Astronomy & Astrophysics manuscript no. output
December 11, 2024

©ESO 2024

Uncertainty estimation for time series classification

Exploring predictive uncertainty in transformer-based models for variable stars

M. Cádiz-Leyton[1,2⋆], G. Cabrera-Vives[1,3,4,5⋆⋆], P. Protopapas[2], D. Moreno-Cartagena[1,2,3], C. Donoso-Oliva[3,5], and I. Becker[4,5,6]

[1] Department of Computer Science, Universidad de Concepción, Edmundo Larenas 219, Concepción, Chile
[2] John A. Paulson School of Engineering and Applied Science, Harvard University, Cambridge, MA, 02138
[3] Center for Data and Artificial Intelligence, Universidad de Concepción, Edmundo Larenas 310, Concepción, Chile
[4] Millennium Institute of Astrophysics (MAS), Nuncio Monseñor Sotero Sanz 100, Of. 104, Providencia, Santiago, Chile
[5] Millennium Nucleus on Young Exoplanets and their Moons (YEMS), Chile
[6] Department of Computer Science, Pontificia Universidad Catolica de Chile, Macul, Santiago 7820436, Chile

December 11, 2024

ABSTRACT

Context. Classifying variable stars is key for understanding stellar evolution and galactic dynamics. With the demands of large astronomical surveys, machine learning models, especially attention-based neural networks, have become the state-of-the-art. While achieving high accuracy is crucial, enhancing model interpretability and uncertainty estimation is equally important to ensure that insights are both reliable and comprehensible.

Aims. We aim to enhance transformer-based models for classifying astronomical light curves by incorporating uncertainty estimation techniques to detect misclassified instances. We tested our methods on labeled datasets from MACHO, OGLE-III, and ATLAS, introducing a framework that significantly improves the reliability of automated classification for the next-generation surveys.

Methods. We used Astromer, a transformer-based encoder designed for capturing representations of single-band light curves. We enhanced its capabilities by applying three methods for quantifying uncertainty: Monte Carlo Dropout (MC Dropout), Hierarchical Stochastic Attention (HSA), and a novel hybrid method combining both approaches, which we have named Hierarchical Attention with Monte Carlo Dropout (HA-MC Dropout). We compared these methods against a baseline of deep ensembles (DEs). To estimate uncertainty estimation scores for the misclassification task, we selected Sampled Maximum Probability (SMP), Probability Variance (PV), and Bayesian Active Learning by Disagreement (BALD) as uncertainty estimates.

Results. In predictive performance tests, HA-MC Dropout outperforms the baseline, achieving macro F1-scores of 79.8 ± 0.5 on OGLE, 84 ± 1.3 on ATLAS, and 76.6 ± 1.8 on MACHO. When comparing the PV score values, the quality of uncertainty estimation by HA-MC Dropout surpasses that of all other methods, with improvements of 2.5 ± 2.3 for MACHO, 3.3 ± 2.1 for ATLAS and 8.5 ± 1.6 for OGLE-III.

Key words. methods: statistical – methods: data analysis – techniques: photometric – stars: variables: general

1. Introduction

The identification and classification of variable stars is crucial in advancing our understanding of the cosmos. For example, Cepheids and RR Lyrae stars are key rungs on the cosmological distance ladder (Feast et al. 2014; Ngeow 2015). The emergence of next-generation survey telescopes, such as the Vera Rubin Observatory and its Legacy Survey of Space and Time (Ivezić et al. 2019), presents new opportunities for analyzing an abundance of photometric observations. Such comprehensive data, with their corresponding time-stamps (i.e., light curves), are instrumental in detecting new classes of variable stars and uncovering previously unknown astronomical phenomena (Bassi et al. 2021).

Classifying light curves is essential for analyzing variable stars; however, this task presents significant challenges due to heteroskedasticity, sparsity, and observational gaps (Mahabal et al. 2017). Astronomy has transitioned from traditional feature-based analysis to advanced data-driven models enabled by deep learning (Smith & Geach 2023). This evolution is evidenced by the adoption of diverse architectures ranging from multilayer perceptrons (Karpenka et al. 2013) to Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) (Cabrera-Vives et al. 2016; Mahabal et al. 2017; Protopapas 2017; Naul et al. 2018; Carrasco-Davis et al. 2019; Becker et al. 2020; Donoso-Oliva et al. 2021).

Building upon this evolution, self-attention-based models are setting new benchmarks (Donoso-Oliva et al. 2023; Moreno-Cartagena et al. 2023; Pan et al. 2024; Leung & Bovy 2023; Parker et al. 2024). These models address applications from inferring black hole properties (Park et al. 2021), to denoising light curves (Morvan et al. 2022) and specialized multi-band light-curve classification (Pimentel et al. 2022; Cabrera-Vives et al. 2024).

As the complexity of models escalates, ensuring the reliability of classification results becomes increasingly critical. Within the framework of deep neural networks, uncertainty estimation is pivotal, as it not only enhances the confidence in predictions but also proves particularly beneficial in the detection of misclassifications, where uncertain predictions can signal potential errors (Gawlikowski et al. 2023). Traditional techniques such as

⋆ e-mail: mcadiz2018@inf.udec.cl
⋆⋆ e-mail: guillecabrera@inf.udec.cl

Bayesian Neural Networks (BNNs; Blundell et al. 2015), which inherently model uncertainty by approximating a posterior distribution over model parameters, have been effective in astronomy (Möller & de Boissière 2020; Killestein et al. 2021; Ciucă et al. 2021). However, their training stage requires high computational capacity, leading to increased resource use and extended convergence times.

In this work, we present a methodology that combines deep attention-based classifiers with uncertainty estimation techniques for misclassification detection. Our model is based on Astromer, the transformer-based embedding approach proposed by Donoso-Oliva et al. (2023). For our analysis, we implemented three state-of-the-art uncertainty estimation approaches: deep ensembles (DE; Valdenegro-Toro 2019; Ganaie et al. 2022, Monte Carlo Dropout (MC Dropout; Gal & Ghahramani 2016), and Hierarchical Stochastic Attention (HSA; Pei et al. 2022). We additionally propose an alternative approach, Hierarchical Attention with MC Dropout (HA-MC Dropout), which integrates a hierarchical attention structure with dropout activation during the inference stage.

Although DEs are computationally expensive in practice, they establish a robust baseline for capturing uncertainty and facilitate comparisons with other techniques (Lakshminarayanan et al. 2017). Thus, we used DEs as a baseline to compare the performance of MC Dropout, HSA and HA-MC Dropout within a variable star transformer-based classifier. We further assess the uncertainty estimation capabilities of our models through statistical significance tests. Our main contributions are:

1. We present a methodology that fuses deep attention-based classifiers for astronomical time series with uncertainty estimation techniques.
2. We establish through empirical evaluations that the HA-MC Dropout approach outperforms DEs in the context of variable star classification, emphasizing its utility as a primary technique for this application.

