORMANCE

In the Kaggle Competition [6], the evaluation metric is defined in equation 2, which is simply the log loss score.

$$S = - \frac{\sum_{k=1}^K W_k \sum_{j=1}^{J_i} \frac{\theta_{j,k}}{J_i} \ln(P_{jk})}{\sum_{k=1}^K W_k} \quad (2)$$

where, J_i is the number of objects in the class set, K is the number of classes, \ln is the natural logarithm, $\theta_{j,k}$ is 1 if observation (j) belongs to class (k) and 0 otherwise, (P_{jk}) is the predicted probability that observation j belongs to class k. It is also critically experimented through extensive testing that the log loss function which is defined in the Kaggle Competition [6] and overall model accuracy are inversely proportional to each other. The equation which is used to compute the overall accuracy of the model is given in equation 3.

$$Accuracy = \frac{\sum_{i=1}^{14} TruePositives_i}{\sum_{i=1}^{14} (TruePositives_i + FalsePositives_i)} \quad (3)$$

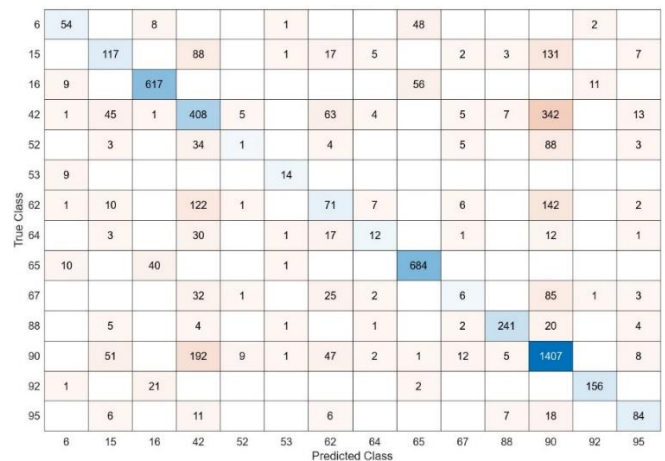
The overall performance comparison along with model training time is presented in table 3.

TABLE 3. Model Performance Metrics

Model	Accuracy (%)	Training Time (sec)
Random Forest [12]	63.0	345
kNN	43.0	3
CART	56.0	6
SVM	57.0	19.5
MLP	60.9	698
Majority Voting	61.0	337
AdaBoost	50.5	18.6
GradBoost	68.7	1228
Bagging	65.4	26.3

It has been revealed that given the consideration of accuracy while ignoring the model computational complexity,

Gradient Boost [7] outperformed all other machine learning algorithms. However, the best optimization was achieved using Bagging Ensemble method wherein decent and comparable accuracy was achieved with substantially less computational complexity. The entire models used and best ones along with the respective training times are highlighted in the table 3. Likewise, the Confusion Matrix for the best model is shown in 2, wherein, it is also established that the most challenging case was regarding Class Label: 52. Since, the class under consideration was unpopulated and the dataset as already established is highly unbalanced, we made use of Synthetic Minority Over-sampling Technique (SMOTE) [18] to balance the class distribution. The results were improved, however, we ensured data integrity, therefore, we have not posted the results achieved after SMOTE.



	6	15	16	42	52	53	62	64	65	67	88	90	92	95
6	54													
15		117												
16			617											
42				408										
52					34									
53						14								
62							122							
64								30						
65									40					
67										32				
88											32			
90												51		
92													21	
95														6

FIGURE 2. Confusion Matrix for Gradient Boost Classifier.

The AUC for 3 x best models has been presented in table 4 whereas the plot for the best model is showcased in figure 3. The Gradient Boost classifier has achieved consistent superior AUC score(s) class-wise once compared with Bagging and Random Forest classifiers. For instance, the AUC for class label 53 is 0.9976 with Gradient Boost; once compared with the scores acquired with Bagging (0.9878) and Random Forest (0.9772).

TABLE 4. Class-wise AUC Values - Comparative Analysis of Best Performers

Class Label	AUC-GB	AUC-Bagging	AUC-RF
6	0.9714	0.9610	0.9228
15	0.875	0.8263	0.8313
16	0.9917	0.9618	0.9421
42	0.8067	0.7839	0.7663
52	0.7263	0.7059	0.6999
53	0.9976	0.9878	0.9772
62	0.7946	0.7601	0.7528
64	0.9288	0.9002	0.8823
65	0.992	0.9617	0.9424
67	0.765	0.7415	0.7239
88	0.9754	0.9450	0.9252
90	0.8741	0.8455	0.8303
92	0.987	0.9669	0.9371
95	0.9584	0.9291	0.9098

Quite understandably, all 3 models struggled with AUC score for Class:52 which faced the imbalance issue in comparison to other classes.

