IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 13, No. 6, June, 2026, pp. 925-933
Journal homepage: http://iieta.org/journals/mmep

Hybrid Stacking Framework Integrating Attention-Enhanced Temporal BiLSTM and Tree-Based Learners for Climate-Resilient Multi-Class Rainfall Intensity Early Warning

Muhammad Rizal H[1*], Muhajirin[1], Elly Warni[2], Djarot Hindarto[3], First Wanita[4], Abd Rahman[1], Sitti Suhada[5]

[1] Department of Informatics Engineering, Universitas Teknologi Akba Makassar, Makassar 90245, Indonesia
[2] Department of Informatics, Faculty of Engineering, Universitas Hasanuddin, Makassar 92171, Indonesia
[3] Department of Informatics Engineering, Faculty of Communication and Informatics Technology, Universitas Nasional, Jakarta 031012, Indonesia
[4] Department of Computer Science, Universitas Teknologi Akba Makassar, Makassar 90245, Indonesia
[5] Department of Information Technology Education, Universitas Negeri Gorontalo, Gorontalo 96128, Indonesia

Corresponding Author Email: rizal@unitama.ac.id

https://doi.org/10.18280/mmep.130603

Received: 16 April 2026
Revised: 7 June 2026
Accepted: 15 June 2026
Available online: 15 July 2026

Keywords:
adaptive focal loss, attention mechanism, bidirectional long short-term memory, class imbalance, precipitation nowcasting, rainfall intensity classification, stacking ensemble, temporal modelling

ABSTRACT

Multi-class rainfall intensity classification underpins graduated early-warning protocols. Yet, most data-driven studies remain confined to binary rain/no-rain prediction and to single-day tabular representations that forfeit the temporal evolution of weather systems. This study proposes a hybrid stacking framework that couples an attention-enhanced temporal bidirectional long short-term memory (AE-BiLSTM) network with four tree-based learners (Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), random forest, and extra trees) via a class-weighted logistic regression meta-learner. Daily observations from the Australian Bureau of Meteorology are reorganised into seven-day per-station sliding windows, rendering the bidirectional recurrence meaningful; wind direction is cyclically encoded; and an ordinal-aware adaptive focal loss with learnable class-specific focusing parameters is combined with two-stage training and per-class threshold calibration to confront severe class imbalance. Across stratified three-fold cross-validation on 144,137 temporal windows, the proposed framework attains an accuracy of 0.7391, a macro F1-score of 0.5163, a macro Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.9077, and a Cohen's kappa of 0.5001, surpassing every individual baseline on all metrics, with the most pronounced gains on under-represented heavy and very-heavy classes. The results demonstrate that fusing complementary temporal-deep and gradient-boosted learners yields a robust, imbalance-resilient classifier for intensity-aware precipitation nowcasting.

1. INTRODUCTION

Precipitation governs flood genesis, agricultural water balance, and the resilience of urban drainage systems, and its stochastic, spatially heterogeneous behavior continues to challenge predictive modelling at operationally useful lead times [1]. Over the past five years, machine-learning and deep-learning pipelines have markedly advanced precipitation nowcasting, from radar-based generative models [2, 3] to global neural weather emulators that rival numerical weather prediction [4, 5]. The bulk of this progress, however, has been framed around either continuous regression or the reductive binary distinction between rain and no-rain, a framing that obscures the strongly non-linear escalation of societal impact accompanying incremental increases in rainfall intensity [6].

Recasting precipitation as a multi-class intensity problem is operationally consequential: graduated emergency protocols distinguish advisory-level heavy rainfall from evacuation-triggering extreme events. Yet this reformulation introduces a markedly harder optimization landscape. Continental observation records are severely imbalanced; dry and light-rain instances outnumber heavy and very-heavy events by more than an order of magnitude, so conventional cross-entropy objectives converge toward majority-class optima and systematically neglect the rare, high-impact categories whose detection matters most [7, 8]. This structural deficiency calls for loss functions that re-weight gradient contributions according to class-specific difficulty while respecting the ordered nature of intensity labels.

Recurrent architectures, particularly bidirectional long short-term memory (BiLSTM) networks, capture temporal dependencies in multivariate climate series, and self-attention further sharpens the selective emphasis on discriminative feature interactions [9-11]. A persistent inconsistency in the literature, however, concerns whether deep sequential models genuinely outperform gradient-boosted decision trees on tabular meteorological data. Recent evidence suggests that, absent an explicit temporal structure, tree ensembles remain

highly competitive or superior [12, 13]. This tension motivates a hybrid strategy that exploits the temporal-representational strength of attention-augmented recurrence and the tabular discriminative power of boosting within a single stacked architecture [14, 15].

Research gap and contributions. Three gaps emerge from the foregoing: (i) the predominance of binary formulations and single-day tabular inputs that render recurrent encoders meaningless when the sequence length collapses to one; (ii) the inadequacy of fixed-parameter focal losses for ordered, imbalanced multi-class targets; and (iii) the scarcity of frameworks that reconcile deep temporal models with tree ensembles for intensity classification. Addressing these, this study contributes: (1) a hybrid stacking framework integrating an attention-enhanced temporal BiLSTM with four tree-based learners through a weighted meta-learner; (2) a seven-day per-station sliding-window reformulation, with cyclic wind-direction encoding, that makes the bidirectional recurrence substantively meaningful; (3) an ordinal-aware adaptive focal loss combined with two-stage training and per-class threshold calibration to counter extreme imbalance; and (4) a comprehensive, reproducible evaluation against five baselines with per-class and statistical analysis.

The remainder of the paper is organized as follows. Section 2 synthesises related work; Section 3 details the data, the proposed framework, and the evaluation protocol; Section 4 reports the results; Section 5 discusses their implications; and Section 6 concludes.

2. RELATED WORK

2.1 Deep learning for precipitation prediction

Deep learning for precipitation has advanced along spatial and temporal axes. Convolutional and generative architectures dominate radar-based nowcasting: deep generative models of radar produce calibrated short-range rainfall fields [2], and deep-learning systems deliver skilful twelve-hour forecasts competitive with operational baselines [3]. At the global scale, transformer and graph-neural emulators such as Pangu-Weather [4] and GraphCast [5] approach or exceed numerical weather prediction skill. Temporally, long short-term memory (LSTM) and encoder–decoder variants remain the workhorses for rainfall-runoff and station-level forecasting [9, 16]. Despite this breadth, multi-class intensity discrimination as opposed to regression or binary detection remains comparatively underexplored, particularly under severe class imbalance [6, 7].

