Int. J. Sci. R. Tech., 2026 3(11)
A Multidisciplinary peer-reviewed Journal
www.ijsrtjournal.com [ISSN: 2394-7063]

Original Article

Multi-Hazard Disaster Prediction and Warning System Using Machine Learning

Kaustubh M. Patil*, Rushikesh M. Patange, Siddhesh A. Nikam, Laxmikant M. Gholave

Department of Computer Applications, D.Y. Patil Institute of MCA and Management, Pune, Maharashtra, India.

ABSTRACT

Disaster early-warning studies often report headline accuracy, which can hide class imbalance, temporal drift and spatial extrapolation error. This paper applies a leakage-aware evaluation protocol to two research modules. The first is flood severity classification among 4,548 recorded flood events from 155 gauges, evaluated with a chronological split (training before 2010, validation 2010–2014, test 2015–2020). The second is a heat-event proxy built from IMD gridded daily maximum temperature (1° grid, 1951–2024). A grid-cell day is positive when it belongs to a run of at least three consecutive hot days, where a hot day exceeds the cell's 1961–1990 calendar-day 90th percentile and its baseline climatological Tmax exceeds 35 °C. The proxy covers 338 eligible cells, April–June target dates and 1991–2024. It is not an official IMD declaration. For flood severity, validation PR-AUC selected logistic regression, which achieved test balanced accuracy 0.6445 and ROC-AUC 0.6483 against a majority-class balanced accuracy of 0.5000. For the heat-event proxy, histogram-based gradient boosting was selected by validation PR-AUC (0.5811). On the 2020–2024 test period it achieved balanced accuracy 88.82%, recall 90.36%, ROC-AUC 0.9548 and PR-AUC 0.5813, but produced 18,415 false positives at threshold 0.50. A simple persistence rule had higher F1 (0.591 against 0.457) and fewer false positives. A spatial hold-out over 51 unseen cells gave balanced accuracy 87.33%. Attributions are associations, not causes, and these results do not establish operational early-warning performance.

Keywords: disaster risk classification, flood severity, heat-event detection, class imbalance, temporal and spatial generalization, gridded temperature, leakage-aware evaluation

INTRODUCTION

Climate-related hazards dominate recorded disaster occurrence. Between 2000 and 2019, 6,681 of 7,348 recorded disasters worldwide were climate-related, meaning meteorological, climatological or hydrological [1]. India is exposed to several of these hazards at once: riverine and flash flooding during the monsoon, recurrent heatwaves in the pre-monsoon season, and multi-season drought. Machine learning (ML) has been applied to all three. Reviews of flood prediction report data-driven gains alongside persistent problems in data quality and evaluation design [2].

Three methodological issues recur in this literature. First, many studies use random train/test splits on autocorrelated, geographically clustered data, which can overstate predictive performance [3]. Second, disaster datasets are typically imbalanced, and accuracy or ROC-AUC alone can give an optimistic picture of minority-class detection [4]. Third, the modeled target is often poorly specified. A model trained to label recorded events is sometimes presented as forecasting occurrence, a distinction that matters for early-warning use.

This paper applies one evaluation protocol to two completed modules. The first is flood severity classification on recorded flood events, where the target separates severe from non-severe events among events that have already occurred. The second is an IMD gridded-Tmax heat-event proxy, where the target marks grid-cell days that belong to a run of at least three consecutive hot days under a fixed 1961–1990 climatological baseline. A third module, drought, is planned and is not evaluated here.

The contributions are as follows.

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

- a common protocol for both targets: target definition, leakage audit, chronological split, validation-based model selection, a single frozen test evaluation, and explicit baselines;
- a chronologically split evaluation of flood severity classification on 4,548 recorded events, in which all three model families are reported on the test partition for transparency;
- a heat-event proxy evaluation on 1,043,130 grid-cell/day rows, with a spatial hold-out over 51 cells and a persistence baseline;
- majority-class and persistence baselines, class prevalence and false-positive counts for every reported result; and
- full model configurations, software versions and code, so that every reported number can be traced to the code and data used to produce it.

