IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151163

# Tamil Nadu District-Level Next-Day Rainfall Prediction Using Explainable Machine Learning and a Two-Stage Hurdle Architecture

**A Soujanya Lakshmi[1], Ruthiksha P[2], Yuvaraj P[3]**

Dept. of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani[1-3]

**Abstract:** This paper presents an end-to-end machine learning framework for forecasting next-day rainfall depth, in millimetres, and rain occurrence across all 38 administrative districts of Tamil Nadu, India. Because precipitation data are strongly zero-inflated (roughly 80% dry days) and right-skewed, a Two-Stage Hurdle Architecture is proposed: a CatBoost classifier gate estimates whether rain will occur (≥ 0.1 mm), and a CatBoost regressor estimates the rainfall depth. Using the TNWeather24 dataset (2019–2024), raw records are aggregated to district-daily values, and 27 leakage-safe features built only from observations up to the 23:59 IST cutoff are used to train the models. Evaluation is strictly chronological (training 2019–2022, validation 2023, unseen test 2024 with 13,870 records) and compares the hurdle model with Random Forest, XGBoost, a single-stage CatBoost regressor, and a historical-climatology baseline. On the 2024 test set the hurdle model reaches a rain-occurrence accuracy of 82.14%, a ROC-AUC of 90.56%, a rain-day recall of 81.80%, and a depth MAE of 0.1393 mm, a 34.2% reduction relative to the baseline. SHAP TreeExplainer provides global and local explanations, and a six-page Streamlit dashboard and 14 passing automated tests support use and verification of the framework.

**Keywords:** rainfall prediction, Tamil Nadu, two-stage hurdle model, explainable machine learning, CatBoost, SHAP, leakage-safe feature engineering, district-level forecasting, TNWeather24

## I. INTRODUCTION

Rainfall depth information is critical for flood warning, water resource management, and agricultural planning, and it is most useful at an administrative granularity that operational workflows often miss. Existing rainfall prediction work is often limited in one or more ways: it formulates the task only as binary rain/no-rain classification, aggregates data only at a regional or state-wide level, risks data leakage through random train/test splits, relies on opaque black-box models, or offers no interactive decision-support interface. In addition, precipitation data exhibit intense zero-inflation (roughly 80% dry days) and positive right-skewness. In single-stage regression, models struggle because most days have 0 mm of rainfall, which pulls all predictions toward small averages.

This paper develops an end-to-end machine learning framework for continuous next-day (t+1) precipitation depth forecasting, in millimetres, and binary rain-occurrence classification across all 38 districts of Tamil Nadu, India. To address zero-inflation, a Two-Stage Hurdle Architecture separates occurrence detection from depth estimation. The framework is built on the TNWeather24 dataset [1], which covers 2019–2024, and encompasses data ingestion and cleaning, leakage-safe feature engineering, a multi-model benchmark, out-of-time validation and testing, SHAP-based explainability, and an interactive six-page Streamlit dashboard. All predictive features use only observations recorded on or before the cutoff time t (23:59 IST). The limitations listed above are presented as a proposed contribution, not as a claim that no prior literature offers these features in combination. This paper makes the following research contributions:

- A Two-Stage Hurdle Architecture that separates rain-occurrence detection (CatBoost classifier: 82.14% accuracy, 90.56% ROC-AUC, 81.80% recall) from depth estimation (CatBoost regressor), reducing the test MAE to 0.1393 mm, a 34.2% reduction relative to the historical-climatology baseline.
- District-level next-day forecasting for all 38 Tamil Nadu districts, evaluated with a strict chronological split (2019–2022 training, 2023 validation, 2024 hold-out test).
- A leakage-safe set of 27 features that use only observations up to the cutoff time t, with lag and rolling features computed on calendar-day indices.
- A benchmark of the hurdle model against Random Forest, XGBoost, a single-stage CatBoost regressor, and a historical-climatology baseline on the unseen 2024 test set.
- SHAP TreeExplainer-based global (beeswarm) and local (waterfall) explanations.
- An interactive six-page Streamlit dashboard and an automated verification suite of 14 unit and integration tests.

