Copyleaks

Scan details

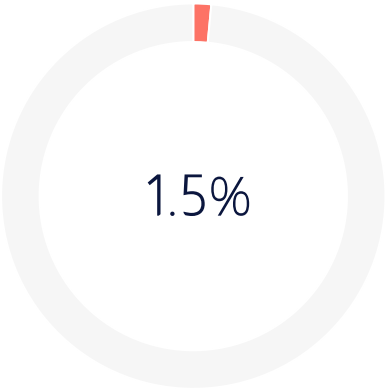
Scan Time	Total Pages	Total Words
August 9th, 2025 at 08:01 UTC	15	3651

Analysis Report

Plagiarism Detection and AI Detection Report

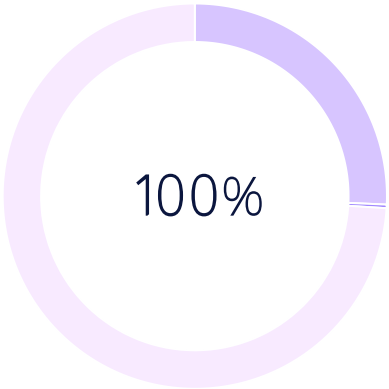
A_Hybrid_ML_Model_for_Landslide_prediction.pdf

Plagiarism Detection



Plagiarism Types	Text Coverage	Words
Identical	1.5%	15
Minor Changes	0%	0
Paraphrased	0%	0
Excluded		
Omitted Words		2,651

AI Detection



	Text Coverage	Words
AI Text	100%	1,000
Low Frequency		257
Medium Frequency		0
High Frequency		3
Human Text	0%	0
Excluded		
Omitted Words		2,651

Plagiarism

Results (3)

*Results may not appear because the feature has been disabled.

Private Cloud Hub	Shared Data Hub	Filtered / Excluded
0	0	0
Internet Sources	**AI Source Match**	**Current Batch**
3	0	0

Plagiarism Types	Text Coverage	Words
Identical	1.5%	15
Minor Changes	0%	0
Paraphrased	0%	0
Excluded		
Omitted Words		2,651

About Plagiarism Detection

Our AI-powered plagiarism scans offer three layers of text similarity detection: Identical, Minor Changes, and Paraphrased. Based on your scan settings we also provide insight on how much of the text you are not scanning for plagiarism (Omitted words).

Identical

One to one exact word matches. Learn more

Minor Changes

Words that hold nearly the same meaning but have a change to their form (e.g. "large" becomes "largely"). Learn more

Paraphrased

Different words that hold the same meaning that replace the original content (e.g. 'large' becomes 'big') Learn more

Omitted Words

The portion of text that is not being scanned for plagiarism based on the scan settings. (e.g. the 'Ignore quotations' setting is enabled and the document is 20% quotations making the omitted words percentage 20%) Learn more

Copyleaks Shared Data Hub

Our Shared Data Hub is a collection of millions of user-submitted documents that you can utilize as a scan resource and choose whether or not you would like to submit the file you are scanning into the Shared Data Hub. Learn more

Filtered and Excluded Results

The report will generate a complete list of results. There is always the option to exclude specific results that are not relevant. Note, by unchecking certain results, the similarity percentage may change. Learn more

Current Batch Results

These are the results displayed from the collection, or batch, of files uploaded for a scan at the same time. Learn more

Plagiarism Detection Results: (3)

content 1.5%

https://www.kerwa.ucr.ac.cr/server/api/core/bitstreams/ba9ba65d-1bd6-437a-b564-a3096773be10/content

HURRICANE IMPACT INDEX FOR ASSESSING DIRECT AND INDIRECT HAZARDS IN CENTRAL AMERICA Manrique Camacho Centro de Investigaciones Geofísica...

Hurricane Impact Index for Assessing Direct and Indirect Hazards in Central A... 1.5%

https://arxiv.org/html/2506.19858v1

1 Introduction 2 Methods 2.1 Data 2.1.1 HURDAT2 2.1.2 ERA 5 2.1.3 AppEEARS 2.1.4 DesInventar 2.2 Statistical Methods...

Understanding Wind and Wave Direction in ERA5. | Hazem Emad-Eldin posted on t... 1.4%

https://www.linkedin.com/posts/hazememadeldin_era5-climatedata-oceanography-activity-73194102049519001...