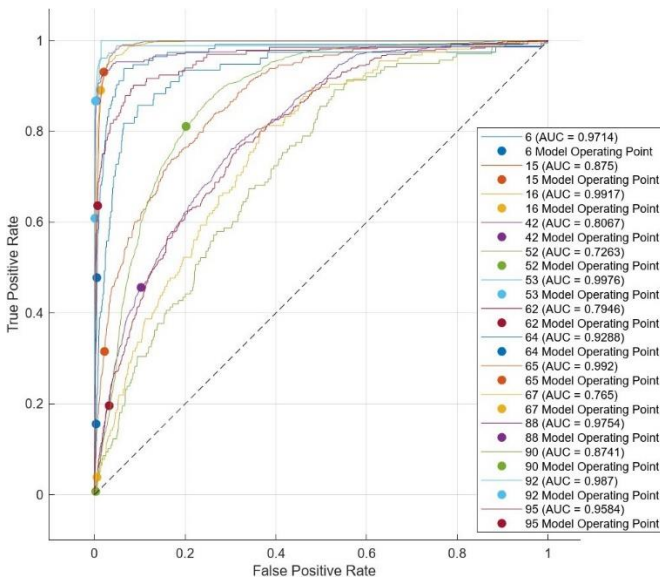


FIGURE 3. ROC Curve for AUC for Gradient Boost Classifier

In figure3, the overall performance of Gradient Boost ensemble is showcased, which is consistent with the findings as presented in table4. Overall AUC is also computed using equation4.

$$\text{Average AUC} = \frac{1}{n} \sum_{i=1}^n \text{AUC}_i \quad (4)$$

It is also established that gradient boost ensemble technique outperformed the other methods, however, considering the trade-off between the accuracy and computational efficiency, Bagging ensemble technique was most suited to perform this classification task.

VI. CONCLUSION

To conclude, an effort was made in this paper to solve a complex classification problem in the absence of domain expertise. Making use of pre-processing and feature engineering the overall classification accuracy was substantially improved. Among the evaluated models, the gradient boosting ensemble achieved the highest performance but also incurred the highest

computational cost. Conversely, the bagging ensemble offered a compelling balance of simplicity and performance. To pursuit the research continuity, improved metrics may be introduced. Additionally, integration of domain knowledge and astrophysical insights into Machine Learning models for enhanced interpretation would also yield considerable positive outcomes. Deep learning models trained on tailored astronomical datasets, may also be suited to perform state-of-the-art classification of the light-curves.

REFERENCES

- [1] "Rubin Observatory | See the Universe in Action." <https://rubinobservatory.org/>. [Accessed 15/12/2023].
- [2] S. Djorgovski, A. Drake, A. Mahabal, M. Graham, C. Donalek, R. Williams, E. Beshore, S. Larson, J. Prieto, M. Catelan, et al., "The Catalina Real-time Transient Survey (CRTS)," arXiv preprint arXiv:1102.5004, 2011.
- [3] "Zwicky Transient Facility Website." <https://www.ztf.caltech.edu/>. (Accessed on 15/12/2023).
- [4] "Time Series Classification Website." <https://www.timeseriesclassification.com/description.php?Dataset=LSST>. (Accessed on 14/12/2023).
- [5] E. E. Ishida, "Machine learning and the future of supernova cosmology," 82019.
- [6] "PLAsTiCC Astronomical Classification | Kaggle." <https://www.kaggle.com/c/PLAsTiCC-2018>. (Accessed on 10/12/2023).
- [7] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," <https://doi.org/10.1214/aos/1013203451>, vol. 29, pp. 1189–1232, 10 2001.
- [8] M. Dai, T. Allam, A. Bahmanyar, R. Biswas, A. Boucaud, L. Galbany, R. Hložek, E. E. O. Ishida, S. W. Jha, D. O. Jones, R. Kessler, M. Lochner, A. A. Mahabal, A. I. Malz, K. S. Mandel, J. R. Martínez-Galarza, J. D. McEwen, D. Muthukrishna, G. Narayan, H. Peiris, C. M. Peters, K. Ponder, and C. N. Setzer, "The photometric lsst astronomical time series classification challenge (plasticc)," (Accessed on 14/12/2023).
- [9] K. Boone, "Avocado: Photometric classification of astronomical transients with gaussian process augmentation," *The Astronomical Journal*, vol. 158, p. 257, 12 2019.
- [10] R. Hložek, A. I. Malz, K. A. Ponder, M. Dai, G. Narayan, E. E. O. Ishida, T. A. Jr, A. Bahmanyar, X. Bi, R. Biswas, K. Boone, S. Chen, N. Du, A. Erdem, L. Galbany, A. Garreta, S. W. Jha, D. O. Jones, R. Kessler, M. Lin, J. Liu, M. Lochner, A. A. Mahabal, K. S. Mandel, P. Margolis, J. R. Martínez-Galarza, J. D. McEwen, D. Muthukrishna, Y. Nakatsuka, T. Noumi, T. Oya, H. V. Peiris, C. M. Peters, J. F. Puget, C. N. Setzer, Siddhartha, S. Stefanov, T. Xie, L. Yan, K.-H. Yeh, and W. Zuo, "Results of the photometric lsst astronomical time-series classification challenge (plasticc)," *The Astrophysical Journal Supplement Series*, vol. 267, p. 25, 7 2023.
- [11] R. Hložek, K. A. Ponder, A. I. Malz, M. Dai, G. Narayan, E. E. O. Ishida, T. Allam, A. Bahmanyar, R. Biswas, L. Galbany, S. W. Jha, D. O. Jones, R. Kessler, M. Lochner, A. A. Mahabal, K. S. Mandel, J. R. Martínez-Galarza, J. D. McEwen, D. Muthukrishna, H. V. Peiris, C. M. Peters, and C. N. Setzer, "Results of the photometric lsst astronomical time-series classification challenge (plasticc)," *The Astrophysical Journal Supplement Series*, vol. 267, p. 25, 12 2020.
- [12] A. M. Khan, M. U. Akram, S. G. Khawaja, and A. S. Khan, "A machine learning technique to classify LSST observed



- astronomical objects based on photometric data," in 2019 6th Swiss Conference on Data Science (SDS), pp. 46–50, IEEE, 2019.
- [13] M. Pal, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing*, vol. 26, pp. 217–222, 1 2005.
- [14] K. Ponder, R. Hlozek, T. Allam, A. Bahmanyar, R. Biswas, K. Boone, M. Dai, L. Galbany, E. Ishida, S. Jha, D. Jones, R. Kessler, M. Lochner, A. Mahabal, A. Malz, K. Mandel, R. Martínez-Galarza, J. McEwan, D. Muthukrishna, G. Narayan, H. Peiris, C. Peters, C. Setzer, A. Boucaud, L. P. Collaboration, L. D. E. S. Collaboration, L. Transients, V. S. S. Collaboration, K. Ponder, R. Hlozek, T. Allam, A. Bahmanyar, R. Biswas, K. Boone, M. Dai, L. Galbany, E. Ishida, S. Jha, D. Jones, R. Kessler, M. Lochner, A. Mahabal, A. Malz, K. Mandel, R. Martínez-Galarza, J. McEwan, D. Muthukrishna, G. Narayan, H. Peiris, C. Peters, C. Setzer,
- [15] A. Boucaud, L. P. Collaboration, L. D. E. S. Collaboration, L. Transients, and V. S. S. Collaboration, "The photometric lsst astronomical time series classification challenge (plasticc): Final results," *AAS*, vol. 235, p. 203.15, 2020.
- [16] R. Kessler, G. Narayan, A. Avelino, E. Bachelet, R. Biswas, P. J. Brown, D. F. Chernoff, A. J. Connolly, M. Dai, S. Daniel, R. D. Stefano, M. R. Drout, L. Galbany, S. González-Gaitán, M. L. Graham, R. Hložek, E. E. O. Ishida, J. Guillochon, S. W. Jha, K. S. Mandel, D. Muthukrishna, A. O'grady, C. M. Peters, J. R. Pierel, K. A. Ponder, A. Prša, S. Rodney, and V. A. Villar, "Models and simulations for the photometric lsst astronomical time series classification challenge (plasticc)," *Publications of the Astronomical Society of the Pacific*, vol. 131, p. 094501, 2019.
- [17] A. I. Malz, R. Hložek, T. Allam, A. Bahmanyar, R. Biswas, M. Dai, L. Galbany, E. E. O. Ishida, S. W. Jha, D. O. Jones, R. Kessler, M. Lochner, A. A. Mahabal, K. S. Mandel, J. R. Martínez-Galarza, J. D. McEwen, D. Muthukrishna, G. Narayan, H. Peiris, C. M. Peters, K. Ponder, and C. N. Setzer, "The photometric lsst astronomical time-series classification challenge plasticc: Selection of a performance metric for classification probabilities balancing diverse science goals," *The Astronomical Journal*, vol. 158, p. 171, 10 2019.
- [18] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 9 1995.
- [19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 6 2002.