 This work is licensed under a Creative Commons Attribution 4.0 International License 

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151163

## II. RELATED WORK

A. Rainfall Prediction Formulation and Zero-Inflated Targets

Many rainfall prediction studies are formulated only as binary rain/no-rain classification, which discards the depth information needed for flood warning and water resource decisions. Others aggregate at a regional or state-wide scale rather than for individual districts, or evaluate with random train/test splits that leak information across time. Opaque black-box models and the lack of an interactive decision-support interface further limit their use by forecasters. Precipitation data are also zero-inflated and positively right-skewed, so a single regression model trained on all days is pulled toward small averages by the many 0 mm days. These observations motivate the district-level, chronologically evaluated, two-stage formulation adopted here. It is presented as a proposed contribution; it is not claimed that no prior study offers these features together.

B. Gradient Boosting and Tree Ensembles

Three tree-based ensemble regressors are benchmarked in this work: Random Forest, XGBoost [2], a scalable tree boosting system, and CatBoost [3], which introduced unbiased boosting with categorical features. CatBoost is also used for both stages of the proposed hurdle model. All models are compared with a naive climatological baseline that represents what a forecast without machine learning would achieve.

C. Explainability and IMD Risk Categories

SHAP [4] decomposes each prediction into individual feature contributions using a formulation rooted in cooperative game theory, and its TreeExplainer variant is applied here to the CatBoost model. SHAP attributions describe statistical feature contributions relative to the model's expected output; they do not establish physical causal relationships between meteorological variables and precipitation processes. In the dashboard, predicted depths are mapped to India Meteorological Department (IMD) risk tiers [5].

## III. SYSTEM ARCHITECTURE

The framework is organized as a set of connected modules that take raw meteorological CSV files to district-level rainfall forecasts, explanations, and a decision-support interface (Fig. 1). Models, hyperparameters, paths, and thresholds are configured centrally in a config/settings.yaml file.

