universe

MDPI

Article

Periodic Variable Star Classification in Imbalanced Data: A Two-Stage Decoupled Approach with Multimodal Contrastive Learning

Qingyu Lu [1,†], Feng Zhang [1,†], Zhiqiang Zou [1,2,3,*], Tianyu Su [1], Xiaohang Zhang [1], A-Li Luo [3,4,5] and Xiao Kong [4]

[1] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; 1224045707@njupt.edu.cn (Q.L.); 1224045740@njupt.edu.cn (T.S.); 1025040820@njupt.edu.cn (X.Z.);
[2] Jiangsu Key Laboratory of Big Data Security and Intelligent Processing, Nanjing 210023, China
[3] University of Chinese Academy of Sciences, Nanjing 211135, China
[4] CAS Key Laboratory of Optical Astronomy, National Astronomical Observatories, Beijing 100101, China; kongx@nao.cas.cn
[5] School of Astronomy and Space Science, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: zouzq@njupt.edu.cn
† These authors contributed equally to this work.

Abstract

Large photometric surveys require classification methods that retain sensitivity to sparsely represented periodic variable-star classes. We present a two-stage decoupled multimodal network (TDMN) for imbalanced periodic-variable-star classification using phase-folded light curves, three-channel morphology images, and global features. Stage 1 organizes the three modalities using proxy contrastive objectives and real-observation class prototypes. Stage 2 retains low-level morphology extractors while adapting higher-level representations and combining class-balanced, logit-adjusted, and cosine-focal experts. Across three training runs on fixed survey-specific splits, TDMN achieves mean macro F1-scores of 0.85 on CRTS, 0.83 on the OGLE-IV 15-subclass task, and 0.87 on the OGLE-IV 6-superclass task. A retrained UPSILoN-feature random forest evaluated on the same test objects yields corresponding mean scores of 0.66, 0.78, and 0.88. Thus, TDMN has higher mean macro F1 on CRTS and the OGLE-IV subclass task, but not on the superclass task. Rotational variables remain difficult, and increased Type II Cepheid recall with TDMN is accompanied by lower precision. The class-wise recovery and purity estimates can support the construction of known periodic-variable samples and the prioritization of candidates for follow-up observations. The study remains a catalogue-based, closed-set classification analysis and assumes an externally supplied period for TDMN; period discovery, unknown-class recognition, and nonperiodic transients lie outside its scope.

Keywords: variable star classification; time-domain astronomy; contrastive learning; imbalanced learning; neural networks

Check for updates

Academic Editor: Lorenzo Iorio

Received: 31 July 2026
Revised: 6 September 2026
Accepted: 11 September 2026
Published: 19 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Variable-star classification is important in time-domain astronomy because the morphology and timescales of variability help distinguish stellar pulsation, eclipses, and rotational modulation [1]. Reliable samples of particular pulsating classes can support distance-scale and stellar-population studies. Large archives from the Wide-field Infrared Survey Explorer (WISE) [2], the Kepler space telescope (Kepler) [3], the Catalina Real-Time

Transient Survey (CRTS) [4], and the Optical Gravitational Lensing Experiment (OGLE) [5] motivate automated candidate classification. Here, we address periodic variable stars with an externally supplied period, not disruptive or explosive transients or the initial detection of variability.

To process these large data volumes, automated variable-star classification has evolved from traditional machine-learning methods that rely on handcrafted feature engineering to deep-learning methods capable of automatic feature extraction. Early studies generally used statistical methods or tools such as Feature Analysis for Time Series (FATS) [6] to extract time- and frequency-domain features from light curves, which were subsequently classified using support vector machines (SVMs) or random forests [7]. To address class imbalance, Zhang et al. [8] investigated an imbalanced-learning framework based on Self-Paced Ensemble (SPE) learning to improve classification performance. However, handcrafted feature extraction generally depends on predefined feature sets and may be limited in low-signal-to-noise or irregularly sampled settings. In recent years, deep-learning methods have increasingly been applied to light-curve classification because of their nonlinear representation capabilities. Aguirre et al. [9] demonstrated that convolutional neural networks (CNNs) can extract discriminative information from light curves with less reliance on handcrafted feature engineering. Szklenar et al. [10] and Subebekova et al. [11] employed CNNs to process two-dimensional light-curve images folded from one-dimensional sequences. Recurrent neural networks (RNNs) and their long short-term memory (LSTM) variants have also been used to model long-range temporal dependencies and dynamic temporal behavior in light curves [12]. In addition, applications of deep learning and large-language-model-based methods to stellar light-curve classification have attracted attention [13].

The long-tailed class distribution [14,15] denotes a sample in which a few classes contain many objects while numerous other classes contain comparatively few; it does not describe the temporal duration of a light curve. Burhanudin et al. [16] studied imbalanced light-curve classification with RNNs, and Hosenie et al. [17] examined augmentation including SMOTE, random sampling from light-curve errors (RASLE), and Gaussian-process fitting. Zhang et al. [8] applied Self-Paced Ensemble learning to seven predefined scalar features—six FATS-derived features together with the catalogue period—using self-paced under-sampling and random-forest base classifiers; this approach does not learn a deep representation that can be separated from its classifier. Kang et al. [18] combined recurrent and convolutional inputs with Gaussian-process augmentation and a bagging-like ensemble, while the representation extractors and Softmax classifier within each base network were optimized jointly. Self-regulated CNNs offer another related deep-learning approach [19]. Thus, these variable-star studies address imbalance primarily through sampling, augmentation, and ensemble prediction rather than through an explicit separation of representation learning from classifier calibration.

The decoupled long-tailed recognition study of Kang et al. [15] showed on visual benchmarks that natural instance-balanced sampling can yield generalizable representations and that subsequent classifier rebalancing can substantially improve tail-class recognition. This finding motivates treating representation organization and classification-boundary bias as related but separable problems. The methodological contribution of TDMN lies in the task-specific integration of these components into a decoupled periodic-variable-star pipeline: Stage 1 uses real-observation prototypes and multimodal proxy contrastive objectives to organize the representation space, whereas Stage 2 retains low-level morphology extractors and recalibrates the decision boundary through sample-adaptive fusion of heterogeneous classifiers.

Accordingly, this study proposes a Two-stage Decoupled Multimodal Network (TDMN) for imbalanced periodic-variable-star classification. Inspired by the separation of representation learning and classifier adjustment in long-tailed recognition [15], representation constraints and classifier recalibration are implemented in separate stages. Stage 1 introduces multimodal proxy contrastive learning [20–22], in which class prototypes constrain temporal, image, and global statistical features so that samples of the same class remain relatively compact in feature space. Stage 2 freezes low-level convolutional feature-extraction layers while retaining the Stage 1 prototype-anchoring constraint and the adaptability of high-level representations and classification heads. To address class-prior bias and difficult samples, class-balanced, logit-adjusted, and cosine-focal classification heads are employed in parallel, and a fusion-weight network adaptively combines their outputs for each sample to recalibrate the classification decision boundary in stages. The principal methodological contributions and experimental findings of this study are as follows:

(1) A two-stage multimodal classification framework is proposed to address two coupled problems in long-tailed periodic-variable-star classification: the tendency of minority-class representations to overlap with majority classes and the bias of classification boundaries toward majority-class priors. The framework decouples representation learning and classifier calibration in stages.
(2) A real-sample-prototype-constrained multimodal proxy contrastive-learning mechanism is developed for imbalanced periodic-variable-star classification. Class prototypes are maintained using real observations, thereby providing clearer class references for minority classes, reducing interference from augmented samples in class-center estimation, and establishing a stable representation basis for subsequent classification-boundary calibration.
(3) Systematic experiments on CRTS and OGLE-IV, including an object-matched comparison with a retrained random forest using the 16 official UPSILoN features, evaluate TDMN under distinct and strongly imbalanced class taxonomies. Across three training runs, TDMN achieves mean macro F1-scores of 0.85 on CRTS, 0.83 on the OGLE-IV subclass task, and 0.87 on the superclass task. The class-wise precision and recall quantify the recovery and purity of known periodic-variable samples, supporting catalogue construction, follow-up prioritization, and subsequent population studies while also revealing unresolved Rot/Ecl/EA confusion.
(4) To facilitate reproducibility, data-access instructions and the model-training, data-processing, evaluation, and visualization code used in this study are publicly available at https://github.com/lqy0718/TDMN (accessed on 10 September 2026).

2. Data and Preprocessing

2.1. Datasets

Two catalogue-derived datasets were used to study classification under different class compositions and observational conditions. CRTS provides an established 11-class benchmark for comparison with earlier imbalanced-learning studies, whereas OGLE-IV supports a separate hierarchical task toward the Magellanic Clouds. The reported results are survey-specific, with training and evaluation performed independently within each survey.

For class counts N_c in a specified sample, we define the imbalance ratio as $\mathrm{IR} = \max_c N_c / \min_c N_c$. There is no universal threshold at which imbalance necessarily degrades a classifier, because total sample size, class overlap, and separability also affect learning [14]. Nevertheless, computer-science studies have used operational severity levels. For example, López et al. [23] treated IR > 9 as a high degree of imbalance in binary classification, while long-tailed visual-recognition benchmarks commonly evaluate

imbalance factors such as 10, 50, and 100 [24]. These are empirical conventions rather than task-independent laws. The ratios below refer to the source benchmark counts in Tables 1 and 2, not to augmented training sets. They are 18,803/147=127.91 for CRTS and 65,232/66=988.36 for OGLE-IV; both exceed 100, the largest benchmark setting cited here. For these multiclass data, however, the maximum-to-minimum ratio summarizes only the two extreme classes and does not capture the full class-frequency shape. Filtering and augmentation can change these ratios and must not be conflated with the catalogue counts.

Table 1. CRTS benchmark counts before filtering for the three-run experiment. RRab, RRc, and RRd denote fundamental-mode, first-overtone, and double-mode RR Lyrae stars, respectively; Blazhko denotes modulated RR Lyrae variability. EA denotes Algol-type detached eclipsing binaries, while Ecl is the benchmark's contact/semi-detached group. LPV denotes long-period variables.

