JIKO (Jurnal Informatika dan Komputer) Accredited KEMDIKTISAINTEK, No.0173/C3/DT.05.00/2025
Vol. 9, No. 2, August 2026, pp. 181-189 p-ISSN: 2614-8897
DOI: 10.33387/jiko.v9i2.11857 e-ISSN: 2656-1948

A RECALL-ORIENTED STACKING ENSEMBLE FOR EXTREME RAINFALL EARLY WARNING IN TERNATE

Achmad Fuad[1], Muhamad Sabri Ahmad[2*], Muhammad Ridha Albaar[3], Yasir Muin[4]
[1234]Department of Informatics Engineering, Faculty of Engineering, Universitas Khairun, Ternate 97719, Indonesia
Email: [1]foeoe09@gmail.com, [*2]msabri@unkhair.ac.id, [3]ridha.albaar@ukhair.ac.id, [4]yasirmuin@unkhair.ac.id

(Received: 09 May 2026, Revised: 11 July 2026, Accepted: 22 July 2026)

Abstract

Extreme-rainfall detection in small tropical islands is difficult because extreme events are rare and false negatives have substantial disaster-mitigation consequences. This study aims to (1) compare imbalance-aware machine-learning and deep temporal models, (2) determine whether an out-of-fold stacking ensemble can maximize sensitivity, and (3) explain model behavior using SHAP. ERA5 hourly data for a representative Ternate grid point were collected from 1 January 2005 to 6 February 2026 and aggregated into 7,707 daily records; only 12 days met the extreme threshold of 75 mm/day. A 30-day sequence was used to predict the next-day class, and three expanding temporal test folds were evaluated. Weighted loss, focal loss, balanced tree learning, validation-based threshold selection, and leakage-controlled stacking were applied. Across 2,229 pooled test days containing four extreme events, the stacking model detected all events (recall 1.000) but produced low precision (0.0023), F2-score 0.0115, specificity 0.2256, and 1,723 false positives. Balanced Random Forest provided the best overall trade-off, with recall 0.750, F2-score 0.0316, specificity 0.7951, and MCC 0.0570. Therefore, stacking is useful as a high-sensitivity screening layer, but it is not yet suitable as a standalone operational warning model. Local observations and additional event samples are required for calibration.

Keywords: *class imbalance, early warning, extreme rainfall, focal loss, SHAP, stacking ensemble*

**Corresponding Author: Muhammad Sabri Ahmad*

1. INTRODUCTION

Extreme rainfall is a major driver of flash flooding and urban inundation, particularly in tropical island environments where short response times, steep terrain, and concentrated settlements increase exposure [1]. Ternate is vulnerable to hydrometeorological disruption, yet local early-warning development is constrained by sparse observations and the rarity of daily rainfall exceeding extreme thresholds. ERA5 provides a temporally consistent reanalysis source that can support long-term model development when local station records are incomplete [2].

The statistical difficulty is not ordinary rainfall prediction but rare-event detection. In the processed series used here, only 12 of 7,707 days reached at least 75 mm/day. Under such severe imbalance, a classifier may achieve high overall accuracy by predicting the majority class while missing the events that matter most [3]. Previous rainfall-classification studies have likewise shown that standard learners can be biased toward non-extreme conditions and require explicit imbalance handling [4], [5].

Deep temporal models offer complementary representations. Long Short-Term Memory (LSTM) networks learn sequential persistence, whereas Transformer attention can model longer-range interactions. Nevertheless, either model can still be poorly calibrated when only a few positive events are available. A stacking architecture is therefore investigated not because an ensemble is automatically superior, but because heterogeneous probability outputs may encode different temporal evidence. The meta-learner can exploit their agreement and disagreement, while validation-based thresholding can explicitly prioritize sensitivity.

This study addresses three questions. First, how do imbalance-aware tree models, weighted neural networks, and focal-loss temporal models compare under chronological evaluation? Second, can a leakage-controlled LSTM-Transformer-XGBoost stack reduce missed extreme events relative to individual learners? Third, what do SHAP explanations reveal about both meta-level model combination and meteorological feature influence?

The contributions are fourfold. The study formulates next-day extreme-rainfall warning as an imbalanced binary time-series problem; applies three

expanding temporal folds instead of a single random split; corrects stacking through valid out-of-fold probability generation and validation-only threshold selection; and reports both sensitivity and false-alarm burden using PR-AUC, F2-score, specificity, MCC, false-alarm ratio, and Brier score. The contribution is therefore a transparent assessment of the operational trade-off, rather than a claim that stacking dominates every baseline..

2. RELATED WORK

Data-driven precipitation nowcasting has progressed from conventional statistical models to machine-learning and deep-learning architectures capable of representing nonlinear atmospheric dynamics [6]. LSTM remains a common sequence model because its gating mechanism controls long-term information flow [7], and recent time-series variants have improved adaptive learning for complex nonstationary signals [8]. Transformer architectures use self-attention to represent long-range dependencies without recurrent computation [9]. In extreme-weather applications, model performance is strongly affected by input resolution, forecast horizon, and the rarity of target events [10].

Imbalance handling is therefore central. Cost-sensitive boosting and distribution-aware objectives can improve minority-event detection, although gains in recall may increase false alarms [11]. XGBoost is widely used because it models nonlinear interactions efficiently and supports class weighting [12]. Indonesian work on imbalanced rainfall classification has also reported that hybrid or voting strategies require careful evaluation beyond overall accuracy [5].