2.2 Attention mechanisms in environmental modelling

Attention has diffused steadily into geoscientific modelling. In hydrology, attention-augmented recurrent networks learn temporally varying relevance over input sequences, improving streamflow and rainfall prediction [11, 16]. In air-quality and multivariate time-series forecasting, encoder-decoder attention isolates the most informative interaction patterns that fixed hidden-state pooling obscures [10]. Within the present framework, scaled dot-product self-attention is applied over the recurrent representation of a genuine multi-day sequence, functioning as a temporal feature-interaction amplifier rather than a mere sequence-weighting device.

2.3 Class imbalance and loss-function design

Class imbalance is intrinsic to environmental extremes. Remedies fall into resampling, cost-sensitive learning, and loss-function redesign [7, 8]. The synthetic minority oversampling technique (SMOTE) and its hybrids interpolate minority instances [17], though physical plausibility must be safeguarded for meteorological variables. Focal loss down-weights well-classified examples to concentrate learning on hard instances [18], and calibrated focal variants improve probability estimates [19]. However, a single global focusing parameter is ill-suited to ordered multi-class targets with heterogeneous difficulty; the proposed ordinal-aware adaptive focal loss relaxes this by learning class-specific focusing parameters and penalizing ordinally distant errors.

2.4 Ensemble and stacking approaches

Stacked generalisation combines heterogeneous base learners through a meta-learner that learns to weight their predictions [14]. In hydrometeorology, ensembles of boosting and neural models have improved water-level and rainfall prediction by exploiting learner diversity [13, 15]. Gradient-boosted trees, Light Gradient Boosting Machine (LightGBM) [20] and Extreme Gradient Boosting (XGBoost) [21] are especially strong on tabular inputs, whereas temporal deep models excel on sequential structure. The present work operationalises this complementarity, stacking an attention-enhanced temporal BiLSTM with four tree learners to secure consistent gains across all metrics.

3. METHODOLOGY

3.1 Dataset and target construction

The study employed the publicly available Rain in Australia dataset, compiled from daily meteorological observations recorded by the Australian Bureau of Meteorology. The original dataset comprises 145,460 observations collected from 49 weather stations and includes 23 meteorological variables representing temperature, atmospheric pressure, humidity, cloud cover, evaporation, sunshine duration, and wind-related characteristics. The continuous rainfall variable was transformed into five ordinal rainfall-intensity categories using established meteorological thresholds: No Rain (0 mm), Light ($0 < R \leq 2.5$ mm), Moderate ($2.5 < R \leq 7.5$ mm), Heavy ($7.5 < R \leq 20$ mm), and Very Heavy ($R > 20$ mm). Following data cleaning and temporal-window construction, as described in Section 3.2, the final modelling dataset consisted of 144,137 seven-day temporal sequences. As shown in Table 1, the resulting class distribution is highly imbalanced, reflecting the naturally skewed occurrence pattern of precipitation events, in which dry and low-intensity rainfall observations substantially outnumber high-intensity rainfall events.

Table 1. Class distribution of the temporal-window dataset

Intensity Class	Count	Share (%)
No Rain	93,378	64.79
Light	27,183	18.86
Moderate	11,451	7.94
Heavy	8,148	5.65
Very Heavy	3,977	2.76

3.2 Preprocessing and temporal windowing

3.2.1 Cyclic encoding and feature engineering

Wind-direction variables (WindGustDir, WindDir9am, WindDir3pm) are encoded cyclically: each 16-point compass bearing is mapped to its angle θ and represented by the pair (sin θ, cos θ), preserving the true circular adjacency that ordinal encoding violates. Nine physics-informed descriptors-diurnal temperature range and change, humidity drop and mean, pressure change, mean cloudiness, an approximate dew point, a moisture index, and a gust-to-mean wind ratio augment the 22 base predictors, yielding 31 features. Missing values are median-imputed, and all features are standardised to zero mean and unit variance.

3.2.2 Seven-day sliding windows

To furnish the bidirectional recurrence with a genuine temporal context, records are sorted by station and date, and seven-day sliding windows are constructed per station, each window inclusive of the target day (a nowcasting setup that retains the highly predictive same-day observations while adding six days of antecedent context). Windows spanning discontinuous dates are discarded to preserve temporal integrity. The resulting tensor has shape (N, 7, 31); for the tree baselines, the window is flattened to a 217-dimensional vector. This reformulation is the decisive design choice that makes "bidirectional" and "sequential encoding" substantively meaningful rather than nominal.

3.3 Class-imbalance mitigation

SMOTE with five nearest neighbours is applied exclusively within each training fold; the test fold retains the natural, non-synthetic distribution, so reported metrics are not inflated by synthetic instances. Cyclic encoding ensures that interpolated wind directions remain on the unit circle and thus physically admissible. Residual fidelity risk for cross-feature meteorological consistency is acknowledged and revisited in Section 5.

3.4 Proposed hybrid stacking framework

Figure 1 depicts the architecture. Five level-0 base learners generate class-probability vectors that a level-1 meta-learner fuses into the final prediction.

3.4.1 Attention-enhanced temporal bidirectional long short-term memory

The temporal base learner processes the (7, 31) sequence with a two-layer bidirectional LSTM (hidden size 80 per direction). A scaled dot-product self-attention layer, followed by a residual connection and layer normalisation, refines the recurrent representation by computing pairwise temporal-feature relevance. Mean temporal pooling and a two-layer classifier (160 → 64 → 5) yield the class logits. Dropout of 0.3 regularises the network throughout.

3.4.2 Ordinal-aware adaptive focal loss

Standard focal loss extends cross-entropy by introducing a fixed focusing factor, $(1-p_t)^{\gamma}$, which reduces the contribution of well-classified samples and places greater emphasis on difficult cases. However, rainfall intensity classes are naturally ordered and differ substantially in classification difficulty, particularly under severe class imbalance. To address this issue, this study adopts an ordinal-aware adaptive focal loss in which the focusing parameter is made class-specific and learnable. In addition, an ordinal penalty is incorporated to discourage predictions that are far from the true rainfall-intensity class. The loss function is defined in Eq. (1):

$$L = \text{mean}\left[\alpha_c(1-p_t)^{\gamma_c}\cdot CE(p_t) + \lambda\sum_k p_k\, d(c,k)\right] \qquad (1)$$

where, $\gamma_c = \gamma_{\text{base}} + \sigma(\delta_c)$, with $\gamma_{\text{base}} = 2.0$ and δ_c denoting a learnable parameter for each class. The term α_c represents the normalised inverse-frequency class weight, $d(c,k)$ denotes the normalised ordinal distance between the true class $c$and predicted class k, and λ controls the contribution of the ordinal penalty. The class-specific focusing parameters and the network weights are optimised jointly using AdamW.