Variable-Star Class	Number
Ecl (contact and semi-detached binary)	18,803
EA (detached binary)	4509
RRab	4325
RRc	3752
Rot (rotational variable)	3636
LPV (long-period variable)	1286
RRd	502
Blazhko	171
ACEP (anomalous Cepheid)	153
Cep-II (Type II Cepheid)	153
δ Scuti	147

Table 2. OGLE-IV sample counts toward the LMC and SMC. ACEP and CEP denote anomalous and classical Cepheids; DSCT denotes δ Scuti stars; ECL denotes eclipsing binaries; RRLYR denotes RR Lyrae stars; T2CEP denotes Type II Cepheids. F, 1O, and 1O2O identify the fundamental, first-overtone, and simultaneous first/second-overtone modes. C/NC denote contact/non-contact systems. BLHER, WVIR, and RVTAU denote BL Herculis, W Virginis, and RV Tauri subclasses. LPVs are not included in this selected OGLE sample.

Variable-Star Class	Number
ECL_NC	65,232
RRLYR_RRAB	34,411
DSCT_SINGLEMODE	17,350
RRLYR_RRC	11,084
CEP_F	5230
CEP_1O	3580
RRLYR_RRD	2811
ECL_C	2737
DSCT_MULTIMODE	716
CEP_1O2O	560
ACEP_F	181
T2CEP_WVIR	123
T2CEP_BLHER	119
ACEP_1O	89
T2CEP_RVTAU	66

CRTS. The selected benchmark contains 37,437 objects in 11 classes. An initial observation-count screening stage uses 36,340 objects, while the three-run comparison uses a separately generated split described in Section 4.4. Both counts refer to different processing stages and are kept distinct. The 11-class label space follows Hosenie et al. [17], who selected these classes from CSDR2. The source catalogue of Drake et al. [4] additionally

lists a separate LMC classical-Cepheid label (LMC-Cep, 10 objects) and a miscellaneous category (298 objects), neither of which belongs to this inherited benchmark; their absence is therefore a property of the benchmark label space rather than a removal by TDMN. In the source catalogue, initial classification used the period together with light-curve, statistical, and color information, followed by catalogue-level verification [4]. The labels used here are inherited rather than established anew by TDMN.

OGLE-IV. Periodic variable-star samples toward the Large Magellanic Cloud (LMC) and Small Magellanic Cloud (SMC) were selected from OGLE-IV. The data were obtained from the official OGLE Collection of Variable Stars download page (https://www.astrouw.edu.pl/ogle/ogle4/OCVS/, accessed on 10 September 2026). The collection reports class-specific catalogue searches based on periodicity and characteristic light-curve morphology; we inherit those published labels and do not repeat their astrophysical adjudication [5]. Six major classes—ACEP, CEP, DSCT, ECL, RRLYR, and T2CEP—and their corresponding 15 subclasses were retained. Compared with CRTS, OGLE provides denser sampling and a different observational depth. This dataset also exhibits a high class-imbalance ratio of approximately 988:1, as shown in Table 2.

2.2. Data Preprocessing

The task assumes that period estimation precedes TDMN. A period is not itself a class label, but its estimation and possible aliases are upstream sources of uncertainty. The scope is classification of preidentified periodic candidates using supplied catalogue periods. Here, "long-term" refers to the survey light-curve baseline. For an object, this baseline is $T_{\mathrm{obs}} = \max(t_i) - \min(t_i)$, and T_{obs}/P is the number of covered cycles; no universal minimum is inferred from the present datasets.

Raw survey light curves are affected by photometric uncertainties, irregular sampling, and outlying observations. The objective of preprocessing is not to oversmooth the observations, but to suppress incidental noise while preserving the period scale, light-curve morphology, and uncertainty information required to distinguish variable-star classes.

First, the known period P is used to map the observation times onto an intra-period time coordinate $[0, P)$, after which the observations are sorted in ascending order of this coordinate. Compared with a representation based only on normalized phase in $[0, 1)$, this procedure preserves the period scale associated with the light-curve morphology and provides additional information for distinguishing variable-star classes with similar morphologies but different period ranges.

The SuperSmoother algorithm [25] is then used to estimate the light-curve baseline. For ordinary classes, an observation is regarded as low quality when its magnitude deviates from the smoothed baseline by more than three times its photometric uncertainty or when its photometric uncertainty is unusually large. RRd and Blazhko variables exhibit genuine multiperiodic or amplitude-modulated behavior; therefore, the preceding outlier-rejection procedure is not applied to these classes, thereby avoiding the misidentification of beating patterns or envelope variations as noise.

For OGLE-IV, the same protective treatment is applied to the catalogue subclasses CEP_1O2O, DSCT_MULTIMODE, and RRLYR_RRD. These class-specific exceptions describe the preprocessing of the labeled benchmark; applying such a rule to an unknown object would require a class-independent alternative.

For the densely sampled OGLE-IV light curves, down-sampling based on 800 phase-grid points is adopted to control sequence length while preserving the dominant phase morphology. After the preceding processing, the raw and smoothed magnitudes are separately mean-aligned for each light curve. Global features, including the period, amplitude, brightness statistics, and Fourier parameters, are then extracted. The clipping, transforma-

tion, and standardization parameters for the global features are determined exclusively from the training set and are subsequently applied to the validation and test sets.

2.3. Dynamic Training-Set Augmentation

The central objective of long-tailed classification is not to expand every class mechanically to an identical size, but to increase the effective training frequency of tail classes while preventing synthetic samples from dominating class representations. Accordingly, a dynamic augmentation strategy based on class size is applied only within the training set; neither repeated samples nor generated samples are included in the validation or test set.

Specifically, the pre-split class count sets the target training count: the multiplier is four below 500 objects, three from 500 to 999, two from 1000 to 2999, and one from 3000 upward. The required supplementation is the target minus the number of real training objects after splitting. Consequently, a multiplier of one does not imply zero augmentation, because some source objects are assigned to validation and test sets. No repeated or generated augmentation is applied to the validation or test set. The three-run CRTS comparison uses a different, median-target rule and is reported separately in Section 4.4.

For each class requiring supplementation, approximately 20% of the additional samples are drawn with replacement from real training light curves, while the remaining samples are generated using Gaussian processes (GPs) and random sampling based on photometric errors (RASLE). For ordinary classes, GP and RASLE are selected with probabilities of 30% and 70%, respectively. GP generation provides smooth light-curve trends in phase space, whereas RASLE perturbs the original light curve according to its photometric uncertainties. Together, the two methods increase the morphological coverage of tail-class training samples without altering the distribution of real observations in the validation and test sets.

Duplication changes an object's training frequency without adding a new morphology. GP/RASLE augmentation instead imposes assumptions about smoothness, sampling, or photometric noise and provides alternative realizations consistent with those assumptions; it does not create independent astronomical observations. An inappropriate generator can distort amplitude, phase structure, or class overlap. The dynamic-versus-uniform experiment below compares supplementation targets; determining the specific contribution of generative augmentation requires a matched comparison with duplication at equal sample counts.

2.4. Three-Channel Phase-Folded Light-Curve Images Incorporating Photometric Uncertainties

To supplement the representation of local morphology and observation quality in one-dimensional sequences, the cleaned phase-folded light curves are rendered as three-channel images. On a black background, the red channel displays the raw observational points, the blue channel displays the smoothed light-curve baseline produced by SuperSmoother, and the green channel displays the corresponding photometric error bars centered on the smoothed baseline. Compared with a conventional grayscale image that retains only phase–magnitude scatter, this representation simultaneously presents the observational dispersion, overall variation trend, and associated uncertainty in the same phase coordinate. The image branch therefore receives both local deviations and their photometric uncertainties rather than relying solely on the locations of the scatter points.

Figure 1 compares conventional phase-folded light curves with the three quantities encoded by the image representation. For each variable-star class, one real sample was selected from the CRTS data and processed by the data-processing scripts used in this study, including data loading, phase folding, SuperSmoother fitting, and outlier rejection. The same eleven CRTS objects are used in both panels, which repeat the phase-folded data

over two cycles, $[0, 2)$. For legibility at normal page scale, the right-hand panels redraw the three encoded quantities on a white background at the same height as the light curves, with larger markers and wider lines. This publication view changes neither the values nor the black-background image tensors supplied to the network.

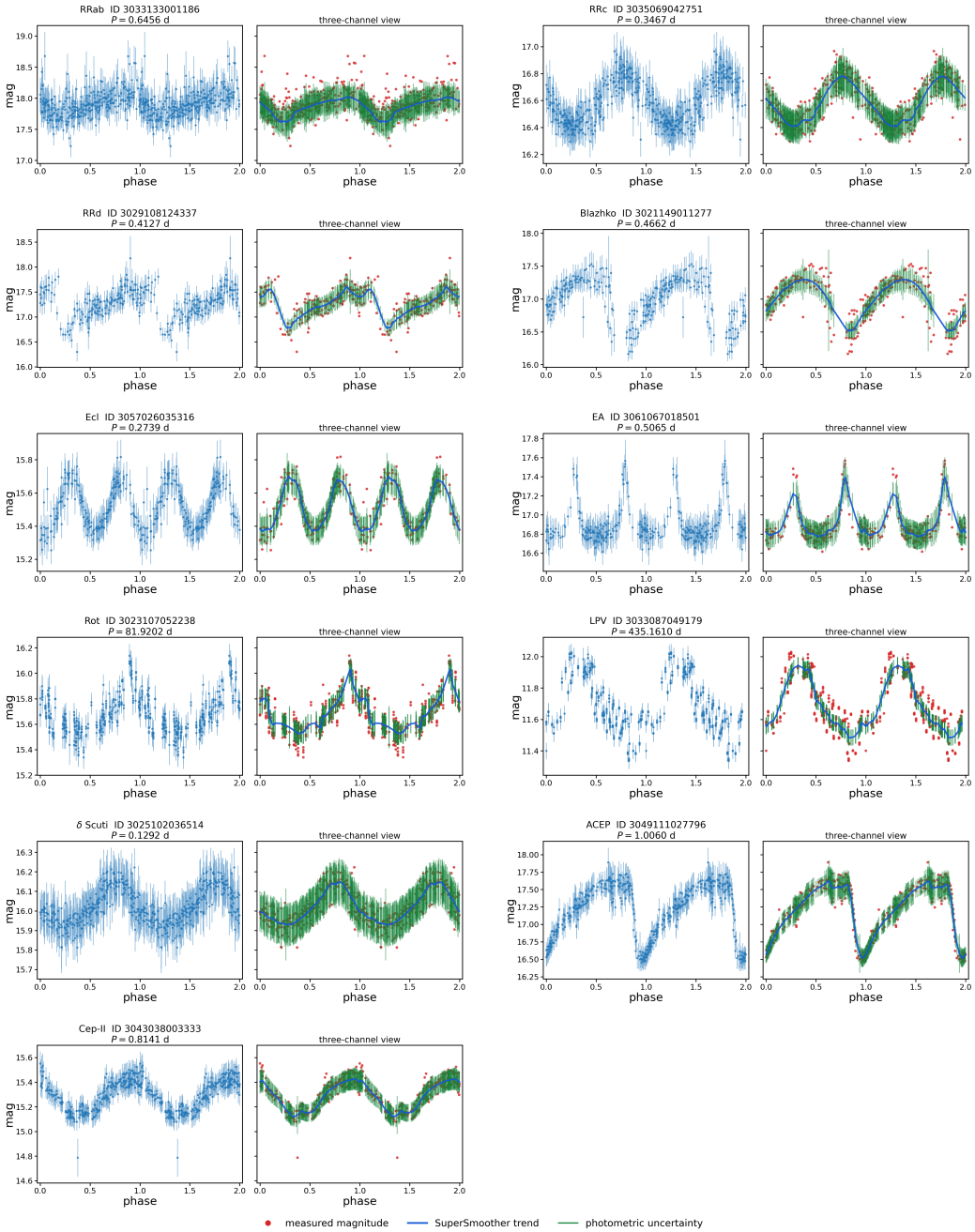


Figure 1. Phase-folded CRTS light curves and a white-background visualization of their three-channel representation. Red denotes measured magnitudes, blue the SuperSmoother trend, and green photometric uncertainties. The panels have equal height and use enlarged markers and lines for legibility; only the display is changed, while the network input encoding remains unchanged.