Stacking combines base-model probabilities through a separate meta-learner. Hydrological studies have used stacking to integrate complementary neural architectures for runoff prediction [13]. A recent JIKO study also demonstrated the methodological value of combining sequence learning, tree boosting, chronological validation, and SHAP-based explanation, although in a regression domain [14]. These studies motivate a heterogeneous ensemble, but they do not guarantee improvement when the positive class is extremely scarce.

For imbalanced classification, precision-recall analysis is more informative than accuracy or ROC curves alone because it focuses on the positive class [15]. SHAP provides additive feature attributions for global and local interpretation [16]. In this study, two explanation levels are kept distinct: SHAP on the XGBoost meta-learner explains how LSTM and Transformer probabilities are combined, whereas meteorological SHAP is computed for the weighted XGBoost baseline and is not presented as an end-to-end explanation of the stack.

2.1 Synthesis and Research Gap

A central distinction in the literature is whether rainfall is treated as a continuous quantity or as a rare hazardous event. Regression-oriented studies optimize average error over all days, whereas an early-warning classifier must separate a very small number of hazardous days from thousands of normal observations. The latter formulation changes both model design and evaluation: a model can have a small average error or high overall accuracy while still missing the events that matter most. Previous extreme-rainfall classification studies have reported this imbalance problem, but many evaluations still rely on a single split or emphasize global accuracy [4], [5], [10]. For the Ternate series, the positive class represents only 0.156% of all daily observations, making minority-class behavior the primary methodological concern rather than a secondary preprocessing issue.

Existing imbalance remedies include class weighting, resampling, cost-sensitive boosting, focal objectives, and ensemble learning [3], [5], [11]. Random oversampling or synthetic interpolation can be useful in ordinary tabular settings, but it may distort chronological dependence when applied indiscriminately to time-series windows. The present study therefore prioritizes cost-sensitive and ensemble methods that preserve temporal order. In parallel, precision-recall analysis is used because the positive-class prevalence is extremely low; ROC-AUC alone can appear favorable even when the absolute number of false alarms remains operationally unacceptable [15]. F2-score, specificity, MCC, and false-alarm ratio are consequently reported together so that sensitivity is not interpreted in isolation.

The unresolved gap is not simply the absence of another hybrid architecture. Rather, it is the lack of leakage-controlled evidence showing whether heterogeneous temporal learners add value under an event count this small. Stacking can combine complementary probability patterns [13], and SHAP can reveal how the meta-learner and meteorological baseline reach their decisions [16], [18], [23]. However, an ensemble contribution is defensible only when it is compared against strong imbalance-aware single models and when identical predictions are ruled out. This study addresses that gap by using valid chronological out-of-fold probabilities, expanding temporal folds, explicit threshold selection, and separate interpretation of the meta-learner and meteorological baseline.

3. RESEARCH METHOD

The revised workflow was designed to prevent temporal leakage, explicitly handle class imbalance, and evaluate whether perfect sensitivity can be achieved without concealing the associated false-alarm cost. Figure 1 summarizes the procedure.

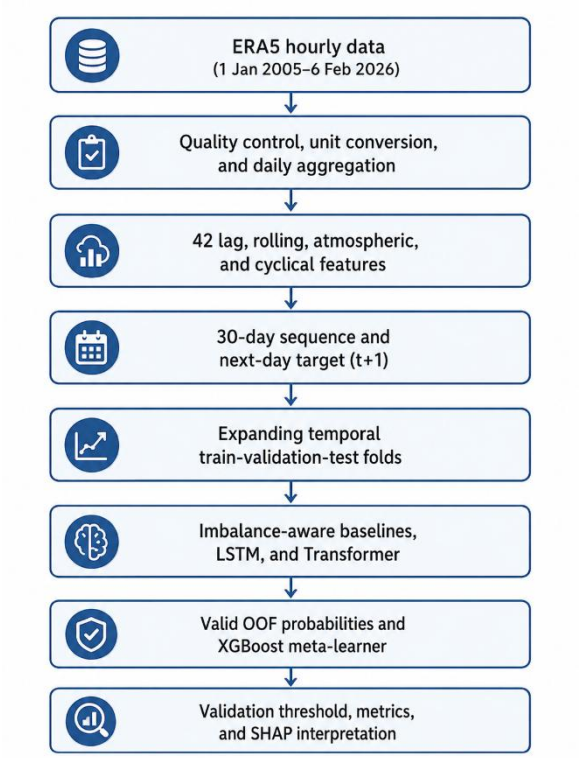

Figure 1. Revised research workflow for next-day extreme-rainfall detection.

3.1 Data Source and Study Definition

Hourly ERA5 single-level data were obtained from the Copernicus Climate Data Store for the grid point 0.25 degrees N, 127.50 degrees E, representing the Ternate study region. The source dataset is available at https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels and is cited through its official DOI [2], [22]. The extracted file contains 184,968 consecutive hourly observations from 1 January 2005 to 6 February 2026, with no duplicated timestamps, missing hours, or missing values.

The variables were total precipitation (tp), 2-m temperature (t2m), 2-m dewpoint temperature (d2m), mean sea-level pressure (msl), surface pressure (sp), 10-m zonal and meridional wind components (u10 and v10), and surface solar radiation downwards (ssrd). Total precipitation was converted from meters to millimeters and summed over 24 hours. Temperature was converted to degrees Celsius, pressure to hPa, and radiation to MJ/m2. Small negative radiation artefacts were clipped to zero.