Section II reviews related work. Section III states the research gap and objectives. Section IV describes the protocol. Section V describes data and methods, Section VI reports results, Sections VII and VIII discuss results and limitations, Section IX outlines future work, Section X states code and data availability, and Section XI concludes.

II. RELATED WORK

Flood prediction with machine learning. Mosavi, Ozturk and Chau reviewed ML approaches to flood prediction and identified neural networks, support vector machines and decision-tree ensembles as dominant families. They also noted that comparisons across studies are difficult because datasets and evaluation protocols differ [2]. Flood-susceptibility studies have compared machine-learning families, including logistic regression, random forest and gradient-boosted trees, and the reported rankings differ across datasets and regions [5], [6], [7]. The ranking should therefore be established for each study rather than assumed.

Heat waves and extreme temperature. Rohini, Rajeevan and Srivastava analyzed heat-wave variability over India using gridded daily temperature data and an Excess Heat Factor [8]. The IMD gridded dataset used here is a daily 1° product for India developed from station data [9]. The three-consecutive-day requirement, the April–June season and the five-day centered window for the 90th-percentile threshold follow [8]. The 1961–1990 reference period appears in that study's excess-heat index; its use for the threshold applied here is our choice. The 35 °C gate is also our choice and is not taken from that study (Section V-C).

Rohini and Rajeevan evaluated TIGGE ensemble forecasts for Indian heat-wave events [10]. That is a forecasting study. The present task is retrospective classification of grid-cell days and is not evidence of operational forecast skill.

Drought prediction with machine learning. A recent review reported rapid growth in drought ML research since 2013, with forecasting and monitoring dominating and impact assessment and explainable AI underexplored [11]. An earlier review catalogued the model families and drought indices used as targets [12]. These reviews inform the planned drought module, which is not evaluated here.

Evaluation under imbalance, shift and extrapolation. Balanced accuracy, the mean of per-class recall, corrects for the optimistic bias of accuracy on imbalanced data [13]. Precision–recall analysis is more informative than ROC analysis when the positive class is rare [4]. Blocked validation is recommended where temporal or spatial dependence exists [3]. Area-of-applicability methods delineate the predictor-space region in which a model's error can be assumed to hold [14]. We use a spatial hold-out but do not estimate the area of applicability.

III. RESEARCH GAP AND OBJECTIVES

Existing hazard-ML studies often (i) report single-split or random-split performance, (ii) omit majority-class baselines and prevalence, (iii) do not separate research evaluation from deployment, and (iv) describe severity classifiers or threshold-based labels as occurrence forecasts. Few studies apply one protocol across more than one hazard. The objectives are:

- O1. evaluate flood severity classification on recorded events with a chronological split and validation-based model selection;
- O2. evaluate an IMD Tmax-based heat-event proxy whose predictors are restricted to observations available by the issue time, on a temporal test period;
- O3. assess generalization across time and, for the heat proxy, across held-out spatial blocks; and

- O4. report baselines, class prevalence and error burden, and identify model-associated features with caveats.

IV. PROPOSED EVALUATION PROTOCOL

The protocol has four stages, applied identically to each module. Target definition uses real historical records and states the target in terms the data can support. Feature engineering and a leakage audit restrict predictors to information available at an explicit issue time. Chronological evaluation selects a model on a validation period and then performs a single frozen test evaluation. Explainability reports attributions as associations and keeps the research model separate from any deployment. The protocol is hazard-agnostic at the level of procedure and hazard-specific at the level of target and features.

V. DATA AND METHODS

Table I summarizes the datasets, targets and evaluation status.

Table I: Summary of Datasets, Targets and Evaluation Status

Module	Source	Period	Unit	Size	Target	Status
Flood	Recorded flood events, 155 gauges	≈1965–2020	Event	4,548 events	Severe (1) vs. non-severe (0)	Complete; chronological split
Heat	IMD gridded daily Tmax, 1° grid	1951–2024 archive; 1991–2024 targets	Grid-cell/day	1,043,130 rows; 338 cells	Tmax-based heat-event proxy (≥3 consecutive hot days)	Complete; not an official declaration
Drought	Not documented	–	–	–	–	Planned; not evaluated

A. *Flood Data and Protocol*

The flood dataset contains 4,548 recorded flood events from 155 gauges, covering approximately 1965–2020. The events come from the INDOFLOODS database [19], [20], which combines long-term station discharge observations with official warning and danger water levels. The database does not cover the Ganga and Brahmaputra basins [20]. The target is binary: Flood (0) and Severe Flood (1). The dataset contains no non-flood examples. The task is therefore severity discrimination among events and does not predict whether a flood will occur.