3. Two-Stage Decoupled Classification Method

3.1. Long-Tailed Classification Problem

The long-tailed distribution of periodic variable stars gives rise to two interrelated problems. On the one hand, majority classes provide denser optimization signals, causing the light-curve morphologies and statistical features of minority classes to become

entangled with those of morphologically similar majority classes. On the other hand, differences in class frequency bias the classifier toward majority-class priors; even when minority classes have acquired somewhat separable representations, their decision regions may still be compressed. The former is a representation-quality problem, whereas the latter is a classification-boundary problem. When both are addressed under the same end-to-end cross-entropy objective, the frequency bias of the classifier can interfere with the organization of feature space, making the two problems difficult to control separately.

Accordingly, a two-stage decoupled multimodal network is constructed for long-tailed periodic-variable-star classification. The model takes as input a one-dimensional light-curve sequence x_{seq}, a three-channel phase-folded image x_{img}, and global features x_{feat} comprising the period, amplitude, brightness statistics, and Fourier parameters.

Stage 1 centers on proxy contrastive learning constrained by real-sample prototypes and does not introduce a cross-entropy classification loss. By pulling multimodal samples toward the corresponding class prototypes and pushing samples from different classes away from nonmatching prototypes, the model first establishes a representation space characterized by intraclass compactness and interclass separation. The prototypes are updated using only real observations to reduce the influence of training-set augmentation samples on the estimated centers of minority classes. Stage 2 loads the Stage 1 parameters, freezes low-level morphology-extraction layers such as convolutional and pooling layers, and introduces three heterogeneous classification experts—class-balanced, logit-adjusted, and cosine-focal—while retaining the adaptability of high-level representations. Their predictions are combined through sample-adaptive multi-head fusion. Consequently, Stage 1 primarily alleviates the entanglement of minority-class representations, whereas Stage 2 primarily corrects decision-boundary bias induced by majority-class priors.

The overall two-stage classification framework is illustrated in Figure 2.

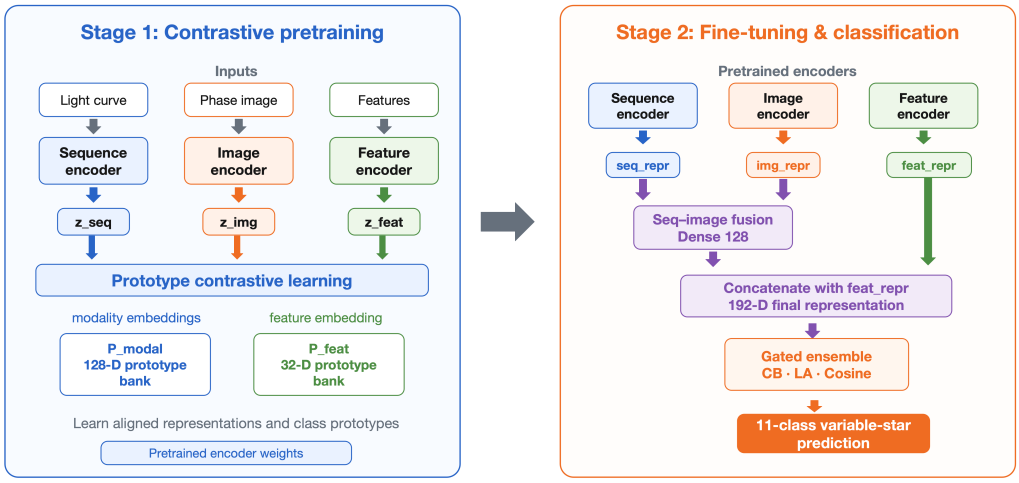


Figure 2. Two-stage multimodal classification framework of TDMN. Externally supplied period estimation and phase folding precede the displayed classification stages.

3.2. Stage 1: Multimodal Representation Learning Constrained by Real-Sample Prototypes

3.2.1. Multimodal Feature Encoding and Fusion

To capture the heterogeneous physical information contained in variable-star light curves, a three-branch parallel feature-extraction network is constructed from a temporal encoder, an image encoder, and a global-feature encoder. The three encoders extract information from the one-dimensional light-curve sequence, two-dimensional phase-folded image, and global statistical features, respectively, producing a temporal representation H_{seq}, an image representation H_{img}, and a global statistical representation H_{feat}. After the

corresponding projection or fusion operations, these representations serve as shared inputs to proxy contrastive learning in Stage 1 and classifier recalibration in Stage 2.

The detailed architecture of the three input branches and their fusion is shown in Figure 3.

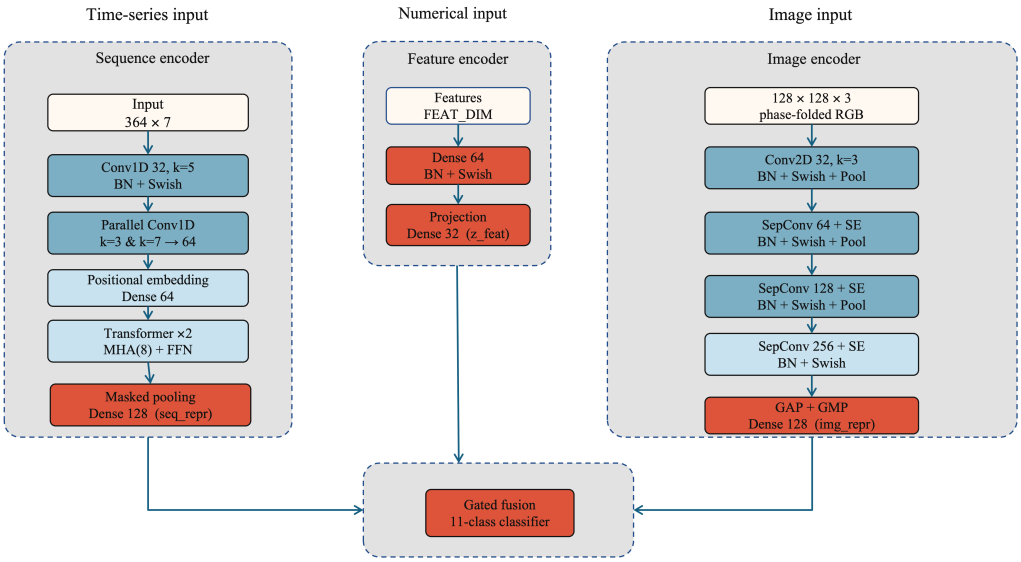


Figure 3. TDMN neural network model. White boxes denote inputs, blue boxes denote feature-extraction operations, and orange boxes denote dense, projection, aggregation, fusion, and output operations. Dashed outlines group the encoder branches. The colors are schematic and do not encode quantitative values.

The temporal branch preserves local variations and period-scale information in the light curve. Its inputs comprise, in sequence, the raw magnitudes, smoothed magnitudes, photometric uncertainties, sine and cosine phase encodings normalized over the valid phase range, the unnormalized phase-time coordinate, and a valid-observation mask. The two types of phase information describe periodic continuity and preserve the actual period scale, respectively, thereby avoiding the loss of period length as discriminative observational information when only normalized phase in $[0, 1)$ is used. The model first takes the raw magnitudes, smoothed baseline, and photometric uncertainties as local convolutional inputs and extracts morphological variations on different temporal scales using parallel one-dimensional convolutions with kernel sizes of 3 and 7. The mapped phase positional encoding is then added to the convolutional features, which are passed through a two-layer masked multi-head self-attention network. The attention mask prevents padded positions from participating in feature interactions among valid observations. Finally, mask-constrained average pooling and max pooling aggregate the temporal information to produce the 128-dimensional temporal representation H_{seq}. This branch enables the model to attend simultaneously to local fluctuations, overall morphology, and period scale, providing additional discriminative information for classes with similar morphologies but different typical period ranges.

The two-dimensional image branch takes the three-channel phase-folded image as input: the red channel represents raw observational points, the blue channel represents the smoothed light-curve baseline, and the green channel represents photometric error bars. Thus, the image is not a simple rendering of a one-dimensional curve; instead, it simultaneously displays, within the same phase coordinate, the extent to which actual observations deviate from the smoothed trend and the photometric reliability of those deviations. A lightweight convolutional network first extracts local image patterns, after which depth-wise separable convolutions reduce the number of parameters. Squeeze-and-Excitation (SE)

modules [26] are introduced at each convolutional stage to adaptively recalibrate channel responses. By jointly encoding photometric uncertainties and light-curve morphology, this branch allows the network to reference the corresponding observational uncertainties when interpreting local scatter or morphological details and produces the 128-dimensional image representation H_{img}.

The global-feature branch takes the period, amplitude, brightness statistics, Fourier parameters, and related quantities as input and maps them through fully connected layers to the 64-dimensional statistical representation H_{feat}. CRTS uses 17-dimensional global features, whereas OGLE-IV uses 7-dimensional global features. This branch supplements global periodic and statistical information that is difficult to represent robustly in local morphology images and finite-length sequences.