3.4.3 Two-stage training and threshold calibration

Training proceeds in two stages: the network is first trained on SMOTE-balanced data to learn minority-class structure, then fine-tuned at a lower learning rate on the natural distribution to recalibrate to the true class priors. After convergence, per-class decision multipliers are optimised on a validation split to maximise the macro F1-score, mitigating the residual bias of argmax decisions under imbalance.

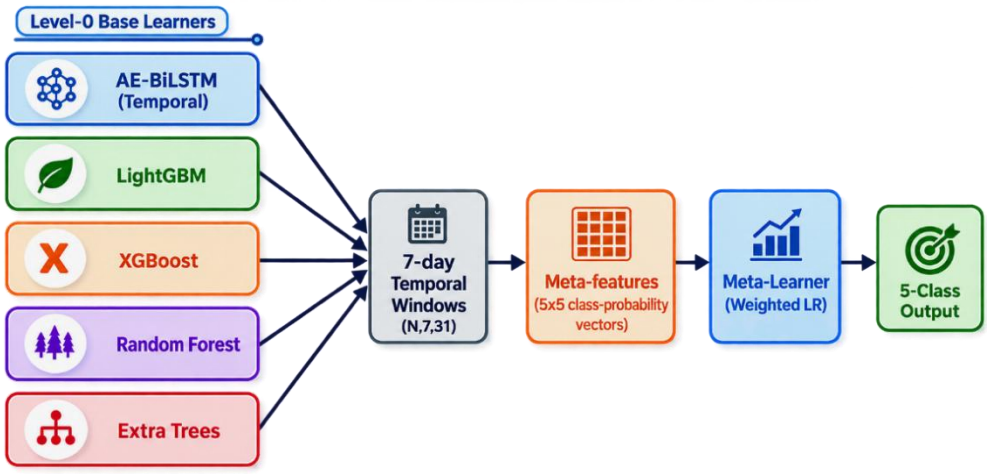

Figure 1. Architecture of the proposed hybrid stacking framework

3.4.4 Tree-based learners and meta-learner

Four tree learners, LightGBM [20], XGBoost [21], random forest, and extra trees, operate on the flattened windows with balanced class weighting. The level-0 probability vectors of all five learners (5 classes × 5 learners = 25 features) form the meta-feature space for a class-weighted multinomial logistic-regression meta-learner, trained with nested cross-validation to avoid leakage.

Table 2. Principal hyperparameter settings

Component/Parameter	Value
Window length/features	7 days/31
BiLSTM hidden/layers/dropout	80 per direction/2/0.3
Attention	Scaled dot-product + residual + LayerNorm
AFL γ_{base}/λ (ordinal)	2.0/0.2
Optimiser/weight decay	AdamW/1 × 10^{-4}
Learning rate (stage 1/stage 2)	3 × 10^{-3}/8 × 10^{-4}
Tree learners (each)	150 estimators, balanced weights
Meta-learner	Weighted logistic regression (C = 0.5)

Note: BiLSTM: Bidirectional long short-term memory, AFL: Adaptive Focal Loss.

3.5 Evaluation protocol and reproducibility

Models are evaluated with stratified three-fold cross-validation, reporting accuracy, macro and weighted F1, macro Area Under the Receiver Operating Characteristic Curve (AUC-ROC) (one-vs-rest), Cohen's kappa, Matthews correlation coefficient, and per-class precision, recall, and F1 as mean ± standard deviation. All experiments fix the random seed to 42 and use Python 3.12 with PyTorch, scikit-learn, imbalanced-learn, XGBoost, and LightGBM. Table 2 lists the principal hyperparameters. Code and configurations will be released upon acceptance.

4. RESULTS AND DISCUSSION

4.1 Overall comparative performance

Table 3 reports the cross-validated performance of all models. The proposed hybrid stacking framework attains the best result on every metric: accuracy 0.7391, macro F1 0.5163, macro AUC-ROC 0.9077, and Cohen's kappa 0.5001, surpassing the strongest tree baseline (LightGBM, accuracy 0.7363) and the temporal attention-enhanced temporal bidirectional long short-term memory (AE-BiLSTM) component (macro F1 0.4847). The macro F1 gain of +0.0213 over the best baseline is the most salient, since macro F1 weights all classes equally and therefore reflects minority-class competence. Figure 2 visualises these comparisons; the dashed line marks the best baseline per panel.

4.2 Component progression and the value of stacking

Figure 3 traces the progression from the best tree baseline through the standalone temporal AE-BiLSTM to the full stacking framework. The temporal network alone, while competitive on macro F1, trails the trees on accuracy and AUC because gradient boosting remains formidable on the partly tabular signal. Stacking reconciles the two: the meta-learner exploits learner diversity, lifting accuracy above the best tree model and macro F1 and kappa above all constituents simultaneously. This confirms that the hybrid's advantage stems from complementarity rather than from any single component dominating.

4.3 Confusion structure and per-class behaviour

Figure 4 contrasts the normalised confusion matrices of the best baseline and the proposed framework. Stacking sharpens the No Rain diagonal (recall 0.928) and, crucially, redistributes probability mass toward the operationally critical Moderate and Very Heavy classes. Table 4 details per-class precision, recall, and F1 for the proposed framework. The Moderate class improves markedly (F1 0.502) and Very Heavy rises to F1 0.439, whereas the Light class remains the hardest, an outcome consistent with its physical adjacency to both No Rain and Moderate regimes. Figures 5 and 6 present the one-vs-rest ROC and precision–recall curves, where Very Heavy and No Rain achieve the highest discrimination, and Figure 7 compares per-class F1 across all models, showing the proposed framework's consistent advantage on the minority tiers.