Nine event-outcome variables measured during or after each event were excluded from predictors: Peak Flood Level, Peak FL Date, Num Peak FL, Peak Discharge Q, Peak Discharge Date, Flood Volume, Event Duration, Time to Peak and Recession Time. Fifteen candidate geomorphological columns with at least 80% missing values were removed during cleaning, leaving 46 geomorphological features. The research feature set has 58 features: ten antecedent precipitation totals (T1d–T10d), the 46 geomorphological features, the event month, and a monsoon-season indicator.

Events were split chronologically by start year. Training contains 3,251 events before 2010 from 88 gauges. Validation contains 613 events from 2010–2014 at 74 gauges. Test contains 684 events from 2015–2020 at 131 gauges. Of the test gauges, 70 (333 events) also appear in training and 61 (351 events) do not. Median imputation and standardization were fit on the training partition only. The decision threshold was fixed at 0.50 for all models, and no threshold tuning was performed.

B. *Heat-Event Data*

The heat experiment uses IMD gridded daily maximum temperature on a 1° grid, with 31×31 nominal points over 7.5–37.5° N and 67.5–97.5° E, for 1951–2024 (27,029 days) [15]. The values are gridded estimates, not station observations. The documented missing-value sentinel 99.9 was converted to missing. No other values were changed or filtered. Of 25,974,869 cell-day slots, 36.55% hold valid values, so 63.45% are missing. Of the 961 nominal cells, 362 have at least one valid value and 599 are missing throughout. The number of cells with at least one valid value changes over time. It was 334–340 in 1951–1968, rose to 358 in 1969, and varied between 341 and 362 afterwards. We did not investigate the step near 1969. The archive's minimum valid value is −2.06 °C, and no quality control beyond the sentinel was applied.

A cell is eligible when it has valid data and, for every March–July centered five-day window in the 1961–1990 baseline, at least 120 of the 150 nominal values are valid. A total of 338 cells satisfy this rule and form the study domain.

C. *Heat-Event Target*

For each eligible cell and calendar day, a five-day centered window over 1961–1990 provides the 90th-percentile threshold and the climatological mean Tmax. A day is hot when its observed Tmax exceeds the 90th-percentile threshold and the climatological mean exceeds 35 °C. The 35 °C condition applies to the baseline mean, not to target-day Tmax. February 29 is mapped to the February 28 calendar position. A grid-cell day is positive when it belongs to at least one run of three or more consecutive hot days. Membership is determined using observations up to two days after the target date, so the label is not available at issue time. We call this a heat-event proxy. It is not an official IMD declaration and not a reconstruction of station or subdivision declarations.

D. *Issue Time, Predictors and Splits*

The issue time is 23:59 IST on day D, and the target is D+1. The archive has no publication timestamps, so predictors end at target−2 (D−1), a conservative one-day buffer. The 19 predictors are: grid latitude and longitude; target month and the sine and cosine of the target day of year; the latest observed Tmax (target−2) and lagged Tmax at 3 to 8 days before the target; trailing 3-, 5- and 7-day means and 3- and 7-day maxima; the latest Tmax minus the local 90th percentile; and the number of consecutive hot days ending at target−2, capped at seven. No predictor uses the target day or any later observation. The climatology is fixed over 1961–1990 and predates every target year.

A row is included only when all seven input days (target−8 to target−2), all five label-window days (target−2 to target+2) and baseline support are valid. Excluded rows are not assigned negative labels. Target dates are April–June, 1991–2024, giving 1,043,130 rows. Training covers 1991–2014, validation 2015–2019 and test 2020–2024. Model selection uses the validation partition only, and the test partition is evaluated once for the selected model.