3.2.2. Multimodal Proxy Contrastive Constraint

Under long-tailed training, minority-class samples rarely occur together within the same batch, and pairwise sample comparisons alone cannot readily establish stable class references for them. Accordingly, separate sets of class prototypes, P_{modal} and P_{feat}, are maintained for the temporal–image modalities and the global statistical modality, respectively. The temporal and image representations are mapped by their respective projection heads into 128-dimensional spaces and share P_{modal}, whereas the global statistical representation is mapped by its projection head into a 32-dimensional space and uses P_{feat}. Let C denote the number of classes, y the ground-truth label of a sample, z_m the projected representation from modality m, and P_c the prototype of class c in the corresponding prototype set. After L_2 normalization of the projected vectors and class prototypes, the proxy contrastive loss for modality m is defined in Equation (1) as

$$L_m = -\log \frac{\exp(z_m \cdot P_y/\tau)}{\sum_{c=1}^{C} \exp(z_m \cdot P_c/\tau)}. \tag{1}$$

Here, τ is the temperature coefficient, which is gradually annealed from 0.2 to 0.08 during training. Rather than directly learning class probabilities, this loss organizes the feature space around class prototypes: samples from the same class are drawn toward the corresponding prototype, whereas samples from different classes are separated from nonmatching prototypes. For classes containing few samples, this explicit class reference reduces the risk that their representations will be absorbed by neighboring majority classes.

The loss scales of the three modalities may differ. To prevent a modality with a larger loss magnitude early in training from persistently dominating optimization, dynamic weights are computed from the losses in the current batch, with gradients detached during weight computation. For $m \in \{\text{seq}, \text{img}, \text{feat}\}$, the dynamic weight is given by Equation (2), and the multimodal proxy contrastive objective is defined in Equation (3):

$$w_m = \frac{L_m}{\sum_{k\in\{\text{seq,img,feat}\}} L_k + \epsilon}, \tag{2}$$

$$L_{\text{PCL}} = \sum_m w_m^{\gamma} L_m, \tag{3}$$

where $\epsilon = 10^{-9}$ is a numerical-stability term and $\gamma = 0.5$. The dynamic weights adjust only the relative contributions of the branch losses, while backpropagation continues to act on the representation of each corresponding modality. Their purpose is not to alter the class prior, but to prevent representation learning before multimodal fusion from becoming biased toward a single input source.

3.2.3. Real-Sample Prototype Initialization and Updating

Here, a class prototype corresponds to the reference position of real observations of a given variable-star class in feature space, and its reliability directly affects how effectively proxy contrastive learning constrains tail classes. If augmented samples are included in estimating class centers, statistical biases introduced during generation may accumulate in the prototypes and become amplified when minority-class samples are scarce. Therefore, before Stage 1 training begins, the prototypes are warm-started using only real training samples. The mean of the temporal and image projections is used as the joint temporal–image projection; its within-class mean and the within-class mean of the global statistical projection are calculated separately and normalized to initialize P_{modal} and P_{feat}.

During training, the prototypes continue to be updated from real samples in the current batch using an exponential moving average (EMA). Let $\bar{z}_c$ denote the mean projection of real samples from class c, and let the momentum coefficient be $\mu = 0.995$. The prototype is updated according to Equation (4):

$$P_c^{(t)} = \frac{\mu P_c^{(t-1)} + (1-\mu)\bar{z}_c}{\left\| \mu P_c^{(t-1)} + (1-\mu)\bar{z}_c \right\|_2}. \tag{4}$$

When a class has no real samples in the current batch, its historical prototype remains unchanged. This mechanism allows augmented samples to participate in network-parameter learning and thereby increase training support for tail classes, but prevents them from directly changing prototype locations. It therefore controls sample-count supplementation separately from class-center estimation. The output of Stage 1 is not a final classifier, but relatively stable multimodal representations and class anchors for decision-boundary recalibration in Stage 2.

3.3. Stage 2: Classifier Recalibration for Long-Tailed Bias

3.3.1. Classifier Adaptation Under Representation-Preservation Constraints

After Stage 1, the model has acquired multimodal representations referenced to real-sample prototypes; however, representability does not necessarily imply correct partitioning. In long-tailed data, even when minority-class observations lie near appropriate class regions, they may still be assigned to majority classes because of majority-class priors. Stage 2 therefore does not relearn the entire light-curve morphology, but instead focuses on classification-boundary recalibration. After loading the weights learned in Stage 1, the model freezes low-level morphology-extraction layers, including one-dimensional convolutions, two-dimensional convolutions, depthwise separable convolutions, and max-pooling layers, to preserve the learned local light-curve structures. The remaining unfrozen high-level representations, global-feature encoding, and fusion modules continue to adapt to the classification objective. Thus, the method adopts a staged, partially frozen decoupling strategy rather than fixing the entire representation and training only the classification heads.

Stage 2 is driven primarily by classification losses while retaining the prototype-anchoring constraint from Stage 1. Let L_{CB}, L_{LA}, L_{CF}, and L_{fuse} denote the class-balanced expert loss, logit-adjusted expert loss, cosine-focal expert loss, and fused-classification loss, respectively, and let L_{anchor} denote the mean of the three prototype-anchoring losses. The total training loss is given by Equation (5):

$$L_{\text{stage2}} = L_{\text{CB}} + L_{\text{LA}} + L_{\text{CF}} + L_{\text{fuse}} + \lambda_{\text{anchor}} L_{\text{anchor}}. \tag{5}$$

The baseline configuration uses $\lambda_{\text{anchor}} = 0.1$. This term penalizes departures from Stage 1 prototypes while permitting classifier adaptation; a smaller value weakens anchoring, while a larger value emphasizes representation preservation.

3.3.2. Heterogeneous Classification Heads for Long-Tailed Classification

The three classification experts share the 192-dimensional joint representation H but address boundary problems in long-tailed classification from different perspectives. The purpose is not to assign a predetermined classifier to each class, but to allow the model to retain three complementary bases for prediction: sample-count compensation, class-prior modeling, and difficult-morphology discrimination.

The Class-Balanced Head uses class weights based on the effective number of samples [24] and assigns relatively greater cross-entropy weights to tail-class samples. In the implementation, the corresponding hyperparameter is $\beta = 0.99$, and the weights are normalized across classes. This term addresses the domination of the training loss by the cumulative contribution of majority classes, ensuring that tail-class misclassifications influence parameter updates in proportion to their sample sizes.

The Logit-Adjusted Head corrects the output logit for class c using the training-set class prior π_c [27], as given by Equation (6):

$$l_c^{\text{LA}} = l_c + \alpha_{\text{LA}} \log(\pi_c + 10^{-9}). \tag{6}$$

The baseline coefficient is $\alpha_{\text{LA}} = 0.3$, while $\alpha_{\text{LA}} = 0$ removes the explicit prior term. In CRTS, adjusted logits are used for this branch's training loss and unadjusted logits for prediction; in OGLE, the adjustment is retained for prediction. The coefficient controls the strength of prior-dependent adjustment and is held fixed in the reported runs.

The Cosine Focal Head first applies L_2 normalization to the joint features and class weights, generates classification logits from cosine similarities, and uses a scale factor of 20 to amplify relative interclass differences. It then applies Focal Loss with a focusing parameter of 2. The cosine formulation emphasizes the relative similarity between a sample and each class direction, while the focal term increases the loss weight of samples with low current prediction probabilities. Their combination makes this expert more attentive to difficult variable-star samples with similar morphologies and strongly overlapping boundaries, without allowing the large number of easily classified majority-class samples alone to determine the update direction.

The three experts do not merely apply repeated weighting to the same loss: the Class-Balanced Head focuses on imbalanced sample counts, the Logit-Adjusted Head explicitly addresses class priors, and the Cosine Focal Head focuses on difficult boundary samples. Together, they provide a multiview recalibration of decision bias under a long-tailed distribution.

For intuition, the class-balanced loss changes the cost of a tail-class error, the logit-adjusted loss changes the class-dependent score offset during fitting, and focal weighting emphasizes observations that the current model finds difficult, regardless of whether their class is rare.

3.3.3. Sample-Adaptive Multi-Head Fusion

Class imbalance does not affect every variable-star sample in the same manner. For some samples, the main difficulty is classification bias caused by class priors; for others, morphological similarity places them closer to a classification boundary. Combining the three classification heads in fixed proportions would make it difficult to adjust each branch's contribution to the classification difficulty of an individual sample. Therefore, a fusion-weight network comprising two fully connected layers takes the joint representation H as

input and generates adaptive Softmax weights for the Class-Balanced, Logit-Adjusted, and Cosine Focal Heads. Let $g(H)$ denote the unnormalized output of this network. The fusion weights are obtained from Equation (7):

$$r = \mathrm{Softmax}(g(H)), \qquad r_{\mathrm{CB}} + r_{\mathrm{LA}} + r_{\mathrm{CF}} = 1. \tag{7}$$

Let p_{CB}, p_{LA}, and p_{CF} denote the class probabilities output by the three classification heads, with corresponding fusion weights r_{CB}, r_{LA}, and r_{CF}. The final class probabilities are computed using Equation (8):

$$p_{\mathrm{final}} = r_{\mathrm{CB}} p_{\mathrm{CB}} + r_{\mathrm{LA}} p_{\mathrm{LA}} + r_{\mathrm{CF}} p_{\mathrm{CF}}. \tag{8}$$

All three classification heads participate in the prediction for every sample. The fusion-weight network continuously adjusts their contributions to the current sample rather than selecting only one head. The final predicted class is the class with the highest probability in p_{final}.

4. Experiments and Results

4.1. Experimental Settings and Evaluation Metrics

4.1.1. Computing Environment and Hyperparameter Settings

Both stages use AdamW with weight decay 10^{-4}, gradient clipping at 1.0, and cosine learning-rate decay with warm-up. CRTS uses a batch size of 256 and initial learning rates 4×10^{-4} and 10^{-4}. OGLE scales these rates in proportion to its configured global batch size relative to 256. The three-run comparison uses stage-wise epoch limits of 300/120 for CRTS and 300/150 for OGLE. The coefficients $\lambda_{\mathrm{anchor}} = 0.1$ and $\alpha_{\mathrm{LA}} = 0.3$ are the fixed baseline settings, and the reported uncertainty summarizes variation across the recorded training runs.