This work is organized as follows: Sect. 2 details the attention-based architecture used, outlines the methods employed, and formally defines the evaluation techniques used in our research. Section 3 presents the experimental setup, including dataset descriptions and the generation of misclassification instances. Our findings are then introduced in Sect. 4, and the paper concludes with Sect. 5, which summarizes the key outcomes and insights derived from our work.

2. Methods

This section presents the methodology adopted in our work, focusing on the Astromer transformer-based model and its integration of uncertainty estimation techniques. We detail the four principal approaches: deep ensembles, Monte Carlo Dropout, hierarchical stochastic attention and the hierarchical attention with MC dropout to assess predictive uncertainty. Additionally, we explain the three types of uncertainty estimates employed for the misclassification task and outline the methodology for quality evaluation.

2.1. Transformer-based model for light curves

Our work uses an architecture inspired by Astromer, which is an encoder-decoder model derived from the principles of the Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. 2019). Unlike BERT, which is specifically designed for natural language processing (NLP) tasks, Astromer is adapted for capturing embedding representations of astronomical light curves, serving as an automatic feature extractor for these time series.

The model includes several key components: positional encoding (PE), two self-attention blocks, and a long-short-term memory network classifier (LSTMs; Hochreiter & Schmidhuber 1997). It processes single-band time series, each consisting of $L = 200$ observations. These observations are characterized by a vector of magnitudes, $\boldsymbol{x} \in \mathbb{R}^L$, and corresponding timestamps, $\boldsymbol{t} \in \mathbb{R}^L$. The PE component encodes each timestamp using trigonometric functions, similar to those defined in Vaswani et al. (2017); however, it operates in the time domain of the light curve, rather than using index positions. Simultaneously, the magnitude data is fed into a feedforward (FF) neural network. For each point in the light curve, both the PE and the FF neural network output a vector of dimensionality $d_x = 256$. The outputs from the PE and the magnitudes FF neural network are added to produce $\boldsymbol{X} \in \mathbb{R}^{L \times d_x}$. This 200×256 matrix serves as the input representation for the standard transformer self-attention mechanism.

Transformers are composed of multiple layers of self-attention heads. A self-attention head transforms $\boldsymbol{X}$ into query, key, and value matrices ($\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$) by using trainable weights $\boldsymbol{W_Q}, \boldsymbol{W_K}, \boldsymbol{W_V} \in \mathbb{R}^{d_x \times d_k}$:

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W_Q}, \quad \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W_K}, \quad \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W_V}, \tag{1}$$

where d_k is the dimension of the output of the attention head for each observation.

Attention weights $\alpha_{i,j}$ represent how much token i attends token j. These weights are determined by calculating the dot product between the query and key vectors, $\boldsymbol{q}_i = \boldsymbol{W_Q}^\top \boldsymbol{x}_i$ and $\boldsymbol{k}_j = \boldsymbol{W_K}^\top \boldsymbol{x}_j$, and passing them through a softmax function in order to normalize them. Here, $\boldsymbol{x}_i$ represents the i-th row of $\boldsymbol{X}$. To stabilize the gradients during training, the dot product is divided by the square root of the dimension of the key vectors $\sqrt{d}$. In other words,

$$\begin{aligned} \boldsymbol{\alpha}_i &= \text{softmax}\left(\frac{\boldsymbol{q}_i^\top \boldsymbol{k}_1}{\sqrt{d_k}}, \frac{\boldsymbol{q}_i^\top \boldsymbol{k}_2}{\sqrt{d_k}}, \cdots, \frac{\boldsymbol{q}_i^\top \boldsymbol{k}_L}{\sqrt{d_k}}\right), \\ &= \text{softmax}\left(\boldsymbol{K}\boldsymbol{q}_i\right). \end{aligned} \tag{2}$$

The output $\boldsymbol{h}_i$ for each head is computed by applying the normalized attention weights to the corresponding values $\boldsymbol{v}_i = \boldsymbol{W_V}^\top \boldsymbol{x}_i$:

$$\boldsymbol{h}_i = \sum_j \alpha_{i,j} \boldsymbol{v}_j. \tag{3}$$

In other words, the output of each head is the sum of the value vector representation of the input weighted by the attention payed to each of these vectors. This is usually written as

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{4}$$

where the softmax is applied row-wise.

After calculating the outputs $\boldsymbol{h}_i$ of each head, they are concatenated $[\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{\#\text{heads}}]$ to form the final output of the multi-head attention block. Astromer has two blocks, each consisting of four heads with 64 neurons. To mitigate overfitting, it incorporates five dropout layers with a rate of 0.1. The output from this encoder is then fed into a two-layer LSTM network, which uses a softmax activation funcxtion to perform classification.

We initialized our encoders with pre-trained weights provided by the authors, who pretrained Astromer using 1,529,386 R-band light curves from the Massive Compact Halo Object survey (MACHO; Alcock et al. 2000). Along with the classifier, we apply uncertainty quantification techniques to the encoder to enhance reliability in variable star classification. The model variants are further explained in subsequent sections.

2.2. Deep Ensembles

Using BNNs require significant computational resources, primarily due to the complex tuning necessary to achieve a consistent learning progress. In contrast, Lakshminarayanan et al. (2017) presented deep ensembles as a scalable and viable non-Bayesian alternative.

Consider a training dataset $\mathcal{D} = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{N}$, where each $\boldsymbol{x_i} \in \mathbb{R}^D$ is a D-dimensional feature vector. For a classification task, the labels y_i are assumed to be one of K classes, i.e., $y_i \in \{1, \ldots, K\}$. Now, consider a predictive model (such as a neural network) that, for a given input $\boldsymbol{x}$ outputs a prediction $\hat{y} = f_\theta(\boldsymbol{x})$, where $\boldsymbol{\theta}$ are the parameters of the model. When estimating uncertainties for a new data point $\boldsymbol{x}^*$, we aim at modeling the predictive distribution as:

$$p(y^*|\boldsymbol{x}^*, \mathcal{D}) = \int p(y^*|\boldsymbol{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\theta. \tag{5}$$

This integral is often intractable and approximate inference is typically applied. Hence, one alternative for uncertainty estimation is the use of deep ensembles. In our work, the strategy used involves training T independent neural network models, each with parameters $\{\boldsymbol{\theta}_t\}_{t=1}^{T}$. The predictive outcome for a new test point $\boldsymbol{x^*}$ is estimated by averaging the outputs from all T models:

$$y^* \approx \frac{1}{T} \sum_{t=1}^{T} f_{\theta_t}(\boldsymbol{x}^*). \tag{6}$$

Although this method is computationally intensive, it is easier to tune and can yield good performance while capturing predictive uncertainty.