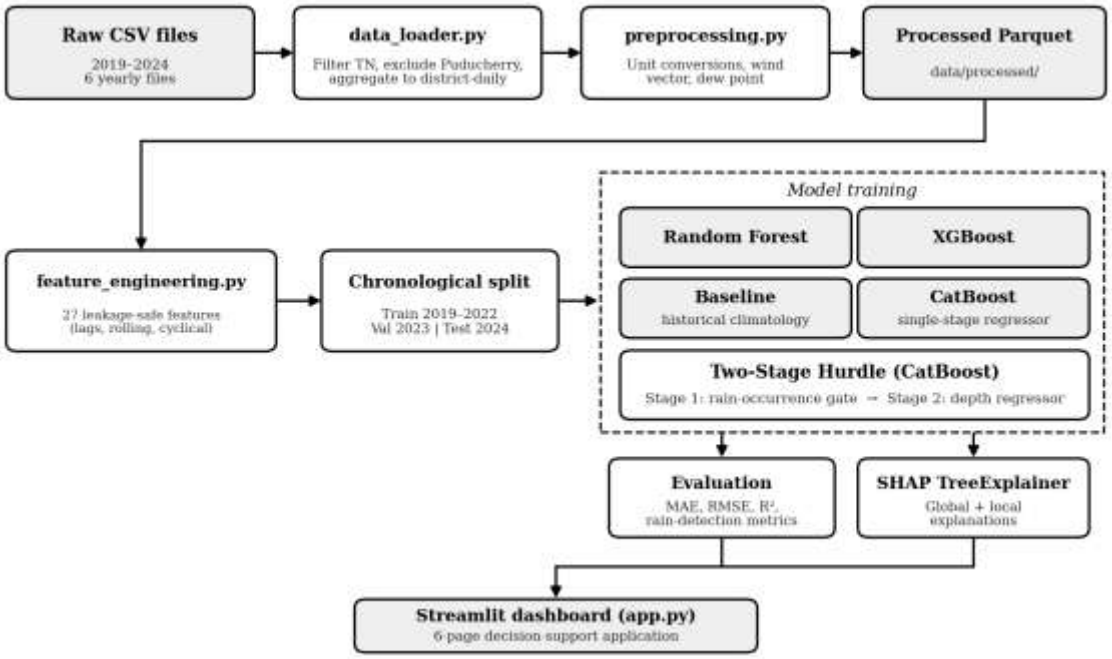



Fig. 1. System architecture of the proposed rainfall prediction framework.

A. Data Ingestion, Aggregation, and Preprocessing

The data_loader.py module reads the six yearly CSV files, filters the records to Tamil Nadu, excludes the Puducherry union territory records, and aggregates the observations from 629 pincodes and 508 stations to daily district averages for

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151163

the 38 districts. A data_validation.py module performs schema, calendar-gap, and physical-plausibility checks. The preprocessing.py module converts units (temperature from Kelvin to degrees Celsius and pressure from Pa to hPa), computes wind speed and direction from the 10 m U and V components, and derives the dew point and the temperature–dew point depression with the Magnus approximation (Section IV). The processed district-daily data are stored in Parquet format, with a CSV copy.

B. Operating Modes

Mode A, which is implemented, performs historical next-day forecasting using only information available at the prediction cutoff time t = 23:59 IST: features come from day t, and the target is the actual rainfall on day t+1. Mode B, which is planned, would couple the models with live numerical weather prediction (WRF/GFS) forecasts from the IMD for real-time operational prediction.

C. Baseline and Single-Stage Models

The baseline, implemented as BaselineRainfallRegressor, represents a naive climatological forecast made without machine learning. It supports two modes: doy_normal, which predicts the historical mean target for each (district_code, day_of_year) pair, and persistence, which predicts the most recent day's rainfall (rain_lag_1d). The doy_normal mode is used for the reported results, and baseline predictions are clipped to be non-negative. The single-stage Random Forest, XGBoost, and CatBoost regressors are configured with the hyperparameters in Table I.

| Random Forest | XGBoost | CatBoost |
|---|---|---|
| n_estimators = 100 | n_estimators = 150 | iterations = 200 |
| max_depth = 12 | max_depth = 6 | depth = 6 |
| min_samples_split = 5 | learning_rate = 0.05 | learning_rate = 0.05 |
| min_samples_leaf = 2 | subsample = 0.8 | l2_leaf_reg = 3.0 |
| n_jobs = 4 | colsample_bytree = 0.8 | random_seed = 42 |
| random_state = 42 | reg_alpha = 0.1 | — |
| — | reg_lambda = 1.0 | — |

TABLE I. SINGLE-STAGE REGRESSOR HYPERPARAMETERS

D. Chronological Split and Expanding-Window Cross-Validation

Data are split strictly in chronological order: 2019–2022 for training (54,264 records), 2023 for validation (10,944 records), and 2024 for the hold-out test (13,870 records). The 2024 test set is completely unseen during all training and model selection decisions. Before final training, an expanding-window cross-validation with four windows (Table II) simulates deployment over time. Cross-validation uses lighter hyperparameters (Random Forest: 50 trees with depth 10; XGBoost: 100 iterations; CatBoost: 120 iterations) to speed up iteration, whereas final training on the full 2019–2022 set uses the configurations in Table I.

| Window | Train Years | Validation Year |
|---|---|---|
| 1 | 2019 | 2020 |
| 2 | 2019–2020 | 2021 |
| 3 | 2019–2021 | 2022 |
| 4 | 2019–2022 | 2023 |