The main performance estimates use seeds 42, 142, and 242 on one fixed split per survey, with model selection based on validation macro F1. The CRTS split is approximately 7:2:1, and the OGLE split is approximately 80:10:10, as detailed in Section 4.4. The CRTS ablation configurations are evaluated using the same three seeds and fixed test objects and are reported separately in Section 4.5. To examine sensitivity to the data-allocation ratio, the same CRTS pipeline was also trained with the same three seeds on a separate fixed 6:1:3 split; these results are reported separately in Section 4.2.3 and are not pooled with the main 7:2:1 estimates.

4.1.2. Evaluation Metrics

Because overall accuracy under a long-tailed distribution can readily obscure performance on minority classes, and following evaluation protocols commonly adopted in imbalanced variable-star classification studies [8,18], the following metrics were selected to provide a comprehensive evaluation of model performance.

Macro F1-score: The unweighted mean of the F1-scores across all classes gives minority and majority classes equal weight in the metric and is therefore an important measure of imbalanced-classification performance. For class c, let TP_c, FP_c, and FN_c denote the numbers of true positives, false positives, and false negatives, respectively. Precision describes the proportion of samples predicted as class c that truly belong to that class, recall describes the proportion of true samples from class c that are correctly identified, and the F1-score is the harmonic mean of precision and recall, as defined in Equations (9)–(11):

$$\mathrm{Precision}_c = \frac{TP_c}{TP_c + FP_c}. \tag{9}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}. \tag{10}$$

$$F1_c = \frac{2\,\text{Precision}_c\,\text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}. \tag{11}$$

Balanced accuracy (macro recall): This metric is the macro-average of recall across all classes and reflects the model's mean detection capability across classes. For C classes, it is defined in Equation (12):

$$\text{Macro-Recall} = \frac{1}{C}\sum_{c=1}^{C}\text{Recall}_c\,. \tag{12}$$

Geometric mean of class recalls (G-Mean): This metric summarizes whether recall is maintained across all classes. For C classes, it is defined in Equation (13):

$$\text{G-Mean} = \left(\prod_{c=1}^{C}\text{Recall}_c\right)^{1/C}. \tag{13}$$

Unlike balanced accuracy, which is the arithmetic mean of the class recalls, G-Mean penalizes a very low recall in any individual class more strongly.

4.2. Experimental Results on the CRTS Dataset

As the principal validation dataset in this study, CRTS contains 11 classes of periodic variable stars and exhibits a pronounced long-tailed distribution. The overall classification performance of TDMN was evaluated on CRTS and compared with that of existing methods for imbalanced variable-star classification. This experiment primarily assesses the overall effectiveness of the complete two-stage framework under a long-tailed distribution.

4.2.1. Overall Classification Performance

Across the three training runs, TDMN achieves mean macro F1 of 0.85, balanced accuracy of $(89.50 \pm 0.72)\%$, and G-Mean of 0.88 ± 0.0058 on 3623 CRTS test objects. The same runs yield mean accuracy of $(89.89 \pm 0.23)\%$. Zhang et al. [8] report a G-Mean of 0.86 under their published definition and protocol. On the object-matched test cohort, the retrained UPSILoN-feature RF—a random forest using the 16 official UPSILoN features, as defined in Section 4.4—obtains mean macro F1 of 0.66, balanced accuracy of $(64.02 \pm 1.83)\%$, and G-Mean of 0.55 ± 0.0467. Thus, TDMN provides the stronger aggregate result for this CRTS task.

Table 3 summarizes the published reference results and the object-matched CRTS comparison.

Table 3. Published reference results and object-matched three-run results on CRTS. The TDMN and UPSILoN-feature RF rows give mean ± sample standard deviation over seeds 42, 142, and 242 on the same 3623 test objects; the latter uses the features of Kim and Bailer-Jones [28]. Their Macro F1 and balanced-accuracy entries are based on the same three-run results. Their G-Means are computed from the same test predictions. Macro F1 and G-Mean means are rounded to two decimals; standard deviations retain four decimals. The other literature values are reproduced as reported under their respective definitions and protocols.

Experiment	Macro F1-Score	Balanced Accuracy	G-Mean
Zhang et al. [8]	–	78.0%	0.86
Kang et al. [18]	0.75	86.2%	–
UPSILoN-feature RF [28]	0.66 ± 0.0122	$(64.02 \pm 1.83)\%$	0.55 ± 0.0467
TDMN (our method)	0.85 ± 0.0054	$(89.50 \pm 0.72)\%$	0.88 ± 0.0058

4.2.2. Analysis of Minority-Class Recognition

The three-run class-specific comparison in Table A3 shows that aggregate performance does not imply an improvement for every minority class. On Rot, TDMN has higher mean precision but lower recall and F1 than the UPSILoN-feature RF baseline. On Cep-II, TDMN raises recall while lowering precision, with nearly unchanged mean F1.

Figure 4 summarizes the three 7:2:1 CRTS runs by averaging their row-normalized confusion matrices. RRab, RRc, Blazhko, Ecl, EA, LPV, and δ Scuti have mean recalls above 0.94. RRd has mean recall 0.79, with most errors assigned to other RR Lyrae subclasses. Rot remains the most difficult class, with mean recall 0.44 and 0.49 of its objects assigned to Ecl, whereas Cep-II has mean recall 0.96. These class-level results are consistent with the aggregate balanced accuracy and the difficult-class comparison reported in Appendix A.

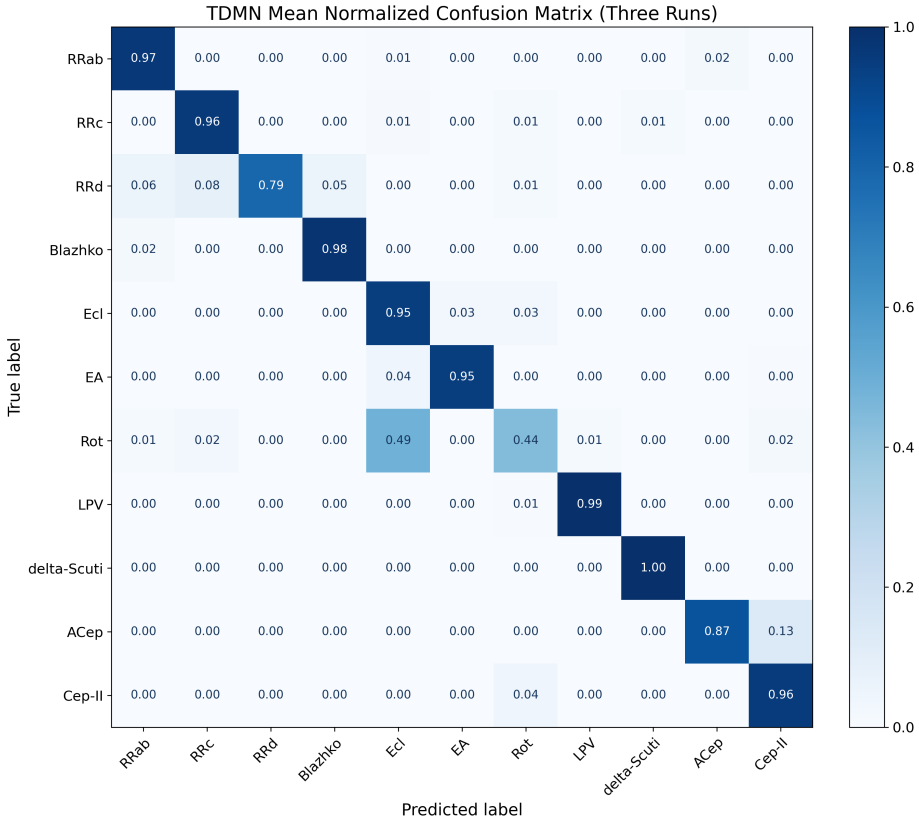


Figure 4. Mean row-normalized CRTS confusion matrix over training seeds 42, 142, and 242 on the same 3623 test objects. Rows denote true classes and columns denote predicted classes.

4.2.3. Sensitivity to the Train/Validation/Test Ratio

The split-ratio experiment uses the same CRTS data-processing and TDMN training pipeline for both ratios. For each ratio, one class-wise split generated with split seed 42 is held fixed across training seeds 42, 142, and 242. Table 4 reports the mean and sample standard deviation over the three training runs.

Table 4. Sensitivity of CRTS performance to the train/validation/test ratio. Each row uses one fixed class-wise split and training seeds 42, 142, and 242. Values are mean ± sample standard deviation; Macro F1 means are rounded to two decimals, and percentage quantities are reported to three decimal places.

Split	Test Objects	Macro F1-Score	Balanced Accuracy	Accuracy
7:2:1	3623	0.85 ± 0.0054	$(89.500 \pm 0.720)\%$	$(89.890 \pm 0.230)\%$
6:1:3	10,850	0.82 ± 0.0041	$(87.400 \pm 0.830)\%$	$(89.260 \pm 0.090)\%$

Changing from 7:2:1 to 6:1:3 assigns fewer objects to training and validation and more objects to testing. Mean macro F1 decreases from 0.85 to 0.82, balanced accuracy decreases by 2.100 percentage points, and overall accuracy decreases by 0.630 percentage points. The larger change in the macro-averaged measures suggests that recognition across classes, particularly the less frequent classes, is more sensitive to this allocation than the majority-dominated overall accuracy. A smaller training set provides fewer real examples for representation and classifier learning, while a smaller validation set can make model selection less stable; conversely, the larger test set provides an evaluation over more objects. However, changing the ratio also changes which objects are assigned to each subset, so the observed differences cannot be attributed to training-set size alone and do not establish a universally optimal split. The 7:2:1 split remains the main reporting protocol, while the 6:1:3 results serve as a separate sensitivity analysis.

4.3. Validation on an Internal OGLE-IV Split

CRTS and OGLE-IV differ in observational passbands, sampling cadences, photometric uncertainties, and class composition. To determine whether the proposed pipeline is applicable only under CRTS data conditions, training and evaluation were repeated using OGLE-IV's own training, validation, and test splits.

4.3.1. Validation Setup

For the OGLE-IV experiments, the two-stage decoupled network architecture and multimodal input configuration were retained, without introducing survey-specific network layers. Because OGLE-IV has a more complex physical hierarchy, two levels of evaluation tasks were designed:

Fine-grained subclass recognition (15-class task): This task includes physically similar subclasses such as CEP_1O, CEP_F, and RRLYR_RRAB.

Coarse-grained superclass recognition (6-class task): This task aggregates the 15 subclasses into six fundamental physical superclasses, including CEP, DSCT, and ECL.