Before daily aggregation, the raw file contained 184,968 hourly records from 1 January 2005 00:00 to 6 February 2026 23:00. Quality control found no duplicated timestamps, missing hours, or missing values in the required variables, and every retained calendar day contained 24 hourly observations. Total precipitation was clipped at zero to remove impossible negative numerical artifacts and converted from metres to millimetres. Temperature and dewpoint were converted from Kelvin to degrees Celsius, pressure from pascals to hectopascals, and solar radiation from joules per square metre to megajoules per square metre per day. These checks ensure that the subsequent class imbalance is a property of the event definition rather than a consequence of missing or irregular timestamps.

The spatial representation should also be interpreted carefully. ERA5 is a gridded reanalysis product rather than a local rain-gauge record [2], [22]. The selected point provides a consistent long-term atmospheric series for the Ternate study region, but it does not reproduce neighbourhood-scale orographic rainfall around the volcanic island. Accordingly, the experiment is positioned as a computational case study of rare-event warning from reanalysis data. Local station, radar, or satellite observations are required before operational deployment or claims of site-specific hydrometeorological accuracy.

$$R\_d = 1000 \times \mathrm{sum}(tp\_h),\ h = 1,\ldots,24 \quad (1)$$

A day was labelled extreme when daily rainfall was at least 75 mm. The binary target for the following day was defined as y(t+1)=1 when R(t+1) was at least 75 mm and y(t+1)=0 otherwise. This produced 7,707 daily observations, of which only 12 were extreme (0.156%).

3.2 Feature Engineering and Sequences

Forty-two predictors were generated. They included current daily rainfall; rainfall lags at 1, 2, 3, 7, and 14 days; rolling mean, sum, maximum, and rainy-day count over 3-, 7-, and 14-day windows; daily atmospheric means and ranges; wind speed; dewpoint depression; one-day pressure and dewpoint changes; and cyclical month and day-of-year features. Constant latitude and longitude values were excluded from model input. A 30-day input sequence ending on day t was used to predict the class on day t+1, which prevents use of future information.

The predictors were organized into four groups. The first group represented short-term rainfall persistence through current rainfall, five lags, and rolling summaries. The second described moisture and thermal state using temperature, dewpoint, dewpoint depression, and daily ranges. The third represented circulation and surface forcing using pressure tendencies, zonal and meridional wind, wind speed, and solar radiation. The fourth encoded seasonality through sine and cosine transformations of day-of-year and month. This grouping allows the later SHAP analysis to distinguish antecedent wetness from broader atmospheric and seasonal signals.

Causal alignment was enforced explicitly. For each sample, the input tensor contained the 30-day window ending on day t, while the target represented whether rainfall on day t+1 reached the extreme threshold. Rolling statistics were calculated only with observations available on or before day t. The first 14 days were removed because of the longest lag and rolling window, and the first 30 eligible feature rows were then used only as historical context for sequence construction. This arrangement prevents the target day or any future observation from entering the predictors.

$$WS = \mathrm{sqrt}(u10\text{^}2 + v10\text{^}2) \quad (2)$$

3.3 Expanding Temporal Evaluation

Three expanding train-validation-test folds were used. The test periods did not overlap, allowing pooled evaluation without counting the same day twice. Standardization was fitted only on each outer training fold and then applied to its validation and test data. Table 1 reports the resulting class distribution after feature construction and 30-day sequencing.

The outer folds were selected to test generalization across distinct climate periods. Fold 1 trained through 2017, validated on 2018-2019, and tested on 2020-2021. Fold 2 extended training through 2019, validated on 2020-2021, and tested on 2022-2023. The final fold trained through 2021, validated on 2022-2023, and tested from 2024 to 6 February 2026. The three test windows contained one, two, and one extreme days, respectively. Although this remains a small event sample, pooling the non-overlapping test windows is more informative than reporting performance from a final holdout containing only one positive case.

A second chronological layer was used inside each outer training set to generate out-of-fold probabilities for stacking. Early observations that could not receive a valid out-of-fold prediction were stored as missing and excluded from meta-training rather than being replaced by zero. This detail is important because a fabricated zero can be interpreted by the meta-learner as genuine evidence of a non-extreme event. All standardization parameters, class weights, neural-network fitting, and meta-features were estimated without using the outer validation or test labels.

Table 1. Class distribution in the expanding temporal folds.

Fold	Split	N	Extreme	Rate
Fold 1	Train	4704	6	0.128%
	Validation	730	2	0.274%
	Test	731	1	0.137%
Fold 2	Train	5434	8	0.147%
	Validation	731	1	0.137%
	Test	730	2	0.274%
Fold 3	Train	6165	9	0.146%
	Validation	730	2	0.274%
	Test	768	1	0.130%

3.4 Imbalance-Aware Models

The baselines were class-weighted Logistic Regression, class-weighted Random Forest, weighted XGBoost, and Balanced Random Forest. The XGBoost positive weight used the square root of the negative-to-positive ratio and was capped at 50 to avoid the extreme overcorrection observed with the raw ratio. Balanced Random Forest repeatedly undersampled the majority class within its trees.

The LSTM contained 32 and 16 recurrent units, dropout of 0.20, and a 16-unit dense layer. The Transformer projected inputs to 32 dimensions, used two attention heads, a 64-unit feed-forward block, global average pooling, and a 16-unit dense layer. Each architecture was trained with weighted binary cross-entropy and with focal loss (alpha=0.75; gamma=2.0). Early stopping monitored validation PR-AUC rather than recall at a fixed threshold.