We use DEs as the baseline uncertainty estimation technique. To calculate statistics over our uncertainty estimates, we trained ten ensembles, each consisting of ten deterministic models. These models were independently trained with different random seeds, which are used to initialize model parameters to different starting values. Additionally, we trained each model with different training chunks of the complete dataset. This approach introduces variability in the learning process, promoting diverse model behaviors and mitigating overfitting, thereby enhancing the robustness of our uncertainty estimates. The procedure was conducted for each survey.

2.3. Monte Carlo Dropout (MC Dropout)

Dropout is a regularization technique used to prevent overfitting by deactivating neurons during the training stage (Srivastava et al. 2014). While initially aimed at mitigating overfitting, Gal & Ghahramani (2016) provided a theoretical framework for its application in uncertainty quantification. They demonstrated that Monte Carlo Dropout works as an approximation of Bayesian inference in deep Gaussian processes. This is achieved by sampling neuron activations from a Bernoulli distribution across all hidden layers of the neural network during both training and inference stages.

MC Dropout uses dropout during inference to approximate the predictive distribution of Eq. 5. This adaptation allows MC Dropout to mimic an ensemble of diverse models. Each stochastic pass temporarily deactivates a random subset of neurons, and the expected value of the output y^* for a given input $\boldsymbol{x}^*$ is calculated by averaging the outputs from the model $f_{\theta_t}(\boldsymbol{x}^*)$ over all samples, as shown in Eq. 6.

Consider the neural network layer i, which receives the output vector $\boldsymbol{x}_{i-1}$ from the preceding layer as its input. When dropout is applied with probability p, the output of layer i can be defined as:

$$\boldsymbol{x}_i = \sigma(\boldsymbol{x}_{i-1}|\boldsymbol{W}_i, \boldsymbol{M}_i), \tag{7}$$

where σ represents the activation function, $\boldsymbol{W}_i$ denotes the weights of layer i, and $\boldsymbol{M}_i$ are the variational parameters that modulate these weights during training and inference. The weights $\boldsymbol{W}_i$ are given by:

$$\boldsymbol{W}_i = \boldsymbol{M}_i \cdot \mathrm{diag}([z_{i,j}]_{j=1}^{K_i}), \qquad z_{i,j} \sim Bernoulli(1 - p_i), \tag{8}$$

here $z_{i,j} = 0$ indicates that the neuron j in layer $i - 1$ is dropped as an input to layer i.

This method offers advantages over DEs as it avoids the need to train multiple models while retaining the ability to estimate uncertainties by using the variability of subnetworks within a single model. The authors proposed applying this method to all hidden neural network layers. In adapting MC Dropout for transformers, following the methodology described by Shelmanov et al. (2021), we applied dropout to all layers in the model, including those within the attention blocks. This strategy captures uncertainty at multiple levels of abstraction by introducing stochasticity throughout the entire network, ensuring that diverse model behaviors are accounted for, leading to more robust uncertainty estimates than if dropout were applied only to the output layer.

2.4. Hierarchical Stochastic Attention (HSA)

Pei et al. (2022) proposed Hierarchical Stochastic Attention, an approach for producing probabilistic outputs rather than deterministic ones by injecting stochasticity through the Gumbel-softmax distribution (Jang et al. 2017). This uncertainty estimation method was applied to NLP tasks, achieving competitive predictive performance while allowing for uncertainty estimation. HSA is composed of two hierarchical stochastic self-attention mechanisms: stochasticity over the self-attention heads and over a set of learnable centroids.

Stochastic self-attention replaces the traditional softmax activation function of the attention heads with the Gumbel-softmax distribution, which approximates samples from a categorical distribution with class probabilities $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ and associated logits values $\log(\theta_i)$. The Gumbel-softmax distribution considers a temperature τ, and samples $\tilde{\boldsymbol{y}} \sim \mathcal{G}(\boldsymbol{\theta}, \tau)$ are computed as

$$\tilde{y}_i \quad = \quad \frac{\exp\left(\left(\log(\theta_i) + g_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\left(\log(\theta_j) + g_j\right)/\tau\right)}, \tag{9}$$

where g_i represents i.i.d. samples from the Gumbel distribution, defined as $g_i = -\log(-\log(u))$, where u is uniformly drawn from the interval [0,1]. This approximation facilitates discrete choice sampling and gradient-based optimization needed on neural networks and that can not be performed using a categorical distribution. Furthermore, using the temperature τ, the Gumbel-softmax distribution can be smoothly annealed into a categorical distribution: as $\tau \to 0$, the Gumbel-softmax samples are identical to the categorical distribution.

In stochastic self-attention, logits are computed as the logarithm of the dot product between the query and the key vectors, and the attention weights are calculated during a forward pass as follows:

$$\hat{\alpha}_{i,j} = \frac{\exp\left(\left(\log(\boldsymbol{q}_i^\top \boldsymbol{k}_j) + g_{i,j}\right)/\tau\right)}{\sum_{l=1}^{L}\exp\left(\left(\log(\boldsymbol{q}_i^\top \boldsymbol{k}_l^\top) + g_{i,l}\right)/\tau\right)}, \quad (10)$$

$$\hat{\alpha}_i \sim \mathcal{G}(\boldsymbol{K}\boldsymbol{q}_i, \tau), \quad (11)$$

where $g_{i,j}$ represents i.i.d. samples from the Gumbel distribution. Here, the Gumbel-softmax distribution has a similar role to the traditional softmax by normalizing scores across all keys k, but allows to sample from it and estimate gradients efficiently.

The second component of HSA forces heads to pay stochastic attention to a set of centroids. Instead of directly attending to each key, HSA employs the Gumbel-softmax distribution to group keys around c learnable centroids $\boldsymbol{C} \in \mathbb{R}^{d_k \times c}$, where each centroid $\boldsymbol{c}_j$ represents the j-th column of $\boldsymbol{C}$ and matches the dimension of each key head. The model starts by stochastically paying attention to the centroids, and, then, a new set of keys $\{\tilde{\boldsymbol{k}}_i\}_{i=1}^{L}$ are calculated by weighting each centroid by this attention:

$$\tilde{\alpha}_i \sim \mathcal{G}\left(\boldsymbol{C}^\top \boldsymbol{k}_i, \tau\right), \quad (12)$$

$$\tilde{\boldsymbol{k}}_i = \sum_j \tilde{\alpha}_{i,j}\boldsymbol{c}_j. \quad (13)$$

Here, $g_{i,j}$ are again sampled from the Gumbel-softmax distribution. The new keys of Eq. 13 are used to calculate the stochastic attention weights of Equation 11 which are used to combine the values $\{\boldsymbol{v}_i\}_{i=1}^{L}$ as:

$$\hat{\alpha}_i \sim \mathcal{G}\left(\tilde{\boldsymbol{K}}\boldsymbol{q}_i, \tau\right), \quad (14)$$

$$\boldsymbol{h}_i = \sum_{j=1}^{L} \hat{\alpha}_{i,j}\boldsymbol{v}_j, \quad (15)$$

where $\tilde{\boldsymbol{k}}_i$ is represented as the i-th row of matrix $\tilde{\boldsymbol{K}} \in \mathbb{R}^{L \times d_k}$. Notice the hierarchical nature of HSA: the method employs the Gumbel-softmax distribution twice; initially to generate the new keys $\{\tilde{\boldsymbol{k}}_i\}_{i=1}^{L}$, and a second time to compute the outputs for each attention block $\{\boldsymbol{h}_i\}_{i=1}^{L}$.