TABLE II. EXPANDING-WINDOW CROSS-VALIDATION

E. Interactive Dashboard

The Streamlit dashboard (app.py) provides six pages: (1) Overview, with executive metrics, the 38-district coverage summary, data validation status, and out-of-time benchmark charts; (2) Rainfall Prediction, with interactive forecasting by district that displays the rainfall depth in mm, an IMD risk badge, the live rain likelihood (%), and a local SHAP waterfall chart; (3) District Analytics, with 38-district comparative trends, historical time series, and geographical error breakdowns; (4) Explainable AI (SHAP), with global feature-importance beeswarm and summary charts; (5) Rainfall Risk and IMD Alerts, with the IMD risk tiers (No Rain, Light, Moderate, Heavy, Very Heavy, Extreme) and an interactive depth-risk simulator; and (6) Model Performance, with benchmark tables, the two-stage metric cards (82.1% accuracy, 90.6% AUC), and a residuals plot.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151163

F. Automated Software Verification Suite

All 14 unit and integration tests pass (pytest tests/). test_preprocessing.py (4 tests) covers unit conversions (K→°C, Pa→hPa), the accuracy of the Magnus dew point formula, and wind vectors. test_features.py (4 tests) covers district lag isolation, safety across missing calendar dates, and the integrity of the cyclical encoding. test_no_data_leakage.py (6 tests) covers the absence of target leakage, verification of the chronological split, and non-negative prediction bounds.

## IV. DATASET AND FEATURE ENGINEERING

The primary dataset is TNWeather24, a historical weather dataset for Tamil Nadu and Puducherry (2019–2024) published on Mendeley Data [1] and provided as six yearly CSV files (Weather_Final_data_YYYY_imputed.csv). Table III summarizes it.

| Attribute | Detail |
|---|---|
| Name | TNWeather24: A Historical Weather Dataset for Tamil Nadu and Puducherry (2019–2024) |
| Source | Mendeley Data (contributor: Balaji V, 2024) |
| DOI | 10.17632/nwdg922wmx.2 |
| License | CC BY 4.0 (Creative Commons Attribution 4.0 International) |
| Coverage | 2019–2024 |
| Spatial scope | 629 pincodes and 508 stations aggregated to daily district averages across all 38 Tamil Nadu districts |

TABLE III. TNWEATHER24 DATASET SUMMARY

The raw meteorological variables provided in each file are listed in Table IV.

| Variable Code | Description | Variable Code | Description |
|---|---|---|---|
| TMP_2m | 2 m air temperature (K) | TCDCRO_atmc | Total cloud cover (%) |
| RH_2m | Relative humidity at 2 m (%) | LCDC_atmc | Low cloud cover (%) |
| PRES_sfc | Surface pressure (Pa) | MCDC_atmc | Mid-level cloud cover (%) |
| PRMSL_msl | Mean sea-level pressure (Pa) | HCDC_atmc | High cloud cover (%) |
| UGRD_10m | U-component of wind at 10 m (m/s) | APCP_sfc | Accumulated surface precipitation (mm) |
| VGRD_10m | V-component of wind at 10 m (m/s) | VIS_2m | Visibility (m) |
| GUST_10m | Wind gust at 10 m (m/s) | | |

TABLE IV. RAW METEOROLOGICAL VARIABLES

Data quality verification produced the following findings, all of which are handled in the pipeline:

- Puducherry exclusion: 23,034 records belonging to the Puducherry union territory were isolated and removed.
- Calendar gaps: date gaps were detected in 2021 (August 1–10) and 2023 (February, March, April, and November). Lags use calendar-day index spacing, so missing dates never falsely appear consecutive.
- Zero duplicates: overlapping records at year boundaries were deduplicated, leaving no duplicate district-date records.