The weighting strategy was deliberately moderated. Using the full negative-to-positive ratio would assign several hundred times more weight to each positive sample and can force a neural classifier to label most days as extreme. The revised pipeline therefore used the square root of the class ratio, capped at 50, for weighted XGBoost and weighted binary cross-entropy. Focal loss with alpha=0.75 and gamma=2.0 was evaluated separately to reduce the contribution of easy majority examples while retaining difficult minority windows. Balanced Random Forest and EasyEnsemble supplied comparison points based on repeated majority-class subsampling without disrupting chronological order before sequence generation.

3.5 Corrected Stacking Procedure

The proposed stack used focal-loss LSTM and Transformer models as base learners and XGBoost as the meta-learner. Event-aware chronological inner folds generated out-of-fold probabilities on the outer training data. Positions without a valid OOF prediction were stored as missing and excluded, rather than being incorrectly treated as zero probabilities. The meta-features were LSTM probability, Transformer probability, their mean, absolute difference, maximum, and minimum. Final base learners were then refitted on the complete outer training fold before producing validation and test probabilities.

This design answers the reviewers' concern that the original ensemble and base models produced identical decisions. The revised stack is methodologically distinct and learns from model disagreement; its benefit is evaluated empirically rather than assumed.

The meta-feature vector contained six quantities: LSTM probability, Transformer probability, their mean, absolute disagreement, maximum, and minimum. Mean probability summarizes consensus, absolute disagreement measures complementarity, and the extrema allow the meta-learner to recognize cases in which only one temporal model produces a strong warning. XGBoost was selected as the meta-learner because it can model nonlinear interactions among these probability descriptors while remaining interpretable through TreeSHAP. The final base learners were retrained on the complete outer training set after OOF generation, and their probabilities were then transformed into meta-features for outer validation and testing.

3.6 Threshold Selection and Metrics

Decision thresholds were selected only from the validation fold. Candidate thresholds were first required to provide recall of at least 0.75 and specificity of at least 0.50, after which F2-score was maximized. If no candidate satisfied both constraints, specificity was relaxed while the recall constraint was

retained. This policy reflects a screening context but exposes when high recall requires excessive alarms.

Evaluation included PR-AUC, ROC-AUC, precision, recall, F1-score, F2-score, specificity, balanced accuracy, Matthews correlation coefficient (MCC), false-positive rate, false-alarm ratio (FP/(TP+FP)), and the Brier score. RMSE of rainfall was not used because the study predicts a binary event rather than a continuous rainfall amount. The Brier score directly supplies the mean squared error of predicted event probabilities:

The validation rule formalizes the intended risk preference rather than fixing an arbitrary threshold such as 0.01. Candidate thresholds were screened for the predefined recall and specificity constraints, and the F2-score was used because it weights recall more heavily than precision without ignoring false alarms. When no threshold satisfied both constraints, the algorithm preserved the recall requirement and recorded the specificity relaxation transparently. The Brier score complements discrimination metrics by measuring the mean squared error of predicted probabilities, thereby addressing probabilistic calibration without treating the binary task as rainfall regression.

$$BS = (1/N) \times sum((p\_i - y\_i)^2) \quad (3)$$

3.7 SHAP Interpretation

SHAP was used for diagnosis, not as a performance metric. Meta-learner SHAP quantified how the six probability-combination features influenced stacking. A separate TreeSHAP analysis on the weighted XGBoost baseline ranked meteorological predictors and was used to assess whether the learned relationships were physically plausible. Keeping these analyses separate avoids attributing tabular baseline explanations to the end-to-end stack.

4. RESULT AND DISCUSSION

The results of the research are based on a logical **4.1 Pooled Model Performance**

The three non-overlapping test periods contained 2,229 days and four extreme events. Table 2 reports pooled metrics for the principal configurations. The revised experiment no longer produced identical results for LSTM, Transformer, and stacking. The stacking model achieved the highest recall, whereas Balanced Random Forest provided the strongest overall discrimination and alarm-control trade-off.

Pooled prevalence was 4/2,229, or approximately 0.18%. This value provides the no-skill reference for precision-recall analysis. Consequently, even the highest observed precision of 0.0065 remains below 1%, although it is about 3.6 times the event prevalence. Reporting only relative improvement would therefore overstate practical utility. The revised presentation retains both relative discrimination metrics and absolute TP, FP, and FN counts so that the operational meaning of each result is visible.

The weighted linear and boosting baselines were selective but insensitive. Logistic Regression and weighted XGBoost each detected only one of four extreme days, with recall 0.250 and specificity near 0.77. Weighted binary cross-entropy did not resolve the problem for the neural models: LSTM-WBCE and Transformer-WBCE also achieved recall 0.250 and produced negative MCC values. Focal loss improved LSTM and Transformer recall to 0.500, indicating that reducing the influence of easy majority examples helped minority detection, but the corresponding precision and specificity remained low.

Balanced Random Forest produced the clearest standalone compromise. It detected three of four extreme days, achieved ROC-AUC 0.7840, F2-score 0.0316, specificity 0.7951, and MCC 0.0570, and issued far fewer alarms than the temporal ensembles. The weighted Logistic model had a slightly larger PR-AUC of 0.0057, but its recall was only 0.250. This apparent discrepancy is possible because average precision evaluates probability ranking across thresholds, whereas the reported recall and F2-score use the threshold selected on each validation fold. Both threshold-free and threshold-dependent metrics are therefore needed.