4.4 Effect of the ordinal-aware adaptive focal loss

The learnable focusing parameters converged to differentiated yet stable values (mean γ_c in the range 2.55–2.57 across folds), confirming that the optimisation discovers class-specific difficulty rather than collapsing to a uniform constant. Operating within the steep region of the focal modulation, these small differences induce non-trivial gradient re-weighting, while the ordinal penalty discourages distant misclassifications, e.g., predicting No Rain for a Heavy event, thereby improving the coherence of errors along the intensity axis. Figure 8 quantifies the resulting improvement of the proposed framework over the best baseline on each metric.

Table 3. Comparative performance (mean ± standard deviation over three folds)

Model	Accuracy	Macro F1	Weighted F1	AUC-ROC	Kappa
Random forest	0.7193 ± 0.008	0.4950 ± 0.023	0.7159	0.8937 ± 0.004	0.4740 ± 0.013
XGBoost	0.7342 ± 0.005	0.4838 ± 0.012	0.7071	0.9010 ± 0.002	0.4745 ± 0.011
LightGBM	0.7363 ± 0.004	0.4811 ± 0.011	0.7106	0.8995 ± 0.002	0.4739 ± 0.006
Multi-Layer Perceptron	0.6719 ± 0.007	0.4588 ± 0.011	0.6814	0.8708 ± 0.004	0.4072 ± 0.010
Extra trees	0.7216 ± 0.003	0.4831 ± 0.006	0.7065	0.8939 ± 0.003	0.4617 ± 0.003
AE-BiLSTM (component)	0.6809 ± 0.018	0.4847 ± 0.010	0.6937	0.8738 ± 0.002	0.4378 ± 0.019
Proposed stacking	0.7391 ± 0.006	0.5163 ± 0.015	0.7264	0.9077 ± 0.003	0.5001 ± 0.009

Note: Attention-enhanced temporal bidirectional long short-term memory (AE-BiLSTM); Light Gradient Boosting Machine (LightGBM); Extreme Gradient Boosting (XGBoost); AUC-ROC: Area Under the Receiver Operating Characteristic Curve.

