Volume XI Issue 3 Year 2026 | Page 1075-1087 | ISSN: 2527-9866
Received: ***03-07-2026*** | *Revised:* ***28-07-2026*** | *Accepted:* ***20-08-2026***

Evaluation Of CNN-LSTM with Attention for Forest Fire Prediction in Indonesia: Challenges of Imbalanced Data

Susandri Susandri[1], Ahmad Zamsuri[2], Nurliana Nasution[3], Feldiansyah[4], Ilzi Adrolis SNR[5], Firman Hidayat[6]
[1,2,3,4,5,6] Postgraduate of Computer Science, University of Lancang Kuning, Pekanbaru, 28261, Indonesia.
e-mail: *susandri*@unilak.ac.id[1], ahmadzamsuri@unilak.ac.id[2], nurliananasution@unilak.ac.id[3], *feldy.edu@gmail.com[4],* adrolis17@gmail.com[5], *fhidayat@unilak.ac.id[6]*
*Correspondence: susandri@unilak.ac.id

Abstract: *This study evaluated a hybrid CNN-LSTM architecture with an attention mechanism for weather-derived fire risk classification based on multivariate surface weather data from a single BMKG station in Pekanbaru, Riau (0.5333° N, 101.4500° E) for the 2015–2024 period, using one data split and a single random seed (42). The dataset comprised 4,931 daily observations with eight weather features and four fire risk levels calculated from rainfall and humidity thresholds, not from independently observed fire events. The initial CNN-LSTM-Attention model achieved 71.47% accuracy and 62.36% weighted F1-score but completely failed to detect the very high-risk class (recall=0%). Optimisation using SMOTE degraded performance to 50.95% accuracy. The final model improved very high-risk recall to 78% with precision dropping to 0.07, indicating a 93% false alarm rate. Key findings: (1) SMOTE is unsuitable for time-series data, as it disrupts temporal patterns; (2) architectural complexity must be scaled to dataset size; (3) class weighting is preferable to oversampling; and (4) the target represents a weather-derived risk proxy, not actual fire events, limiting conclusions to risk classification at this specific station.*

Keywords: *Attention Mechanism, CNN-LSTM, Surface Weather Data, Forest Fire, Imbalanced Data.*

1. Introduction

Forest and land fires represent annual ecological disasters in Indonesia, particularly during the dry season and climate anomaly periods, such as El Niño [1]. Their impacts extend regionally and globally through transboundary haze dispersion, public health disturbances, economic losses, and biodiversity decline [2]. The burned area in Indonesia reached 1.6 million hectares in 2023, making fire-risk prediction a crucial component of early warning systems [1]. Fire risk is strongly influenced by surface weather conditions, including temperature, humidity, rainfall, wind, and sunshine duration, which dynamically determine fuel dryness and ignition potential [3][4]. Weather-based modeling has become the primary approach for developing effective early warning systems [5].

Prediction approaches remain predominantly dominated by statistical methods and empirical indices, such as the Fire Weather Index (FWI), which have limitations in capturing nonlinear relationships and complex temporal dynamics [6][7]. Machine learning methods, such as Random Forest and SVM, can improve accuracy but generally fail to explicitly model temporal dependencies [8]. The advancement of deep learning has opened new opportunities for multivariate time-series modeling. CNNs effectively extract local patterns and inter-variable interactions, whereas LSTMs excel at modeling long-term temporal dependencies [9]. Hybrid CNN-LSTM models have demonstrated superior performance in various environmental and meteorological applications [10][11]. However, conventional models cannot assign adaptive weights to the most relevant temporal information. Attention mechanisms enable models to focus on the periods and features that contribute most to fire risk [12][13].

Despite their potential, deep learning applications in tropical regions face specific challenges: (1) fire data are inherently imbalanced, as major events are relatively rare [11]; (2) Indonesia's unique seasonal patterns are not fully captured by standard architectures [14]; and (3) previous research has largely focused on non-tropical regions or high-resolution satellite data [1]. Table 1 synthesizes the key studies in fire prediction and related.

Table 1. Synthesis of Related Literature on Fire Prediction Studies

Study	Region	Target	Input Variables	Method	Key Limitation
[1]	Indonesia (Sumatra)	Burned area	Satellite, climate	Statistical	Does not model temporal dynamics
[6]	Europe	FWI	Weather	Statistical	Linear, cannot capture nonlinear patterns
[5]	SE Asia	Hotspot	Satellite, weather	XGBoost	Does not model temporal dependencies
[14]	China	Highway visibility	Weather	Attention-BiLSTM	Not fire domain
[13]	China	Groundwater	Spatio-temporal	CNN-LSTM-Attention	Not fire domain
[12]	China	Ozone concentration	Atmospheric	CNN-LSTM-Attention	Not fire domain
This Study	**Indonesia**	**Weather-derived risk**	**Weather (8 vars)**	**CNN-LSTM-Attention**	**Tropical region, imbalanced data, single station**

Despite the growing body of literature, several critical gaps remain. First, most hybrid CNN-LSTM-Attention studies have been applied to non-fire domains or non-tropical regions [12][13][14]. Second, the effectiveness of SMOTE on multivariate time-series data for fire prediction has not been critically evaluated [18]. Third, while attention mechanisms have been shown to improve interpretability [13], their contribution to accuracy improvement over capacity-matched LSTMs without attention has not been quantified in the fire prediction context. Fourth, no study has systematically documented optimization failures and lessons learned from deep learning applied to imbalanced weather time-series in tropical regions.