The stacking model occupied the opposite end of the trade-off. It detected all four events, but the selected thresholds classified 77.48% of pooled test days as extreme. Its ROC-AUC of 0.3656 and PR-AUC of 0.0016 show that the probability ranking was weak even though the final threshold yielded recall 1.000. This result demonstrates why recall cannot be interpreted without specificity, alert rate, and false-alarm ratio. It also supports the revised claim that stacking is a high-sensitivity screening experiment rather than the best overall or operationally ready model.

ROC-AUC and PR-AUC convey different aspects of this problem. ROC-AUC measures ranking across true-positive and false-positive rates and can remain moderate because the number of true negatives is very large. PR-AUC instead compares precision directly with recall and is therefore more sensitive to the rarity of extreme days. The gap between Balanced Random Forest ROC-AUC 0.7840 and PR-AUC 0.0052 illustrates this point: the model ranks some extreme windows relatively well, but a practical threshold still produces hundreds of false alarms.

Probability calibration presents another caution. Pooled Brier scores were 0.1656 for Balanced Random Forest, 0.0042 for LSTM-Focal, and 0.0039 for Stacking-XGB. The apparently smaller neural and stacking Brier scores do not imply superior warning decisions because the unweighted Brier score is dominated by the overwhelming number of non-extreme days. A model assigning small probabilities to most observations can obtain a low average squared error while still providing poor minority-class ranking. Brier score is therefore interpreted alongside PR-AUC and event-detection counts rather than as an independent winner-selection criterion.

Table 2. Pooled performance across three non-overlapping temporal test folds.

Model	PR-AUC	ROC-AUC	Precision	Recall	F2	Specificity	MCC
Logistic-W	0.0057	0.6818	0.0019	0.250	0.0094	0.7681	0.0018
XGBoost-W	0.0029	0.6196	0.0019	0.250	0.0093	0.7658	0.0016
Balanced RF	0.0052	0.7840	0.0065	0.750	0.0316	0.7951	0.0570
LSTM-WBCE	0.0019	0.4308	0.0009	0.250	0.0045	0.5133	-0.0200
Transformer-WBCE	0.0018	0.3558	0.0008	0.250	0.0039	0.4261	-0.0277
LSTM-Focal	0.0036	0.6803	0.0026	0.500	0.0127	0.6553	0.0138
Transformer-Focal	0.0016	0.3328	0.0014	0.500	0.0067	0.3420	-0.0141
Soft Voting	0.0020	0.4591	0.0019	0.500	0.0096	0.5393	0.0033
Stacking-XGB	0.0016	0.3656	0.0023	1.000	0.0115	0.2256	0.0229

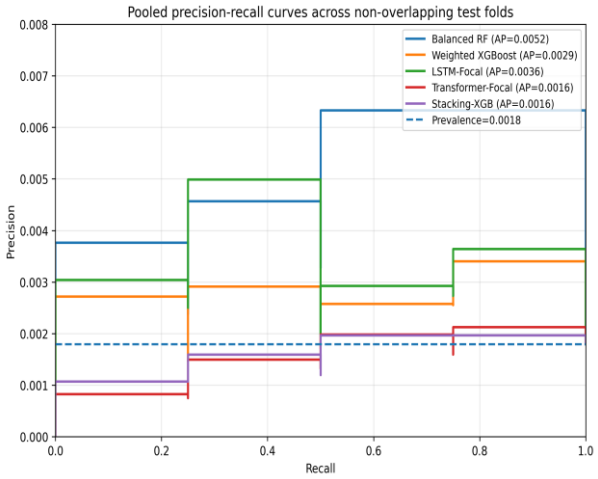

Figure 2. Pooled precision-recall curves for representative imbalance-aware models (y-axis zoomed to highlight low-precision differences under severe class imbalance)

Balanced Random Forest achieved the highest pooled F2-score (0.0316), specificity (0.7951), balanced accuracy (0.7725), and MCC (0.0570), while detecting three of four extreme days. Its precision remained low at 0.0065, but it reduced false alarms substantially relative to the temporal ensembles. The pooled PR curves in Figure 2 confirm that all models operate close to the severe class-prevalence baseline, illustrating the difficulty of learning from only a few independent events. For readability, the y-axis is zoomed to 0–0.008 so that small precision differences remain visible.

Stacking detected all four extreme days, giving recall 1.000 and zero false negatives. However, PR-AUC was 0.0016, precision was 0.0023, specificity was 0.2256, and the false-alarm ratio was 0.9977. Consequently, recall alone would provide a misleading impression of model quality. The result supports stacking only as a high-sensitivity candidate-generation layer.

Alert rate gives the most direct operational comparison. Balanced Random Forest flagged 20.59% of pooled test days, LSTM-Focal flagged 34.50%, and Stacking-XGB flagged 77.48%. Even the balanced tree model would require substantial confirmation before public use, but its workload is much lower than that of the stack. The ranking of models therefore changes with the decision objective: stacking is preferred only when the cost of one missed event is assumed to exceed the cost of a very large number of alarms.

Because the pooled evaluation contains only four positive days, confidence intervals or significance tests would be unstable and potentially misleading. A difference of one detected event changes pooled recall by 0.25, while a single positive in Fold 1 changes fold recall from 0 to 1. The revised analysis consequently emphasizes transparent counts, fold-specific behavior, and consistency across complementary metrics. This reporting choice is more informative for the present sample than claiming statistically significant superiority from a very small number of hazardous events.

4.2 Temporal Robustness Across Folds

The pooled values summarize all non-overlapping test observations, but fold-level results reveal how strongly conclusions depend on individual extreme events. The comparison below contrasts the two models that best illustrate the sensitivity-selectivity trade-off. Balanced Random Forest produced the strongest aggregate balance, whereas Stacking-XGB was the only configuration that detected every extreme day across all folds.