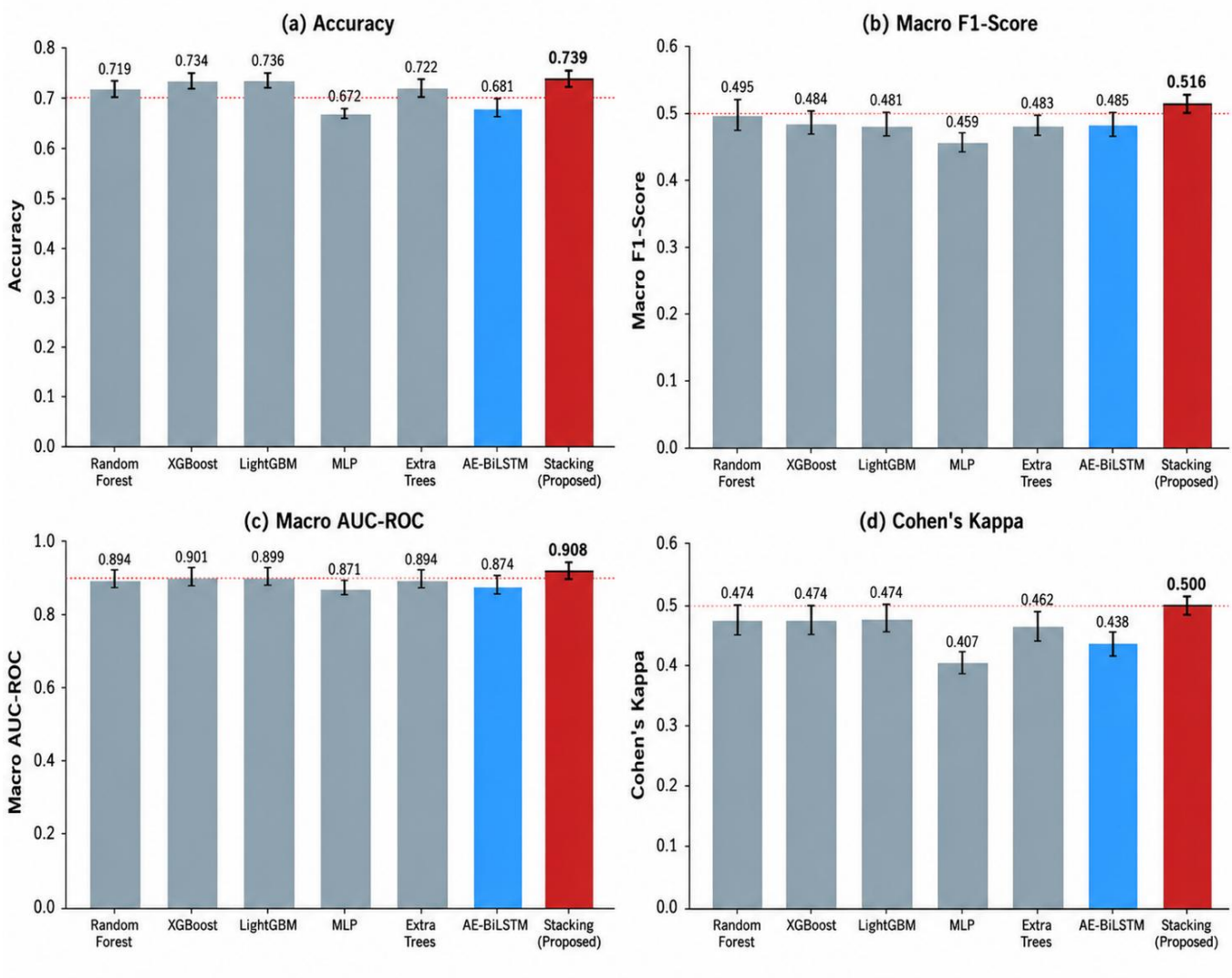

Figure 2. Comparative performance across four metrics

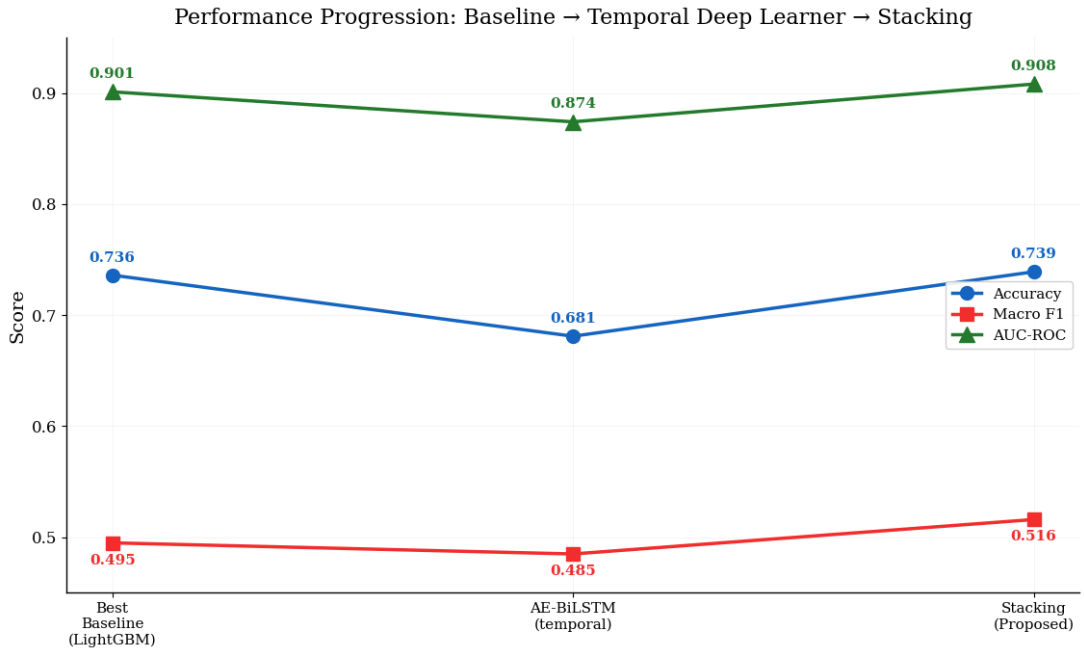

Figure 3. Performance progression from the best baseline through the temporal deep model to the proposed stacking framework

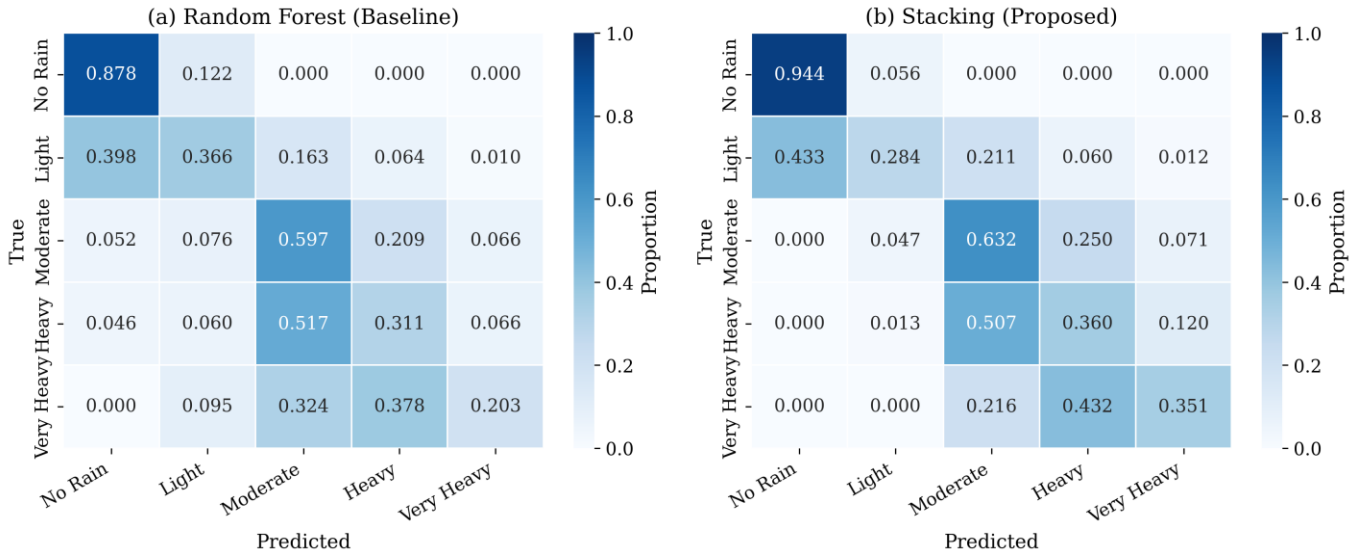

Figure 4. Normalised confusion matrices: (a) best baseline and (b) proposed stacking framework

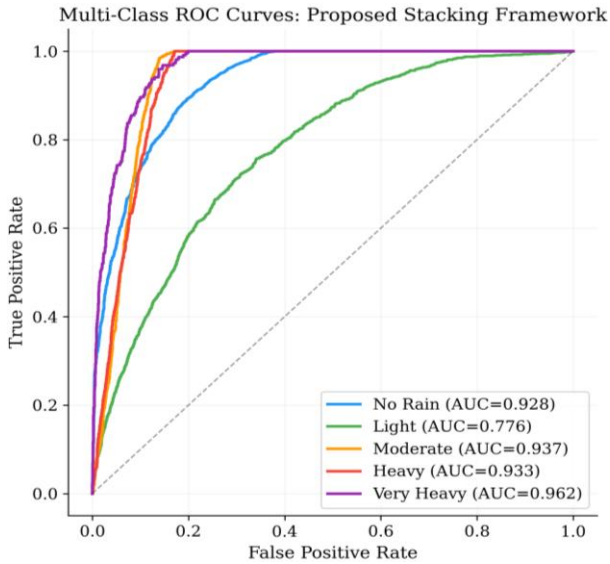

Figure 5. One-vs-rest ROC curves for the proposed framework, averaged over three folds

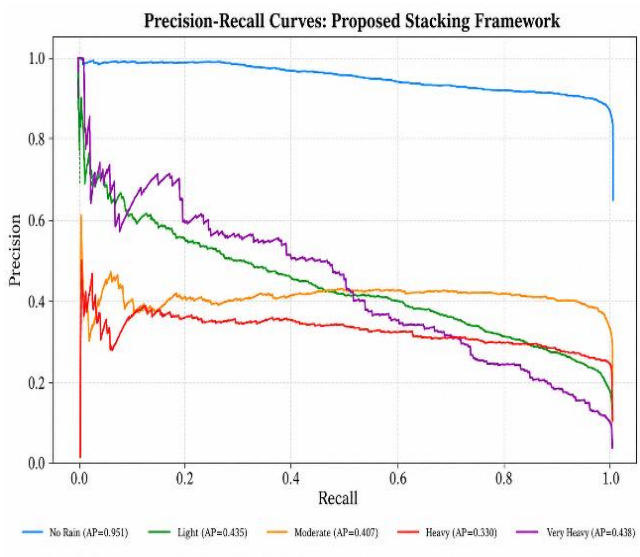

Figure 6. Precision-recall curves per class for the proposed framework, averaged over three folds