4.3.2. Validation Results

Using the same three runs summarized in Table 5, TDMN achieves mean macro F1 of 0.83 and balanced accuracy of $(84.01 \pm 1.89)\%$ on 15 subclasses, compared with 0.78 and $(73.31 \pm 0.34)\%$ for the retrained UPSILoN-feature RF. After mapping subclass predictions to the six superclasses, TDMN obtains mean macro F1 of 0.87 and balanced accuracy of $(87.95 \pm 3.55)\%$, whereas the UPSILoN-feature RF obtains 0.88 and $(83.21 \pm 0.32)\%$. The relative result therefore depends on the evaluation level: TDMN has higher mean macro F1 on the subclass task, but not on the superclass task. All runs use the same 14,181 test objects, with model selection performed on the validation set. This internal evaluation does not constitute cross-survey transfer; the complete aggregate comparison is reported in Section 4.4.

4.4. Comparison with a Retrained UPSILoN-Feature Random Forest

We use the official UPSILoN feature extractor [28] and retrain a random forest separately for the CRTS 11-class and OGLE 15-subclass label spaces. This baseline is denoted UPSILoN-feature RF; it is not the pretrained UPSILoN classifier, whose taxonomy and training data differ. For each survey, TDMN and the UPSILoN-feature RF use the same fixed test objects, and seeds 42, 142, and 242 are repeated training runs rather than independently redrawn splits. Model and checkpoint selection is performed on the validation set before evaluation on 3623 CRTS or 14,181 OGLE test objects. Because TDMN uses externally supplied periods and multimodal inputs whereas UPSILoN estimates periods internally and applies its own preprocessing, the experiment compares two complete pipelines

on matched test cohorts. Implementation, run-level, and bootstrap details are provided in Appendix A.

Table 5. Three-run results, reported as mean ± sample standard deviation over seeds 42, 142, and 242. TDMN denotes our method; the UPSILoN-feature RF denotes the retrained random forest using the features of Kim and Bailer-Jones [28]. Accuracy and balanced accuracy (BA) are percentages. Macro F1 means are rounded to two decimals; standard deviations retain four decimals. Both pipelines use the same test objects within each survey.

Task	Method	Accuracy (%)	BA (%)	Macro F1
CRTS, 11 classes	TDMN (our method)	89.89 ± 0.23	89.50 ± 0.72	0.85 ± 0.0054
	UPSILoN-feature RF	85.00 ± 0.02	64.02 ± 1.83	0.66 ± 0.0122
OGLE, 15 subclasses	TDMN (our method)	98.11 ± 0.16	84.01 ± 1.89	0.83 ± 0.0324
	UPSILoN-feature RF	96.65 ± 0.04	73.31 ± 0.34	0.78 ± 0.0019
OGLE, 6 superclasses	TDMN (our method)	99.40 ± 0.19	87.95 ± 3.55	0.87 ± 0.0484
	UPSILoN-feature RF	98.65 ± 0.03	83.21 ± 0.32	0.88 ± 0.0008

TDMN has higher mean macro F1 on CRTS (0.85 versus 0.66) and the OGLE subclass task (0.83 versus 0.78), whereas the UPSILoN-feature RF is slightly higher on the OGLE superclass task (0.88 versus 0.87). For OGLE superclasses, TDMN nevertheless has higher mean balanced accuracy. The relative conclusion therefore depends on the survey taxonomy and metric; macro F1 accounts for both false-positive assignments and missed members of every class, whereas balanced accuracy averages class recalls. Paired bootstrap intervals in Appendix A are positive for all three CRTS runs but not for every OGLE subclass run. Evidence for an advantage is therefore consistent on CRTS but mixed for the OGLE subclass task.

The difficult CRTS classes reinforce this qualification. TDMN has lower mean Rot F1 than the UPSILoN-feature RF (0.5510 versus 0.5854). For Cep-II, TDMN raises recall from 0.7556 to 0.9556 but lowers precision from 0.6551 to 0.5540, leaving mean F1 effectively unchanged (0.7009 versus 0.7012). These class-level weaknesses coexist with substantially higher aggregate CRTS macro F1 and balanced accuracy for TDMN. A higher result for one class can accompany lower performance elsewhere, so aggregate and class-specific metrics answer complementary questions. Thus, the aggregate CRTS gain does not resolve Rot or Cep-II, while their individual results alone do not characterize performance over all 11 classes. Full precision, recall, support, and run-level results are reported in Appendix A.

4.5. CRTS Ablation Study Analysis

The CRTS ablation comparison comprises five experiments corresponding to the principal components of the proposed method. Exp 1 is the end-to-end cross-entropy baseline; Exp 2 uses the modified augmentation strategy; Exp 3 incorporates multimodal proxy-contrastive learning; Exp 4 introduces two-stage training with a single class-balanced head; and Exp 5 is the complete TDMN with sample-adaptive multi-head fusion. Each experiment was trained with seeds 42, 142, and 242 on the fixed CRTS split, with model selection performed on the validation set and final evaluation on the held-out test set.

Exp 1 obtains a mean macro F1-score of 0.8197 ± 0.0069. Exp 2 yields a similar value of 0.8223 ± 0.0110, indicating that the modified augmentation strategy alone produces only a limited change in aggregate performance. After proxy-contrastive learning is introduced, Exp 3 increases the mean macro F1-score to 0.8379 ± 0.0013. This improvement is consistent with the use of class prototypes and contrastive constraints to organize the multimodal representation space.

Exp 4 achieves the highest mean macro F1-score, 0.8626 ± 0.0152, compared with 0.8379 ± 0.0013 for Exp 3. The complete TDMN also achieves a higher mean macro F1-score

than Exp 3, reaching 0.8458 ± 0.0054, together with a balanced accuracy of $(89.50 \pm 0.72)\%$. The performance of both two-stage configurations therefore supports the suitability of separating representation learning from classification-boundary adjustment for the long-tailed CRTS task. Although Exp 4 has the highest mean macro F1-score, the complete TDMN shows smaller run-to-run variation while maintaining a similar balanced accuracy.

Table 6 summarizes the results of the five CRTS ablation configurations.

Table 6. Three-run comparison of CRTS ablation configurations on the fixed data split. Values are mean ± sample standard deviation over seeds 42, 142, and 242.

Experiment	Data Aug.	Proxy Contrastive	Two-Stage Training	Adaptive Multi-Head	Macro F1	Balanced Accuracy
Exp 1 (Baseline)	–	–	–	–	0.8197 ± 0.0069	$(89.32 \pm 0.51)\%$
Exp 2	✓	–	–	–	0.8223 ± 0.0110	$(88.15 \pm 1.15)\%$
Exp 3	✓	✓	–	–	0.8379 ± 0.0013	$(89.49 \pm 0.58)\%$
Exp 4	✓	✓	✓	–	0.8626 ± 0.0152	$(89.44 \pm 1.75)\%$
Exp 5 (TDMN)	✓	✓	✓	✓	0.8458 ± 0.0054	$(89.50 \pm 0.72)\%$

Two additional comparisons examine the data and image representations used by the model. Dynamic augmentation gives macro precision, recall, and F1 of 0.83, 0.84, and 0.83 for the five low-support classes, whereas uniform augmentation gives 0.71, 0.88, and 0.78, respectively. The dynamic strategy therefore provides a better precision–recall balance for these classes. In the image branch, three-channel phase images achieve a macro F1-score of 0.66, compared with 0.60 for grayscale phase images. This result supports encoding the observations, smoothed trend, and photometric uncertainties in separate channels.

4.6. Feature-Space Visualization

To examine changes in representation structure before and after the two-stage training procedure, the Stage 1 proxy-contrastive projected representations and Stage 2 multimodal fused representations of test-set samples were separately projected into two dimensions using t-distributed stochastic neighbor embedding (t-SNE). The two figures correspond to high-dimensional representations with different dimensionalities and semantic levels, and t-SNE was fitted independently for each figure.

Figure 5 presents the Stage 1 proxy-contrastive projected representations. Classes including RRab, RRc, RRd, and δ Scuti have already formed relatively compact local regions in the projected space, indicating that the proxy contrastive constraint anchored by real-sample prototypes can establish relatively well-defined class references for some minority classes. Meanwhile, the regions occupied by Rot and by Ecl and EA still overlap substantially, and Cep-II remains locally close to neighboring classes, indicating that Stage 1 representation learning alone cannot completely eliminate confusion between morphologically similar classes.

Figure 6 presents the multimodal fused representations after Stage 2 recalibration. These representations are formed by concatenating the temporal–image fused features with the global statistical features and are used as the shared input to the three classification heads. RRab, RRc, and δ Scuti retain relatively separable local structures in this qualitative projection, whereas Rot continues to overlap with regions occupied by Ecl and EA and Cep-II remains close to neighboring classes. Figures 5 and 6 visualize representations from a single CRTS training run. The projections are qualitative, are not averaged over runs, and are interpreted separately from the three-run class-wise recall and error analysis.

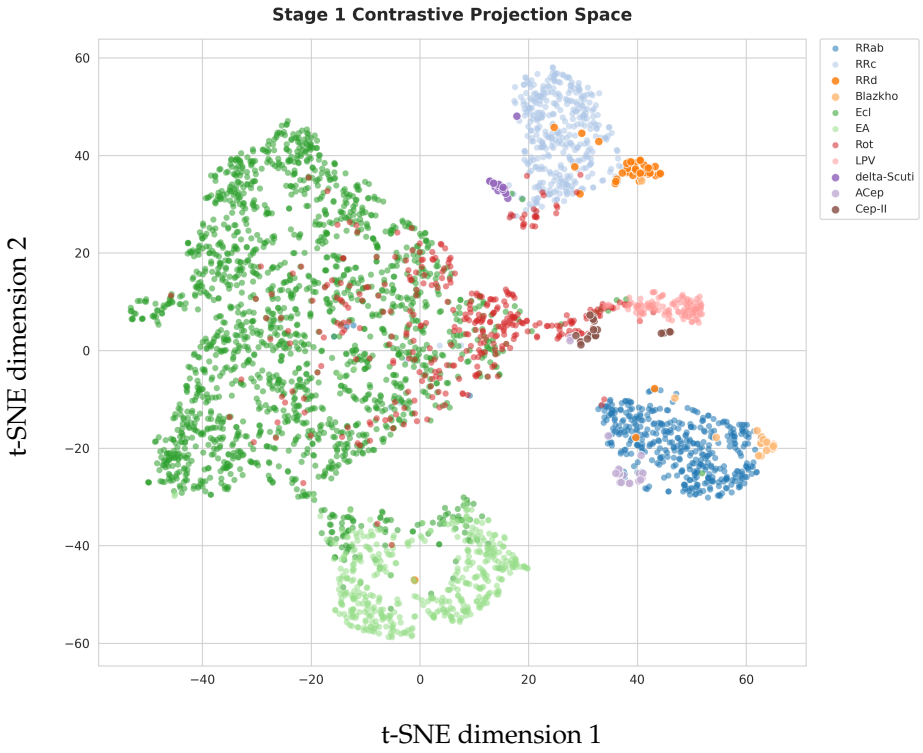


Figure 5. Qualitative t-SNE visualization of Stage 1 projected representations from a single CRTS training run; the projection is not averaged over runs. Axes are embedding coordinates without physical units.