Fold 1 - Balanced Random Forest: recall 0.000, F2-score 0.0000, specificity 0.9192, and 59 false positives; Stacking-XGB: recall 1.000, F2-score 0.0070, specificity 0.0342, and 705 false positives.

Fold 2 - Balanced Random Forest: recall 1.000, F2-score 0.0549, specificity 0.7637, and 172 false positives; Stacking-XGB: recall 1.000, F2-score 0.0183, specificity 0.2637, and 536 false positives.

Fold 3 - Balanced Random Forest: recall 1.000, F2-score 0.0217, specificity 0.7066, and 225 false positives; Stacking-XGB: recall 1.000, F2-score 0.0103, specificity 0.3716, and 482 false positives.

Fold 1 demonstrates the instability created by a single positive test case. Balanced Random Forest maintained specificity of 0.9192 but missed that event, whereas stacking detected it only by issuing 705 false alarms and reducing specificity to 0.0342. In Folds 2 and 3, both models detected all extreme days, yet Balanced Random Forest retained substantially higher specificity and F2-score. Thus, the additional sensitivity of stacking is concentrated in the earliest test fold and is accompanied by a very large warning burden.

The selected thresholds also differed by model family. Balanced Random Forest used thresholds of approximately 0.620, 0.444, and 0.496 across the three

folds. Stacking required much smaller thresholds of approximately 0.0039, 0.0020, and 0.0053. These values indicate that the stacked probabilities are not well calibrated and that a small numerical increase can trigger many alarms. The Brier score and validation trade-off therefore remain necessary complements to recall.

The fold analysis supports two conclusions. First, the corrected stack is behaviorally distinct from both base learners and no longer reproduces their decisions. Second, distinct behavior does not establish superiority. Balanced Random Forest is the more stable standalone model, while stacking represents the maximum-sensitivity boundary of the evaluated design space. This distinction is retained throughout the revised discussion and conclusion.

4.3 Why Use an Ensemble?

The purpose of the ensemble is now explicitly diagnostic. A single Balanced Random Forest provided the best aggregate trade-off, whereas stacking provided the strictest sensitivity. The models therefore answer different operational priorities: Balanced Random Forest limits alarm volume, while stacking minimizes missed events. Across the pooled test periods, stacking decisions differed from LSTM on 1,262 days and from Transformer on 1,225 days, confirming that the corrected meta-learner no longer duplicates its base models.

Table 3. Pooled detection counts and false-alarm burden.

Model	TP	FP	FN	FAR
Balanced RF	3	456	1	0.9935
Stacking-XGB	4	1723	0	0.9977

Table 3 and this comparison directly address why one model is not sufficient for every warning objective. Nevertheless, the current stack cannot be claimed as universally superior because its extra sensitivity is purchased with an operationally excessive false-alarm burden.

4.4 Final-Fold Confusion Matrix and Threshold Behavior

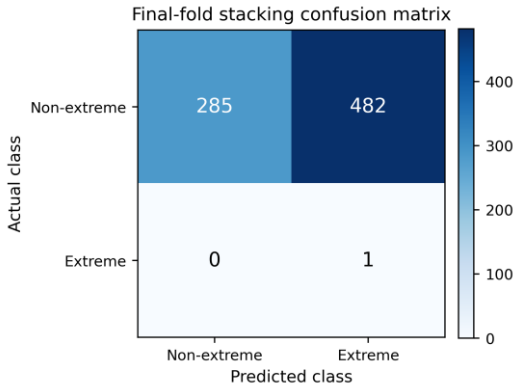

Figure 3. Confusion matrix of the stacking model for the final test fold (2024-6 February 2026).

Figure 3 contains one actual extreme day. The stack detected it, but also classified 482 of 767 non-extreme days as extreme, leaving only 285 true negatives. The figure therefore does not show that an extreme prediction was actually non-extreme; it shows that many predicted extreme days were false positives. The corresponding final-fold recall was 1.000 and specificity was 0.3716.

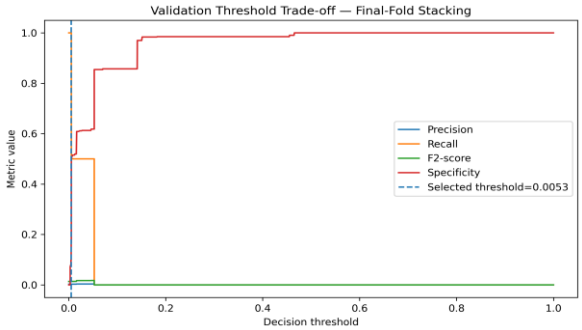

Figure 4. Validation threshold trade-off for the final-fold stacking model.

The validation threshold was approximately 0.0053. No candidate threshold simultaneously achieved the required recall and specificity, so the predefined policy retained the recall constraint and relaxed specificity. Figure 4 makes this limitation visible: increasing the threshold improves specificity, but recall changes abruptly because the validation set contains only two positive days.

4.5 Error Pattern and Warning Burden

Most stacking errors were false positives rather than false negatives. This pattern is consistent with the optimization objective: the model was encouraged to identify weak precursors of rare events, and the selected threshold retained sensitivity when specificity constraints could not be met. Because extreme days are sparse, many non-extreme windows share rainfall lags, high dewpoint, wind, or seasonal conditions similar to the few positive windows. The classifier therefore learns a broad hazardous-condition envelope rather than a narrow extreme-event boundary.