For the spatial experiment, a deterministic 5×5-degree block pattern withheld 51 cells. The model was trained on pre-2020 rows of the remaining cells (287 cells, 754,806 rows) and evaluated on 2020–2024 rows of the withheld cells (21,272 negative and 1,933 positive rows).

E. *Models and Selection*

Three model families were compared in each module: logistic regression (LR), random forest (RF) [16] and gradient boosting [17]. All models were implemented in scikit-learn [18], with seed 42. Configurations were fixed in advance and were not tuned by search.

- Flood: LR with max_iter=2000 and balanced class weights; RF with 500 trees and balanced weights; GradientBoostingClassifier with 300 estimators, depth 5, learning rate 0.05 and subsample 0.8.
- Heat: LR with max_iter=1500, balanced weights and LBFGS, with a training-fitted StandardScaler; RF with 150 trees, maximum depth 18, minimum leaf size 5, square-root feature sampling and balanced-subsample weights; HistGradientBoostingClassifier with 250 iterations, depth 5, learning rate 0.05, minimum leaf size 20, L2 penalty 1.0 and balanced class weights.

Gradient boosting in both modules uses scikit-learn implementations (GradientBoostingClassifier for flood and HistGradientBoostingClassifier for heat).

In both modules the selected model is the one with the highest validation PR-AUC (average precision), since PR-AUC is more informative than ROC-AUC under low positive rates [4]. Scores are decision scores at threshold 0.50, not calibrated probabilities.

F. *Metrics, Baselines and Explainability*

We report accuracy, balanced accuracy, precision, recall, F1, ROC-AUC and PR-AUC [13]. A majority-class baseline is reported for each experiment. For the heat proxy, two persistence rules are also reported. Both were fixed in advance and require no fitting. Rule P1 flags a day when the target−2 day was hot. Rule P2 flags a day when Tmax at target−2 exceeded its local 90th percentile. The ROC-AUC and PR-AUC of P1 are computed from the hot-streak count and those of P2 from the anomaly, so they rank days rather than report the binary rule.

Explainability is limited to two methods. For flood, we report standardized LR coefficients fitted on the training partition. For heat, we report permutation importance on the validation partition, measured as the drop in PR-AUC after shuffling each feature (three repeats). Neither method establishes causation.

VI. EXPERIMENTAL RESULTS

A. *Flood Results*

Validation. Table II reports validation metrics at threshold 0.50. LR had the highest PR-AUC (0.6402), balanced accuracy (0.6902), recall (0.5903) and F1 (0.6063). RF had the highest validation ROC-AUC (0.7433). Selection used the pre-specified rule, validation PR-AUC, and LR was selected.

Table II: Flood Validation Comparison (2010–2014; 613 Events; Threshold 0.50)

Model	Acc.	Bal. Acc.	Prec.	Recall	F1	ROC-AUC	PR-AUC
Logistic regression (selected)	0.7162	0.6902	0.6233	0.5903	0.6063	0.7104	0.6402
Random forest	0.6721	0.6253	0.5739	0.4449	0.5012	0.7433	0.6161
Gradient boosting	0.6966	0.6320	0.6541	0.3833	0.4833	0.7311	0.6225
Majority baseline	0.6297	0.5000	–	–	–	0.5000	0.3703

Selection rule: highest validation PR-AUC. Non-selected rows are reported for completeness.

Test. On the test partition (684 events, 318 severe), the validation-selected LR achieved balanced accuracy 0.6445, ROC-AUC 0.6483 and PR-AUC 0.6210 (Table III). Its confusion matrix is TN = 283, FP = 83, FN = 154 and TP = 164. Its ROC-AUC fell from 0.7104 on validation to 0.6483 on test, which indicates degradation under temporal shift. The test partition was scored for all three models in the same run after selection. The non-selected models were not chosen on the test set. Even so, gradient boosting had the highest test ROC-AUC (0.7248) and PR-AUC (0.6772), and LR had the highest test balanced accuracy. Because the test partition had been inspected descriptively in earlier work, this comparison is descriptive and not confirmatory.

