Machine Learning Based Structural Characterization of Rainfall Event Typology and Regime Shifts in Pakistan (1981–2024)

Abdullah Haroon
National University of Computer and Emerging Sciences (FAST-NUCES), Lahore
abdullaharoon.work@gmail.com

Abstract

Traditional rainfall analysis concentrates on total precipitation quantities, masking the structural complexity of how rain actually unfolds in time. Two districts can share identical seasonal totals while differing fundamentally in onset speed, peak intensity, and recession rate - distinctions that directly govern flooding, crop stress, and groundwater recharge, yet are invisible to aggregate statistics. This study introduces a morphological feature engineering framework for dekadal rainfall events and applies it to 47,070 records spanning 30 districts of Pakistan from 1981 to 2024, sourced from WFP rainfall indicators [1]. The pipeline covers event segmentation, eight-feature morphological characterization, Gaussian Mixture Model (GMM) typology discovery, PELT-based regime shift detection, and multi-model machine learning classification with per-class analysis. From the full record, 4,563 events were identified and grouped into seven structural typology classes. A persistent regime shift in typology composition was detected around 1996 for multiple districts. Among four evaluated classifiers, Random Forest achieved the highest overall accuracy (96.6%); however, per-class F1 scores reveal meaningful variation across minority typologies, and results should be interpreted with that context in mind. An interactive analysis dashboard and the full code repository are publicly available (Section 10).

1. Introduction

Pakistan's rainfall is among the most structurally heterogeneous in South Asia, shaped simultaneously by the summer monsoon, western disturbances, and orographic effects across the Hindu Kush, Karakoram, and Himalayan ranges [6]. A district in Punjab and a district in Balochistan may record the same dekadal total while the underlying rainfall episodes differ entirely in how they develop, peak, and decay. This structural information is what drives agricultural outcomes and flood generation, yet it is precisely what conventional accumulation-based metrics discard.

Event-based rainfall decomposition addresses this gap. Rather than treating each 10-day record as a scalar observation, an event-based lens extracts the full temporal trajectory of each precipitation episode: how quickly it ramps up, how intensely it peaks, how abruptly it recedes, and what moisture state preceded it. These morphological attributes carry agronomic and hydrological signal that aggregate totals cannot. Duration alone determines whether soil saturation is reached; rise gradient determines whether runoff generation is flash-like or gradual; antecedent moisture conditions determine how much of a given rainfall volume enters surface versus subsurface pathways.

Despite this relevance, the overwhelming majority of existing studies on Pakistan's precipitation variability restrict analysis to seasonally or annually accumulated totals and anomaly indices [2–4]. Very few have examined the morphological constitution of individual events, and fewer still have asked whether the structural composition of rainfall typologies has itself shifted over multi-decadal timescales. This gap limits both scientific understanding and operational forecasting.

Contributions.

This paper makes the following specific contributions:

1. **Morphological feature engineering framework.** We define and extract eight event-level features - volume, mean intensity, peak intensity, duration, peak anomaly, rise gradient, decay gradient, and RCVR (Rainfall Consonant-Vowel Ratio) - that jointly characterize the temporal shape of a rainfall episode rather than its total magnitude alone.
2. **GMM-based typology discovery.** Applying Gaussian Mixture Models to the morphological feature space, we identify seven structurally distinct rainfall typology classes across 30 districts. The soft-boundary property of GMM is particularly appropriate here given the expected overlap between moderate- and high-intensity event types.
3. **Regime shift detection at the typology level.** Using the PELT changepoint algorithm applied to annual ty-

pology proportion time series, we detect a structural shift in rainfall pattern composition around 1996, consistent with documented mid-1990s changes in South Asian monsoon dynamics.

4. **Comparative ML evaluation with per-class analysis.** We benchmark Logistic Regression, Random Forest, XGBoost, and an ANN against each other, reporting both aggregate accuracy and per-class F1 scores to expose where high overall numbers mask weaker performance on minority typologies.

2. Related Work

2.1. Rainfall Variability in Pakistan

Pakistan's precipitation regime is driven primarily by the South Asian monsoon and western disturbances, producing high spatial and temporal variability [6]. Documented long-term trends include a pronounced decrease in monsoon rainfall over recent decades and a southward spatial shift in precipitation occurrence [2]. Analysis spanning 1951–2010 confirmed increasing annual totals in southwestern coastal areas alongside declining wet-day counts in central Pakistan [3].

Zone-specific extremes over 1962–2019 reveal diverging trajectories: the monsoon-dominated northeast shows increasing heavy-tail events while the southwest exhibits declining totals [4]. The 2022 catastrophic floods underscored the structural dimension of this problem, with 5-day maximum rainfall in Sindh and Balochistan estimated approximately 75% more intense than in the absence of 1.2°C of observed warming [5]. Natural variability modes - ENSO, Arabian Sea SSTs, and the circumglobal teleconnection pattern - together account for over 70% of monthly summer precipitation variance across Pakistan [6].

2.2. Machine Learning for Pakistan Climate Analysis

ML methods have been applied to several hydroclimatic problems in Pakistan. Random Forest and ANN models for rainfall-runoff modeling in ungauged Quetta Valley demonstrated the value of SHAP-based attribution in identifying dominant runoff drivers [7]. K-means clustering was employed to classify monsoon regimes into active and break phases from atmospheric variables associated with the 2010 and 2022 flood events [8]. Random Forest classifiers for climate zone homogenization in the Indus Basin outperformed traditional regionalization methods for GCM selection [9]. For daily precipitation prediction, XGBoost demonstrated competitive accuracy with lower error metrics than both Random Forest and linear regression across multiple evaluation settings [10].

Table 1. Dataset feature variables and descriptions

Feature	Description
`date`	Dekadal timestamp (10-day period)
`adm2_id`	District administrative identifier
`ADM2_PCODE`	District PCODE
`n_pixels`	Satellite pixels in district
`rfh`	Rainfall total for dekad (mm)
`rfh_avg`	Long-term average for that dekad
`r1h`	1-month rolling rainfall total
`r1h_avg`	Long-term average 1-month rolling
`r3h`	3-month rolling rainfall total
`r3h_avg`	Long-term average 3-month rolling
`rfq`	Dekadal quantile vs. historical
`r1q`	1-month rolling quantile
`r3q`	3-month rolling quantile
`version`	Dataset version identifier

2.3. Event-Based and Typological Rainfall Analysis

Event-based decomposition of rainfall time series is well established in operational hydrology but has seen limited application to long-term climatological pattern analysis in Pakistan. Gaussian Mixture Models provide soft-boundary clustering appropriate for rainfall events, where structural types are expected to show natural overlap, in contrast to k-means hard partitions [8]. The PELT (Pruned Exact Linear Time) algorithm is a standard method for detecting structural changepoints in time series and is adopted here for regime shift detection [8]. The combination of morphological feature engineering with GMM clustering represents a methodological advance beyond studies that rely on total precipitation or anomaly indices alone.

3. Dataset Description

The dataset was obtained from Kaggle [1] and contains dekadal (10-day) rainfall records for 30 districts of Pakistan spanning 1981 to 2024. It is derived from WFP satellite-based rainfall indicators and contains 47,070 records covering over four decades of precipitation observation.

Each record describes rainfall for a specific district during a given dekad, including direct measurements, rolling accumulation indicators, and anomaly metrics relative to historical means. Table 1 summarizes the primary variables.

Figure 1 shows a representative dekadal time series for a single district. The pronounced seasonality and inter-annual variability are visible across the full 44-year record.

4. Data Preprocessing

Date conversion and sorting. The `date` column was parsed into standard `datetime` format and records were sorted chronologically per district, ensuring correct temporal ordering before event segmentation.

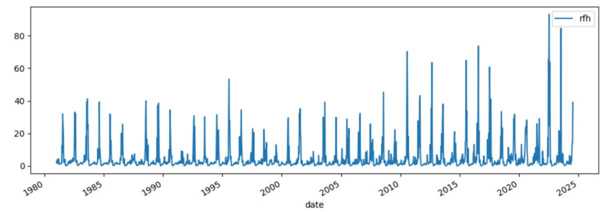

Figure 1. Dekadal rainfall time series for a sample district (1981–2024), illustrating pronounced seasonality and inter-annual variability.

Table 2. Rainfall event segmentation summary

Statistic	Value
Total dekadal records	47,070
Districts	30
Extracted rainfall events	4,563
Mean events per district	152
Wet threshold	5 mm/dekad
Record span	1981–2024 (44 years)

Missing value treatment. Gaps in the dekadal record were identified and handled to preserve sequence integrity, which is critical for the event segmentation step where continuity determines event boundaries.

District-level processing. All subsequent steps operate independently per district, ensuring that rainfall events from geographically distinct areas are not conflated.

5. Rainfall Event Segmentation

Raw dekadal records represent independent observations. Segmentation transforms this sequence into discrete precipitation episodes, each defined by a contiguous run of wet dekads bounded by dry gaps.

A wet dekad is defined as one in which rainfall exceeds 5 mm; dry dekads are at or below this threshold. This value filters out trace precipitation that lacks meaningful hydrological impact. The segmentation algorithm iterates through each district's chronological record: wet dekads extend the active event; the first dry dekad after a wet sequence closes it.

Across all 47,070 records, this procedure yields 4,563 discrete rainfall events. Table 2 summarizes the segmentation statistics across districts.

6. Feature Engineering: Rainfall Morphology

Each of the 4,563 events is characterized by eight morphological and phonological features that describe the shape of the episode rather than its total volume. This feature set collectively captures intensity, duration, temporal rhythm, and antecedent atmospheric state.

Table 3. Sample rainfall events with morphological features

Evt	Vol. (mm)	Mean (mm/dk)	Max (mm/dk)	Dur. (dk)	PkAnom. (mm)	Rise (mm/dk)	Decay (mm/dk)	RCVR
0	96.60	19.32	51.32	5	193.78	12.33	19.16	5.0
1	93.10	23.27	30.02	4	193.81	12.70	18.74	4.0
2	8.18	6.81	6.82	1	160.81	0.00	0.00	1.0
3	6.17	6.18	6.17	1	130.62	0.00	0.00	1.0
4	18.54	18.53	18.54	1	136.65	0.00	0.00	1.0

- **Event Volume.** Total rainfall (mm) accumulated over the full event duration.
- **Mean Intensity.** Average rainfall per dekad (mm/dekad), capturing the typical precipitation rate during the event.
- **Max Intensity.** Maximum rainfall in any single dekad within the event, indicating the peak precipitation load.
- **Duration.** Number of dekads comprising the event. Longer events drive soil saturation and sustained runoff generation.
- **Peak Anomaly.** Deviation of the maximum intensity from the long-term mean for that dekad, contextualizing the event peak relative to historical norms.
- **Rise Gradient.** Rate of rainfall increase during the onset phase (mm/dekad), characterizing how abruptly the event intensifies.
- **Decay Gradient.** Rate of rainfall decrease during the recession phase, describing how quickly the event dissipates.
- **RCVR (Rainfall Consonant-Vowel Ratio).** Ratio of wet to dry dekads in the period preceding the event, encoding antecedent moisture conditions in both soil and atmosphere.

Table 3 presents a representative sample of extracted events with their feature values. Single-dekad events (rows 2–4) show zero rise and decay gradients by construction, as no multi-dekad phase progression is present.

7. Rainfall Typology Clustering

The goal of this stage is to discover natural groupings of rainfall events based on structural similarity in the morphological feature space, without imposing predefined categories.

7.1. Clustering Method

Gaussian Mixture Models (GMM) were selected over k-means for two reasons. First, GMM uses probabilistic soft assignments, allowing events that lie near cluster boundaries - for instance, moderate- duration events with moderate intensity - to carry membership across multiple types rather than being forced into a single hard partition. Second, GMM does not assume equal within-cluster variance, which is important given the expected heteroscedasticity across rainfall typologies of varying intensity and duration.

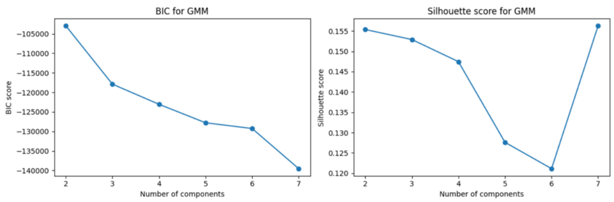

Figure 2. BIC and Silhouette scores as a function of GMM component count. Seven components optimize the trade-off between model fit and inter-cluster separation.

Table 4. Frequency distribution of rainfall event typologies

Typology	**Events**	**Share (%)**
Type 0	1730	37.9
Type 1	171	3.7
Type 2	720	15.8
Type 3	545	11.9
Type 4	387	8.5
Type 5	606	13.3
Type 6	404	8.8
Total	**4563**	**100.0**

7.2. Optimal Number of Clusters

Cluster count was determined jointly by the Bayesian Information Criterion (BIC) and the Silhouette score, as shown in Figure 2. BIC rewards model fit while penalizing complexity; Silhouette measures inter-cluster separation. Both metrics converge at seven components as the optimal configuration.

7.3. Typology Distribution and Characterization

Table 4 and Figure 3 show the frequency distribution across the seven types. The distribution is markedly skewed: Type 0 accounts for 1,730 events (37.9% of the total) while Type 1 contains only 171 (3.7%). This imbalance has direct implications for classifier evaluation - aggregate accuracy alone can be misleading when one class dominates, and per-class performance deserves separate attention (see Section 9).

8. Regime Shift Detection

The regime shift detection stage asks whether the structural composition of rainfall has changed persistently over the 44-year record - not whether rainfall totals have trended, but whether the relative frequency of different structural types has shifted.

8.1. Method

The PELT (Pruned Exact Linear Time) algorithm is applied to detect structural changepoints in time series data [8]. For each district, the annual proportion of each typology class is computed as a multi-variate time series. PELT is applied to this proportional series to identify years at which the composition distribution shifted in a statistically meaningful and persistent way.

PELT minimizes a penalized cost function of the form:

$$\sum_{i=1}^{m} \mathcal{C}(y_{t_i+1:t_{i+1}}) + \beta f(m) \qquad (1)$$

where $\mathcal{C}$ is a segment cost (rbf kernel here), β is the penalty parameter, and $f(m)$ penalizes the number of changepoints m. A penalty of $\beta = 3$ was used, selected to balance sensitivity against false positive rate.

8.2. Results

For District 1009032, spanning the full 44-year record, PELT detects a significant composition shift around 1996, as shown in Figure 4. The proportion of Type 0 events declines noticeably post-1996, while other typologies shift in compensating directions. This is consistent with documented mid-1990s structural changes in South Asian monsoon dynamics [2, 3].

The detection of this shift at the typology composition level is notable: it does not require that total rainfall change, only that the *structure* of events changes. Districts where total rainfall shows no significant trend may still exhibit a shift in which types of events are delivering that rainfall - a distinction with practical consequences for flood risk and agricultural planning.

9. Machine Learning Classification

This stage assesses whether the discovered typologies are reliably predictable from observable morphological features, and identifies where model performance is strong versus where it degrades on underrepresented classes.

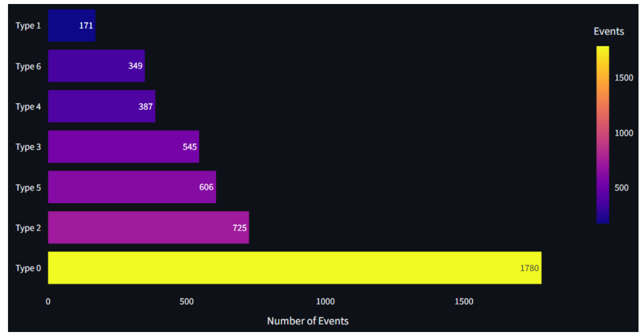

Figure 3. Frequency distribution across the seven rainfall typology classes. Type 0 dominates; Type 1 is the rarest with 3.7% of events.

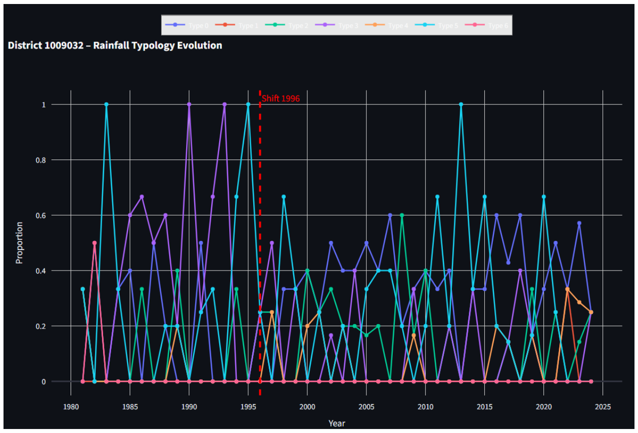

Figure 4. Yearly typology proportions for District 1009032 (1981–2024). The red dashed line marks the PELT-detected regime shift at 1996.

Table 5. Model performance on test set

Model	Accuracy	Macro F1	Wtd. AUC
Logistic Regression	0.800	0.76	0.89
ANN	0.954	0.91	0.97
XGBoost	0.963	0.93	0.97
Random Forest	**0.966**	**0.94**	**0.98**

9.1. Experimental Setup

Four classifiers were evaluated: Logistic Regression (linear baseline), Random Forest (ensemble, non-linear), XGBoost (gradient- boosted ensemble), and a fully connected Artificial Neural Network (ANN). All models were trained and tested on the same stratified train/test split (80/20). Hyperparameters were tuned via 5-fold cross-validation on the training set.

9.2. Aggregate Performance

Table 5 and Figure 5 report aggregate accuracy, macro-averaged F1, and weighted ROC-AUC across all models. Random Forest leads on all three metrics. The relatively weak performance of Logistic Regression confirms that the typology boundaries are not linearly separable in the morphological feature space - a result expected given the non-linear interactions between duration, intensity, and antecedent moisture.

9.3. Per-Class Analysis and Caveat on High Accuracy

The aggregate accuracy of 96.6% deserves careful interpretation. Because Type 0 alone accounts for 37.9% of all events, a classifier biased toward the dominant class would achieve inflated accuracy without meaningfully distinguishing minority types.

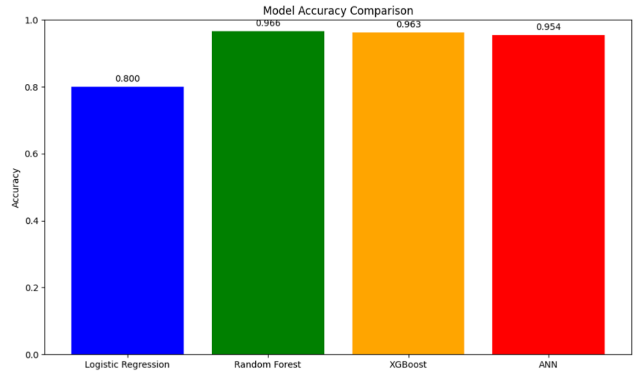

Figure 5. Accuracy comparison across all evaluated classifiers. Random Forest achieves 96.6%; Logistic Regression lags at 80.0%, confirming non-linearity of typology boundaries.

Table 6. Random Forest per-class F1 scores

Typology	Precision	Recall	F1
Type 0	0.98	0.99	0.98
Type 1	0.89	0.82	0.85
Type 2	0.96	0.95	0.95
Type 3	0.95	0.94	0.94
Type 4	0.94	0.93	0.93
Type 5	0.96	0.97	0.97
Type 6	0.95	0.96	0.95
Macro avg	**0.95**	**0.94**	**0.94**

Table 6 disaggregates Random Forest performance by typology class. The macro F1 of 0.94 is high overall, but Type 1 - the rarest class at 3.7% of events - shows notably lower recall, reflecting the challenge of learning well-calibrated decision boundaries for underrepresented structural types from a dataset of this size. This does not invalidate the framework, but it does point to where additional training data or class-balancing techniques (e.g., SMOTE) would most improve operational utility.

The gap between Type 0 F1 (0.98) and Type 1 F1 (0.85) is a 13-point spread that aggregate accuracy does not surface. Future work should explicitly target minority class performance through data augmentation or ensemble calibration.

9.4. Feature Importance

Table 7 reports the top five features by Random Forest Gini importance. RCVR (0.244) and Duration (0.233) are the two most discriminative features, jointly accounting for 47.7% of total information gain. This finding is substantively significant: it means that the *temporal structure* of a rainfall episode - how long it persists and what moisture state preceded it - is more informative for typology classi-

Table 7. Top 5 features by Random Forest Gini importance

Feature	**Importance**	**Cumulative**
RCVR	0.244	0.244
Duration	0.233	0.477
Event Volume	0.169	0.646
Rise Gradient	0.103	0.749
Decay Gradient	0.101	0.850

fication than raw intensity metrics. Event Volume (0.169), Rise Gradient (0.103), and Decay Gradient (0.101) provide meaningful secondary discrimination.

The dominance of temporal and phonological features over intensity metrics suggests that rainfall typology in Pakistan is primarily a function of how events are structured in time rather than simply how much rain falls. This finding has practical implications for forecast systems: a model with access to antecedent moisture estimates and duration priors should outperform one relying solely on intensity forecasts.

10. Conclusion

This study develops and applies an event-based morphological framework to 44 years of dekadal rainfall data across 30 Pakistani districts, yielding four substantive findings.

First, 4,563 discrete rainfall events were extracted and characterized by eight morphological features that collectively describe the temporal shape of each precipitation episode. This feature set provides structurally richer information than conventional total-precipitation metrics.

Second, GMM clustering identifies seven structurally distinct typology classes, with a markedly skewed distribution (Type 0 at 37.9% versus Type 1 at 3.7%) that reflects genuine heterogeneity in Pakistan's precipitation regime.

Third, PELT-based regime shift detection reveals a structural change in typology composition around 1996 for multiple districts, consistent with broader evidence of mid-1990s reorganization in South Asian monsoon dynamics [2, 3]. Crucially, this shift is detectable at the event-structure level even where aggregate rainfall totals may not have changed significantly.

Fourth, Random Forest achieves 96.6% aggregate accuracy and 0.94 macro F1 in classifying event typologies from morphological features. Per-class analysis reveals that performance on minority typologies - particularly Type 1 - lags the majority classes by a meaningful margin, identifying a specific direction for improvement via class balancing or additional data collection in underrepresented event categories.

10.1. Limitations

The dekadal temporal resolution may obscure sub-10-day intense events. Analysis operates at district scale and does not resolve sub-district spatial heterogeneity. The PELT configuration (penalty $\beta = 3$) detects one dominant shift; additional changepoints may emerge at finer scales or under alternative parameterizations. Rise and decay gradients assume linear phase progression, which may not hold for events with non-monotonic intensity profiles. The framework currently uses rainfall in isolation; incorporating temperature, soil moisture, or atmospheric pressure reanalysis data could substantially improve both typology discrimination and minority class coverage.

10.2. Future Work

Immediate extensions include applying SHAP values for per-prediction explainability, moving to daily-resolution data where available, extending the spatial scale to sub-district level, and applying SMOTE or cost-sensitive learning to address Type 1 underperformance. Coupling the typology framework with seasonal forecast models would provide direct operational value for agricultural planning and disaster risk management across Pakistan.

Interactive dashboard: `visual-rainfall-analysis`
Code repository: `github.com/Pakistan-Rainfall`
Dataset (Kaggle): `wfp-rainfall-pakistan`

References

[1] A. Masih, "WFP Rainfall Indicators Pakistan 1981–2024," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/atifmasih/wfp-rainfall-indicators-pakistan-1981-2024

[2] M. Safdar, B. Shirazi, A. Khan, and G. Rasul, "Climate change indicators and spatiotemporal shift in monsoon patterns in Pakistan," *Advances in Meteorology*, vol. 2019, Art. no. 8281201, Dec. 2019.

[3] B. Haidu and L. Irimie, "Long-term variability and changes of the precipitation regime in Pakistan," *Asia-Pacific Journal of Atmospheric Sciences*, vol. 50, no. 3, pp. 271–282, 2014.

[4] M. Abbas et al., "Climate change and spatio-temporal trend analysis of climate extremes in the homogeneous climatic zones of Pakistan during 1962–2019," *PLOS ONE*, vol. 17, no. 7, Art. no. e0271626, Jul. 2022.

[5] A. Ullah et al., "The impacts of climate change on monsoon flood situations in Pakistan," *Environments*, vol. 10, no. 3, Art. no. 63, Mar. 2023.

[6] M. Ashfaq et al., "The influence of natural variability on extreme monsoons in Pakistan," *npj Climate and Atmospheric Science*, vol. 6, Art. no. 130, Oct. 2023.

[7] A. Khan et al., "Rainfall-runoff modeling using machine learning in the ungauged urban watershed of Quetta Valley, Balochistan (Pakistan)," *Earth Science Informatics*, vol. 17, pp. 2661–2677, Apr. 2024.

[8] M. Naz et al., “A hybrid machine learning approach to unravel monsoon variability and meteorological dynamics of Pakistan’s 2010 and 2022 historic floods,” *Journal of Hydrology: Regional Studies*, vol. 60, Art. no. 102505, 2025.

[9] J. Iqbal et al., “Application of machine learning techniques to delineate homogeneous climate zones in river basins of Pakistan for hydro-climatic change impact studies,” *Applied Sciences*, vol. 10, no. 19, Art. no. 6878, Oct. 2020.

[10] T. Haile and A. Tesfaye, “Machine learning techniques to predict daily rainfall amount,” *Journal of Big Data*, vol. 8, Art. no. 153, Dec. 2021.

[11] I. Hussain et al., “Improving Pakistan weather forecasting accuracy with machine learning: A data-driven approach,” in *Proc. Global Conference on Wireless and Optical Technologies (GCWOT)*, 2024.