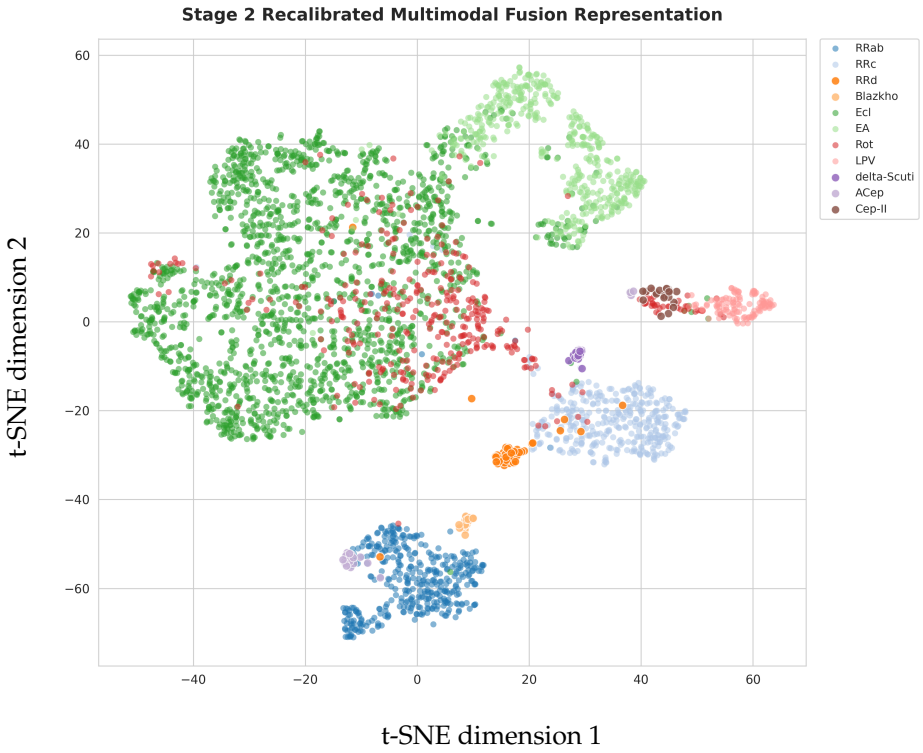


Figure 6. Qualitative t-SNE visualization of Stage 2 fused representations from a single CRTS training run; the projection is not averaged over runs. The embedding was fitted independently of Stage 1, so coordinate locations and apparent inter-cluster distances are not directly comparable between the two plots.

These projected overlaps are consistent with the difficulty of separating smooth periodic morphologies when amplitudes are low or survey sampling and photometric noise obscure local detail. The t-SNE views are qualitative and do not establish the physical cause of an individual error. The three-run confusion matrix confirms Rot–Ecl as the principal remaining class-level confusion, while the less frequent Cep-II errors are assigned mainly to Rot.

5. Discussion

Within an astrophysical survey workflow, TDMN serves as a catalogue-classification tool for known periodic-variable populations. Its class-wise recall and precision quantify, respectively, the recovery and purity of the resulting candidate samples, rather than reducing performance to a majority-dominated overall accuracy. The Cep-II results illustrate that higher recovery can be accompanied by lower precision, while the persistent Rot errors show that improved aggregate metrics do not resolve every difficult class. By applying a consistent classifier to catalogue-scale light curves, the procedure can reduce repeated manual screening and help prioritize candidates for multiband or spectroscopic follow-up. More complete and purer samples of known pulsators, eclipsing binaries, and rotational variables can then support population comparisons, period–luminosity studies, and investigations of stellar evolution, although the present work does not perform those physical analyses.

The scientific scope is limited by the adopted data and labels. CRTS and OGLE represent different selected populations, and the absence of classical Cepheids from the CRTS benchmark does not imply their absence from the survey. TDMN also assumes periodic candidates and externally estimated periods. Because it is a closed-set classifier, an object outside the training taxonomy may be assigned an available label; extension to additional classes requires representative labeled examples, revised class prototypes and output heads, retraining, and independent evaluation. The present experiments therefore do not demonstrate period discovery, unknown-class rejection, or discovery completeness in an unlabeled survey.

The class-level errors also define practical priorities. Rot is most often confused with Ecl, and sparse sampling or weak modulation may obscure the morphology needed to distinguish rotational variability from shallow eclipses. These explanations are observational hypotheses rather than object-by-object physical diagnoses. SuperSmoother, adopted from prior work [25], guides outlier rejection and supplies a trend to the encoders; smoothing may reduce noise but may also weaken shallow eclipses or rotational modulation. Since the current comparisons do not isolate these two roles, a dedicated control is still needed before attributing a measured Rot/Ecl/EA improvement to smoothing.

The three training seeds quantify variation between fitted models on fixed test objects, not uncertainty across independently sampled survey populations. The small support of several tail classes therefore remains important when interpreting their scores. Future work will examine period errors, preprocessing and augmentation controls, and coefficient sensitivity under matched settings. A longer-term objective is to connect raw survey time streams to variability detection, period inference, and uncertainty-aware classification. Self-supervised or unsupervised representation learning and astronomical foundation-model approaches [13], combined with open-set detection and expert validation, may help identify unfamiliar light-curve morphologies without treating them automatically as new stellar classes.

6. Conclusions

To address majority-class bias in periodic-variable-star classification under long-tailed distributions, a two-stage decoupled multimodal network is proposed for imbalanced periodic-variable-star classification, targeting the two interrelated difficulties whereby minority-class representations become entangled with majority classes and classification boundaries become biased toward majority-class priors. In Stage 1, class prototypes are maintained using real observations, and proxy contrastive learning constrains the projected representations of light-curve sequences, three-channel phase images, and global statistical features, thereby establishing relatively stable class references for minority classes. In Stage 2, high-level representations and classifiers are adapted while retaining low-level morphology-extraction capabilities, and classification decisions under long-tailed conditions are recalibrated through sample-adaptive fusion of the Class-Balanced, Logit-Adjusted, and Cosine Focal Heads.

The main results summarize three training runs on fixed survey-specific splits, with the repeated CRTS ablation configurations reported separately. The TDMN mean macro F1 is 0.85 on CRTS, 0.83 on OGLE subclasses, and 0.87 on OGLE superclasses. The corresponding UPSILoN-feature RF means are 0.66, 0.78, and 0.88; Table 5 reports the run-to-run standard deviations. Thus, TDMN improves mean macro F1 on CRTS and OGLE subclasses, but the UPSILoN-feature RF has higher mean macro F1 on OGLE superclasses despite lower mean balanced accuracy. Rot F1 remains lower with TDMN, while the Cep-II recall gain is accompanied by reduced precision. The observed benefits therefore depend on class, metric, and survey taxonomy. These results characterize the specified classification pipelines within the selected catalogue label spaces, rather than an end-to-end discovery system.

The confusion matrix and feature-space visualizations show that several RR Lyrae subclasses and classes such as δ Scuti form relatively separable local structures; nevertheless, confusion remains between Rot and Ecl or EA, with the remaining Cep-II errors assigned mainly to Rot. In particular, the lower Rot F1-score is a current weakness rather than a problem resolved by the aggregate improvement. Improving Rot discrimination from Ecl and EA is therefore a specific priority for future work. This should include object-level examination of the errors and evaluation of whether self-supervised pretraining, multiband photometry, or auxiliary information such as spectra and colors can preserve the weak morphology needed for this distinction.

Author Contributions: Conceptualization, Q.L., F.Z., Z.Z., A.-L.L. and X.K.; methodology, Q.L. and F.Z.; software, Q.L.; validation, Q.L., F.Z. and T.S.; formal analysis, Q.L., Z.Z., T.S. and A.-L.L.; investigation, Q.L. and Z.Z.; resources, Z.Z., A.-L.L. and X.K.; data curation, Q.L., T.S., X.Z. and X.K.; writing—original draft preparation, Q.L. and X.K.; writing—review and editing, Q.L., F.Z., Z.Z. and A.-L.L.; visualization, Q.L.; supervision, F.Z., Z.Z., T.S., X.Z., A.-L.L. and X.K.; project administration, F.Z., Z.Z., A.-L.L. and X.K.; funding acquisition, Z.Z., T.S., X.Z. and A.-L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant numbers 12473104 and 12273075) and the National Key Research and Development Program of China (grant number 2025YFF0510602).

Data Availability Statement: Publicly available datasets were analyzed in this study. The CRTS and OGLE-IV data are available from the public survey archives described in Section 2.1 and cited in this article [4,5]. Data-access instructions and the model-training, data-processing, evaluation, and visualization code supporting the findings of this study are publicly available at https://github.com/lqy0718/TDMN (accessed on 10 September 2026).

Conflicts of Interest: The authors declare no conflicts of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Implementation and Run-Level Details for the UPSILoN-Feature RF Comparison

The official UPSILoN extractor provides 16 features: amplitude, half-light amplitude ratio, kurtosis, period, phase CUSUM, phase eta, Fourier phase combinations ϕ_{21} and ϕ_{31}, quartile ratio, Fourier amplitude ratios R_{21} and R_{31}, Shapiro–Wilk statistic, skewness, two slope percentiles, and Stetson K. Random-forest class weights are recomputed per bootstrap sample (`balanced_subsample`). The validation grid comprises 24 configurations: 300 or 600 trees, maximum depth unlimited/20/40, minimum leaf size 1/2, and the number of candidate features set to the square root of the feature count or 10. Nonfinite values are median-imputed within a pipeline fitted only on training data. Validation macro F1 selects the UPSILoN-feature RF configuration. TDMN similarly uses validation macro F1 to select either the best individual Stage 2 checkpoint or the average of its three highest-ranked checkpoints.