All 27 predictive features use observations recorded strictly on or before the cutoff time t (23:59 IST) and are grouped in Table V. District identity is integer-encoded alphabetically (0–37), and the mapping is saved to district_mapping.json

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151163

and reloaded in training, evaluation, and the dashboard to prevent encoding mismatches. Mid-level cloud cover is loaded and aggregated from the raw data but is not used as a feature.

| Category | Features | Rationale |
|---|---|---|
| Calendar and season | month, day_of_year, sin/cos of day_of_year, sin/cos of month | Northeast monsoon peak (Oct–Dec) versus Southwest monsoon |
| Spatial identity | district_code (0–37) | Alphabetical integer encoding of each district |
| Thermal and moisture | temp_celsius, dew_point_celsius, temp_dew_depression, relative_humidity | Magnus saturation deficit ($T - T_d$) measuring moisture availability |
| Dynamic and pressure | pressure_hpa, wind_speed_ms, wind_direction_deg | Synoptic surface pressure and wind advection |
| Cloud and visibility | total_cloud_cover, low_cloud_cover, high_cloud_cover, visibility_m | Optical thickness and convective condensation indicators |
| Precipitation lags | rain_lag_1d, rain_lag_2d, rain_lag_3d, rain_lag_7d | Past rainfall: autocorrelation and persistence |
| Rolling summaries | rain_roll_mean_3d, rain_roll_mean_7d, rain_roll_max_7d, rain_roll_std_7d, rain_any_last_3d | Past 3-day and 7-day rainfall accumulation trends |

TABLE V. THE 27 INPUT FEATURES BY CATEGORY

The lag features rain_lag_1d, rain_lag_2d, rain_lag_3d, and rain_lag_7d correspond to the rainfall on days t, t−1, t−2, and t−6. Each district's series is reindexed to a full calendar daily frequency before lags and rolling statistics are computed; gap days are filled with NaN, and rows in which temp_celsius is NaN are dropped, so that missing dates cannot corrupt the rolling statistics. The binary indicator rain_any_last_3d, despite its name, is computed from rain_roll_max_7d > 0.1 mm and therefore represents a 7-day lookback. Wind speed is $\sqrt{(u^2 + v^2)}$, and the wind direction follows the meteorological FROM convention, atan2(−u, −v).

The target for district d on day t is the rainfall on the following day (1), and a non-negativity constraint is applied to predictions at inference (2). The Magnus approximation (3)–(4) gives the dew point $T_d$, where T is the air temperature in °C, RH is the relative humidity in %, a = 17.62, and b = 243.12 °C; the feature temp_dew_depression is $T - T_d$.

$$Y(d, t) = \text{Rainfall}(d, t + 1) \tag{1}$$

$$\hat{Y} = \max(0, \hat{Y}_{raw}) \tag{2}$$

$$\gamma = \ln(RH/100) + aT / (b + T) \tag{3}$$

$$T_d = b\gamma / (a - \gamma) \tag{4}$$

## V. TWO-STAGE HURDLE ARCHITECTURE

The proposed architecture resolves the difficulty of zero-inflated rainfall by separating occurrence detection from depth estimation. Stage 1 is a CatBoost classifier that acts as an occurrence gate: it estimates whether rain of at least 0.1 mm will occur and outputs the probability P(Rain) (5). Stage 2 is a CatBoost regressor that estimates the rainfall depth in millimetres (6). The two outputs are combined into the final prediction (7): when P(Rain) is at least 0.48, the regression estimate is used as it is; otherwise it is scaled down by a factor of 0.15, rather than being set to zero.

$$P(\text{Rain}) = \sigma(f_{clf}(X)) \tag{5}$$

$$Y_{reg} = \max(0, f_{reg}(X)) \tag{6}$$

$$Y = Y_{reg} \text{ if } P(\text{Rain}) \geq 0.48, \text{ else } 0.15 \cdot Y_{reg} \tag{7}$$

 This work is licensed under a Creative Commons Attribution 4.0 International License 

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151163

Here X is the feature vector of Section IV, σ is the logistic function, and $f_{clf}$ and $f_{reg}$ are the classifier and regressor, respectively. In the dashboard, the gate probability is shown to the user as the live rain likelihood (%), alongside the predicted depth and the IMD risk badge.