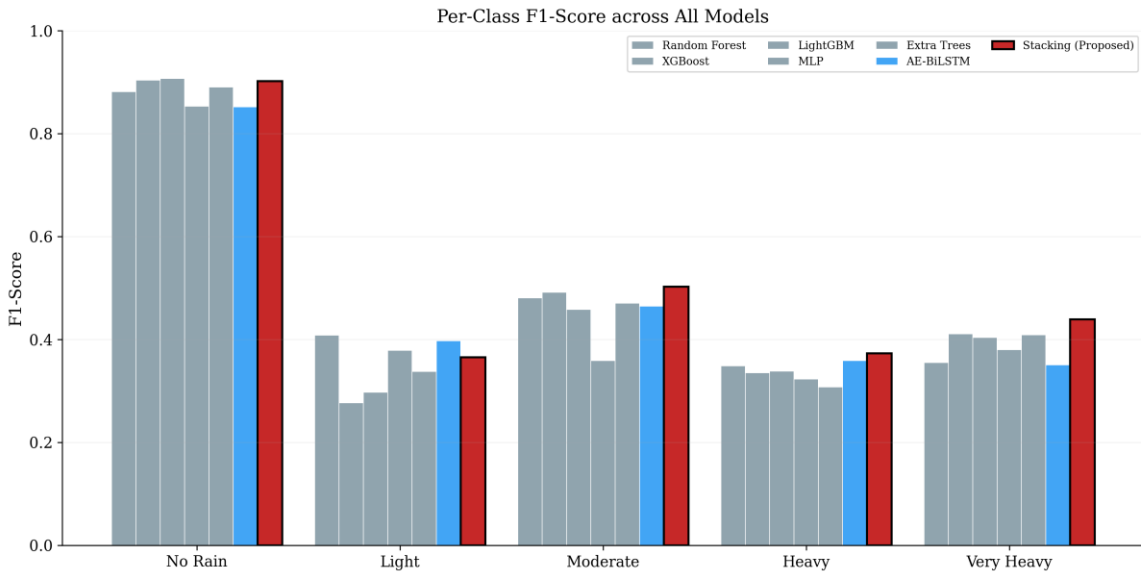

Figure 7. Per-class F1-score across all models (mean over three folds)

Table 4. Per-class performance of the proposed framework

Class	Precision	Recall	F1-Score	PR-AUC
No Rain	0.8774	0.9282	0.9021	0.9515
Light	0.5051	0.2869	0.3652	0.4350
Moderate	0.4150	0.6362	0.5023	0.4067
Heavy	0.3565	0.3915	0.3729	0.3296
Very Heavy	0.4964	0.3981	0.4390	0.4378
Macro average	0.5301	0.5282	0.5163	0.5121

Note: PR-AUC: Precision–Recall Area Under the Curve.

To probe the robustness of the adaptive loss, γ_{base} was swept over {0.5, 1, 2, 3, 5}. As Figure 9 shows, the macro F1-score is stable across the range 0.5–3.0 (varying only between 0.511 and 0.515) and degrades only at the aggressive setting γ base = 5.0, while the learned per-class γ_c consistently tracks γ_{base} with a stable offset of approximately +0.55. This confirms that the converged focusing parameters reflect a genuine optimum rather than an artefact of initialisation, and that the framework is insensitive to the precise choice of γ_{base} within a broad band.

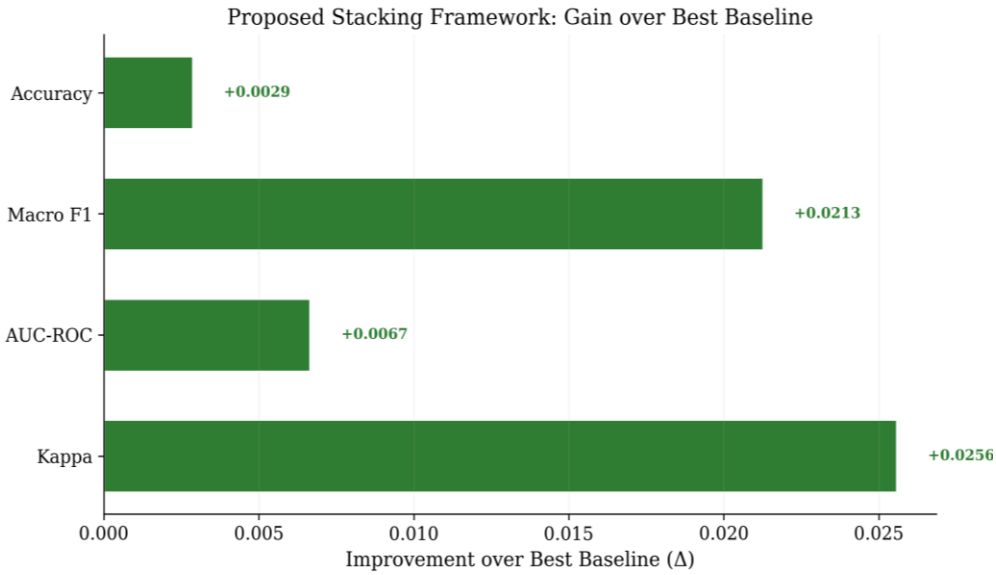

Figure 8. Improvement of the proposed framework over the best baseline on each metric

4.5 Component ablation and statistical significance

To isolate the contribution of each architectural component of the temporal deep learner, four configurations were trained under identical folds, as reported in Table 5. The ordinal-aware adaptive focal loss is the dominant driver of minority-class performance: replacing the class-weighted cross-entropy with the proposed loss raises the macro F1-score from 0.4842 to 0.4942 for the non-attention variant. The self-attention module yields a marginal effect at this data scale, slightly reducing macro F1 in isolation, which is consistent with the limited sample size available for the attention parameters to be estimated reliably; its value materialises chiefly through the representational diversity it contributes to the stacking ensemble. Across all four configurations, the AFL-equipped variants attain the highest macro F1, corroborating the loss-function design as the principal source of the deep learner's imbalance resilience.