Table III: Flood Frozen Test Comparison (2015–2020; 684 Events; Threshold 0.50)

Model	Acc.	Bal. Acc.	Prec.	Recall	F1	ROC-AUC	PR-AUC
Logistic regression (selected)	0.6535	0.6445	0.6640	0.5157	0.5805	0.6483	0.6210
Random forest	0.6126	0.5957	0.6532	0.3553	0.4603	0.6666	0.6380
Gradient boosting	0.6404	0.6206	0.7500	0.3396	0.4675	0.7248	0.6772
Majority baseline	0.5351	0.5000	–	–	–	0.5000	0.4649

Descriptive: all three models were scored after validation-based selection; the test partition had been inspected in earlier work.

Gauge subgroups. On the 333 events from gauges seen in training, balanced accuracy is 0.6296 and ROC-AUC is 0.6834. On the 351 events from unseen gauges, balanced accuracy is 0.6580 and ROC-AUC is 0.6245. These subgroup results are exploratory. The unseen group has higher balanced accuracy but lower ROC-AUC, and the group sizes do not support a claim of geographic generalization.

Coefficients. The largest standardized LR coefficients are for third-order stream length (−2.88), stream order (+2.49), maximal flow length (−1.93), first-order stream length (+1.67), number of fifth-order streams (+1.62), circularity ratio (+1.59), infiltration number (−1.31), fifth-order stream length (−1.26), third–fourth bifurcation ratio (+1.22) and number of second-order streams (−1.15). Many stream-network features are strongly correlated, so individual coefficients are unstable. They describe this fitted model and are not independent effects.

B. *Heat-Event Results*

Table IV gives the partition sizes and class counts. The positive rate falls from 9.58% in validation to 5.84% in test.

Table IV: Heat-event Partitions and Class Counts

Partition	Target years	Rows	Positive	Negative	Positive rate
Training	1991–2014	735,866	44,746	691,120	6.08%
Validation	2015–2019	153,529	14,714	138,815	9.58%
Test	2020–2024	153,735	8,976	144,759	5.84%
Total	1991–2024	1,043,130	68,436	974,694	6.56%
Spatial hold-out (51 cells)	2020–2024	23,205	1,933	21,272	8.33%

Validation and selection. Table V reports validation results at threshold 0.50. Validation PR-AUC selected gradient boosting (0.5811), ahead of RF (0.5515) and LR (0.5366). The selected model was then scored once on the test partition, where it achieved balanced accuracy 88.82%, recall 90.36%, precision 30.58%, ROC-AUC 0.9548 and PR-AUC 0.5813. Its accuracy is 87.46%, below the all-negative baseline of 94.16%. That is expected at a 5.84% positive rate and does not indicate poor detection of events. The model produces 18,415 false positives, which is 12.72% of the 144,759 negative rows. LR and RF were not scored on the test partition.

Table V: Heat-event Results at Threshold 0.50

Setting	Model	Acc.	Bal. Acc.	Prec.	Recall	F1	ROC-AUC	PR-AUC
Validation 2015–2019	Gradient boosting (selected)	81.82%	85.83%	33.46%	90.80%	48.90%	0.9313	0.5811
Validation 2015–2019	Random forest	88.42%	83.69%	44.10%	77.83%	56.30%	0.9255	0.5515
Validation 2015–2019	Logistic regression	79.62%	84.00%	30.68%	89.41%	45.68%	0.9168	0.5366
Validation 2015–2019	All-negative baseline	90.42%	50.00%	0.00%	0.00%	0.00%	0.5000	0.0958
Test 2020–2024	Gradient boosting (frozen)	87.46%	88.82%	30.58%	90.36%	45.69%	0.9548	0.5813
Test 2020–2024	Persistence P1 (hot at target−2)	94.43%	82.49%	51.74%	68.97%	59.13%	0.8265	0.4033
Test 2020–2024	Persistence P2 (Tmax > local q90)	87.39%	79.34%	27.38%	70.22%	39.39%	0.8917	0.2657
Test 2020–2024	All-negative baseline	94.16%	50.00%	0.00%	0.00%	0.00%	0.5000	0.0584
Spatial 2020–2024 (51 cells)	Gradient boosting	82.16%	87.33%	31.05%	93.53%	46.62%	0.9486	0.6417
Spatial 2020–2024 (51 cells)	All-negative baseline	91.67%	50.00%	0.00%	0.00%	0.00%	0.5000	0.0833

P1 and P2 are fixed rules; their ROC-AUC and PR-AUC rank days using the hot-streak count and the anomaly, respectively. Test false positives for the model: 18,415; for P1: 5,775.