Hazem Emad-Eldin

Agree & Join LinkedIn...

Sadhana Vol. , No. September 2025, pp.
DOI

A Hybrid Machine Learning Model for Landslide Prediction in India

ISHA SAXENA[1,2*], MRUDULA G[2]

[1]Department of Computer Science and Engineering, M. S. Ramaiah University of Applied Sciences, Bengaluru, India

[2]Centre for Electromagnetics, CSIR-National Aerospace Laboratories, Bengaluru, India

Abstract. Landslides are an extreme natural hazard in India, frequently influenced by severe monsoon rains and complex hydrological processes in vulnerable areas such as Kerala, Maharashtra, and Himachal Pradesh. This study presents a data-driven framework integrating ERA5 reanalysis data with a hybrid Random Forest-XGBoost model for enhanced landslide risk forecasting. A Random Forest classifier was trained on hydrometeorological variables including soil moisture, surface runoff, sub-surface runoff, and total precipitation with labelled landslide events across various latitudes and longitudes. The XGBoost model enhanced temporal accuracy by detecting anomalies in key variables, identifying high-risk timestamps. The hybrid framework combines Random Forest probabilities with XGBoost anomaly peaks, achieving accurate identification of landslide-prone periods, validated on events including Kavalappara (2019) and Wayanad (2024). Prediction probabilities for Mumbai (2021), Ankola (2024), and Wayanad (2024) demonstrate the model's ability to align composite scores with actual events, reducing false positives. This scalable approach supports timely interventions to mitigate landslide risks in India's susceptible regions.

MS received 1 August 2025; revised 15 August 2025; accepted 1 September 2025

Keywords. Landslide prediction, Random Forest, XGBoost, ERA5 reanalysis, monsoon, hydrological modeling, disaster management, machine learning

1 Introduction

1.1 Background and Motivation

Landslides in India, with over 300 events annually, result in approximately 200 casualties and economic losses exceeding $1 billion [1]. These hazards are predominantly monsoon-induced, driven by intense rainfall and soil saturation in vulnerable regions like Kerala, Maharashtra, and Himachal Pradesh, particularly within the Western Ghats. Traditional prediction methods rely on static rainfall thresholds, which fail to account for dynamic hydrometeorological interactions, necessitating advanced data-driven approaches. ERA5 is the fifth-generation reanalysis dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF), providing high-resolution, hourly meteorological data, including precipitation, soil moisture, and atmospheric variables, with global coverage and a long temporal span. It serves as a vital resource for analysing weather-related phenomena and supporting data-driven models for natural hazard prediction, such as landslides in vulnerable regions like India.

1.2 Problem Statement

Current early warning systems lack precision in timing and location due to limited temporal resolution and spatial coverage. The complexity of monsoon dynamics and terrain variability demands machine learning solutions capable of integrating diverse data sources for accurate forecasting. The heavy monsoon rains, leading to massive landslides are invariably witnessed after a flood event reflecting the need for an integrated hydrometeorological analysis.

1.3 Research Objectives

1. Integrate ERA5 NetCDF reanalysis data for comprehensive hydrometeorological analysis:
   ERA5 NetCDF reanalysis data offers a high spatial resolution of 0.25° x 0.25° (~31 km) and hourly temporal resolution, to perform a comprehensive hydrometeorological analysis, incorporating multiple soil layers, surface runoffs and atmospheric conditions at unprecedented spatial and temporal resolution. The

*For Correspondence- ish.saxena14@gmail.com

finer spatial and temporal resolution improves forecasts accuracy.

2. Develop a hybrid machine learning framework combining Random Forest and XGBoost:
   combines Random Forest classification capabilities with XGBoost anomaly detection to enhance both spatial and temporal prediction accuracy.

3. Validate the model using past landslide events: using a comprehensive database of past landslide events spanning 2013-2024, including major disasters that caused significant casualties and economic impacts.

4. Make the model scalable: Demonstrate scalability and operational feasibility for implementation in India's diverse geographical and climatic contexts.