## VI. EXPERIMENTAL RESULTS

The framework is implemented in Python 3.11+ using NumPy, Pandas, SciPy, scikit-learn, XGBoost, CatBoost, SHAP, Streamlit, Plotly, Matplotlib, PyYAML, Joblib, PyArrow, and pytest. Depth forecasts are evaluated with MAE, RMSE, and $R^2$ (8)–(10) on the unseen 2024 test set (13,870 records). Rain occurrence is evaluated with accuracy, precision, recall, and ROC-AUC. Because roughly 80% of days are dry, accuracy is interpreted together with precision, recall, and ROC-AUC.

$$MAE = (1/n)\ \Sigma\ |\hat{y}_i - y_i| \tag{8}$$

$$RMSE = \sqrt{[(1/n)\ \Sigma\ (\hat{y}_i - y_i)^2]} \tag{9}$$

$$R^2 = 1 - \Sigma\ (y_i - \hat{y}_i)^2\ /\ \Sigma\ (y_i - \bar{y})^2 \tag{10}$$

A. Depth Forecasting Performance

Table VI reports the out-of-time performance on the 2024 test set, and Fig. 2 compares MAE and RMSE. The Two-Stage Hurdle model ranks first on MAE, with 0.1393 mm. Relative to the baseline MAE of 0.2118 mm, the MAE reductions are 34.2% for the hurdle model, 31.6% for single-stage CatBoost, 30.6% for XGBoost, and 29.0% for Random Forest. The hurdle model also improves on single-stage CatBoost by 3.9% in MAE (0.1393 versus 0.1449 mm). Its RMSE (0.3216 mm) and $R^2$ (0.4232) are, however, slightly worse than those of single-stage CatBoost (0.3189 mm and 0.4329), which has the best RMSE and $R^2$ among the models. The baseline reaches an $R^2$ of only 0.0046, confirming that climatological averages provide almost no predictive power for individual-day forecasting.

| Rank | Algorithm | MAE (mm) | RMSE (mm) | $R^2$ | Rain Precision | Rain Recall | Rain Detection Accuracy |
|---|---|---|---|---|---|---|---|
| **1** | **Two-Stage Hurdle (CatBoost)** | **0.1393** | **0.3216** | **0.4232** | **74.95%** | **81.80%** | **82.14%** |
| 2 | CatBoost Regressor (Single) | 0.1449 | 0.3189 | 0.4329 | 57.14% | 10.39% | — |
| 3 | XGBoost Regressor | 0.1469 | 0.3192 | 0.4318 | 52.94% | 11.69% | — |
| 4 | Random Forest Regressor | 0.1503 | 0.3292 | 0.3958 | 72.73% | 10.39% | — |
| 5 | Baseline (Historical Climatology) | 0.2118 | 0.4226 | 0.0046 | 0.00% | 0.00% | — |

TABLE VI. BENCHMARK ON THE UNSEEN 2024 TEST SET (13,870 RECORDS)

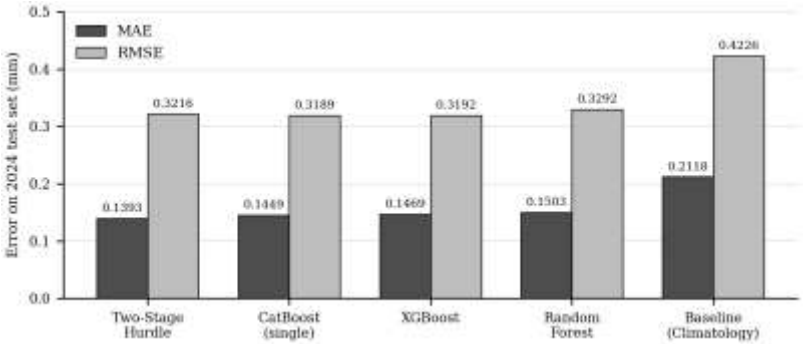



Fig. 2. Test-set MAE and RMSE of the five models (2024).