Comparison with persistence. At threshold 0.50, rule P1 has higher accuracy (94.43%), precision (51.74%) and F1 (0.591), and it produces fewer false positives (5,775) than the model. The model has higher recall (90.36% against 68.97%), ROC-AUC (0.9548 against 0.8265) and PR-AUC (0.5813 against 0.4033). Rule P2 has ROC-AUC 0.8917 and PR-AUC 0.2657. The learned model therefore ranks days better and detects

more events, but at the default threshold it does not dominate the persistence rule on F1 or false-positive burden.

Spatial hold-out. On the 51 withheld cells in 2020–2024, balanced accuracy is 87.33%, recall 93.53%, precision 31.05%, ROC-AUC 0.9486 and PR-AUC 0.6417. Accuracy is 82.16%, below the all-negative baseline of 91.67%. Balanced accuracy differs from the temporal test by less than 1.6 points. This is one spatial partition with 51 cells, so it does not establish generalization to other regions.

Feature importance. Table VI reports validation permutation importance. Hot-streak length (PR-AUC drop 0.302) and the latest Tmax anomaly relative to the local 90th percentile (0.188) dominate. No other feature exceeds 0.057, and several contribute nothing or slightly negative values. Both dominant features are built from the same hot-day definition used for the labels, so their importance partly reflects how the target is constructed (Section VII).

Table VI: Validation Permutation Importance for the Selected Heat Model (Drop in PR-AUC; Top Features)

Feature	PR-AUC drop	Std.
hot_streak_days_cap7	0.3023	0.0030
latest_tmax_minus_local_q90_c	0.1883	0.0032
trailing_3d_tmax_max_c	0.0568	0.0026
tmax_lag_3d_before_target_c	0.0503	0.0022
target_doy_sin	0.0367	0.0018
latest_observed_tmax_c	0.0323	0.0018
tmax_lag_8d_before_target_c	0.0257	0.0021

Three repeats on validation rows; the remaining 12 features each contribute at most 0.009.

DISCUSSION

The two targets have different evaluation profiles. The flood severity task is harder by the measures reported here: test balanced accuracy is 0.6445 and ROC-AUC is 0.6483, with a substantial drop from validation. The heat proxy has high ROC-AUC and balanced accuracy, but its precision is low at the default threshold. The two targets also differ in prevalence (46.5% severe among flood test events against 5.84% positive heat test days), so their PR-AUC values and balanced accuracies are not directly comparable [4].

For the heat proxy, accuracy alone would mislead. The all-negative baseline is more accurate than the model but detects no events. A fair summary is a trade-off: the model detects about 90% of target-positive test days at the cost of about 18,400 false positives. Whether that trade-off is acceptable depends on the relative cost of missed events and false alarms. That is an application decision outside this study.

The heat proxy has a structural caveat. Its strongest predictors encode the hot-day definition used for the labels. A high ROC-AUC therefore partly reflects the model recovering its own target definition from recent observations, not an independent physical signal. The persistence rules show the same effect: P1, which uses only the target definition, already reaches balanced accuracy 0.8249 and ROC-AUC 0.8265. The learned model adds value in ranking, precision-recall behavior and recall, but this study does not show that it captures physical processes beyond the definition.

For flood severity, the coefficients are consistent with hydrological intuition about stream-network structure. They are, however, associations inside a recorded-event dataset and should not be used to infer which catchment features cause severe floods. The validation-selected LR was not the best model on the test ranking metrics. The validation rule is a prespecified basis for the choice, and the test comparison should not be used to re-select the model.