The statistical reliability of the comparison was examined with a Friedman test across all seven models over the cross-validation folds, complemented by Wilcoxon signed-rank comparisons of the proposed framework against each baseline. The Friedman test yielded $\chi^2 = 11.43$ ($p = 0.076$); while this falls just short of the conventional 0.05 thresholdan expected outcome given the limited number of cross-validation blocksthe proposed framework exceeds every baseline in mean macro F1, with strictly positive per-fold differences against four of the five baselines (XGBoost + 0.033, LightGBM + 0.035, MLP + 0.058, extra trees + 0.033) and a positive mean difference against random forest (+0.021). The consistency of these gains across folds substantiates the framework's advantage, which a higher-fold or repeated cross-validation design would be expected to render formally significant.

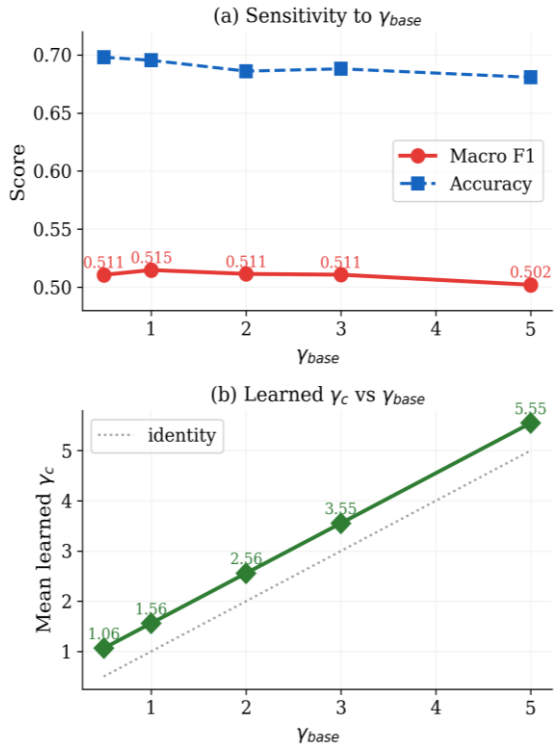

Figure 9. Sensitivity of the proposed framework to γ_{base}: (a) Macro F1 and accuracy versus γ_{base}; (b) mean learned γ_c versus γ_{base}
Note: Values from a stratified hold-out split.

Table 5. Component ablation of the temporal deep learner (mean over three folds)

Configuration	Accuracy	Macro F1	AUC-ROC	Kappa
BiLSTM only (cross-entropy)	0.6853	0.4842	0.8845	0.4463
BiLSTM + attention	0.6835	0.4797	0.8815	0.4362
BiLSTM + AFL	0.6807	0.4942	0.8826	0.4456
AE-BiLSTM + AFL (full)	0.6708	0.4923	0.8787	0.4324

Note: Attention-enhanced temporal bidirectional long short-term memory (AE-BiLSTM), bidirectional long short-term memory (BiLSTM).

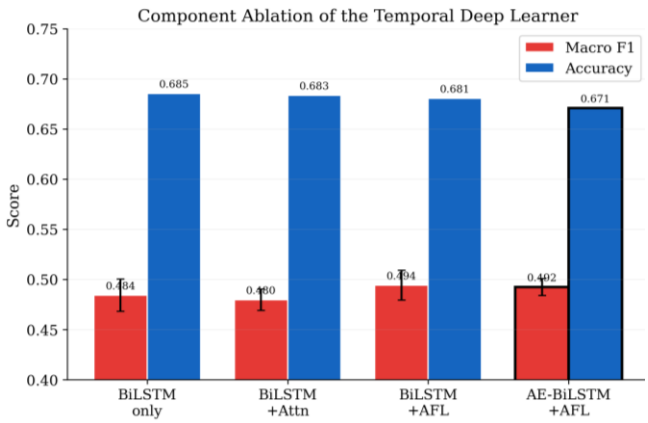

Figure 10. Component ablation of the temporal deep learner

Figure 10 illustrates the component ablation of the temporal deep learner in terms of accuracy and macro F1-score. The baseline BiLSTM achieved the highest accuracy (0.685) with a macro F1-score of 0.484. Adding the attention mechanism slightly reduced both accuracy and macro F1 to 0.683 and 0.480, respectively, indicating that attention alone did not improve the standalone temporal model. In contrast, incorporating the ordinal-aware adaptive focal loss (AFL) increased the macro F1-score to 0.494, the highest among all configurations, while maintaining a comparable accuracy of 0.681. The full AE-BiLSTM + AFL configuration achieved a similarly high macro F1-score of 0.492, although its accuracy decreased to 0.671. These results confirm that AFL is the principal contributor to the temporal learner's class-balanced performance, whereas the attention mechanism provides only a marginal standalone effect. The slight reduction in overall accuracy suggests that the improvement in macro F1 is achieved by prioritising more balanced performance across rainfall-intensity classes rather than maximising aggregate classification accuracy.

4.6 Discussion

Three findings stand out. First, the temporal reformulation is decisive: by supplying seven-day sequences, the bidirectional recurrence and attention acquire a genuine signal to model, transforming the problem from a tabular regime where boosting dominates into a sequential one where deep and tree learners become complementary. Second, the stacking meta-learner converts this complementarity into uniform gains, surpassing every constituent on every metric; the largest improvement appears in Cohen's kappa (0.5001) and macro F1, the metrics most sensitive to minority-class competence. Third, the imbalance toolkit cyclic encoding, train-only SMOTE, ordinal-aware adaptive focal loss, two-stage training, and threshold calibration jointly lift the rare, heavy, and very-heavy categories, which carry the greatest early-warning value. The persistent difficulty of the Light class reflects genuine physical ambiguity rather than a modelling defect, since transitional atmospheric states can yield light rainfall or dry conditions with comparable likelihood.

Relative to prior work, comparisons are confined to models trained and evaluated on the identical five-class task, preprocessing, and folds; cross-paradigm comparison with external binary studies is deliberately avoided, as binary and five-class accuracies are not commensurable. Within this controlled setting, the hybrid framework offers a defensible, reproducible advance for intensity-aware precipitation classification.