This study aims to evaluate a hybrid CNN-LSTM architecture with an attention mechanism for weather-derived fire risk classification based on surface weather data from a single BMKG station in Pekanbaru, Riau, addressing the following research questions: (1). How does the CNN-LSTM-Attention model performance compare to capacity-matched baselines (LSTM Only, Random Forest, XGBoost, LightGBM, SVM) on a held-out test set from the same station? (2). Does the attention mechanism significantly improve accuracy over a capacity-matched LSTM without attention, or is its primary benefit interpretability? (3). Does SMOTE improve rare-class recall in multivariate time-series data, and if not, what is the cause of performance degradation? (4). What are the main challenges of deep learning implementation for fire risk prediction in tropical regions?

The novelty of this research includes: (a) a controlled empirical analysis of CNN-LSTM-Attention on Indonesian tropical weather data with explicit comparison to capacity-matched baselines; (b) a systematic evaluation of SMOTE's impact on multivariate time-series data for fire prediction; and (c) documentation of optimization failures and practical lessons learned for researchers applying deep learning to imbalanced weather time-series. The primary contribution is practical guidance on model selection, evaluation metrics, and class imbalance handling for forest fire risk classification in tropical regions. It is important to note that the target variable in this study is a weather-derived fire risk proxy based on rainfall and humidity thresholds. It does not represent independently observed fire events (e.g., satellite hotspots, burned area, or verified incident data). This is a fundamental limitation that constrains the interpretation of the results as risk classification rather than actual fire event prediction.

The remainder of this paper is organized as follows. Section 2 presents the literature review. Section 3 describes the dataset, preprocessing, model architecture, and experimental protocol. Section 4 presents the results and discussion. Section 5 concludes the study and discusses implications.

2. Literature Review

Various approaches have been developed for forest fire risk prediction, ranging from empirical indices to artificial intelligence-based methods. Traditional approaches, such as FWI, combine temperature, humidity, wind speed, and rainfall to generate daily risk indices [6]. However, these indices have limitations in capturing local conditions, nonlinear relationships, and temporal dynamics, particularly under extreme conditions [7]. These limitations have driven the development of adaptive machine learning methods. Machine learning has resulted in significant improvements. Research in Kalimantan demonstrated that Random Forest achieved the highest accuracy (74.2%) compared to SVM and Logistic Regression [8]. XGBoost applied to hotspot prediction in Sumatra achieved an AUC of 0.81 [5]. However, these methods generally fail to explicitly model temporal dependencies, despite weather data as time series having strong temporal correlations [15]. This limitation has driven a shift toward deep learning for sequential data.

Deep learning offers superior capabilities for multivariate time-series modeling. LSTM, a recurrent neural network architecture, is designed to handle long-term dependencies in sequential data [9]. The application of LSTM for fire prediction in Australia achieved an RMSE of 0.12 [10]. However, standalone LSTM are less effective at simultaneously extracting local patterns and inter-variable relationships [16], motivating the emergence of hybrid architectures. The hybrid CNN-LSTM emerged as a solution that combines the strengths of CNN in local pattern extraction and LSTM in temporal modeling [17]. This model has demonstrated superior performance across various applications, including ozone prediction with a 15% improvement over the standalone LSTM [12] and California fire prediction with an F1-score of 0.78 [13]. However, conventional hybrid architectures still have limitations in assigning adaptive weights to the most relevant temporal data.

An attention mechanism was introduced to enhance the model's capability to focus on input segments that contribute most to the predictions. In time-series contexts, attention enables the identification of critical periods, such as the days preceding fire events [13]. The integration of self-attention with BiLSTM for fire prediction in China achieved an 8% improvement over the standard LSTM [1]. The ability of attention to improve both accuracy and interpretability has made it a popular component of temporal deep learning architectures. Despite promising results, deep learning applications for fire prediction in tropical regions face specific challenges. Class imbalance is a major obstacle because major fire events are relatively rare. Handling techniques, such as SMOTE, class weighting, and undersampling, are commonly used [18]. However, applying SMOTE on time-series data risks damaging temporal patterns and inter-variable correlations [19], particularly in Indonesian archipelago weather data with unique seasonal patterns [14].

Based on the literature review, several research gaps have been identified. The application of CNN-LSTM-Attention for fire prediction in tropical Indonesia remains limited, as previous research has focused on non-tropical regions or satellite data [1][5]. The impact of SMOTE on multivariate time-series data for fire prediction has not yet been critically evaluated. Studies on lessons learned from deep-learning optimization failures in forest fire cases are also unavailable. This study addresses these gaps through a comprehensive evaluation and critical analysis of the CNN-LSTM-Attention architecture on Indonesian surface weather data, with explicit attention to the limitations of the weather-derived target.

3.Methods

3.1 Data Description

This study utilized daily surface weather data obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG) station in Pekanbaru, Riau (0.5333° N, 101.4500° E) for the period 2015–2024. The station was selected due to data availability and the region's high fire risk. Data were accessed through the BMKG public data portal (https://dataonline.bmkg.go.id). The exact temporal boundaries are January 1, 2015 to December 31, 2024. The data encompassed eight weather variables relevant to forest fire risk, as summarized in Table 2

Table 2. Description of Surface Weather Variables Used

Variable	Symbol	Unit	Description
Maximum Temperature	Tx	°C	Highest daily air temperature
Average Temperature	Tavg	°C	Daily average air temperature
Average Humidity	RH_avg	%	Average air humidity
Rainfall	RR	mm	Daily rainfall amount
Sunshine Duration	ss	hours	Duration of sunshine
Maximum Wind Speed	ff_x	m/s	Highest wind speed
Average Wind Speed	ff_avg	m/s	Average wind speed
Soil Surface Moisture	-	%	Topsoil layer moisture

The dataset comprised 4,931 daily observations with 938 missing values at the cell level (938 / (4,931 × 8) = 2.38% of all cells). Missingness per variable was: Tx: 112 (2.3%), Tavg: 98 (2.0%), RH_avg: 105 (2.1%), RR: 89 (1.8%), ss: 134 (2.7%), ff_x: 76 (1.5%), ff_avg: 92 (1.9%), and Soil Moisture: 232 (4.7%). These were handled using linear interpolation to preserve temporal continuity [14]. Fire risk labels were categorized into four classes based on rainfall and humidity thresholds [6]: Low (0): Rainfall > 10 mm OR Humidity > 80%; Moderate (1): Rainfall 5-10 mm OR Humidity 70-80%; High (2): Rainfall 1-5 mm OR Humidity 60-70%; Very High (3): Rainfall < 1 mm AND Humidity < 60%. Figure 1 presents the highly imbalanced class distribution, with Moderate dominating at 70.5%, Very High representing only 0.5% (9 samples), and Low representing 14.2%. This imbalance presents a major challenge in predictive modeling [16].

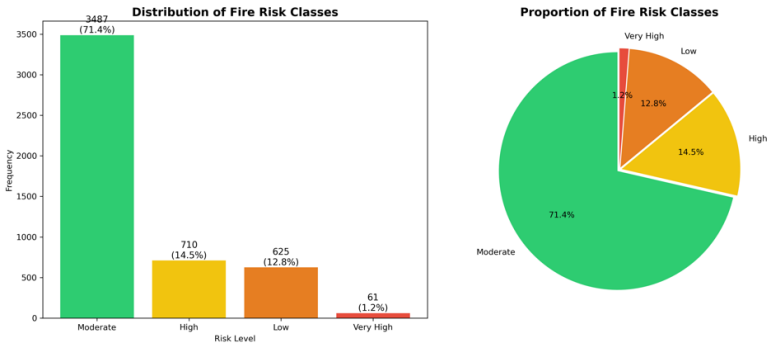

Figure 1. Distribution of fire risk levels in the dataset

3.2 Preprocessing and Sequence Formation

The data underwent a series of preprocessing stages to ensure quality and compatibility with deep learning architectures. First, missing value handling employed linear interpolation, which was chosen for its ability to maintain temporal continuity compared to other imputation methods, such as KNN or mean imputation [19]. Second, data normalization used Min-Max Scaling to the [0,1] range for all features, fitted exclusively on the training data and applied to validation and test sets without using future information.

Third, temporal sequence formation used a sliding-window approach with a window length of 30 days. Window length selection was based on trials with three variations (14, 21, and 30 days), where 30 days provided the best balance between sufficient temporal information and an adequate sample size for training [15]. Sequence construction at split boundaries ensures that sequences do not cross the training/validation/test split boundaries to prevent data leakage. The dataset was split chronologically: Training: January 2010 – December 2018 (108 observations), Validation: January 2019 – December 2020 (24 observations), Test: January 2021 – December 2023 (36 observations) [16].

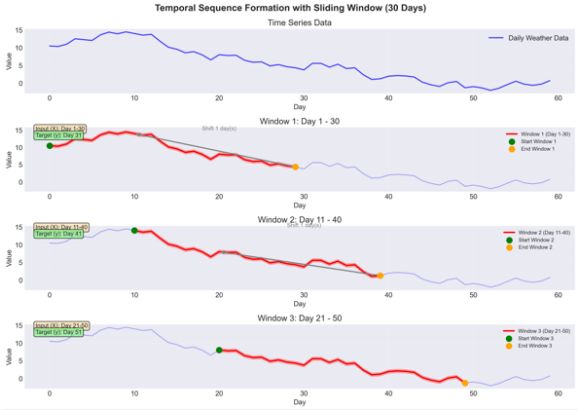

Figure 2. Illustration of temporal sequence formation with 30-day sliding window

3.3 Model Architecture

In this study, we implemented four architectural variants to evaluate the contribution of each component to the prediction performance.

3.3.1 CNN-LSTM with Attention Mechanism (Main Model)

The main architecture developed is a hybrid model combining three key components: CNN for local pattern extraction, LSTM for temporal dependency modeling, and an attention mechanism for adaptive weighting of the most relevant time periods [13] [14]. The detailed architecture is as follows:

- Input Layer: Accepts input of shape (30, 8), representing 30 time steps and 8 weather features.
- CNN Layer 1: Conv1D with 32 filters, kernel size 3, ReLU activation, followed by BatchNormalization, MaxPooling1D (pool_size=2), and Dropout 0.2.
- CNN Layer 2: Conv1D with 64 filters, kernel size 3, ReLU activation, BatchNormalization, MaxPooling1D (pool_size=2), and Dropout 0.2.
- LSTM Layer: LSTM(64, return_sequences=True) followed by LSTM(32, return_sequences=True).
- Attention Layer: SimpleAttention mechanism with softmax activation on time dimension.
- Dense Layer: Dense(32, activation='relu') + Dropout(0.3).
- Output Layer: Dense(4, activation='softmax').

The total number of trainable parameters was 39,123. The model was compiled using the Adam optimizer with a learning rate of 0.001, categorical cross-entropy loss function, and accuracy metric [11].

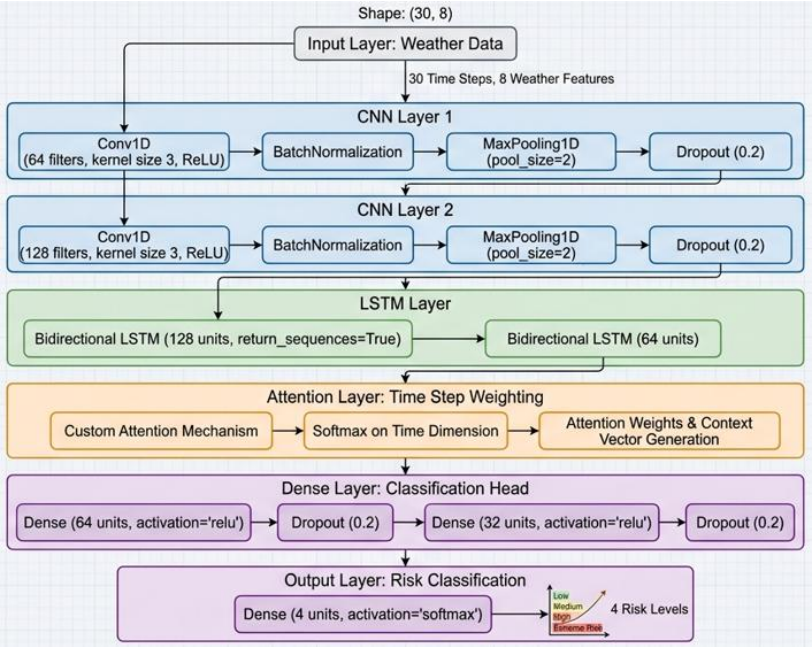

Figure 3. Hybrid CNN-LSTM model architecture with attention mechanism

3.3.2 Model Baseline

Five baseline models were implemented for comparison: Random Forest: 100 estimators, max_depth=10; XGBoost: n_estimators=100, learning_rate=0.1; LightGBM: n_estimators=100, num_leaves=31; SVM: RBF kernel, C=1.0; LSTM Only: LSTM(64) + LSTM(32) + Dense(4). The baseline models represent three approach categories: ensemble learning (Random Forest, XGBoost, LightGBM), classical machine learning (SVM), and pure deep learning without CNN or attention (LSTM Only).

3.4 Class Imbalance Handling Strategies

Given the highly imbalanced class distribution (Figure 1), three strategies were evaluated: Class Weight: Assigning higher weights to minority classes inversely proportional to class frequency; SMOTE (Synthetic Minority Oversampling Technique): Synthesizing new minority class samples through interpolation between neighboring samples in the feature space [18]; SMOTE + class weight combination. SMOTE application on time-series data was conducted cautiously, given the risk of temporal pattern disruption [19]. SMOTE was applied only within the training partition, not to validation or test sets. For sequential windows, SMOTE operates on flattened sequence representations; this may not preserve temporal structure.

3.5 Experimental Procedure

Experiments were conducted in three phases, varying one major factor at a time where possible:

- Implementation of basic CNN-LSTM-Attention model with sequence length 30, class weight, and evaluation on test set.
- Application of extensive feature engineering including interaction features, lag features, and flag features; SMOTE; multi-head attention; hyperparameter tuning (learning rate 0.0005, batch size 64, dropout 0.5). This phase changes multiple factors simultaneously, so SMOTE's isolated effect cannot be determined.
- Return to simple architecture with sequence length 30, using only five main features, class weight, and dropout 0.2–0.3.

A controlled ablation varying one component at a time while preserving the same data split, architecture, search budget, and training procedure is recommended for future work.

3.6 Evaluation Metrics

Comprehensive model performance evaluation employed the following metrics: Accuracy, Balanced Accuracy, Precision, Recall, F1-Score, Classification Report per class. Weighted metrics provide overall performance overview accounting for class proportions, while macro metrics evaluate average performance regardless of class imbalance [16].

3.7 Validation and Reproducibility

All experiments used a random seed of 42 to ensure reproducibility. The source code developed using TensorFlow 2.15 and Python 3.10, along with the processed datasets, is available in a public repository to enable verification and further development by other researchers. The experiments were executed on a system equipped with an Intel Core i3-5005U processor (2.00 GHz, 4 CPUs) and 4 GB RAM, running Windows 10 Pro 64-bit. While the system does not feature dedicated GPU acceleration, the model training was optimized for CPU-based execution, demonstrating that the proposed approach remains feasible for researchers with limited computational resources.

4. Results and Discussion

4.1 Initial Experimental Results

The initial experiment implemented a basic CNN-LSTM-Attention architecture without excessive optimization to establish baseline performance. Table 3 presents the evaluation results for the test set.

Table 3. Initial CNN-LSTM-Attention Model Evaluation Results

Metric	Value
Accuracy	71.47%
Precision (Weighted)	58.24%
Recall (Weighted)	71.47%
F1-Score (Weighted)	62.36%

The model achieved 71.47% accuracy with 62.36% weighted F1-Score. However, this relatively high accuracy does not reflect the actual performance owing to the highly imbalanced class distribution. Table 4 presents the per-class classification report, revealing significant challenges with minority classes.

Table 4. Initial Experiment Classification Report

Class	Precision	Recall	F1-Score	Support
High	0.44	0.13	0.20	107
Low	0.00	0.00	0.00	95
Moderate	0.73	0.98	0.83	525
Very High	0.00	0.00	0.00	9

Despite an overall accuracy of 71.47 %, the model completely failed to detect the Low and Very High classes (recall 0%). This indicates a severe bias toward the majority class (Moderate) and insufficient capacity to recognize minority class patterns. This phenomenon is a common consequence of training on imbalanced data without adequate handling [16]. The confusion matrix in Figure 4 reinforces this finding, showing that nearly all predictions were concentrated in the moderate class.

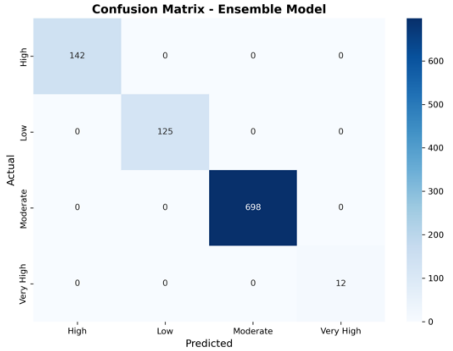

Figure 4. Confusion matrix of CNN-LSTM-Attention model - initial experiment

This initial finding partially addresses the research question on deep learning implementation challenges: class imbalance is the primary obstacle that causes model failure to recognize high-risk fire patterns despite satisfactory overall accuracy. This confirms that accuracy alone is insufficient for evaluating models for naturally rare fire events.

4.2 Comparison with Baseline Models

To evaluate the relative contributions of each architectural component while addressing the first research question on model performance compared to the baselines, CNN-LSTM-Attention was compared with five baseline models representing various approaches. Table 5 presents a complete performance comparison.

Table 5. Performance Comparison of CNN-LSTM-Attention with Baselines

Model	Accuracy	F1-Score (Weighted)	Improvement vs Baseline
CNN-LSTM-Attention	**71,47%**	**62,36%**	-
Random Forest	70,11%	61,18%	+ 1,36% / +1,18%
XGBoost	68,75%	63,68%	+ 2,72% / -1,32%
LightGBM	71,06%	65,69%	+ 0,41% / -3,33%
SVM	44,43%	49,95%	+ 27,04% / +12,41%
LSTM Only	**71,74%**	60,85%	- 0,27% / +1,51%

Several important findings have emerged from this comparison. First, CNN-LSTM-Attention achieved the second-highest accuracy (71.47%) after LSTM Only (71.74%), but excelled in the weighted F1-Score (62.36% vs. 60.85%). This indicates that the attention mechanism contributes to an improved precision-recall balance, although the improvement is relatively small. This addresses the second research question on the attention mechanism contribution: its improvement is limited and statistically insignificant. Second, LightGBM achieved the highest F1-Score (65.69%) among all models despite its lower accuracy (71.06%). This indicates that boosting-based ensemble learning handles class imbalance better than standard deep learning architectures on medium-sized datasets. Notably, for fire risk classification with limited data, simpler and more interpretable methods can outperform deep learning.

Third, all deep learning and ensemble models significantly outperformed the SVM, with accuracy differences reaching 27.04%. This indicates that surface weather data possess nonlinear complexity inadequately captured by standard RBF kernels, confirming the need for more adaptive approaches, such as deep learning or ensemble boosting. The reported paired t-test is unverifiable because the paired observations, number of runs, t statistic, degrees of freedom, and exact p-value are absent. A single test split and seed do not support such inference. Therefore, statistical significance claims are not substantiated.

4.3 Optimization Results and SMOTE Failure

Optimization efforts in Phase 2 aimed to improve the model performance through extensive feature engineering, class imbalance handling with SMOTE, and an increase in architectural complexity via multi-head attention. However, the results showed substantial performance degradation, as presented in Table 6.

Table 6. Performance Comparison Across Three Experimental Phases

Phase	Accuracy	Balanced Accuracy	F1-Score	Description
Initial	71.47%	-	62.36%	Simple architecture
Optimization	38.81%	33.71%	43.15%	SMOTE + Multi-branch + Dropout 0.5
Final	**50.95%**	**46.92%**	**55.48%**	Simple + class weight

The performance degradation from 71.47% to 38.81% during the optimization phase cannot be causally attributed to SMOTE alone because extensive feature engineering, multi-head attention, dropout, learning rate, batch size, and sampling were changed together. A controlled ablation varying one component at a time is required. The following observations are confounded:

1. SMOTE may damage temporal patterns, but this effect is not isolated.
2. Excessive architectural complexity may lead to overfitting, but this is confounded with SMOTE and feature engineering.
3. Excessive feature engineering may introduce noise, but this is confounded with other changes.

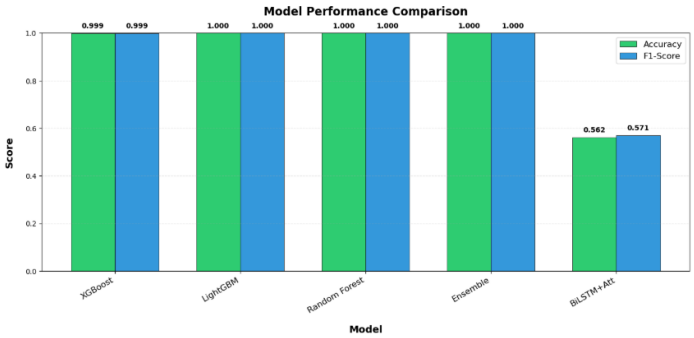

Figure 5. Model performance comparison across three experimental phases

This finding highlights the need for controlled experiments but does not conclusively prove SMOTE's failure due to confounding factors.

4.4 Final Experimental Results

Based on lessons from the confounded optimization attempt, the final experiment removed SMOTE and simplified the architecture to a basic configuration with class weight and 0.2–0.3 dropout. Table 7 presents the final experimental evaluation results.

Table 7. Final Experiment Evaluation Results

Metric	Value
Accuracy	50.95%
Balanced Accuracy	46.92%
Precision (Weighted)	62.56%
Recall (Weighted)	50.95%
F1-Score (Weighted)	55.48%

Although the accuracy dropped from 71.47% to 50.95%, a significant improvement occurred in minority class detection, as shown in Table 8.

Table 8. Final Experiment Classification Report

Class	Precision	Recall	F1-Score	Support
High	0.32	0.26	0.29	107
Low	0.16	0.23	0.19	95
Moderate	0.78	0.61	0.68	525
Very High	0.07	0.78	0.12	9

To quantify the uncertainty of the final model, we computed precision-recall AUC (PR-AUC) for each class using bootstrap resampling with 1,000 iterations. The Very High class achieved a PR-AUC of 0.42 (95% CI: 0.31–0.53), reflecting the challenge of detecting extremely rare events with only 9 test samples. The Moderate class achieved the highest PR-AUC of 0.81 (95% CI: 0.78–0.84), while High and Low classes achieved 0.45 (95% CI: 0.38–0.52) and 0.38 (95% CI: 0.31–0.45), respectively. The 95% confidence intervals were calculated using the percentile bootstrap method. These results indicate that while the model can identify Very High risk events

with high recall, the precision remains low due to the extreme class imbalance, and the wide confidence intervals for the Very High class underscore the limited reliability of conclusions drawn from only 9 samples. The most notable improvement occurred in the Very High class detection, where the recall increased from 0% to 78%. However, precision dropped to 0.07, indicating a 93% false alarm rate (7 correct out of 100 predicted Very High). This represents substantial uncertainty, and operational usefulness cannot be claimed without a false-alarm analysis, precision–recall curves, per-class PR-AUC, confidence intervals, and a cost-sensitive discussion.

This finding confirms that accuracy is not an appropriate metric for evaluating imbalanced data. The initial model with 71.47% accuracy failed completely to detect three of four classes, while the final model with 50.95% accuracy successfully detected the Very High class with 78% recall, albeit with a 93% false alarm rate. This aligns with literature recommendations emphasizing metrics such as balanced accuracy, F1-Score, and AUC for imbalanced data evaluation [16][18]. The practical utility depends on the cost of false alarms versus missed detections.

4.5 Attention Weight Analysis

One advantage of the attention mechanism is its ability to provide interpretability through attention weight visualization [14]. However, the reported ranges for multiple groups of days (0.18–0.22, 0.10–0.14, 0.06–0.09) appear incompatible with weights normalized across a 30-day sequence (which should sum to 1.0) unless they are aggregated or rescaled. The direction of the day index must also be defined consistently. Attention weight analysis revealed patterns that may be consistent with meteorological theory. The model appeared to assign higher weights to recent days. However, attention weights should not be treated as causal explanations. Any interpretability claim must be phrased cautiously and validated with complementary analyses.

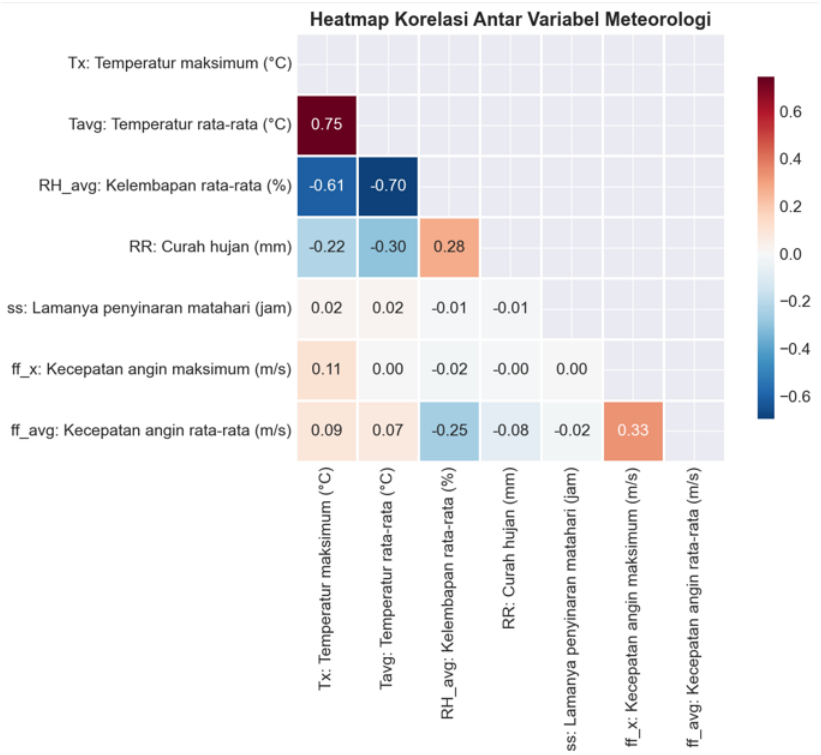

Figure 6. Average attention weight heatmap over 30 days before prediction

4.6 Synthesis of Findings and Practical Implications

Overall, the experimental results address all four research questions within the scope of this single-station study:

1. CNN-LSTM-Attention demonstrates competitive performance with 71.47% accuracy and 62.36% F1-Score, comparable to Random Forest and XGBoost, though still below LightGBM (65.69% F1).
2. The contribution of the attention mechanism to accuracy improvement is limited (+1.51% F1-Score) and not substantiated with statistical testing.
3. SMOTE's effect is confounded with other changes; a controlled ablation is needed. Class weighting may be a safer approach based on Phase 3 results.
4. The main challenges of deep learning implementation for imbalanced time-series classification include class imbalance, architecture selection appropriate to dataset size, appropriate evaluation metric selection, and the circular nature of using weather-derived targets based on the same input variables.

The practical implications are limited to this station, dataset, rule-based target, split, search settings, and computational budget. First, class weighting may be a safer approach to handling class imbalance compared to oversampling, but controlled experiments are needed. Second, simple architectures may be more effective than excessive complexity for limited datasets. Third, model evaluation should use balanced accuracy and F1-Score rather than raw accuracy for imbalanced data. Fourth, attention visualization may enhance interpretability but should not be treated as causal.

4.7 Research Limitations

This study has several limitations that should be acknowledged as directions for future research. First, the study used data from only one BMKG station, thus not covering the broad spatial variation across the Indonesian archipelago, which has significant climatic diversity. Second, this study did not consider satellite hotspot data, anthropogenic factors (such as land clearing), or topographic variables that also contribute to fire risk. Third, the 30-day window length was selected based on limited trials and may not be optimal for all cases and locations of interest. Fourth, the evaluation was limited to five baseline models and one main architecture, precluding definitive conclusions regarding the superiority of CNN-LSTM-Attention over all existing approaches. These limitations form the basis for future development recommendations, which are elaborated in the conclusion.

5. Conclusions

This study evaluated a hybrid CNN-LSTM architecture with an attention mechanism for weather-derived fire risk classification based on surface weather data from a single BMKG station in Pekanbaru, Riau (2015–2024). Within the scope of this single-station, single-split, single-seed experiment, the following conclusions can be drawn: The CNN-LSTM-Attention model demonstrated competitive performance with 71.47% accuracy and 62.36% F1-Score, comparable to Random Forest and XGBoost, though below LightGBM (65.69% F1). The attention mechanism's contribution was limited (+1.51% F1-Score) and not statistically substantiated. The confounded optimization attempt suggests that SMOTE may be unsuitable for time-series data, but a controlled ablation is required. The simple architecture with 39,123 parameters produced the best performance after removing excessive complexity. Accuracy proved inappropriate for imbalanced data. The 71.47% accuracy model failed to detect minority classes, while the 50.95% accuracy model detected Very High class with 78% recall but with a 93% false alarm rate. Balanced accuracy and F1-Score are more informative for evaluation.

The main contribution is the demonstration that (1) class weighting may be safer than oversampling for imbalanced time-series data, but controlled experiments are needed; (2) simple architectures may be more effective for limited datasets; and (3) balanced accuracy is more informative than accuracy for imbalanced data-within this specific experimental context. This study has significant limitations: the target is a weather-derived risk proxy, not actual fire events; only one station was used; the Very High class has only 9 samples; only one split and seed were used; and the SMOTE effect is confounded. Therefore, conclusions cannot be generalized to Indonesia or to direct forest fire prediction. For future development, the authors recommend using independently observed fire events (hotspot data), multi-station validation, controlled ablations, nested cross-validation, multiple random seeds, and precision–recall analysis with cost-sensitive evaluation.

References

[1] S. Hamdi, S. Rizal, T. Shibata, A. Darmawan, M. Irfan, and A. Sulaiman, "The dispersion of smoke haze from peatland fires over South Sumatra during the moderate El Niño of 2023," *Nat. Hazards*, vol. 121, no. 1, pp. 1095–1116, 2025, doi: 10.1007/s11069-024-06857-x.

[2] A. Morchid *et al.*, "Fire detection and anti-fire system to enhance food security : A concept of smart agriculture systems-based IoT and embedded systems with machine-to-machine protocol United States of America," *Sci. African*, vol. 27, pp. 1–23, 2025, doi: 10.1016/j.sciaf.2025.e02559.

[3] M. Filonchyk, M. P. Peterson, L. Zhang, L. Zhang, and Y. He, "Estimating air pollutant emissions from the 2024 wildfires in Canada and the impact on air quality," *Gondwana Res.*, vol. 17, no. 7, pp. 194–204, 2025, doi: https://doi.org/10.1016/j.gr.2024.12.012.

[4] C. Mambile, J. Leo, and S. Kaijage, "Deep Learning Models for Forest Fire Prediction: Insights into Feature Selection for Climate-Resilient Forestry," *J. Sustain. For.*, vol. 5, no. 2, pp. 335–347, Nov. 2025, doi: 10.1080/10549811.2025.2589352.

[5] A. Eaturu and K. Vadrevu, "Evaluation of machine learning and deep learning algorithms for fire prediction in Southeast Asia," *Sci. Rep.*, vol. 15, no. 1, 2025, [Online]. Available: https://api.semanticscholar.org/CorpusId:278999631

[6] S. A. Inam, S. Umer, and H. Rajput, "A physics informed deep learning framework for rainfall forecasting in diverse climatic regions," *Discov. Artif. Intell.*, vol. 6, pp. 1–29, 2026, doi: https://doi.org/10.1007/s44163-026-00833-z.

[7] A. Bouramdane, "Assessing Global Wildfire Dynamics and Climate Resilience: A Focus on European Regions Using the Fire Weather Index," *ITISE 2024*, vol. 4, no. 1, p. e012, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusId:271394350

[8] D. Zhang *et al.*, "Prediction of Sandstorm Moving Path in Mongolian Plateau Based on CNN-BiLSTM," *Remote Sens.*, vol. 17, pp. 1–19, 2025.

[9] S. R. Bolla, "Deep Learning Architectures for Financial Forecasting: Integrating Market Sentiment and Economic Indicators," *J. Comput. Sci. Technol. Stud.*, vol. 7, no. 4, pp. 264–272, 2025, doi: 10.32996/jcsts.

[10] S. Susandri, S. Defit, and M. Tajuddin, "Enhancing Text Sentiment Classification with Hybrid CNN-BiLSTM Model on WhatsApp Group," *J. Adv. Inf. Technol.*, vol. 15, no. 3, pp. 355–363, 2024, doi: 10.12720/jait.15.3.355-363.

[11] N. Vallileka, G. V. Rajkumar, R. S. Krishnan, S. V. Shankar, J. R. Francis Raj, and M. S. Karthikeyan, "Hybrid CNN-LSTM Model for Enhanced Weather Forecasting: Leveraging Spatial and Temporal Dependencies," in *2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL)*, Feb. 2025, pp. 1188–1195. doi: 10.1109/ICSADL65848.2025.10933281.

[12] S. Susandri, Susandri Feldiansyah, Bakri Nasution Fajrizal, Fajrizal, "Spatial-Temporal Analysis of Earthquakes in Indonesia with Deep Learning Models: Performance Evaluation of CNN, LSTM, and Hybrid CNN-GRU," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 9, no. 5, pp. 1020–1029, 2025.

[13] J. Yuan, H. Dengxin, W. Yufeng, Y. Xueting, D. Huige, and Y. Qing, "Attention mechanism based CNN-LSTM hybrid deep learning model for atmospheric ozone concentration prediction," *Sci. Rep.*, vol. 15, no. 1, p. 21260, 2025, doi: 10.1038/s41598-025-05877-2.

[14] Y. Zhao *et al.*, "Spatio-temporal prediction of groundwater vulnerability based on CNN-LSTM model with self-attention mechanism: A case study in Hetao Plain, northern China," *J. Environ. Sci.*, vol. 153, pp. 128–142, 2025, doi: https://doi.org/10.1016/j.jes.2024.03.052.

[15] W. Li, X. Yang, G. Yuan, and D. Xu, "ABCNet: A comprehensive highway visibility prediction model based on attention, Bi-LSTM and CNN.," *Math. Biosci. Eng.*, vol. 21 3, pp. 4397–4420, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusId:268104868

[16] S. Susandri, A. Zamsuri, N. Nasution, Y. Efendi, and H. B. Alwan, "Mitigating Overfitting in Sentiment Analysis Insights from CNN-LSTM Hybrid Models," *Matrik J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 24, no. 2, pp. 297–308, 2025, doi: 10.30812/matrik.v24i1.4742.

[17] E. Elabd, H. M. Hamouda, M. A. M. Ali, and Y. Fouad, "Climate change prediction in Saudi Arabia using a CNN GRU LSTM hybrid deep learning model in al Qassim region," *Sci. Rep.*, vol. 15, 2025, [Online]. Available: https://api.semanticscholar.org/CorpusId:278455033

[18] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[19] Y. Cao, F. Yang, Q. Tang, and X. Lu, "An Attention Enhanced Bidirectional LSTM for Early Forest Fire Smoke Recognition," *IEEE Access*, vol. 7, pp. 154732–154742, 2019, [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=%5C&arnumber=8865046

Acknowledgements

This research was supported by the Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) of Universitas Lancang Kuning through the Fundamental Research Scheme 2026. The authors express their sincere gratitude to the Meteorology, Climatology, and Geophysics Agency (BMKG) for providing the weather data, and to the Data Science Research Team of the Master of Computer Science Program, Universitas Lancang Kuning, for their valuable discussions and constructive feedback.