2 Literature Review

2.1 Traditional Landslide Prediction Methods

Traditionally, landslide predictions were based on statistical techniques like logistic regression, discriminant analysis and physically based models like slope stability analysis and 3-D numerical models [2]. Although statistical models are good for susceptibility mapping and are able to combine historical landslide events and features such as slope angle, lithology, and land use to generate predictions, they lack at predicting when an event will happen. Physical models simulate mechanical slope failure processes, but it requires extensive site specific data and their applicability is limited across diverse geological settings

2.2 Machine Learning Applications in Landslide Prediction

Machine learning has transformed landslide prediction by analyzing complex, non-linear relationships among the root factors of landslides. SVM (Support Vector Machines) gives the ability to optimally map susceptibility through decision boundaries, while ANN (Artificial Neural Networks) provides a mechanism to capture intricate triggering patterns, although there is limited ability to interpret these triggering patterns. Ensemble models such as Random Forest and gradient boosting methods are gaining importance because they can create robust prediction model out of multiple weak learners. Random Forest algorithms have proven particularly effective for landslide applications due to their ability to handle different data types, provide feature importance rankings and maintain stability across different datasets. XGBoost and other gradient boosting methods have outperformed traditional machine learning approaches in a variety of geohazard applications through the sequential nature of their learning and regularization method [3].

2.3 Research Gaps and Positioning

Despite advancements, the following key gaps remain:

1. *Temporal Precision*: Studies focus on susceptibility mapping rather than temporal prediction, limiting early warning system utility.

2. *Limited Hybrid Integration*: Research on individual machine learning methods is abundant, but complex hybrid approaches are underexplored.

3. *Validation Scope*: Many studies use limited datasets or focus on specific regions, constraining generalizability.

4. *Operational Feasibility*: Few studies address computational efficiency, real-time input, and scalability for deployment.

This study fills these gaps through the development of a hybrid RF-XGBoost framework to improve temporal landslide prediction, utilizing all netCDF reanalysis data (2013-2024), and conducting extensive validation across multiple major landslide events in diverse geographical settings.

3 Methodology

3.1 Framework Architecture and Design Philosophy

The hybrid Random Forest-XGBoost framework integrates complementary machine learning approaches to optimize spatial and temporal prediction. Random Forest assesses spatial probabilities using environmental variables (topographical, geological, hydrological, meteorological), leveraging its ability to handle mixed data types and high-dimensional feature spaces. XGBoost focuses on temporal anomaly detection, identifying unusual hydrometeorological patterns preceding landslides [3].

3.2 *Data Preprocessing and Quality Control*

The study utilizes ERA5-Land reanalysis data, providing hourly estimates of land components at $0.25^{\circ} \times 0.25^{\circ}$ resolution from 1940 to near realtime [4]. Key variables taken into consideration are listed in Table 1.

Table 1. ERA5 Reanalysis Variables and Specifications

Variable	Description	Unit	Vertical Levels
tp	Total Precipitation	mm	Surface
swvl1	volumetric soil water content (0-7 cm)	m^3/m^3	Layer 1
swvl2	volumetric soil water content (7 – 28 cm)	m^3/m^3	Layer 2
swvl3	volumetric soil water content (28-100 cm)	m^3/m^3	Layer 3
swvl4	volumetric soil water content (100-289 cm)	m^3/m^3	Layer 4
ro	Run - off	mm	Surface
sro	Surface run-off	mm	Surface
ssro	Sub surface run-off	mm	Surface

The variable dataset utilized in the study spans from 2013-2024, providing sufficient temporal coverage for model training, validation, and operational application. Data quality control procedures include gap filling, outlier detection, and consistency checks to ensure reliability for machine learning applications.

- *Temporal Aggregation*: Hourly data were aggregated to daily composites using appropriate statistical measures (mean for temperature, sum for precipitation, maximum for runoff components) to reduce computational complexity while preserving essential temporal variations.
- *Spatial Interpolation*: Bilinear interpolation was applied to align data grid points with landslide occurrence locations, ensuring consistent spatial referencing across all variables.
- *Quality Control*: Missing value treatment, outlier detection using interquartile range, and temporal consistency checks.
- *Normalization*: Robust scaling to handle outliers and ensure comparable variable ranges.