VIII. LIMITATIONS AND THREATS TO VALIDITY

Target validity. The flood target measures severity among recorded events, not occurrence. The heat target is a gridded-Tmax construct. It is not an official IMD heatwave declaration, not a station-level measure and not an operational warning.

Criterion sourcing. The three-day requirement, the April–June season and the five-day window for the 90th percentile follow [8]. The 35 °C gate and the

1961–1990 baseline applied to the threshold are our choices. We have not tested sensitivity to them.

Data. The heat data are gridded estimates with 63.45% missing cell-days. The number of valid cells changes with a step near 1969, whose cause we did not investigate. No quality control beyond the sentinel was applied. The baseline is fixed at 1961–1990 and may not represent current climate.

Timing. The heat label uses observations up to two days after the target date. Predictors stop at target−2. The archive has no release timestamps, so the cutoff is conservative but not verified against release times.

Leakage and independence. The flood test partition was inspected descriptively before model freezing, so it is not fully independent. Gauges overlap between training and test partitions. The non-selected flood models were scored on test after selection, and the heat predictors encode the label definition, as discussed above.

Implementation. Results use a single seed, and no repeated splits or confidence intervals are reported. Model outputs are not calibrated.

Generalization. The spatial hold-out covers 51 cells and one partition scheme. Flood gauge subgroups are exploratory. The flood data exclude the Ganga and Brahmaputra basins, so the results do not cover those basins.

Explainability. Coefficients and permutation importance are associations. Correlated features make individual attributions unstable. SHAP was not used.

Scope. Two of the three planned hazard modules are evaluated. Findings may not extend to drought. Deployment is outside the scope of this paper.

IX. FUTURE WORK

- 1) Complete the drought module under the same protocol, starting with data acquisition, target definition and a leakage audit.
- 2) Pre-specify and run a new evaluation of all candidate models on a fresh test period. Do not rescore unselected models on a partition already inspected.
- 3) Test sensitivity to the 35 °C gate and to the baseline period, including 1991–2020.
- 4) Investigate the step in valid cell counts near 1969 and validate the proxy against station-level IMD data.
- 5) Estimate the area of applicability of the heat model [14] and compare it with the spatial results.
- 6) Evaluate threshold selection under explicit cost assumptions, and calibrate outputs if probabilistic interpretation is needed.
- 7) Repeat experiments over multiple seeds and splits, and report uncertainty intervals.

X. CODE AND DATA AVAILABILITY

The heat experiment uses the public IMD gridded daily Tmax archive [15]. The flood events come from the INDOFLOODS database [19], [20], which is openly available for academic and research use on Zenodo; the database and its associated article are cited as its terms require. Code, configuration files and intermediate outputs needed to reproduce the reported numbers are available at https://github.com/kaustubhp23/AI-Disaster-Prediction-System. The experiments used scikit-learn 1.9.1, pandas 3.0.5 and NumPy 2.5.3.

CONCLUSION

This paper presented a leakage-aware evaluation protocol and applied it to two modules. For flood severity classification on 4,548 recorded events, the validation-selected logistic regression achieved test balanced accuracy of 0.6445 and ROC-AUC of 0.6483. For the IMD gridded heat-event proxy, histogram-based gradient boosting achieved temporal-test balanced accuracy of 88.82%, recall of 90.36% and ROC-AUC of 0.9548, with 18,415 false positives at threshold 0.50. A persistence rule had higher F1 and fewer false positives, and the learned model's advantage lies mainly in ranking and event detection. The spatial hold-out gave balanced accuracy of 87.33% over 51 cells. The heat target partly reproduces its own definition, and the proxy is not an official declaration. These results support a careful research claim and do not establish operational early-warning performance.

REFERENCES

1. United Nations Office for Disaster Risk Reduction (UNDRR), "The human cost of disasters: An

overview of the last 20 years (2000–2019)," Geneva, Switzerland, 2020.