We implemented the HSA approach in the Astromer attention mechanism, which captures predictive uncertainties through multiple forward passes during the inference stage, similar to the MC Dropout approach.

2.5. Hierarchical Attention with MC Dropout (HA-MC)

We propose a combination of MC Dropout and the Hierarchical Attention strategy for quantifying uncertainty in Astromer. This approach incorporates learnable centroids $\boldsymbol{c}_i$ in the attention blocks without applying the Gumbel-softmax distribution, but instead adding MC Dropout to the attention blocks. This allows us to achieve a regularization effect through the key head's attention to the centroids, while adding stochasticity without the use of the Gumbel-softmax distribution. The approach is defined as follows:

$$\tilde{\alpha}_i = \text{softmax}\left(\frac{\boldsymbol{C}^\top \boldsymbol{k}_i}{\sqrt{d_k}}\right), \quad (16)$$

$$\tilde{\boldsymbol{k}}_i = \sum_j \tilde{\alpha}_{i,j}\boldsymbol{c}_j, \quad (17)$$

where d_k is the original scaling factor, and $\boldsymbol{C}$ is the matrix with the centroids as columns from Eq. 13. The final attention scores are then computed using:

$$\hat{\alpha}_i = \text{softmax}\left(\frac{\tilde{\boldsymbol{K}}\boldsymbol{q}_i}{\sqrt{d_k}}\right), \quad (18)$$

$$\boldsymbol{h}_i = \sum_j \hat{\alpha}_{i,j}\boldsymbol{v}_j, \quad (19)$$

In this context, stochasticity is introduced by activating the dropout layers within each attention block during the inference stage.

2.6. Uncertainty Estimates

When a model is trained using only a maximum likelihood approach, the softmax activation function is prone to generate overconfident predictions (Guo et al. 2017). To explore and quantify the uncertainty inherent in our model's predictions, we employ uncertainty estimates (UEs). These estimates are not designed to assess the quality of the uncertainty quantification but to measure the extent and variability of uncertainty itself.

For the MC Dropout model variants, each UE is calculated by conducting T forward pass inference runs with dropout activated. In the case of HSA, the variation in the T inference runs is obtained through samples drawn from the Gumbel-softmax distribution. For the deep ensembles baseline, T corresponds to the number of independent models. Using these T inference runs, we calculate the following uncertainty estimates:

- Sampled Maximum Probability (SMP):

  $$1 - \max_{c \in C} \overline{p}(y = c|x), \quad (20)$$

  where $\overline{p}(y = c|x) = \frac{1}{T}\sum_{t=1}^{T} p_t(y = c|x)$ is the average probability of each class c across T forward passes for a given input x (being $p_t(y = c|x)$ the probability of class c at the t-th forward pass inference run). The SMP provides an intuitive measure of confidence by considering the maximum mean predicted probability across classes.
- Probability Variance (PV; Gal et al. 2017):

  $$\frac{1}{C}\sum_{c=1}^{C}\left(\frac{1}{T}\sum_{t=1}^{T}\left(p_t(y = c|x) - \overline{p}(y = c|x)\right)^2\right). \quad (21)$$

  This is the variance averaged over all C classes. PV assesses how consistently the model predicts the same class probabilities across different inference runs, providing an insight into the predictive stability.
- Bayesian Active Learning by Disagreement (BALD; Houlsby et al. 2011):

  $$-\sum_{c=1}^{C}\overline{p^c}\log(\overline{p^c}) + \frac{1}{T}\sum_{t=1}^{T}\sum_{c=1}^{C} p_t^c \log(p_t^c), \quad (22)$$

  where $\overline{p^c} = \overline{p}(y = c|x)$, and $p_t^c = p_t(y = c|x)$. The first term is the entropy of the average predictions across the ensemble of models, whereas the second term calculates the mean entropy of individual predictions from each model within the ensemble, reflecting the average uncertainty inherent in each model's predictions. BALD, unlike the other two uncertainty estimates, is entropy-based which can be interpreted as a measure of total uncertainty (Depeweg et al. 2018; Malinin & Gales 2018).

2.7. Evaluating uncertainty estimation via misclassification detection

A well-calibrated uncertainty estimation model should exhibit high uncertainty for incorrect predictions while maintaining low uncertainty for correct ones. To evaluate the performance of uncertainty estimation, we define a misclassification detection task. Although all of our models were trained as multiclass classifiers, we reformulate the evaluation during the testing phase as a binary classification problem focused on identifying misclassifications. This approach is aligned with the work done by Shelmanov et al. (2021) and Vazhentsev et al. (2022) for the NLP tasks.

Specifically, we construct new binary instances $\tilde{e}_i$ as follows:

$$\tilde{e}_i = \begin{cases} 1, & y_i \neq \hat{y}_i, \\ 0, & y_i = \hat{y}_i, \end{cases} \tag{23}$$

where y_i is the true label, and $\hat{y}_i$ is the original predicted label. The new instances $\tilde{e}_i$ indicate whether the model made a mistake in predicting the label of the variable source.

To measure the quality of UE, we compute the Receiver Operating Characteristic Area Under the Curve (ROC AUC) scores based on the binary labels $\tilde{e}_i$ and their corresponding UE scores. We use the UEs values as discrimination values to build the ROC curve, which plots the true positive rate (TPR) against the false positive rate (FPR) across various thresholds (Swets 1988). For a specific UE threshold, the TPR and FPR are calculated as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

and

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}},$$

where:

- TP are the number of instances that have an UE higher than the threslhold, and $\tilde{e}_i = 1$ (missclassified instances with high uncertainty),
- TN are the number of instances that have an UE lower than the threslhold, and $\tilde{e}_i = 0$ (correctly classified instances with low uncertainty),
- FP are the number of instances that have an UE higher than the threslhold, and $\tilde{e}_i = 0$ (correctly classified instances with high uncertainty),
- FN are the number of instances that have an UE lower than the threslhold, and $\tilde{e}_i = 1$ (missclassified instances with low uncertainty).

We evaluated statistical significance of our results using the Wilcoxon-Mund Whitney (WMW) test. The WMW test, a nonparametric alternative to the t-test, evaluates data based on ranks rather than assuming normality or equal variances. This test ranks all observations from two independent samples, X and Y, and calculates the test statistic, U, using the ranks from the smaller sample. The null hypothesis states that X and Y have identical distributions. Significant deviations in U suggests differing distributions, with statistical significance indicating disparities in central tendencies (Fay & Proschan 2010; Pett 2015).