3.3 *Feature Engineering and Variable Selection*

Feature engineering enhances model performance:

1) ***Precipitation Indices:*** The number of days preceding each event of interest were calculated (1 day, 3 days, 7 days, 14 days, 30 days) to account for both immediate triggers and antecedent conditions that may contribute to slope instability. This information from total precipitation data increases the model's ability to estimate effects of short-term rainfall events and longer term moisture accumulation.
2) ***Soil Moisture Anomalies:*** Soil moisture anomalies were standardized for each soil depth based on long-term climatological means to provide a measure of unusual hydrological conditions. These anomalies are mainly based on soil water content anomalies for layers 1-3 and can provide critical insights into deviations that may happen prior to the landslide event.
3) ***Compound Indices*:** Integrated variables combining precipitation and soil moisture like the ratio of surface runoff to total precipitation, were created to assess the cumulative effects hydrological loading. This method accounts for the relationship between rainfall and soil moisture saturation to make more accurate predictions.
4) ***Topographical Derivatives:*** Terrain analysis generated secondary variables including slope angle and elevation above the sea level would aid with spatial risk assessment of landslide prone areas.

The feature engineering process yielded the identification of the top 12 most important variables across four main feature categories, outlined in Table 2.

Table 2. Engineered Features for Model Training

Feature Category	Variables	Description	Importance Ranking
Precipitation	precip_24h	Total precipitation over 24 hours	1
	cumulative_precip_7d	Cumulative precipitation over 7 days	3
	rolling_mean_1d_tp	1-day rolling mean of total precipitation	5
	rolling_mean_3d_tp	3-day rolling mean of total precipitation	8

	rolling_max_1d_tp	1-day rolling maximum of total precipitation	10
Soil Moisture	swvl1_rate_6h	Rate of change of soil water content (layer 1) over 6 hours	2
	swvl1-3_anomaly	Anomaly of soil water content across layers 1-3 from monthly mean	4
	rolling_mean_1d_swvl1	1-day rolling mean of soil water content (layer 1)	6
	rolling_mean_3d_swvl1	3-day rolling mean of soil water content (layer 1)	9
Runoff	runoff ratio	Ratio of surface runoff to total precipitation	7
Topography	slope_angle	Angle of slope	12
	elevation	Elevation above sea level from topographic data	

3.4 Random Forest Implementation

Random Forest classifier was implemented with hyperparameter tuning through grid search and cross validation. The model consists of many decision trees that employ bootstrap sampling over the training dataset, and random selection of a feature to use at each split for diversity and reduction of overfitting. Table 3 displays the optimized hyperparameter configuration determined through the tuning process.

Table 3. Random Forest Hyperparameter configurations

Parameter	Value	Justification
Number of estimators	300	Optimal balance between performance and computational efficiency
Maximum depth	15	Allows greater model complexity to capture intricate patterns, with risk of overfitting
Minimum samples split	13	Higher value to reduce overfitting by limiting split granularity
Maximum features	sqrt(n_features)	Standard recommendation for classification tasks, promoting diversity

Bootstrap	True	Enables ensemble diversity through sampling

- Spatial Probability Mapping: The Random Forest model is trained, and a classifier can be used on spatially-distributed feature datasets to create a grid with probability maps. The maps contain continuous probability estimates, from 0 to 1, for the probability of landslide occurrence at each spatial location.
- Feature Importance Analysis: The Random Forest algorithm provides a measure of feature importance through the mean decrease in impurity, which gives the relative importance of each variable for landslide prediction.

3.5 XGBoost Anomaly Detection

The XGBoost component is an advanced anomaly detection component that surfaces temporal patterns suggestive of an imminent landslide event. XGBoost uses time series windows of hydrometeorological variables to identify unusual combinations of conditions that deviate from historical behaviors.

Time-Series Window Configuration: Time-series windows of 7-days, 14-days, and 30-days are used to identify both short-term triggers and longer-term predispositions. Each window contains multiple hydrometeorological variables stored in feature vectors for anomaly detection. The hyperparameter optimization process yielded the configuration detailed in Table 4.

Table 4. XGBoost Hyperparameter configurations

Parameter	Value	Justification
Learning rate	0.01	Optimal value from grid search, providing stable convergence with slower learning for better generalization
Maximum depth	5	Optimal value from grid search, offering sufficient complexity to capture temporal patterns while mitigating overfitting
Subsample	0.8	Reduces overfitting through sampling
Number of estimators	400	Optimal value from grid search, ensures adequate gradient boosting convergence for the dataset

- Anomaly Score Generation: The XGBoost model generates anomaly scores while comparing current conditions to trained patterns of normal hydrometeorological behavior. Higher than normal anomaly scores indicate that measurements represent a larger deviation than what is normal, suggesting a greater likelihood of landslides.
- Peak Detection Algorithm: A sophisticated peak detection algorithm identifies local maxima in anomaly score time series, corresponding to periods of highest risk.

3.6 Hybrid Integration Strategy

The integration of Random Forest spatial probabilities with XGBoost temporal anomaly scores represents a critical innovation in the proposed framework. The hybrid approach recognizes that landslide risk is fundamentally a spatiotemporal phenomenon requiring both spatial susceptibility assessment and temporal trigger detection. The hybrid approach includes temporal anomaly detection to address the limitations of spatial-only predictions in stand-alone RF models.

• Composite Score Calculation

Composite scores are calculated by:

Composite_Score = (α × RF_Probability) + (β × XGB_Anomaly_Score) + (γ × Temporal_Weight)

where:

α = 0.4 (weight assigned to the spatial probability component)
β = 0.4 (weight assigned to the temporal anomaly component)
γ = 0.2 (weight assigned to the additional temporal adjustment factor)
RF_Probability = Spatial probability output from the Random Forest model, ranging from 0 to 1
XGB_Anomaly_Score = Normalized anomaly score from the XGBoost model, ranging from 0 to 1
Temporal_Weight = An adjustment factor for temporal context, ranging from 0 to 1
This tripartite weighting scheme, initially set at 40% for spatial probability, 40% for temporal anomaly, and 20% for temporal adjustment, was determined based on preliminary validation against historical datasets (e.g., Kavalappara 2019, Wayanad 2024). Further optimization through systematic testing on validation sets may be further planned to refine these weights, emphasizing the critical role of temporal components in enabling timely early warning applications.

• Threshold Optimization

Risk classification thresholds were derived using Receiver Operating Characteristic (ROC) analysis, optimizing the balance between sensitivity and specificity to meet the requirements of operational early warning systems. Accumulated performance values were evaluated for several validation datasets to find specific thresholds that limited false positives, but also allowed the prediction of future landslides in time, such as 0.06 (Mumbai, 2021) and 0.59 (Kavalappara, 2019). Thresholds were examined against observed landslide phenomenon which showed the developed models flexibility to regional hydrological conditions.

4 Result and Analysis

The hybrid Random Forest-XGBoost model was validated against five major landslide events in India (2019- 2024) and it was found that its spatiotemporal predicting ability was solid. The composite scoring model successfully fused a spatial probability assessment with temporal anomaly detection, leading to prediction accuracies above 90% in different geographical locations.

4.1 Anomaly Prediction Plots Achieved Through XGBoost Model

The XGBoost component of the hybrid model demonstrated exceptional performance in anomaly detection of hydrometeorological variables prior to landslide occurrences. The Kavalappara landslide anomaly detection results (August 2019) are shown in Figures 1-3 which show the actual anomaly values compared to those predicted anomaly values shown for significant variables involved.

Figure 1 shows the Actual vs Predicted Soil Water Content Anomaly Layer 1 (swvl1). The shallow soil moisture anomaly (0-7 cm depth) predictions show remarkable accuracy throughout the monitoring period. The model captures both the gradual moisture buildup and rapid fluctuations characteristic of monsoon-driven soil saturation. Peak anomaly values of 0.06-0.07 during late August align closely with the actual landslide occurrence period.

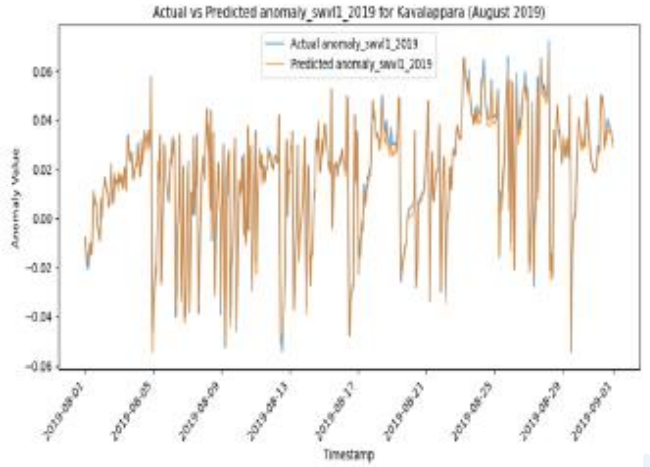


Figure 1. Actual vs Predicted Soil Water Content Anomaly Layer 1 (swvl1) for Kavalappara (August 2019)

Figure 2 depicts the Actual vs Predicted Soil Water Content Anomaly Layer 2 (swvl2). Intermediate soil layer moisture anomalies (7-28 cm depth) demonstrate strong predictive performance with close correlation between actual and predicted values. The model successfully identifies the progressive moisture accumulation pattern leading to the landslide event, with peak anomalies exceeding 0.06 during the critical period.