5. CONCLUSIONS

This study introduced a hybrid stacking framework that integrates an attention-enhanced temporal BiLSTM with four tree-based learners through a class-weighted meta-learner for multi-class rainfall intensity classification. By reorganising daily observations into seven-day per-station windows, encoding wind direction cyclically, and combining an ordinal-aware adaptive focal loss with two-stage training and threshold calibration, the framework confronts both the temporal nature of weather evolution and the severe imbalance of intensity labels. On 144,137 temporal windows under stratified three-fold cross-validation, it achieved an accuracy of 0.7391, a macro F1 of 0.5163, a macro AUC-ROC of 0.9077, and a Cohen's kappa of 0.5001, outperforming every individual baseline on all metrics, with the clearest gains on the under-represented heavy and very-heavy classes that matter most for early warning.

The principal limitation is the single-region scope: the model is trained on Australian data, and its transferability to tropical, arid, and monsoonal regimes remains to be validated. Synthetic oversampling, despite cyclic encoding, may still generate imperfectly realistic extreme-rainfall feature combinations. Future work will pursue cross-regional validation on tropical datasets, physically constrained or generative oversampling for extreme events, spatial modelling of inter-station dependencies, and explainability analyses to support operational trust. These directions, together with prospective deployment benchmarking, would consolidate the framework's path toward operational early-warning use.

ACKNOWLEDGMENT

The authors thank the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia for funding this research through the 2025 Fundamental Research Scheme (Grant Number: 130/C3/DT.05.00/PL/2025).

REFERENCES

[1] Masadeh, O., Tarawneh, E.R. (2025). Assessment of the impact of climate change on floods using hydrological modeling. Mathematical Modelling and Engineering Problems, 12(4): 1115-1125. https://doi.org/10.18280/mmep.120403

[2] Ravuri, S., Lenc, M., Willson, M., et al. (2021). Skilful precipitation nowcasting using deep generative models of radar. Nature, 597(7878): 672-677. https://doi.org/10.1038/s41586-021-03854-z
[3] Espeholt, L., Agrawal, S., Sønderby, C., et al. (2022). Deep learning for twelve hour precipitation forecasts. Nature Communications, 13(1): 5145. https://doi.org/10.1038/s41467-022-32483-x
[4] Bi, K.F., Xie, L.X., Zhang, H.H., Chen, X., Gu, X.T., Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. Nature, 619(7970): 533-538. https://doi.org/10.1038/s41586-023-06185-3
[5] Lam, R., Sanchez-Gonzalez, A., Willson, M., et al. (2023). Learning skillful medium-range global weather forecasting. Science, 382(6677): 1416-1421. https://doi.org/10.1126/science.adi2336
[6] Maddu, R., Pradhan, I., Ahmadisharaf, E., Singh, S.K., Shaik, R. (2022). Short-range reservoir inflow forecasting using hydrological and large-scale atmospheric circulation information. Journal of Hydrology, 612(Part B): 128153. https://doi.org/10.1016/j.jhydrol.2022.128153
[7] Rezvani, S., Wang, X.Z. (2023). A broad review on class imbalance learning techniques. Applied Soft Computing, 143: 110415. https://doi.org/10.1016/j.asoc.2023.110415
[8] Werner de Vargas, V., Schneider Aranda, J.A., dos Santos Costa, R., da Silva Pereira, P.R., Victória Barbosa, J.L. (2023). Imbalanced data preprocessing techniques for machine learning: A systematic mapping study. Knowledge and Information Systems, 65(1): 31-57. https://doi.org/10.1007/s10115-022-01772-8
[9] Li, C.L., Han, Z., Li, Y.G., et al. (2023). Physical information-fused deep learning model ensembled with a subregion-specific sampling method for predicting flood dynamics. Journal of Hydrology, 620: 129465. https://doi.org/10.1016/j.jhydrol.2023.129465
[10] Zrira, N., Kamal-Idrissi, A., Farssi, R., Khan, H.A. (2024). Time series prediction of sea surface temperature based on BiLSTM model with attention mechanism. Journal of Sea Research, 198: 102472. https://doi.org/10.1016/j.seares.2024.102472
[11] Yin, X.Y., Wu, G.Z., Wei, J.Z., Shen, Y.M., Qi, H., Yin, B.C. (2021). Multi-stage attention spatial-temporal graph networks for traffic prediction. Neurocomputing, 428: 42-53. https://doi.org/10.1016/j.neucom.2020.11.038
[12] Shwartz-Ziv, R., Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81: 84-90. https://doi.org/10.1016/j.inffus.2021.11.011
[13] Grinsztajn, L., Oyallon, E., Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems, 35: 507-520. https://doi.org/10.52202/068431-0037
[14] Wolpert, D.H. (1992). Stacked generalization. Neural Networks, 5(2): 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1
[15] Ganaie, M.A., Hu, M.H., Malik, A.K., Tanveer, M., Suganthan, P.N. (2022). Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence, 115: 105151. https://doi.org/10.1016/j.engappai.2022.105151
[16] Yin, H.L., Zhang, Z.W., Wang, F.D., Zhang, Y.N., Xia, R.L., Jin, J. (2021). Rainfall-runoff modeling using LSTM-based multi-state-vector sequence-to-sequence model. Journal of Hydrology, 598: 126378. https://doi.org/10.1016/j.jhydrol.2021.126378
[17] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321-357. https://doi.org/10.1613/jair.953
[18] Lin, T.Y., Goyal, P., Girshick, R., He, K.M., Dollár, P. (2020). Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2): 318-327. https://doi.org/10.1109/TPAMI.2018.2858826
[19] Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.K. (2020). Calibrating deep neural networks using focal loss. Advances in Neural Information Processing Systems, 33: 15288-15299. https://proceedings.neurips.cc/paper_files/paper/2020/hash/aeb7b30ef1d024a76f21a1d40e30c302-Abstract.html.
[20] Ke, G.L., Meng, Q., Finley, T., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30: 3146-3154. https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.
[21] Chen, T.Q., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, pp. 785-794. https://doi.org/10.1145/2939672.2939785

NOMENCLATURE

AE-BiLSTM	Attention-Enhanced Temporal Bidirectional Long Short-Term Memory
LightGBM	Light Gradient Boosting Machine
XGBoost	Extreme Gradient Boosting
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BiLSTM	Bidirectional Long Short-Term Memory
LSTM	Long Short-Term Memory
SMOTE	Synthetic Minority Oversampling Technique
AFL	Adaptive Focal Loss
MLP	Multi-Layer Perceptron
LR	Logistic Regression
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve