FORECASTING SOLAR FLARES WITH THE EVEREST TRANSFORMER

Antanas Žilinskas
Imperial College London
antanas.zilinskas.research@gmail.com

Robert N. Shorten
Imperial College London
r.shorten@imperial.ac.uk

Jakub Mareček
Czech Technical University in Prague, Faculty of Electrical Engineering
jakub.marecek@fel.cvut.cz

ABSTRACT

Solar flare forecasting is a rare-event time-series problem characterised by severe class imbalance, long-range temporal dependencies, and the need for calibrated probabilities and tail-aware behaviour. We present EVEREST, a compact Transformer forecaster trained with auxiliary objectives that improve calibration and tail sensitivity while retaining a single-head inference path. EVEREST integrates (i) a single-query attention bottleneck, (ii) an evidential Normal–Inverse–Gamma head on logits, (iii) an extreme-value head based on Generalized Pareto exceedances, and (iv) a lightweight precursor head for anticipatory supervision. On SHARP–GOES, EVEREST achieves strong True Skill Statistic (TSS) and low Expected Calibration Error (ECE) without inference overhead.

1 INTRODUCTION

Rare, high-impact events in multivariate time series arise in domains such as space weather and industrial monitoring. Learning to forecast these events is difficult because (i) positives are extremely rare, (ii) weak precursors are distributed across long contexts, and (iii) operational pipelines require calibrated probabilities and tail-aware behaviour at high alert thresholds.

We present EVEREST, a compact Transformer forecaster trained with *training-only* auxiliary objectives that regularise calibration and tail sensitivity while keeping inference unchanged. The model uses a single-query attention bottleneck to aggregate long contexts, and three auxiliary heads used only during training: an evidential Normal–Inverse–Gamma (NIG) head on logits, an EVT head that fits a Generalized Pareto distribution (GPD) to logit exceedances, and a lightweight precursor head providing anticipatory supervision. At test time, EVEREST uses only a standard classification head, so deployment cost matches a vanilla Transformer of comparable size.

On SHARP–GOES solar-flare forecasting, EVEREST achieves strong skill across horizons and flare thresholds (e.g., TSS $0.973/0.970/0.966$ for $\geq$C at 24/48/72 h; TSS $0.907/0.936/0.966$ for $\geq$M5) while maintaining low calibration error (e.g., M5–72 h ECE $= 0.016$). We also report cross-domain transfer to SKAB using the same EVEREST design and objectives.

2 RELATED WORK

Time-series Transformers. Transformers are widely used for time series, with recent work improving long-context modelling via patching (Nie et al., 2023), frequency/decomposition modules (Zhou et al., 2022), and modern convolutional alternatives (Luo & Wang, 2024). EVEREST uses a lightweight single-query attention bottleneck for task-conditioned temporal aggregation.

Solar-flare forecasting. Prior SHARP–GOES forecasting pipelines span recurrence-based predictors, convolutional models, and flare-specific transformer hybrids (Liu et al., 2019; Sun et al.,

2022; Abduallah et al., 2023). These systems provide the direct published baselines in our headline comparison.

Calibration and uncertainty. Calibration is essential for high-stakes forecasting. Post-hoc temperature scaling improves marginal calibration (Guo et al., 2017), while ensembles and deterministic uncertainty surrogates can better capture epistemic effects at higher cost (Lakshminarayanan et al., 2017; van Amersfoort et al., 2020). Evidential learning provides closed-form uncertainty estimates without sampling (Sensoy et al., 2018; Amini et al., 2020); we use an NIG head on the logit as a training-time regulariser.

Tail risk and EVT. Extreme value theory models distribution tails via peaks-over-threshold and Generalized Pareto exceedances (Coles, 2001; de Haan & Ferreira, 2006). We adapt this idea as an *auxiliary training loss* on logit exceedances to focus learning signal on high-risk predictions.

3 METHOD

Task. Each example is a window $X \in \mathbb{R}^{T \times F}$ with label $y \in \{0, 1\}$ indicating whether an event occurs within a fixed horizon. The model outputs a logit l and probability $\hat{p} = \sigma(l)$; alerts use a threshold τ.

Architecture. EVEREST is a compact Transformer encoder followed by a single-query attention bottleneck that pools hidden states into a vector z. A small MLP maps z to four heads: one primary classification head (used at inference) and three *training-only* auxiliaries (evidential, EVT, precursor). At inference, auxiliary heads are discarded and the prediction is computed from the primary classification head only, so runtime cost is unchanged. In the reference configuration, EVEREST has approximately 0.81M parameters; we use $d = 128$ embeddings, $L = 6$ encoder layers, $H = 4$ attention heads, FFN width 256, and dropout $p = 0.20$.

Single-query attention bottleneck. Let h_t denote the final encoder state at time t. We compute

$$\alpha_t = \mathrm{softmax}_t(w^\top h_t), \qquad z = \sum_{t=1}^{T} \alpha_t h_t,$$

where w is learned. This concentrates capacity on informative precursors while adding only $\mathcal{O}(Td)$ compute.

Training objective (auxiliaries only at training time). We optimise

$$\mathcal{L} = \lambda_f \mathcal{L}_{\text{focal}} + \lambda_e \mathcal{L}_{\text{evid}} + \lambda_t \mathcal{L}_{\text{evt}} + \lambda_p \mathcal{L}_{\text{prec}}.$$

$\mathcal{L}_{\text{focal}}$ addresses imbalance. With focusing parameter γ, the focal loss is

$$\mathcal{L}_{\text{focal}} = -\tfrac{1}{N} \sum_i \Big[(1 - \hat{p}_i)^\gamma y_i \log \hat{p}_i + \hat{p}_i^\gamma (1 - y_i) \log(1 - \hat{p}_i) \Big],$$

and we anneal γ from $0 \to 2$ over the first 50 epochs. $\mathcal{L}_{\text{evid}}$ predicts Normal–Inverse–Gamma parameters (μ, v, α, β) over the scalar logit and minimises the closed-form evidential objective, acting as a training-time regulariser of logit uncertainty. $\mathcal{L}_{\text{evt}}$ predicts GPD parameters (ξ, σ) and maximises a GPD likelihood on logit exceedances above a high *batchwise* quantile u (90% by default), fitting a GPD to exceedances $\{l_i - u : l_i > u\}$. The precursor head is a second binary classifier branching from z; $\mathcal{L}_{\text{prec}}$ applies binary cross-entropy with the same label y, providing anticipatory supervision during training. Only the *relative ratios* of λ values matter (scale invariance of the composite objective); in the reference configuration we use $(\lambda_f, \lambda_e, \lambda_t, \lambda_p) = (0.8, 0.1, 0.1, 0.05)$. At test time we output $\hat{p}$ from the classification head only. Appendix A.1 summarises the auxiliary heads and Appendix A.5 reports their empirical effects.

Computational footprint. At the reference configuration, EVEREST has approximately 8.14×10^5 parameters and 1.66×10^7 FLOPs per window; the six-layer backbone accounts for more than 97% of both.

Table 1: TSS comparison on SHARP–GOES. Published baseline values are taken from the cited papers; EVEREST values are mean (s.d.) over five seeds.

Method	Horizon	≥C	≥M	≥M5
Liu et al. (2019)	24h	0.612	0.792	0.881
Sun et al. (2022)	24h	0.756	0.826	–
Abduallah et al. (2023)	24h	0.835	0.839	0.818
	48h	0.719	0.728	0.736
	72h	0.702	0.714	0.729
EVEREST	24h	**0.973** (0.001)	**0.898** (0.011)	**0.907** (0.025)
	48h	**0.970** (0.001)	**0.920** (0.007)	**0.936** (0.021)
	72h	**0.966** (0.001)	**0.906** (0.012)	**0.966** (0.024)

4 Experimental Setup

Solar flares (SHARP–GOES). We follow the SHARP–GOES protocol (Abduallah et al., 2023): SHARP parameters are aligned to GOES flare labels with standard quality masks (e.g., $|\mathrm{CMD}| \leq 70^\circ$ and `QUALITY`=0) applied before windowing. We use the same nine SHARP parameters and window construction for 24/48/72 h horizons. To prevent leakage, we use the same `HARPNUM`-stratified train/validation/test split throughout; train and held-out test counts for all nine tasks are reported in Appendix A.2 (Table 2). All preprocessing (normalization, cadence handling, and label alignment) follows that setup to ensure 1:1 comparability.

SKAB transfer. For cross-domain transfer we also evaluate on the Skoltech Anomaly Benchmark (SKAB) (Filonov et al., 2020) using fixed-length windows, chronological splits, and the same training recipe. The transfer protocol and baseline comparison are summarised in Appendix A.6.

Metrics and thresholds. We report discrimination via True Skill Statistic (TSS), and probabilistic accuracy via Brier score and Expected Calibration Error (ECE; 15 equal-frequency bins). For completeness, TSS is defined as

$$\mathrm{TSS} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} - \frac{\mathrm{FP}}{\mathrm{FP+TN}}.$$

Operating thresholds are selected on the validation split by grid search over $\tau \in \{0.10, \ldots, 0.90\}$ (step 0.01) using a balanced score (40% TSS, 20% F1, 15% precision, 15% recall, 10% specificity), and then kept fixed for test evaluation. For operational sensitivity to asymmetric costs, we also evaluate a cost–loss sweep, but omit it here for brevity.

Statistical protocol. We train five random seeds per task and report mean performance with 95% confidence intervals computed via a 10^4-draw bootstrap on the held-out test set, stratified by active-region identifier to prevent leakage across active regions.

Training setup. We train in PyTorch using automatic mixed precision (AMP), AdamW, cosine-decayed learning rate, and gradient-norm clipping (1.0), with the focal schedule and composite objective from §3. The reference loss weights are $\boldsymbol{\lambda} = (0.8, 0.1, 0.1, 0.05)$.

5 Results

SHARP–GOES headline performance. EVEREST achieves strong skill across horizons and flare thresholds on SHARP–GOES, including TSS $0.973/0.970/0.966$ for ≥C at 24/48/72 h and TSS $0.907/0.936/0.966$ for ≥M5. For context, Table 1 compares against published SHARP–GOES forecasters spanning recurrent models, convolutional architectures, and flare-specific networks (Liu et al., 2019; Sun et al., 2022; Abduallah et al., 2023). Full bootstrapped metrics, held-out split counts, and contextual baseline deltas are reported in Appendix A.3.

Calibration and tail behaviour. Alongside strong TSS, EVEREST maintains low calibration error, measured by Brier score and Expected Calibration Error (ECE; 15 equal-frequency bins). On

the most imbalanced task (M5–72 h) we obtain ECE $= 0.016$, indicating that predicted probabilities closely track empirical frequencies under the same evaluation protocol used for the headline metrics. These results support the use of training-time evidential and EVT regularisation to improve reliability in rare-event, high-risk regimes while retaining a single-head inference path. Consistent with their intended roles, ablations attribute gains primarily to temporal focusing (attention bottleneck), tail emphasis (EVT exceedance regularisation), and improved calibration (evidential logit regularisation), with the precursor auxiliary shaping anticipatory representations while remaining unused at inference. Appendix A.4 shows a representative reliability diagram and Appendix A.5 reports the leave-one-component-out study.

Transfer to SKAB. Using the same EVEREST design and objectives, EVEREST reaches mean TSS $= 0.964$ and F1 $= 98.16\%$ on SKAB, suggesting the recipe is not limited to a single scientific domain. Appendix A.6 summarises the transfer protocol, confidence intervals, and comparison against published SKAB baselines.

6 Conclusion

EVEREST is a compact Transformer forecaster for rare-event time series that improves calibration and tail sensitivity through *training-only* auxiliary objectives, while retaining a single-head inference path. On SHARP–GOES, it achieves strong TSS with low ECE, and it transfers to SKAB using the same EVEREST design and objectives. Overall, these results suggest that tail-aware and evidential regularisation can make Transformer time-series models more decision-relevant in high-stakes rare-event settings without increasing deployment cost.

References

Y. Abduallah, X. Wang, W. Xu, B. Zhang, Y. Zheng, and S. E. Gibson. Operational prediction of solar flares using a transformer-based framework. *Scientific Reports*, 13(1):13665, 2023. doi: 10.1038/s41598-023-40884-1. URL https://www.nature.com/articles/s41598-023-40884-1.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14927–14937. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/aab085461de182608ee9f607f3f7d18f-Paper.pdf.

Jean Paul A. Audibert, Pietro Michiardi, Frédéric Guyard, Stéphane Marti, and Maria A. Zuluaga. USAD: UnSupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*, pp. 3395–3404, New York, NY, USA, 2020. ACM. doi: 10.1145/3394486.3403392.

Md Abul Bashar and Richi Nayak. Tanogan: Time series anomaly detection with generative adversarial networks. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1778–1785, 2020. doi: 10.1109/SSCI47803.2020.9308512. URL https://arxiv.org/abs/2008.09567. Also available as arXiv:2008.09567.

Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. Springer, 2006.

Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. Deeplog: Anomaly detection and diagnosis from system logs through deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)*, pp. 1285–1298, New York, NY, USA, 2017. ACM. doi: 10.1145/3133956.3134015.

Pavel Filonov, Andrey Lavrentyev, and Andrey Vorontsov. Skab: Skoltech anomaly benchmark. https://github.com/waico/SKAB, 2020. GitHub repository.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. 2017. URL https://arxiv.org/abs/1706.04599.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Hui Liu, Chang Liu, J. T. L. Wang, and Haimin Wang. Predicting solar flares using a long short-term memory network. *The Astrophysical Journal*, 877(2):121, 2019. doi: 10.3847/1538-4357/ab1b3c.

Donghao Luo and Xue Wang. Moderntcn: A modern pure convolution structure for general time series analysis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vpJMJerXHU.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023. OpenReview: arXiv:2211.14730.

Daehyung Park, Yejin Hoshi, and Charles C. Kemp. A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder. *IEEE Robotics and Automation Letters*, 3(3):1544–1551, 2018. doi: 10.1109/LRA.2018.2801475.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. 2018. URL https://arxiv.org/abs/1806.01768.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*, pp. 1907–1915, New York, NY, USA, 2019. ACM. doi: 10.1145/3292500.3330672.

Pengchao Sun, Wei Dai, Weiqi Ding, Song Feng, Yanmei Cui, Bo Liang, Zeyin Dong, and Yunfei Yang. Solar flare forecast using 3d convolutional neural networks. *The Astrophysical Journal*, 941(1):1, 2022. doi: 10.3847/1538-4357/ac9e53.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 15 (6):1201–1214, 2022. doi: 10.14778/3514061.3514067. URL https://vldb.org/pvldb/vol15/p1201-tuli.pdf.

Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. 2020. URL https://arxiv.org/abs/2003.02037.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhou22g.html.

Table 2: SHARP–GOES train and held-out test counts for the nine forecast tasks.

Task	Train Pos	Train Neg	Test Pos	Test Neg
C-24h	244,968	218,217	31,897	15,878
C-48h	316,149	301,714	40,987	21,573
C-72h	356,219	350,853	46,066	25,663
M-24h	13,989	449,196	1,368	46,407
M-48h	16,709	601,154	1,775	60,785
M-72h	18,505	688,567	2,131	69,598
M5-24h	2,125	461,060	104	47,671
M5-48h	2,255	615,608	104	62,456
M5-72h	2,375	704,697	104	71,625

Table 3: Bootstrapped performance of EVEREST on the held-out test set. Thresholds are the task-specific optima from the balanced scoring rule.

Task	TSS	Precision	Recall	Brier	ECE
C-24h	0.973 ± 0.001	0.994 ± 0.000	0.986 ± 0.001	0.015 ± 0.000	0.049 ± 0.000
C-48h	0.970 ± 0.001	0.993 ± 0.000	0.984 ± 0.001	0.017 ± 0.000	0.054 ± 0.000
C-72h	0.966 ± 0.001	0.992 ± 0.000	0.982 ± 0.001	0.018 ± 0.000	0.052 ± 0.000
M-24h	0.898 ± 0.011	0.728 ± 0.016	0.908 ± 0.011	0.011 ± 0.000	0.037 ± 0.001
M-48h	0.920 ± 0.007	0.772 ± 0.010	0.928 ± 0.007	0.009 ± 0.000	0.029 ± 0.000
M-72h	0.906 ± 0.012	0.834 ± 0.015	0.911 ± 0.012	0.010 ± 0.000	0.033 ± 0.001
M5-24h	0.907 ± 0.025	0.686 ± 0.033	0.908 ± 0.025	0.003 ± 0.000	0.031 ± 0.000
M5-48h	0.936 ± 0.021	0.713 ± 0.035	0.937 ± 0.021	0.002 ± 0.000	0.020 ± 0.000
M5-72h	0.966 ± 0.024	0.727 ± 0.053	0.966 ± 0.024	0.002 ± 0.000	0.016 ± 0.000

A Supplementary Material

A.1 Auxiliary Heads and Reference Configuration

The primary classification head outputs the scalar logit l and probability $\hat{p} = \sigma(l)$ used for all alerts at inference time. The evidential head predicts Normal–Inverse–Gamma parameters (μ, v, α, β) over the logit and minimizes the closed-form evidential objective from Amini et al. (2020); in EVEREST it is used only as a training-time regulariser of logit uncertainty. The EVT head predicts GPD parameters (ξ, σ) for logit exceedances above a high batchwise quantile u (90% by default). For logits $\{l_i\}$ in a mini-batch, exceedances are formed as $\{l_i - u : l_i > u\}$ and optimized with a GPD likelihood plus a small stability regulariser on (ξ, σ). The precursor head is a second binary classifier attached to the shared latent vector z; it uses the same label y through a BCE loss and is discarded at inference.

Unless stated otherwise, the reference setting is $(d, L, H, \mathrm{FFN}, p) = (128, 6, 4, 256, 0.20)$ with loss weights $(\lambda_f, \lambda_e, \lambda_t, \lambda_p) = (0.8, 0.1, 0.1, 0.05)$ and focal γ annealed from 0 to 2 over the first 50 epochs.

A.2 SHARP–GOES Split Counts

We use the HARPNUM-stratified train/validation/test split described in the experimental protocol. The validation partition is disjoint from both train and test and is used only for threshold selection and early stopping. Table 2 reports the train and held-out test counts for the nine forecast tasks used in the headline evaluation.

A.3 Headline Metrics and Baseline Deltas

Table 3 reports bootstrapped performance on the held-out test set for all nine solar-flare tasks. Table 4 reports absolute TSS deltas relative to the strongest published SHARP–GOES baseline (Abduallah et al., 2023). Since those baseline values are quoted from prior work rather than rerun on the identical split, we present them as contextual comparisons rather than paired significance tests.

Table 4: Absolute TSS deltas relative to the strongest published SHARP–GOES baseline (Abduallah et al., 2023).

Task	Baseline TSS	EVEREST TSS	ΔTSS
C-24h	0.835	0.973 (0.001)	+0.138
M-24h	0.839	0.898 (0.011)	+0.059
M5-24h	0.818	0.907 (0.025)	+0.089
C-48h	0.719	0.970 (0.001)	+0.251
M-48h	0.728	0.920 (0.007)	+0.192
M5-48h	0.736	0.936 (0.021)	+0.200
C-72h	0.702	0.966 (0.001)	+0.264
M-72h	0.714	0.906 (0.012)	+0.192
M5-72h	0.729	0.966 (0.024)	+0.237

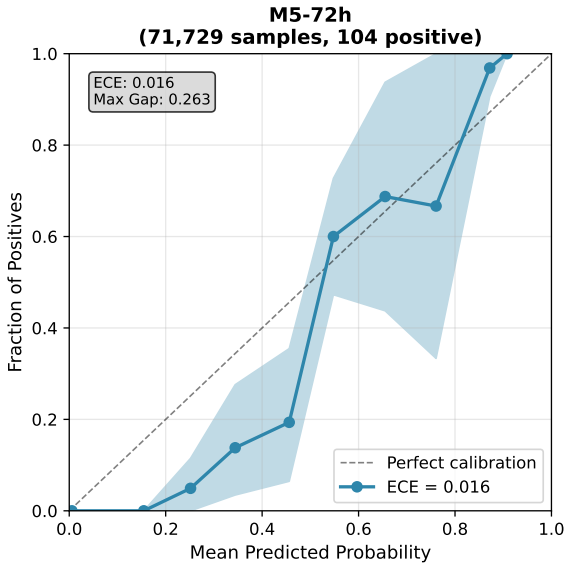

Figure 1: Reliability diagram for the M5–72 h task (15 equal-frequency bins). The dashed line indicates perfect calibration.

A.4 Calibration

A.5 Ablation Study

We ran a leave-one-component-out protocol on the hardest solar-flare task (M5–72 h) with five seeds per variant, identical data splits, early stopping at 120 epochs, and bootstrap evaluation (10^4 replicates). Absolute scores are lower here than in the fully tuned headline runs because the ablation protocol fixes a shorter training schedule for all variants.

A.6 SKAB Transfer Details

For SKAB (Filonov et al., 2020) we use fixed-length windows, chronological splits, and the same EVEREST design and training recipe as in the solar-flare setting, with a smaller width/depth configuration to match the dataset scale. Table 7 reports the aggregate EVEREST metrics and Table 8 situates them against published SKAB baselines, including TranAD (Tuli et al., 2022).

Table 5: Ablation results on M5–72 h (mean $\pm$ s.d. over 5 seeds).

Variant	TSS	F1	Brier	ECE	p
Full model	0.746 ± 0.146	0.747	0.0013	0.0110	—
No Evidential head	0.682 ± 0.193	0.626	0.0015	0.0111	< 0.01
No EVT head	0.461 ± 0.369	0.438	0.0039	0.0336	< 0.01
No Evidential + EVT heads	0.640 ± 0.275	0.594	0.0015	0.0115	< 0.01
Mean pooling	0.319 ± 0.319	0.304	0.0229	0.1158	< 0.001
Cross-entropy loss	0.209 ± 0.332	0.195	0.0013	0.0023	< 0.001
No Precursor head	0.096 ± 0.174	0.095	0.0194	0.1105	< 0.001
FP32 training	0.000 ± 0.000	0.000	0.0520	0.2248	< 0.001

Table 6: Effect of removing each component on M5–72 h.

Component Removed	ΔTSS	Rel. Change (%)	p-value
Mixed Precision (AMP)	-0.746	-100	< 0.001
Precursor head	-0.650	-87	< 0.001
Focal loss	-0.537	-72	< 0.001
Attention bottleneck	-0.427	-57	< 0.001
EVT head	-0.285	-38	< 0.001
Evidential head	-0.064	-9	0.004

Table 7: EVEREST averaged across all SKAB valves.

Metric	Precision (%)	Recall (%)	F1 (%)	TSS
EVEREST	97.7 ± 2.9	98.6 ± 3.2	98.2 ± 1.7	0.964 ± 0.028

Table 8: Published SKAB baseline comparison (F1).

Model	Reference	F1 (%)
Isolation F / LOF / related classical methods	Filonov et al. (2020)	65–75
Autoencoder	Filonov et al. (2020)	70–80
CNN/LSTM hybrids	Filonov et al. (2020)	75–85
TAnoGAN	Bashar & Nayak (2020)	79–92
DeepLog	Du et al. (2017)	87–91
LSTM-VAE	Park et al. (2018)	86–93
OmniAnomaly	Su et al. (2019)	88–94
USAD	Audibert et al. (2020)	89–95
TranAD	Tuli et al. (2022)	91–96
EVEREST	—	$\mathbf{98.2 \pm 1.7}$