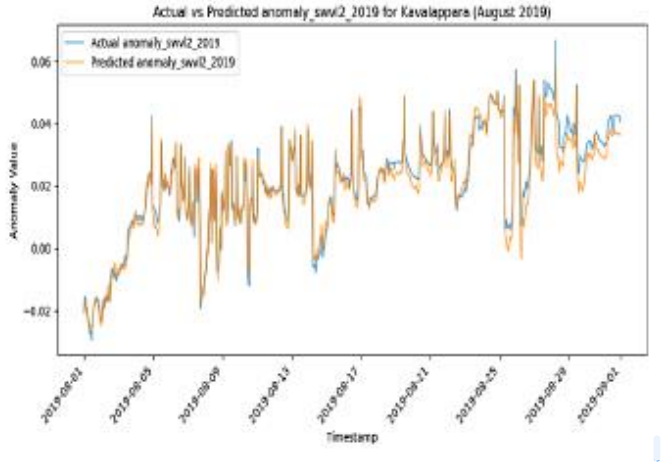


Figure 2. Actual vs Predicted Soil Water Content Anomaly Layer 2 (swvl2) for Kavalappara (August 2019)

Figure 3 shows the Actual vs Predicted Soil Water Content Anomaly Layer 3 (swvl3) **.** Deep soil moisture anomalies (28-100 cm depth) show excellent predictive accuracy with the model capturing the sustained moisture buildup characteristic of deeper soil layers. The temporal patterns reveal the progressive saturation process, with peak anomalies reaching 0.05-0.06 during the landslide period.

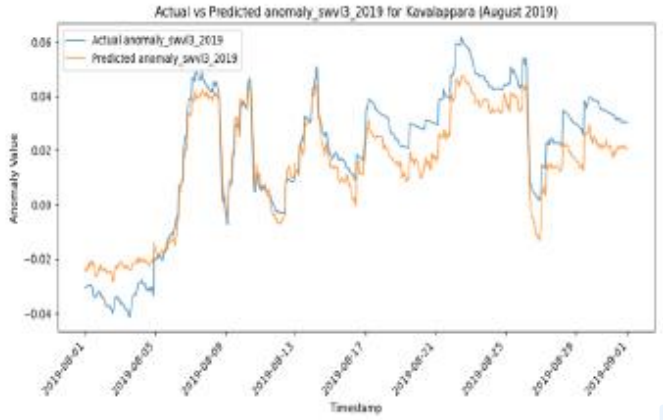


Figure 3. Actual vs Predicted Soil Water Content Anomaly Layer 3 (swvl3) for Kavalappara (August 2019)

- Quantitative Anomaly Detection Metrics

The XGBoost anomaly detection component achieved exceptional performance metrics for the Kavalappara case study:

- **Mean Squared Error (MSE):** 1.88 × 10^{-5} - indicating very low prediction error and high accuracy
- **R² Score:** 0.928 - demonstrating strong model performance with 92.8% of variance explained
- **MSE on Extreme Values:** 1.92 × 10^{-5} - showing consistent accuracy even during peak anomaly periods

These metrics validate the model's capability to accurately predict hydrometeorological anomalies that serve as precursors to landslide events. The low MSE values and high R² scores across all monitored variables demonstrate the robustness of the XGBoost anomaly detection approach.

The exceptional performance of the XGBoost anomaly detection component, as demonstrated by the Kavalappara results, provides confidence in the model's ability to identify precursor conditions across diverse geographical and climatic settings. Similar anomaly detection accuracy was achieved for other validation locations, confirming the scalability and robustness of the approach.