3. Experimental Setup

In this section, we detail our experimental setup, where we evaluate the models using three labeled astronomical catalogs. We balanced the number of samples per class in small chunks for both the training and testing stages. The experiments were conducted on a Nvidia RTX A5000 GPU, our focus was on detecting misclassifications by converting the multiclass task into a binary problem, thus providing a clearly understanding of model performance under varied data scenarios.

3.1. Data and training

In this study, we compared the performance of baseline and the proposed models on three labeled catalogs of variable stars: the Optical Gravitational Lensing Experiment (OGLE-III; Udalski 2003), the Asteroid Terrestrial-impact Last Alert System (ATLAS; Heinze et al. 2018), and MACHO dataset. We considered the classification scheme and filtering methods previously selected by Becker et al. (2020) and utilized by Donoso-Oliva et al. (2023), as detailed in Table 1. These catalogs contain light curves observed through different spectral filters, offering a broad spectrum of data for analysis.

Preserving in-domain integrity during model testing was crucial to our methodology. This principle required evaluating models on data with distributional characteristics similar to those of the training set. Therefore, we avoided combining the test sets from the three catalogs during inference, despite some catalogs sharing classes. This approach highlights the importance of testing models in conditions that mirror their training environment. Although the OGLE-III and MACHO datasets share similar wavelength ranges, the distinct spectral band of the ATLAS catalog indicates the need for a meticulous in-domain evaluation approach.

To emulate scenarios with a small amount of data, we selected 500 samples per class for training and 100 samples per class for test sets from the raw data (see Table 2). A validation set was created by randomly selecting 30% of the training set.

We used ten ensembles for the baseline and ten variants models per approach. A a single test set per survey was used to compare the performance of the models. Consequently, we collected ten predictions for each approach, enabling us to calculate the mean and standard deviation of the samples to conduct significance testing.

For the optimization technique, we chose Adam (Kingma & Ba 2015) with a learning rate of 10^{-3}. The batch size was of 512, and as a regularization technique, we used Early Stopping with a patience of 20 epochs on the validation loss. We used this same hyperparameter settings for each experiment.

4. Results

4.1. Predictive performance

We evaluated the predictive performance using the macro-average of multiclass classification metrics over the three single-band datasets: MACHO, ATLAS and OGLE-III. Table 3 presents the test-sets F1 score, accuracy and precision of the HSA, MC Dropout, HA-MC Dropout and DEs methods.

The DEs baseline serves as a consistent benchmark, achieving macro F1 scores of 68.6/77.8/67.3 on MACHO, ATLAS, and OGLE-III, respectively. The MC Dropout method demonstrates a marginal performance improvement on the ATLAS test-set,

Table 1. Variable stars classes of each survey associated to the corresponding tag.

Dataset	Tag	Class
MACHO[(1)]	Cep 0	Cepheid type I
	Cep 1	Cepheid type II
	EC	Eclipsing binary
	LVP	Long period variable
	RRab	RR Lyrae type ab
	RRc	RR Lyrae type c
ATLAS[(2)]	CB	Close Binaries
	DB	Detached Binary
	Mira	Mira
	Pulse	RR Lyrae, δ-Scuti, Cepheids
OGLE-III[(3)]	EC	Eclipsing binary
	ED	Detached Binary
	ESD	Semi-detached Binary
	Mira	Mira
	OSARG	Small-amplitude red giant
	RRab	RRLyra type ab
	RRc	RRLyra type c
	SRV	Semi-regular variable
	Cep	Cepheid
	DSct	Delta Scuti

References. Classification scheme used by Donoso-Oliva et al. (2023) for the following datasets, listed in the order of their appearance: (1) Alcock et al. (2000); (2) Heinze et al. (2018); (3) Udalski (2003).

Table 2. Data distribution in terms of the number of light curves.

Dataset	Raw data	Training set	Test set
MACHO	21 444	3 000	600
ATLAS	4 719 921	2 000	400
OGLE-III	393 103	5 000	1 000

with a similar macro F1-score and accuracy of 78.8%, indicating limited gains over the baseline.

In contrast, the HSA method significantly improves performance across all datasets, even when tested on datasets with varying numbers of classes (6 classes in MACHO, 4 in ATLAS, and 10 in OGLE-III). This demonstrates its ability to perform well across varying levels of class complexity, reflecting its capacity to capture more intricate patterns. However, it exhibits more variability in terms of standard deviation compared to other approaches, with the variability in performance metrics close to 2%. This increased variability may be attributed to the stochasticity introduced by the Gumbel-softmax distribution, which can impact the predictive performance consistency across different runs. Consequently, while HSA improves overall performance, its stability across different runs may be less consistent than that of other approaches.

The proposed HA-MC Dropout significantly outperforms the other methods achieving higher scores across all metrics. HA-MC Dropout achieves macro F1-scores of 76.6/84.0/79.8 on MACHO, ATLAS, and OGLE-III, respectively, marking a substantial improvement in multiclass classification performance. Particularly in the dataset with 10 classes, OGLE-III, HA-MC Dropout obtained more than a 10% improvement compared to the baseline, with f1-score/accuracy/precision of 79.8/80.5/80.0. These results indicate that the integration of hierarchical attention mechanisms with MC Dropout not only enhances predictive accuracy but also provides a more reliable model with reduced variance in performance metrics. This may be explained by the way stochasticity is injected into the model: unlike HSA that relies on the Gumbel-softmax distribution, HA-MC Dropout utilizes the activation of dropout in the inference stage to estimate uncertainty.

Table 3. Summary of macro average multiclass metrics scores (%) on MACHO, ATLAS and OGLE-III test sets.

Method	Metric	MACHO	ATLAS	OGLE-III
Baseline	F1	68.6±1.7	77.8±2.6	67.3±2.9
	Accuracy	69.5±1.6	77.7±2.5	68.5±2.9
	Precision	69.3±1.7	78.1±2.6	68.6±2.8
MC Dropout	F1	68.0±0.6	78.8±0.7	66.8±1.6
	Accuracy	69.3±0.5	78.8±0.7	68.4±1.4
	Precision	68.4±0.6	79.3±0.8	67.4±1.6
HSA	F1	74.6±1.8	82.1±2.2	75.7±1.9
	Accuracy	75.3±1.8	82.2±2.1	76.7±1.7
	Precision	75.3±1.6	82.9±1.8	77.3±1.3
HA-MC Dropout	F1	**76.6±0.8**	**84.0±1.3**	**79.8±0.5**
	Accuracy	**77.5±0.7**	**84.0±1.3**	**80.5±0.4**
	Precision	**77.1±0.5**	**84.3±1.2**	**80.0±0.3**

Summarizing, two of the methods, HSA and HA-MC Dropout, surpass the DEs baseline. This indicates that combining hierarchical attention with a stochastic component yields improvements in multiclass classification tasks. The comparison presented in Table 3 represents the first stage of our analysis, focusing on the multiclass evaluation to establish a general understanding of the performance enhancements provided by these methods. However, the interpretability and trustworthiness of the results are further supported by the uncertainty estimation, which provides insights into the reliability of our classification outcomes. We extend our analysis by considering the impact of uncertainty estimation on misclassification task in Sect. 4.2, providing a more comprehensive evaluation of the robustness and reliability of the different methods in handling complex classification tasks.