False positives should not automatically be treated as meaningless predictions. Some may represent meteorologically active days that remained below the 75 mm/day label threshold, while others may reflect spatial mismatch between the ERA5 grid and local rainfall. Nevertheless, an operational warning system must account for user fatigue, resource mobilization, and declining trust when alarms occur too frequently. The pooled stacking alert rate of 77.48% is therefore unacceptable for direct public warning even though no pooled extreme day was missed.

A practical architecture would separate screening from confirmation. The high-sensitivity stack could identify candidate periods, after which a more selective model, local station observation, radar estimate, or impact-based rule could confirm escalation. Event-based evaluation should also group consecutive extreme days into episodes and report lead time, warning duration, and alarms per event. These

extensions are proposed as future work rather than being claimed as validated by the present experiment.

4.6 SHAP-Based Diagnostic Interpretation

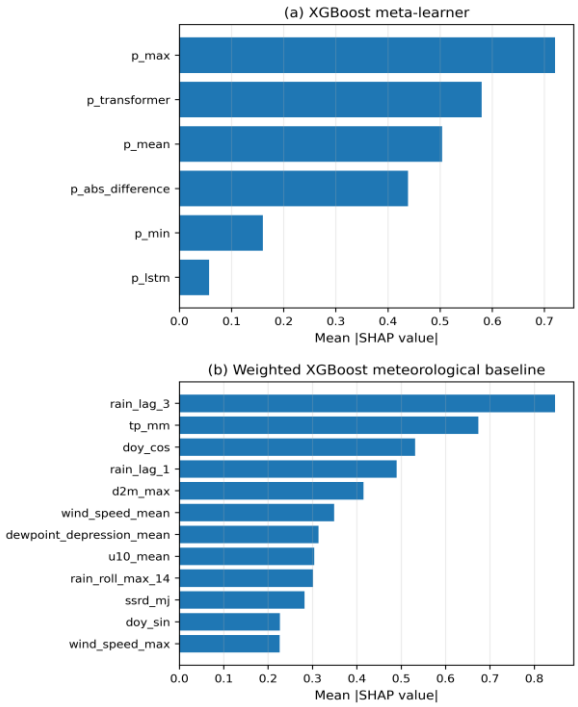

Figure 5. SHAP importance for (a) the stacking meta-learner and (b) the weighted XGBoost meteorological baseline.

At the meta level, the maximum base-model probability was the most influential feature, followed by Transformer probability, mean probability, and absolute disagreement. LSTM probability had the smallest direct contribution. This indicates that the meta-learner primarily reacts to the strongest temporal signal rather than assigning equal importance to both base learners.

For the meteorological baseline, the strongest variables were three-day rainfall lag, current rainfall, seasonal day-of-year position, one-day rainfall lag, maximum dewpoint, wind speed, dewpoint depression, and the 14-day rolling rainfall maximum. These features are consistent with antecedent wetness, atmospheric moisture, circulation, and seasonal forcing. However, Figure 5(b) explains the weighted XGBoost baseline only; it is not presented as an end-to-end explanation of stacking.

4.7 Limitations and Practical Implications

The evaluation is constrained by only 12 extreme days in the full record and four extreme days across the pooled test windows. Several occur in adjacent-day episodes, reducing the number of independent events further. The dataset is a single ERA5 grid-point reanalysis rather than a local rain-gauge, radar, or high-resolution satellite product. Reanalysis uncertainty and spatial mismatch may therefore affect local representativeness [17]-[20].

The observed false-alarm burden means that the stack is not operationally ready. It may serve as an initial screening layer whose candidates are subsequently confirmed using a more selective classifier, local observations, or impact-based information. Future work should perform event-based evaluation, probability calibration, external station validation, and multi-source fusion. Benchmarking should continue to report uncertainty and operational costs rather than only the most favorable metric [21], [23].

The small number of positives also limits conventional significance testing. A one-event change can alter fold-level recall from 0 to 1, and adjacent-day episodes are not fully independent. For this reason, the study reports exact detection counts and fold-specific metrics instead of relying only on averages. The results should be read as evidence about methodological behavior under severe scarcity, not as a definitive estimate of operational performance.

5. CONCLUSION

This study revised extreme-rainfall detection for Ternate as a leakage-controlled, next-day, severely imbalanced classification problem. Weighted baselines, balanced tree learning, focal-loss LSTM and Transformer models, and a corrected out-of-fold XGBoost stack were evaluated across three expanding temporal folds. The revised stacking predictions were no longer identical to those of the base learners and achieved zero missed extreme events in the pooled test periods. However, recall 1.000 was accompanied by precision 0.0023, specificity 0.2256, and 1,723 false positives. Balanced Random Forest offered the best overall trade-off, with F2-score 0.0316, specificity 0.7951, and MCC 0.0570, while detecting three of four events.

The findings therefore do not support a claim that stacking is the best model on every metric. Its defensible role is a high-sensitivity screening layer, while Balanced Random Forest is the stronger standalone trade-off model. Reliable operational deployment requires more extreme-event samples, local validation, improved spatial data, and a confirmation mechanism to reduce false alarms.

In direct response to the review questions, the revised experiment shows why several model families were retained. A single balanced tree model offers the best practical trade-off, while the corrected stack tests the distinct objective of eliminating missed events. The ensemble is therefore not presented as universally superior; its contribution is the transparent characterization of the sensitivity versus false-alarm boundary under chronological, imbalance-aware evaluation.

Future validation should combine ERA5 with gauges, radar or satellite precipitation, use event-based rather than only day-based scoring, assess probability calibration, and test a two-stage confirmation architecture. Until such validation is available, the proposed stack should be used only as a research

screening layer and not as a standalone public-warning mechanism.

6. REFERENCE

[1] B. Dong et al., "City-scale high-resolution flood nowcasting based on high-performance hydrodynamic modelling," *Int. J. Disaster Risk Reduct*., vol. 125, p. 105584, 2025, doi: 10.1016/j.ijdrr.2025.105584.

[2] H. Hersbach et al., "The ERA5 global reanalysis," Q. J. R. Meteorol. Soc., vol. 146, no. 730, pp. 1999-2049, 2020, doi: 10.1002/qj.3803.

[3] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng*., vol. 21, no. 9, pp. 1263-1284, 2009, doi: 10.1109/TKDE.2008.239.

[4] L. Mdegela, E. Municio, Y. De Bock, E. Luhanga, J. Leo, and E. Mannens, "Extreme rainfall event classification using machine learning for Kikuletwa River floods," *Water*, vol. 15, no. 6, p. 1021, 2023, doi: 10.3390/w15061021.

[5] A. Gumilar, S. S. Prasetiyowati, and Y. Sibaroni, "Performance analysis of hybrid machine learning methods on imbalanced data (rainfall classification)," *J. RESTI*, vol. 6, no. 3, pp. 481-490, 2022, doi: 10.29207/resti.v6i3.4142.

[6] S. An, T.-J. Oh, E. Sohn, and D. Kim, "Deep learning for precipitation nowcasting: A survey from the perspective of time series forecasting," *Expert Syst. Appl*., vol. 268, p. 126301, 2025, doi: 10.1016/j.eswa.2024.126301.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput*., vol. 9, no. 8, pp. 1735-1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[8] Y. Wu et al., "Effective LSTMs with seasonal-trend decomposition and adaptive learning and niching-based backtracking search algorithm for time series forecasting," *Expert Syst. Appl*., vol. 236, p. 121202, 2024, doi: 10.1016/j.eswa.2023.121202.

[9] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, pp. 5998-6008.

[10] S. Chkeir, A. Anesiadou, A. Mascitelli, and R. Biondi, "Nowcasting extreme rain and extreme wind speed with machine learning techniques applied to different input datasets," *Atmos. Res*., vol. 282, p. 106548, 2023, doi: 10.1016/j.atmosres.2022.106548.

[11] S. He, Z. Li, and X. Liu, "An improved GEV boosting method for imbalanced data classification with application to short-term rainfall prediction," *J. Hydrol*., vol. 617, p. 128882, 2023, doi: 10.1016/j.jhydrol.2022.128882.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785-794, doi: 10.1145/2939672.2939785.

[13] Y. Xie, W. Sun, M. Ren, S. Chen, Z. Huang, and X. Pan, "Stacking ensemble learning models for daily runoff prediction using 1D and 2D CNNs," *Expert Syst. Appl*., vol. 217, p. 119469, 2023, doi: 10.1016/j.eswa.2022.119469.

[14] M. S. Ahmad, H. Hadiyanto, and R. Sanjaya, "A hybrid deep learning and tree boosting approach for BBCA stock price forecasting with SHAP explainability," *JIKO (Jurnal Informatika dan Komputer)*, vol. 9, no. 1, pp. 68-73, 2026, doi: 10.33387/jiko.v9i1.11102.

[15] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, p. e0118432, 2015, doi: 10.1371/journal.pone.0118432.

[16] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765-4774.

[17] F. J. Hop, R. Linneman, B. Schnitzler, A. Bomers, and M. J. Booij, "Real time probabilistic inundation forecasts using a LSTM neural network," *J. Hydrol*., vol. 635, p. 131082, 2024, doi: 10.1016/j.jhydrol.2024.131082.

[18] F. Zhu et al., "A hybrid process-data driven framework for real-time hydrological forecasting with interpretable deep learning," *J. Hydrol*., vol. 662, p. 134082, 2025, doi: 10.1016/j.jhydrol.2025.134082.

[19] A. B. Upadhyay, S. R. Shah, and R. A. Thakker, "Advanced rainfall nowcasting using 3D convolutional LSTM networks on satellite data," *J. Comput. Math. Data Sci*., vol. 16, p. 100125, 2025, doi: 10.1016/j.jcmds.2025.100125.

[20] F. Piadeh, V. Bakhtiari, and F. Piadeh, "Automated novel real-time framework for rainfall data imputation in flood early warning systems," *Eng. Appl. Artif. Intell*., vol. 164, p. 113348, 2026, doi: 10.1016/j.engappai.2025.113348.

[21] M. Oprea, "A general framework and guidelines for benchmarking computational intelligence algorithms applied to forecasting problems derived from an application domain-oriented survey," *Appl. Soft Comput*., vol. 89, p. 106103, 2020, doi: 10.1016/j.asoc.2020.106103.

[22] Copernicus Climate Change Service, "ERA5 hourly data on single levels from 1940 to present," *Climate Data Store*, 2023, doi: 10.24381/cds.adbb2d47.

[23] S. Bao, C. Yu, and Z. Wan, "Spatial mismatch and attribution analysis of flood risk and resilience in megacities: Insights from a risk-resilience-effect framework and an interpretable XGBoost-SHAP model," *Sustain. Cities Soc*., vol. 131, p. 106758, 2025, doi: 10.1016/j.scs.2025.106758.