For each survey, seeds 42, 142, and 242 reuse one prepared dataset. Alignment of the saved prediction files verifies identical unique object IDs and true labels across methods and runs: 3623 CRTS objects and 14,181 OGLE objects. The OGLE six-class results are obtained by mapping each predicted subclass to its superclass; no separate six-class model is trained. The CRTS preparation partitions eligible real objects at approximately 7:2:1, augments minority training classes to the median real training-class count, and computes class weights and priors from real training objects. OGLE uses an object-ID MD5 assignment at approximately 80:10:10 with bounded minority oversampling in training. These procedures differ from the CRTS augmentation ablation in Section 4.5 and are therefore reported separately.

The comparison concerns the two complete pipelines on matched test objects. TDMN uses externally supplied periods and multimodal inputs as described in Section 2, whereas UPSILoN estimates periods internally and applies its own sigma clipping. The results therefore do not isolate the classifier while holding all input information and preprocessing fixed.

Table A1. Macro F1 by run for TDMN (our method) and the retrained UPSILoN-feature RF of Kim and Bailer-Jones [28]. Aggregate summaries in Table 5 use the underlying unrounded scores.

Task	Method	Seed 42	Seed 142	Seed 242
CRTS, 11 classes	TDMN (our method)	0.84	0.84	0.85
	UPSILoN-feature RF	0.65	0.67	0.65
OGLE, 15 subclasses	TDMN (our method)	0.87	0.80	0.82
	UPSILoN-feature RF	0.79	0.78	0.78
OGLE, 6 superclasses	TDMN (our method)	0.92	0.82	0.86
	UPSILoN-feature RF	0.88	0.88	0.88

For each fitted model pair, a 95% percentile interval was computed from 10,000 paired bootstrap resamples stratified by true class, applying the same object draw to both predictions. These intervals condition on the fitted models, retained cohort, and class counts; they do not include uncertainty from independently redrawing the training, validation, or test populations.

Table A2. Paired class-stratified bootstrap intervals for the macro F1 difference, TDMN (our method) minus the UPSILoN-feature RF [28]. Each interval is the 95% percentile interval from 10,000 resamples of the fixed test objects.

Task	Training Seed	95% Interval
CRTS, 11 classes	42	[0.1490, 0.2424]
	142	[0.1258, 0.2192]
	242	[0.1616, 0.2532]
OGLE, 15 subclasses	42	[0.0359, 0.1343]
	142	[−0.0159, 0.0654]
	242	[−0.0061, 0.0902]

The intervals are positive for all three CRTS runs. For the OGLE subclass task, the interval is positive for seed 42 but includes zero for seeds 142 and 242. The within-test-set evidence is therefore consistent across CRTS runs but not across every OGLE subclass run.

Table A3. Rot and Cep-II classification on CRTS: precision, recall, and F1 averaged over three runs. Support is the number of unique test objects.

Class	Method	Support	Precision	Recall	F1
Rot	TDMN (our method)	361	0.7462	0.4377	0.5510
	UPSILoN-feature RF	361	0.6783	0.5152	0.5854
Cep-II	TDMN (our method)	15	0.5540	0.9556	0.7009
	UPSILoN-feature RF	15	0.6551	0.7556	0.7012

The small Cep-II support makes a single object's classification consequential. OGLE also contains very small test subclasses, including seven ACEP_1O and five T2CEP_RVTAU objects; their TDMN F1-scores range from 0.1818 to 0.6154 and from 0.1852 to 0.7692, respectively, across the three runs. These ranges motivate reporting all repeated runs rather than relying on a single initialization.

References

1. Eyer, L.; Mowlavi, N. Variable stars across the observational HR diagram. *J. Phys. Conf. Ser.* **2008**, *118*, 012010. [CrossRef]
2. Wright, E.L.; Eisenhardt, P.R.M.; Mainzer, A.K.; Ressler, M.E.; Cutri, R.M.; Jarrett, T.; Kirkpatrick, J.D.; Padgett, D.; McMillan, R.S.; Skrutskie, M.; et al. The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance. *Astron. J.* **2010**, *140*, 1868–1881. [CrossRef]
3. Borucki, W.J.; Koch, D.; Basri, G.; Batalha, N.; Brown, T.; Caldwell, D.; Caldwell, J.; Christensen-Dalsgaard, J.; Cochran, W.D.; DeVore, E.; et al. Kepler Planet-Detection Mission: Introduction and First Results. *Science* **2010**, *327*, 977–980. [CrossRef]
4. Drake, A.J.; Djorgovski, S.G.; Catelan, M.; Graham, M.; Mahabal, A.; Larson, S.; Christensen, E.; Torrealba, G.; Beshore, E.; McNaught, R.; et al. The Catalina Surveys Southern periodic variable star catalogue. *Mon. Not. R. Astron. Soc.* **2017**, *469*, 3688–3712. [CrossRef]
5. Udalski, A.; Szymański, M.K.; Szymański, G. OGLE-IV: Fourth Phase of the Optical Gravitational Lensing Experiment. *Acta Astron.* **2015**, *65*, 1–38. [CrossRef]
6. Nun, I.; Protopapas, P.; Sim, B.; Zhu, M.; Dave, R.; Castro, N.; Pichara, K. FATS: Feature Analysis for Time Series. *arXiv* **2015**, arXiv:1506.00010. [CrossRef]
7. Hinners, T.A.; Tat, K.; Thorp, R. Machine Learning Techniques for Stellar Light Curve Classification. *Astron. J.* **2018**, *156*, 7. [CrossRef]
8. Zhang, J.; Zhang, Y.; Kang, Z.; Li, C.; Tao, Y.; Zhao, Y.; Wu, X.-B. A novel approach for variable star classification based on imbalanced learning. *Publ. Astron. Soc. Aust.* **2023**, *40*, e037. [CrossRef]
9. Aguirre, C.; Pichara, K.; Becker, I. Deep multi-survey classification of variable stars. *Mon. Not. R. Astron. Soc.* **2019**, *482*, 5078–5092. [CrossRef]
10. Szklenar, T.; Bodi, A.; Tarczay-Nehez, D.; Vida, K.; Mezo, G.; Szabo, R. Variable Star Classification with a Multiple-Input Neural Network. *Astrophys. J.* **2022**, *938*, 37. [CrossRef]

11. Akhmetali, A.; Namazbayev, T.; Subebekova, G.; Zaidyn, M.; Akniyazova, A.; Ashimov, Y.; Ussipov, N. Classification of Variable Star Light Curves with Convolutional Neural Network. *Galaxies* **2024**, *12*, 75. [CrossRef]
12. Naul, B.; Bloom, J.S.; Perez, F.; van der Walt, S. A recurrent neural network for classification of unevenly sampled variable stars. *Nat. Astron.* **2018**, *2*, 151–155. [CrossRef]
13. Li, Y.-Y.; Bai, Y.; Wang, C.; Qu, M.; Lu, Z.; Soria, R.; Liu, J. Deep Learning and Methods Based on Large Language Models Applied to Stellar Light Curve Classification. *Intell. Comput.* **2025**, *4*, 0110. [CrossRef]
14. Sun, Y.; Wong, A.K.C.; Kamel, M.S. Classification of imbalanced data: A review. *Int. J. Pattern Recognit. Artif. Intell.* **2009**, *23*, 687–719. [CrossRef]
15. Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; Kalantidis, Y. Decoupling Representation and Classifier for Long-Tailed Recognition. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Conference, 25–30 April 2020. [CrossRef]
16. Burhanudin, U.F.; Maund, J.R.; Killestein, T.; Ackley, K.; Dyer, M.J.; Lyman, J.; Ulaczyk, K.; Cutter, R.; Mong, Y.-L.; Steeghs, D.; et al. Light-curve classification with recurrent neural networks for GOTO: Dealing with imbalanced data. *Mon. Not. R. Astron. Soc.* **2021**, *505*, 4345–4361. [CrossRef]
17. Hosenie, Z.; Lyon, R.; Stappers, B.; Mootoovaloo, A.; McBride, V. Imbalance learning for variable star classification. *Mon. Not. R. Astron. Soc.* **2020**, *493*, 6050–6059. [CrossRef]
18. Kang, Z.; Zhang, Y.; Zhang, J.; Li, C.; Kong, M.; Zhao, Y.; Wu, X.-B. Periodic Variable Star Classification with Deep Learning: Handling Data Imbalance in an Ensemble Augmentation Way. *Publ. Astron. Soc. Pac.* **2023**, *135*, 094501. [CrossRef]
19. Perez-Galarce, F.; Martinez-Palomera, J.; Pichara, K.; Huijse, P.; Catelan, M. A self-regulated convolutional neural network for classifying variable stars. *Mon. Not. R. Astron. Soc.* **2025**, *540*, 3263–3280. [CrossRef]
20. Movshovitz-Attias, Y.; Toshev, A.; Leung, T.K.; Ioffe, S.; Singh, S. No Fuss Distance Metric Learning Using Proxies. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 360–368. [CrossRef]
21. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML), Virtual, 13–18 July 2020*; PMLR: Cambridge, MA, USA, 2020; Volume 119, pp. 1597–1607. [CrossRef]
22. Huang, L.; Cai, S.; Zhuang, Y.; Jing, C.; Huang, Y.; Tu, X.; Ding, X. A Two-Stage Contrastive Learning Framework for Imbalanced Aerial Scene Recognition. In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 3518–3522. [CrossRef]
23. López, V.; Triguero, I.; Carmona, C.J.; García, S.; Herrera, F. Addressing imbalanced classification with instance generation techniques: IPADE-ID. *Neurocomputing* **2014**, *126*, 15–28. [CrossRef]
24. Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; Belongie, S. Class-Balanced Loss Based on Effective Number of Samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 9268–9277. [CrossRef]
25. Friedman, J.H. *A Variable Span Smoother*; Technical Report LCS 5; Laboratory for Computational Statistics, Stanford University: Stanford, CA, USA, 1984. [CrossRef]
26. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141. [CrossRef]
27. Menon, A.K.; Jayasumana, S.; Rawat, A.S.; Jain, H.; Veit, A.; Kumar, S. Long-Tail Learning via Logit Adjustment. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Conference, 3–7 May 2021.
28. Kim, D.-W.; Bailer-Jones, C.A.L. A Package for the Automated Classification of Periodic Variable Stars. *Astron. Astrophys.* **2016**, *587*, A18. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.