MALLORN classification

Final project - Machine Learning DV2638

1st Rasmus Eliasson
DIDA, Blekinge Institute of Technology
M.Sc. Eng. in AI & Machine Learning
Karlskrona, Sweden
rael23@sudent.bth.se

2nd Oskar Flodin
DIDA, Blekinge Institute of Technology
M.Sc. Eng. in AI & Machine Learning
Karlskrona, Sweden
osfl22@student.bth.se

I. Introduction

A. Time-Domain Astronomy and the LSST Challenge

Astronomical research is entering a new era in which surveys are discovering more transient objects than ever before, including supernovae and other short-lived phenomena. This shift towards time-domain astronomy introduces new challenges, as the volume of data and the frequency of observations are too large for traditional classification methods.

Classification in astronomy is usually performed using spectroscopic observations, where light intensity is measured as a function of wavelength to identify characteristic emission and absorption features. While spectroscopic classification provides detailed physical information, it is resource-intensive and does not scale to large numbers of transient events.

Therefore, it is necessary to priorities which events are most scientifically interesting. For this reason, classification based on photometric lightcurve measurements of brightness over time becomes essential for enabling efficient and scalable decision-making in modern time-domain surveys such as the Legacy Survey of Space and Time (LSST).

B. Scientific Importance of Tidal Disruption Events

A tidal disruption event (TDE) occurs when a star is ripped apart by a supermassive black hole. The infalling matter produces a bright transient emission that can be observed.

The observation of TDEs is crucial for providing insight into the characteristics of supermassive black holes and for enabling the study of extreme gravitational conditions such as accretion.

TDEs are very rare events, with only about 100 observed cases reported to date [2]. Due to this limited amount of data, drawing robust statistical conclusions and characterising variations among events is challenging.

The Legacy Survey of Space and Time (LSST) has the ability to observe a much larger number of transient events, creating new opportunities to discover additional TDEs. However, follow-up observational resources are limited, making it necessary to prioritise potential TDE candidates.

C. Problem Definition and Learning Task

The problem to be solved is how to efficiently classify whether a transient event is a TDE or not based on photometric lightcurve data.

The learning task is formulated as supervised learning using previously observed and labelled TDE events, where features derived from photometric lightcurves are used to classify whether an event corresponds to a TDE or not.

Therefore, the classification problem is well suited for supervised machine learning, as labelled data are available, making the problem tractable.

D. Scope, Constraints, and Objectives

This project focuses on how well traditional machine learning methods can predict whether a lightcurve corresponds to a TDE or not. No deep learning models are used.

A limitation of the project is the highly skewed class imbalance in the dataset.

The objective of this project is to predict TDEs using traditional machine learning methods.

II. Method

A. Dataset Description

The dataset used in this project is provided by the MALLORN Astronomical Classification Challenge on Kaggle. The data consist of photometric lightcurve observations collected at different times and across multiple filters. In addition, the dataset includes labelled transient events, where each lightcurve is classified as either a TDE or a non-TDE.

B. Lightcurve Representation and Feature Construction

From the data, raw photometric lightcurve samples form irregular time series, while traditional machine learning models require a fixed number of features for representation.

For each lightcurve, per-filter statistical features were computed by aggregating observations over time. These features include the number of observations, time span, summary statistics of flux and flux error, mean signal-to-noise ratio, and peak time/peak flux. The result is a fixed-length feature vector per observed event (one row per `object_id`).

C. Data Pre-processing

Missing values were handled by removing samples with incomplete feature vectors from the training data. The test data were left unchanged to preserve the original data distribution and avoid information leakage during model evaluation. Standard scaling was applied where required, using parameters computed from the training data only.

D. Learning Algorithms

The following learning algorithms were evaluated as baseline models: Neural network, Gaussian Naive Bayes, Logistic Regression, Quadratic Discriminant Analysis, Support Vector Machines with polynomial and sigmoid kernels, Random Forest, and Extreme Gradient Boosting.

In addition, a stacked ensemble model was constructed by combining the predictions of all base models using a Decision Tree classifier as a meta-model. This ensemble approach allows the integration of models with different inductive biases and supports non-linear decision boundaries in the feature space.

E. Parameter Configuration

Baseline model parameters were set to their default configurations, except for the Extreme Gradient Boosting model, where class weighting was applied to account for TDE and non-TDE imbalance, and the Decision Tree classifier used as the meta-model in the stacked ensemble, which was configured with a maximum depth of three. The stacked ensemble was evaluated using five-fold cross-validation.

To ensure reproducibility, the random state was fixed to 42 for all models. Class imbalance was handled within each fold using stratified k-fold cross-validation.

F. Evaluation Procedure and Metrics

Model performance was evaluated using confusion matrices and standard classification metrics derived from classification reports. In addition, the area under the receiver operating characteristic curve (ROC-AUC) was used to assess discriminative performance, particularly in the presence of class imbalance.

To compare the performance of multiple models across cross-validation folds, statistical significance testing was conducted using the Friedman test followed by a post-hoc Nemenyi test. These tests were used to validate whether observed performance differences between models were statistically significant.

III. Results and Analysis

A. Exploratory Feature Analysis

Exploratory data analysis was performed to characterize the statistical properties of the engineered features in both the training and test sets. Feature distributions were examined using histograms and boxplots. These inspections revealed that many features exhibit pronounced skewness, heavy-tailed behavior, and contain outliers. Such characteristics are consistent with the sparse and heterogeneous nature of astronomical time-domain observations, as well as the strong class imbalance in the dataset, with tidal disruption events (TDEs) comprising approximately 5% of the samples.

To assess the discriminative power of individual features, class-conditional distributions were examined by comparing feature values for TDE and non-TDE objects. In addition, scatter plots relating selected features to reference features were analyzed. These visualizations revealed substantial overlap between the two classes across most features. Thus indicating that no single feature provides strong separability between TDEs and non-TDEs.

B. Feature Correlation and Dimensionality Reduction

An interactive feature–feature correlation analysis was performed to identify redundancy and shared information among the features. Several features were found to exhibit moderate to strong correlations, reflecting the aggregation of related statistics derived from the same underlying lightcurve measurements. This suggest that an ensemble approach would be suitable.

To further investigate the structure of the engineered feature space, two-dimensional projections were generated using Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). PCA was used to summarize the global variance structure of the data, while t-SNE was employed to explore local neighborhood relationships.

The PCA projection showed that a large fraction of the variance is captured by the first principal component (x-axis). However no clear linear separation can be observed between TDE and non-TDE. It shows that the minority TDE class remains embedded within the dominant non-TDE population. Thus indicating that directions of maximum variance are not aligned with class-discriminative boundaries.

The t-SNE projection revealed localized clustering patterns, but TDE events did not form a distinct or isolated cluster. Instead, TDEs were dispersed across multiple regions of the feature space and overlapped substantially with non-TDE events. This behavior likely reflects similarities between TDEs and other astrophysical transient phenomena present in the dataset.

C. Model Selection

Altogether, the exploratory analysis and dimensionality reduction indicates that TDE classification is characterized by complex, non-linear structure and significant class overlap. Based on these observations, a stacked ensemble modeling approach was selected.

By combining base models with different inductive biases, the ensemble is able to capture complementary decision patterns and learn more expressive, non-linear decision boundaries. A meta-model was used to combine the predictions of individual base learners, allowing it to adaptively weight their contributions and improve robustness and generalization performance.

D. Feature Importance Analysis

Feature importance was estimated using a Random Forest classifier trained on the full feature set. Relative importance scores were computed based on each feature's contribution to information gain. The 15 least important features were identified as candidates for removal in order to reduce noise and redundancy among the 62 engineered features.

Random Forest models are relatively robust to correlated inputs, making them suitable for this type of heuristic analysis. While the resulting importance scores should not be interpreted as definitive measures of feature relevance, they provide useful guidance for identifying features that contribute minimally to predictive performance.

Applying this feature reduction led to an improvement in validation F1 score, from approximately 0.22 to 0.30, indicating that removing low-importance features helped reduce noise and improve classification performance.

E. Classification Performance

Comparing the full stacked ensemble with the reduced-feature variant reveals a clear improvement in classification performance after removing noisy or low-importance features. In particular, the reduced model achieves a higher F1 score, primarily due to increased precision while maintaining comparable recall. This indicates that feature pruning helps reduce false positives without severely compromising sensitivity to tidal disruption events (TDEs).

Despite this improvement, the overall F1 performance remains modest, with values ranging from approximately 0.22 to 0.30. This limitation is likely driven by strong feature overlap between TDEs and other types of astronomical transients. Such transients can exhibit similar variability statistics and rise–decay behavior, which increases classification ambiguity and can lead to a higher number of false positives.

F. Evaluation of base models and stacking effectiveness

The ROC–AUC curve indicates that the final stacked ensemble ranks positive (TDE) samples higher than non-TDE samples substantially more often than random chance. An AUC of approximately 0.84 suggests that the model captures meaningful discriminative information, even though threshold-dependent metrics such as F1 remain limited by class overlap and imbalance.

Statistical hypothesis testing further supports the suitability of the stacking approach. The Friedman test reveals statistically significant differences in performance among the base models ($p < 0.002$), leading to rejection of the null hypothesis that all models perform equivalently.

The subsequent Nemenyi post-hoc analysis identifies several base-model pairs with significantly different average ranks, confirming heterogeneity in model behavior. Such heterogeneity is a necessary condition for effective stacking, as it allows the ensemble to exploit complementary strengths rather than redundant predictions.

To further assess model complementarity, correlations between out-of-fold (OOF) predicted probabilities were analyzed. OOF predictions are generated on validation folds not seen during training, providing an unbiased estimate of each model's behavior.

The resulting correlation matrix shows generally low to moderate dependence between base models. Since these OOF probabilities serve as inputs to the meta-model, reduced correlation indicates that the base models capture complementary aspects of the data rather than producing highly redundant predictions.

Although some model pairs exhibit moderate correlation, indicating partial redundancy. Sufficient diversity still remains for stacking to be beneficial. Altogether, the statistical tests and OOF correlation analysis demonstrate that the base models exhibit complementary error patterns. Combined with the observed ROC–AUC performance, this provides strong evidence that the stacked ensemble operates as intended, even though absolute performance is constrained by feature overlap in the data.

IV. Conclusions

From this submission distribution, we see that the finished model has classified 1506 TDEs out of 7135. Thus, approximately 21% of the events are classified as TDEs. This is significantly higher than the fraction in the training data, which is only about 5%.

It has also become apparent that the Kaggle dataset is not representative of true astrophysical prevalence. The training data contains 147 labeled TDEs, while the total number of confirmed tidal disruption events globally is only around 134, according to the comprehensive TDE catalog presented in Repeating Flares, X-ray Outbursts and Delayed Infrared Emission [3].

This suggests that the dataset may include simulated, augmented, or highly similar events rather than fully independent observations. While the out-of-fold (OOF) evaluation strategy prevents direct data leakage, reduced effective independence between samples could negatively impact generalization performance. This may help explain the drop in performance observed on the Kaggle validation leaderboard.

As observed in the PCA and t-SNE projections, TDEs are not clearly separable from other transient classes in the engineered feature space. Several non-TDE transients, such as supernovae, can exhibit light-curve properties similar to TDEs, increasing class overlap. Framing the task as a binary classification problem may therefore limit performance. A multi-class formulation that explicitly models additional transient types could allow the model to learn more informative decision boundaries and potentially improve classification performance.

A. Further research

- **Improved feature engineering and pre-processing** Further analysis of feature extraction and pre-processing could improve performance. This includes refining summary statistics, handling missing values more systematically, and evaluating the impact of correlated or redundant features. Additionally, ensuring that any simulated or augmented samples are handled consistently across training and validation folds could help maintain effective sample independence.

- **Deep learning approaches**
  Deep learning models, particularly sequence-based architectures, could potentially improve performance by learning temporal patterns directly from light curves rather than relying on static summary features.
- **Optimized decision thresholds**
  Within this project, a default decision threshold was used. For highly imbalanced problems, selecting an application-specific threshold could provide a better balance between recall and precision, depending on the desired operating point.
- **Multi-class classification**
  Introducing additional transient classes could reduce model confusion by allowing more informative decision boundaries, potentially improving discrimination between TDE and non-TDE events.
- **Include redshift error in training**
  The training data only provides redshift (z) and does not include a redshift uncertainty term (z_err), while the Kaggle test set contains both. To handle this mismatch, redshift uncertainty is approximated at inference time by evaluating the model at perturbed values (z, z + z_err, and z - z_err) and combining the predictions via soft voting. If redshift uncertainty were available in the training data, it could instead be incorporated directly as a feature or modeled probabilistically during training.

V. Contribution

All team members contributed jointly to the different parts of the project, including code implementation, experimentation, evaluation, and report writing. The responsibilities were shared equally throughout the project. Some code from earlier assignments was reused for pre-processing and machine learning models.

References

[1] P. Flach, *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[2] Kaggle, *MALLORN Astronomical Classification Challenge*. [Online]. Available: https://www.kaggle.com/competitions/mallorn-astronomical-classification-challenge/overview. Accessed: Dec. 2025.

[3] A. AUTHOR et al., *Repeating Flares, X-ray Outbursts and Delayed Infrared Emission: A Comprehensive Compilation of Optical Tidal Disruption Events*, arXiv:2506.05476, 2025. [Online]. Available: https://arxiv.org/abs/2506.05476. Accessed: Dec. 2025.