B. Rain Occurrence Detection

On the completely unseen 2024 test data, the Stage 1 gate of the hurdle model achieves a rain-detection accuracy of 82.14%, a ROC-AUC of 90.56%, a precision of 74.95%, and a rain-day recall of 81.80%. Since roughly 80% of days are



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.151163

dry, the precision, recall, and ROC-AUC values, rather than accuracy alone, indicate how well the gate separates rainy from dry days.

C. Explainability with SHAP

SHAP TreeExplainer decomposed the model outputs to identify the top meteorological drivers, listed in Table VII. Rainfall on the most recent day (persistence) is the strongest single predictor, followed by the 7-day cumulative rainfall, which captures sustained monsoon wet spells. The dew point depression, relative humidity, and total cloud cover complete the top five. In addition to these global beeswarm and summary charts, the dashboard shows a local SHAP waterfall chart for each prediction. These contributions are statistical and are not evidence of physical causality.

| Rank | Feature | Interpretation |
|---|---|---|
| 1 | rain_lag_1d | Most recent day's rainfall; strongest single predictor (persistence) |
| 2 | rain_roll_mean_7d | 7-day cumulative rainfall; captures sustained monsoon wet spells |
| 3 | temp_dew_depression | Dew point depression ($T - T_d$); indicates the saturation deficit |
| 4 | relative_humidity | Ambient boundary-layer moisture |
| 5 | total_cloud_cover | Convective potential and cloud presence |

TABLE VII. TOP FIVE GLOBAL PREDICTORS (SHAP)

## VII. ETHICAL CONSIDERATIONS

This study uses a publicly available dataset released under the CC BY 4.0 license [1], consisting of meteorological observations aggregated to the district level. Several limitations should be considered when using the system. First, extreme rainfall events are likely to be underestimated: cloudburst events (> 64.5 mm) make up less than 0.1% of the dataset, and MSE-optimised regression systematically underestimates extreme convective peaks. Second, SHAP explanations describe statistical feature contributions and do not establish physical causal relationships. Third, the system currently operates only in historical Mode A, without live NWP coupling; it predicts only the t+1 lead time; and its models are not retrained online, so performance may degrade as climate patterns shift. The research gap addressed here is presented as a proposed contribution rather than a definitive claim of novelty.

## VIII. CONCLUSION

This paper presented an explainable, leakage-safe framework for district-level next-day rainfall forecasting across the 38 districts of Tamil Nadu, built around a Two-Stage Hurdle Architecture. On the unseen 2024 test year, the hurdle model achieved a rain-occurrence accuracy of 82.14% with a ROC-AUC of 90.56%, and estimated rainfall depth with an MAE of 0.1393 mm, a 34.2% reduction relative to the climatological baseline; single-stage CatBoost retained a slightly better RMSE and $R^2$. SHAP analysis identified recent rainfall, the 7-day cumulative rainfall, and the dew point depression as leading drivers, and the Streamlit dashboard and the 14-test suite make the framework usable and verifiable. Future work will focus on extreme-value loss functions (Tweedie loss or quantile regression) to better capture heavy-tail rainfall events; Mode B integration with live IMD numerical weather prediction forecasts; multi-horizon lead times extending from t+1 to t+3 and t+7; LSTM and Transformer sequence models as additional benchmarks; and online learning with incremental model updates as new weather data arrive.

## ACKNOWLEDGMENT

The authors thank the contributor of the TNWeather24 dataset and Mendeley Data for making the data available under the CC BY 4.0 license.

## REFERENCES

[1] Balaji V, "TNWeather24: A historical weather dataset for Tamil Nadu and Puducherry (2019–2024)," Mendeley Data, V2, 2024, doi: 10.17632/nwdg922wmx.2. [Online]. Available: https://data.mendeley.com/datasets/nwdg922wmx/2
[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. KDD, 2016.
[3] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Proc. NeurIPS, 2018.
[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.

 This work is licensed under a Creative Commons Attribution 4.0 International License 