4.2. Predictive uncertainty

We evaluate the methods in terms of the predictive uncertainty using the misclassification detection task, as described in Sec. 2.7, on the MACHO, ATLAS and OGLE-III test-sets. Table 4 shows the ROC AUC scores for the different UEs. In this context, the ideal classifier is one that aligns uncertainty estimates with the misclassification task: misclassified instances should be associated with high uncertainty, while correctly classified instances should correspond to low uncertainty. We compare the baseline DE method against MC Dropout, HSA, and HA-MC Dropout. The evaluation metric used is the absolute ROC AUC score, which quantifies the ability of the model to discern between missclassifications and correct classifications by using the UEs as discrimination scores. The uncertainty estimates used to calculate each ROC AUC are SMP, PV, and BALD. Specifically, for the baseline, the mean absolute ROC AUC is presented, whereas for the other methods, we report the performance differences relative to the baseline's corresponding uncertainty estimates, highlighting any statistically significant improvements (p-values ≤ 0.05). Note that the results are grouped by dataset. Standard deviations are reported to reflect the variability across ten model iterations. For the baseline, this includes results from ten independent ensemble runs, where each run consists of ten separately trained deterministic models. For MC Dropout, HSA, and HA-MC Dropout, the results are based on ten independently trained models, with $T = 10$ stochastic inference runs performed per object for each model.

The DEs consistently achieve an average ROC AUC exceeding 70% across all uncertainty estimates and datasets. This performance highlights their capability to identify potential errors through probabilistic outputs. MC Dropout excels in capturing predictive uncertainty when using PV and BALD scores for ROC AUC calculation, surpassing the baseline in misclassification detection tasks across all datasets. This is specially noticeable on the OGLE-III dataset, where the incremental percentage differences in ROC AUC relative to the baseline are 4.3 ± 1.6% and 4.2 ± 1.6%, respectively.

Table 4. Uncertainty estimates in the misclassification task on the MACHO, ATLAS and OGLE-III test sets.

Method	UEs	MACHO	ATLAS	OGLE-III
Baseline	SMP	75.6±1.6	85.9±2.1	82.2±1.1
	PV	71.4±2.3	82.0±2.2	73.4±1.7
	BALD	70.0±2.6	80.8±2.4	74.1±1.8
MC Dropout	SMP	0.3±2.0	0.2±1.5	0.0±1.0
	PV	**1.4±2.4**	**1.5±2.3**	**4.3±1.6**
	BALD	**1.8±2.4**	**1.8±2.6**	**4.2±1.6**
HSA	SMP	-1.5±1.8	-1.2±2.1	-0.1±1.0
	PV	0.5±2.4	**2.6±2.2**	**5.9±1.5**
	BALD	-0.3±3.1	**2.3±2.5**	**2.9±1.3**
HA-MC Dropout	SMP	0.3±1.8	0.6±1.8	**2.1±1.2**
	PV	**2.5±2.3**	**3.3±2.1**	**8.5±1.6**
	BALD	**2.1±2.7**	**3.8±2.4**	**6.9±1.6**

Notes. The baseline presents the mean and standard deviation of the absolute ROC AUC scores (%). For all other methods, the values indicate the difference in ROC AUC relative to the baseline, calculated for each uncertainty estimate. Statistically significant improvements (p-values $\leq$ 0.05) over the baseline for the corresponding UE are highlighted in bold font.

Conversely, HSA registers significant improvements with PV and BALD on the ATLAS and OGLE-III datasets. However, it fails to demonstrate an enhancement on the MACHO test set with any of the uncertainty estimates provided. In the case of SMP scores, noticeable differences between the baseline arise where negative values indicate a reduction relative to the baseline.

To address these challenges, we introduced the HA-MC Dropout method, which, as aforementioned, combines the strengths of both MC Dropout and HSA. This hybrid approach significantly outperforms both individual methods, especially in SMP scores, where it rivals or even surpasses the baseline. For instance, it achieves an improvement of 2.1 ± 1.2% on MACHO. The most substantial improvements are observed with PV on OGLE-III, where it reaches 8.5 ± 1.6%, doubling the improvement achieved by MC Dropout. Additionally, with PV, the improvement on MACHO is 2.5 ± 2.3%, and using BALD yields 3.8 ± 2.4% on ATLAS. Consequently, HA-MC Dropout not only mitigates the limitations of its constituent methods but also establishes a new benchmark for balancing predictive accuracy and uncertainty quantification across diverse datasets.

4.2.1. Accuracy-rejection plots

We present a practical application of our misclassification framework by using the accuracy-rejection plots (Nadeem et al. 2009) for MACHO, ATLAS and OGLE-III test sets, as illustrated in Figures 1 (a), (b) and (c). These plots emulate a scenario reflecting a hybrid machine-human behavior, wherein the machine abstains from classifying the most uncertain samples. This approach allows us to visualize accuracy as a function of the rejection rate. In all figures, confidence levels were assessed using the PV score, and the shaded areas represent the standard deviation across ten iterations of each approach. Additionally, each plot includes a zoomed inset focusing on the rejection rate interval from 0.1 to 0.3, providing a detailed comparison of the various methods within this specific range.

As a guideline, in Fig. 1 (a), the accuracy-rejection plot for the MACHO test set suggests that maintaining an accuracy threshold above 80% requires a rejection rate of ∼ 0.15 for MC Dropout and ∼ 0.1 for the DE baseline and HSA. HA-MC Dropout is able to keep a 80% accuracy by rejecting less than the ∼ 5% most uncertain predicted labels in the MACHO dataset. Notably, the HA-MC Dropout method surpasses both the baseline and other techniques for every rejection rate, highlighting its potential. Although HSA shows a marginally higher mean accuracy compared to the baseline, the baseline demonstrates lower variability at a rejection rate of 0.2, indicating higher consistency. Meanwhile, MC Dropout aligns closely with the baseline performance until a rejection rate of ∼ 0.4.

Figure 1 (b) presents the accuracy-rejection plot for the ATLAS test set. Below a rejection rate of ∼ 0.2, HA-MC Dropout shows a performance comparable to the baseline. However, at a rejection rate of ∼ 0.2, the HA-MC Dropout method achieves a mean accuracy of approximately 0.93, surpassing the baseline by about 0.02 points, demonstrating its superior performance.

Finally, Fig. 1 (c) details the accuracy-rejection scenario for the OGLE-III dataset, indicating that both HA-MC Dropout and HSA outperform the baseline, with HA-MC Dropout needing a 20% rejection rate to achieve a mean accuracy score of 0.90. HSA follows closely with an accuracy of 0.87, while the baseline achieves 0.84 and MC Dropout 0.82. This scenario underscores the necessity for expert intervention to maintain high accuracy levels, demonstrating that HA-MC Dropout achieves superior results with minimal expert involvement.

The analysis of these plots presents that, across all datasets evaluated, the performance curves of the MC Dropout approach align with the baseline model for a rejection rate higher than ∼ 0.4. This consistency at a specified threshold highlights the capability of the MC Dropout to maintain accuracy while also providing estimates of uncertainty. Conversely, the HA-MC Dropout method is a better option to other methods in all datasets. Despite the baseline showing lower standard deviation in the OGLE and MACHO datasets, the overall accuracy score consistency across different astronomical datasets highlights the resilience of HA-MC Dropout. This robust performance affirms the cost-efficiency of implementing HA-MC Dropout for estimating uncertainty on

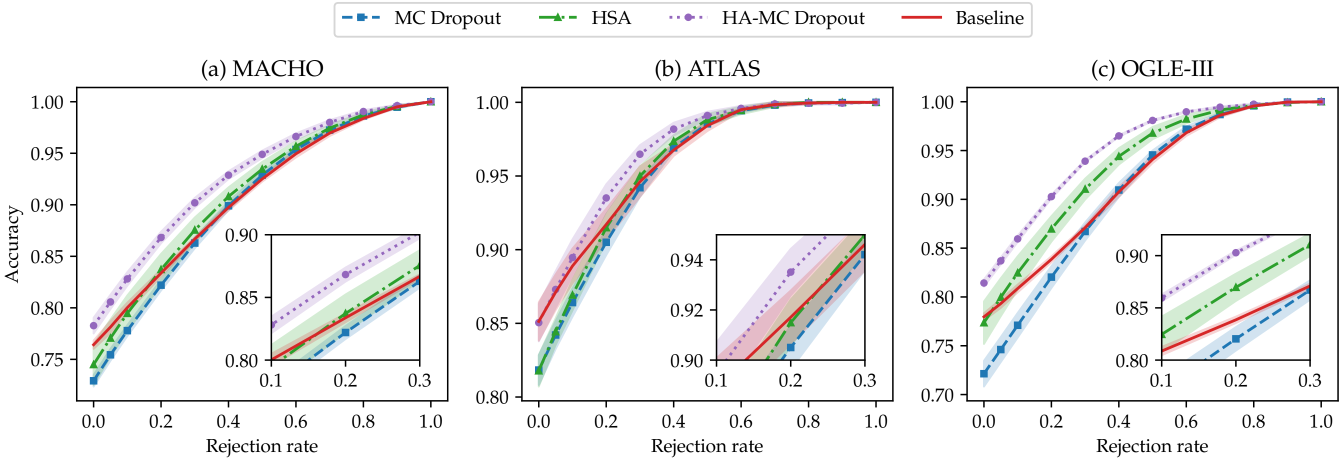

Fig. 1. Figures (a), (b), and (c) presents the accuracy-rejection curves for the MACHO, ATLAS, and OGLE-III datasets, respectively. Techniques compared include MC Dropout (dash-squared line), HSA (dash-triangular line), HA-MC Dropout (dash-dotted line), and the baseline model (solid line). Insets zoom into the lower rejection rate region (0-0.3) to emphasize differences at low rejection levels.

transformer-based classifiers, making them well-suited for real-world applications.

5. Discussion and Conclusions

We have investigated the application of uncertainty estimation techniques to enhance the reliability and interpretability of transformer-based models for light curve classification in the context of variable star analysis. By implementing and evaluating deep ensembles, Monte Carlo Dropout, Hierarchical Stochastic Attention, and our proposed hybrid method, HA-MC Dropout in Astromer, we have demonstrated the potential of these techniques in capturing predictive uncertainty and improving misclassification detection.

Our empirical results highlight that HA-MC Dropout consistently outperforms other methods in terms of predictive accuracy and uncertainty estimation across various datasets. This suggests that integrating hierarchical attention mechanisms with Monte Carlo Dropout offers a powerful approach for enhancing the robustness and reliability of transformer-based models in complex classification tasks. The superior performance of HA-MC Dropout, particularly in scenarios with limited data, highlights its potential for real-world applications in scenarios with limited data and class distribution challenges.

The accuracy-rejection plots provide valuable insights into the practical implications of our work. This plots demonstrate that HA-MC Dropout enables the model to achieve higher accuracy levels with less rejected samples, showcasing its potential for automating the classification process while maintaining high confidence in the results. The consistent performance of MC Dropout across different datasets further reinforces its value as a viable alternative to the other approaches. Hence, it offers a computationally efficient and effective method for uncertainty estimation in transformer-based models.

The findings of this study have significant implications for the future of variable star classification, particularly in the era of next-generation large-scale astronomical surveys such as the LSST. The ability to quantify uncertainty and detect misclassifications will be crucial in ensuring the reliability and interpretability of automated classification systems. The work presented here offers a promising step towards achieving this goal, paving the way for more robust and trustworthy analysis of astronomical light curves.

Future research directions include exploring the application of our work into multi-band light curves model (e.g, Cabrera-Vives et al. 2024). Additionally, considering the impact of different data preprocessing and augmentation strategies on uncertainty estimation could provide valuable insights into improving the performance of the transformer-based model in challenging scenarios. Human feedback for objects classified by the model with low certainty can also be added into a human-in-the-loop framework (see e.g. Richards et al. 2011; Masci et al. 2014; Martínez-Palomera et al. 2018; Ishida et al. 2019; Kennamer et al. 2020; Leoni et al. 2022).

Summarizing, HA-MC Dropout has proven to be competitive against the DEs baseline in three astronomical datasets with different variable star taxonomies. Transformer-based models have established their status as the state-of-the-art across various fields. We emphasize the significance of developing reliable models that can reduce computational expenses when being trained: DEs need to train multiple models, while the MC Dropout strategy uses a single trained model. We believe that the capacity to accurately assess uncertainty can economize human labor while also enhancing confidence in the conclusions derived from these models.

Acknowledgements. The authors acknowledge support from the National Agency for Research and Development (ANID) grants: FONDECYT regular 1231877 (MCL, GCV, DMC, CDO); Millennium Science Initiative Program – NCN2021 080 (GCV, CDO) and ICN12 009 (GCV, CDO, IB).

References

Alcock, C., Allsman, R., Alves, D. R., et al. 2000, The Astrophysical Journal, 542, 281

Bassi, S., Sharma, K., & Gomekar, A. 2021, Frontiers in Astronomy and Space Sciences, 8, 718139

Becker, I., Pichara, K., Catelan, M., et al. 2020, Monthly Notices of the Royal Astronomical Society, 493, 2981

Becker, I., Pichara, K., Catelan, M., et al. 2020, MNRAS, 493, 2981

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. 2015, in International conference on machine learning, PMLR, 1613–1622

Cabrera-Vives, G., Moreno-Cartagena, D., Astorga, N., et al. 2024, Astronomy & Astrophysics, 689, A289

Cabrera-Vives, G., Reyes, I., Förster, F., Estévez, P. A., & Maureira, J.-C. 2016, in 2016 International Joint Conference on Neural Networks (IJCNN), 251–258

Carrasco-Davis, R., Cabrera-Vives, G., Förster, F., et al. 2019, Publications of the Astronomical Society of the Pacific, 131, 108006

Ciucă, I., Kawata, D., Miglio, A., Davies, G. R., & Grand, R. J. 2021, Monthly Notices of the Royal Astronomical Society, 503, 2814

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., & Udluft, S. 2018, in International Conference on Machine Learning, PMLR, 1184–1193

Devlin, J., Chang, M.-W., & Lee, K. 2019, in Proceedings of NAACL-HLT, 4171–4186

Donoso-Oliva, C., Becker, I., Protopapas, P., et al. 2023, Astronomy & Astrophysics, 670, A54

Donoso-Oliva, C., Cabrera-Vives, G., Protopapas, P., Carrasco-Davis, R., & Estévez, P. A. 2021, Monthly Notices of the Royal Astronomical Society, 505, 6069

Fay, M. P. & Proschan, M. A. 2010, Statistics surveys, 4, 1

Feast, M. W., Menzies, J. W., Matsunaga, N., & Whitelock, P. A. 2014, Nature, 509, 342

Gal, Y. & Ghahramani, Z. 2016, in international conference on machine learning, PMLR, 1050–1059

Gal, Y., Islam, R., & Ghahramani, Z. 2017, in International conference on machine learning, PMLR, 1183–1192

Ganaie, M. A., Hu, M., Malik, A., Tanveer, M., & Suganthan, P. 2022, Engineering Applications of Artificial Intelligence, 115, 105151

Gawlikowski, J., Tassi, C. R. N., Ali, M., et al. 2023, Artificial Intelligence Review, 56, 1513

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. 2017, in International conference on machine learning, PMLR, 1321–1330

Heinze, A., Tonry, J. L., Denneau, L., et al. 2018, The Astronomical Journal, 156, 241

Hochreiter, S. & Schmidhuber, J. 1997, Neural computation, 9, 1735

Houlsby, N., Huszár, F., Ghahramani, Z., & Lengyel, M. 2011, arXiv preprint arXiv:1112.5745

Ishida, E., Beck, R., González-Gaitán, S., et al. 2019, Monthly Notices of the Royal Astronomical Society, 483, 2

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, The Astrophysical Journal, 873, 111

Jang, E., Gu, S., & Poole, B. 2017, in 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings (OpenReview.net)

Karpenka, N. V., Feroz, F., & Hobson, M. 2013, Monthly Notices of the Royal Astronomical Society, 429, 1278

Kennamer, N., Ishida, E. E., González-Gaitán, S., et al. 2020, in 2020 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 3115–3124

Killestein, T., Lyman, J., Steeghs, D., et al. 2021, Monthly Notices of the Royal Astronomical Society, 503, 4838

Kingma, D. P. & Ba, J. 2015, in 3rd International Conference on Learning Representations (ICLR), San Diego, California, United States

Lakshminarayanan, B., Pritzel, A., & Blundell, C. 2017, Advances in neural information processing systems, 30

Leoni, M., Ishida, E. E., Peloton, J., & Möller, A. 2022, Astronomy & Astrophysics, 663, A13

Leung, H. W. & Bovy, J. 2023, Monthly Notices of the Royal Astronomical Society, stad3015

Mahabal, A., Sheth, K., Gieseke, F., et al. 2017, in 2017 IEEE symposium series on computational intelligence (SSCI), IEEE, 1–8

Malinin, A. & Gales, M. 2018, in Advances in Neural Information Processing Systems, ed. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett, Vol. 31 (Curran Associates, Inc.)

Martínez-Palomera, J., Förster, F., Protopapas, P., et al. 2018, The Astronomical Journal, 156, 186

Masci, F. J., Hoffman, D. I., Grillmair, C. J., & Cutri, R. M. 2014, The Astronomical Journal, 148, 21

Möller, A. & de Boissière, T. 2020, Monthly Notices of the Royal Astronomical Society, 491, 4277

Moreno-Cartagena, D., Cabrera-Vives, G., Protopapas, P., et al. 2023, in Machine Learning for Astrophysics Workshop, 40th International Conference on Machine Learning (ICML), PMLR 202, Honolulu, Hawaii, USA

Morvan, M., Nikolaou, N., Yip, K., & Waldmann, I. 2022, Machine Learning for Astrophysics, 11

Nadeem, M. S. A., Zucker, J.-D., & Hanczar, B. 2009, in Machine Learning in Systems Biology, PMLR, 65–81

Naul, B., Bloom, J. S., Pérez, F., & van der Walt, S. 2018, Nature Astronomy, 2, 151

Ngeow, C.-C. 2015, Publications of The Korean Astronomical Society, 30, 371

Pan, J.-S., Ting, Y.-S., & Yu, J. 2024, Monthly Notices of the Royal Astronomical Society, 528, 5890

Park, J. W., Villar, A., Li, Y., et al. 2021, in Uncertainty and Robustness in Deep Learning Workshop, 38th International Conference on Machine Learning (ICML), PMLR 139

Parker, L., Lanusse, F., Golkar, S., et al. 2024, Monthly Notices of the Royal Astronomical Society, 531, 4990

Pei, J., Wang, C., & Szarvas, G. 2022, Proceedings of the AAAI Conference on Artificial Intelligence, 36, 11147

Pett, M. A. 2015, Nonparametric statistics for health care research: Statistics for small samples and unusual distributions (Sage Publications)

Pimentel, Ó., Estévez, P. A., & Förster, F. 2022, The Astronomical Journal, 165, 18

Protopapas, P. 2017, in American Astronomical Society Meeting Abstracts# 230, Vol. 230, 104–03

Richards, J. W., Starr, D. L., Brink, H., et al. 2011, The Astrophysical Journal, 744, 192

Shelmanov, A., Tsymbalov, E., Puzyrev, D., et al. 2021, in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 1833–1840

Smith, M. J. & Geach, J. E. 2023, Royal Society Open Science, 10, 221454

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. 2014, The journal of machine learning research, 15, 1929

Swets, J. A. 1988, Science, 240, 1285

Udalski, A. 2003, Acta Astron, 53, 291

Valdenegro-Toro, M. 2019, in Bayesian Deep Learning Workshop, 4th Advances in Neural Information Processing Systems (NeurIPS), Vol. 32, Vancouver, Canada

Vaswani, A., Shazeer, N., Parmar, N., et al. 2017, Advances in neural information processing systems, 30

Vazhentsev, A., Kuzmin, G., Shelmanov, A., et al. 2022, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 8237–8252