2. A. Mosavi, P. Ozturk, and K. Chau, "Flood prediction using machine learning models: Literature review," Water, vol. 10, no. 11, Art. no. 1536, 2018, doi: 10.3390/w10111536.
3. D. R. Roberts et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," Ecography, vol. 40, no. 8, pp. 913–929, 2017, doi: 10.1111/ecog.02881.
4. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, Art. no. e0118432, 2015, doi: 10.1371/journal.pone.0118432.
5. G. Dawson, J. Butt, A. Jones, and P. Fraccaro, "Flood susceptibility mapping at the country scale using machine learning approaches," Appl. AI Lett., vol. 4, no. 4, 2023, doi: 10.1002/ail2.88.
6. R. Madhuri, S. Sistla, and K. Srinivasa Raju, "Application of machine learning algorithms for flood susceptibility assessment and risk management," J. Water Clim. Change, vol. 12, no. 6, pp. 2608–2623, 2021, doi: 10.2166/wcc.2021.051.
7. Y. Wu, Z. Zhang, X. Qi, W. Hu, and S. Si, "Prediction of flood sensitivity based on logistic regression, extreme gradient boosting, and random forest modeling methods," Water Sci. Technol., vol. 89, no. 10, pp. 2605–2624, 2024, doi: 10.2166/wst.2024.146.
8. P. Rohini, M. Rajeevan, and A. K. Srivastava, "On the variability and increasing trends of heat waves over India," Sci. Rep., vol. 6, Art. no. 26153, 2016, doi: 10.1038/srep26153.
9. A. K. Srivastava, M. Rajeevan, and S. R. Kshirsagar, "Development of a high resolution daily gridded temperature data set (1969–2005) for the Indian region," Atmos. Sci. Lett., vol. 10, no. 4, pp. 249–254, 2009, doi: 10.1002/asl.232.
10. P. Rohini and M. Rajeevan, "An analysis of prediction skill of heat waves over India using TIGGE ensemble forecasts," Earth Space Sci., vol. 10, no. 3, Art. no. e2020EA001545, 2023, doi: 10.1029/2020EA001545.
11. G. Ascenso et al., "The maturation of AI in drought science: A review of trends, pitfalls, and priorities," Water Resour. Res., vol. 62, no. 4, Art. no. e2025WR041828, 2026, doi: 10.1029/2025WR041828.
12. K. Sundararajan et al., "A contemporary review on drought modeling using machine learning approaches," Comput. Model. Eng. Sci., vol. 128, no. 2, pp. 447–487, 2021, doi: 10.32604/cmes.2021.015528.
13. K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in Proc. 20th Int. Conf. Pattern Recognit. (ICPR), 2010, pp. 3121–3124, doi: 10.1109/ICPR.2010.764.
14. H. Meyer and E. Pebesma, "Predicting into unknown space? Estimating the area of applicability of spatial prediction models," Methods Ecol. Evol., vol. 12, no. 9, pp. 1620–1633, 2021, doi: 10.1111/2041-210X.13650.
15. India Meteorological Department, "IMD high resolution 1° × 1° gridded daily maximum temperature data, 1951–2024," Climate Monitoring and Prediction Group. [Online]. Available: https://www.imdpune.gov.in/cmpg/Griddata/Max_1_Bin.html. [Accessed: Sep. 30, 2026].
16. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
17. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Statist., vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.
18. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
19. S. K. Kuntla and M. Saharia, "INDOFLOODS: A comprehensive database for flood events in India enhanced with catchment attributes," Bull. Amer. Meteor. Soc., vol. 106, no. 2, pp. E333–E343, 2025, doi: 10.1175/BAMS-D-24-0008.1.
20. S. K. Kuntla and M. Saharia, "INDOFLOODS: A comprehensive database for flood events in India enhanced with catchment attributes," version 1.0, Zenodo, 2025, doi: 10.5281/zenodo.14584654.

HOW TO CITE: Kaustubh M. Patil, Rushikesh *M. Patange*, Siddhesh A. Nikam, Laxmikant *M. Gholave*, Multi-Hazard Disaster Prediction and Warning System Using Machine Learning, Int. J. Sci. R. Tech., 2026, 3 (11), 632-641. https://doi.org/10.5281/zenodo.24215742