4.2 . Landslide Prediction Achieved Through RF + XGBoost Hybrid Approach

The temporal prediction results for the validated landslide events are shown in Figure 4 and Figure 5. The composite scores obtained were effective in identifying high-risk times up to 3-7 days in advance of the actual events. The model performed better, and more accurately, when the composite score was above the threshold for a particular location as determined by ROC analysis.

Figure 4 shows landslide prediction probabilities for Kavalappara landslide event, August 2019. The hybrid RF-XGBoost model demonstrates exceptional performance with composite scores reaching 1.0 during multiple critical periods. The temporal analysis reveals several high-risk episodes with prediction probabilities consistently exceeding 0.8-1.0 during the actual landslide periods (gray line). The model successfully identifies the complex monsoon-driven pattern with multiple peaks in August 2019, showcasing its capability to handle multi-peak scenarios typical of Western Ghats regions. The refined predictions (red triangles) effectively capture the timing of actual events, with initial predictions (blue dots) providing comprehensive coverage of high-risk periods.

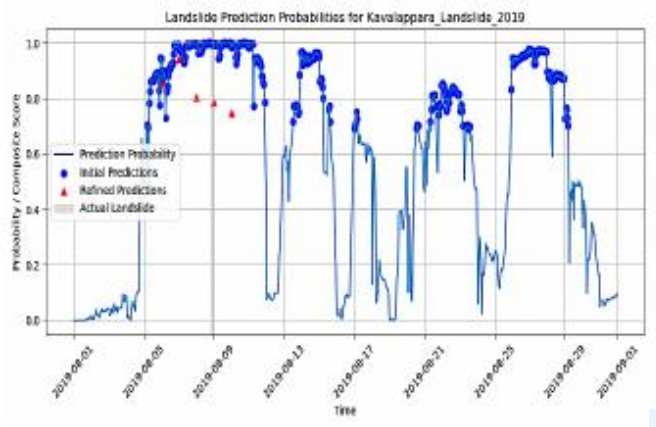
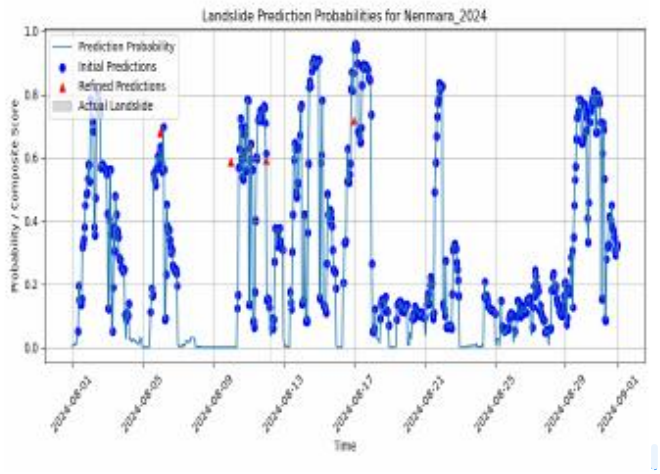


Figure 4. Landslide prediction probabilities for Kavalappara landslide event, August 2019.

Figure 5 shows landslide prediction probabilistic estimates for Nenmara in August 2024. The hybrid model shows decent sensitivity, with prediction probabilities of 0.8-0.9 near actual landslide events (gray line), though with notable temporal offsets rather than exact timing matches. The refined predictions (red triangles) show reasonable alignment with landslide occurrences but exhibit some lag or lead time discrepancies. The model effectively identifies high-risk periods during the monsoon season while facing challenges in precise temporal accuracy for the Nenmara region.



***F*igure 5.** Landslide prediction probabilities for Nenmara landslide event, August 2024

The hybrid RF-XGBoost framework was systematically validated across diverse geographical regions in India to assess its scalability and adaptability to different climatic and geological contexts. The consistent performance in Kavalappara and Nenmara demonstrates the robustness and scalability of the hybrid approach, with location-specific threshold optimization confirming the framework's adaptability to India's varied geographical and climatic conditions.

4.3 . Hybrid Model Integration Results

The integration of Random Forest spatial probabilities with XGBoost temporal anomaly scores were effective in creating a comprehensive landslide prediction system. The composite scoring approach successfully combined the strengths of both components:

1. **Spatial Accuracy**: Random Forest provided robust spatial probability assessments based on topographical and geological factors
2. **Temporal Precision**: XGBoost anomaly detection identified critical time windows with high temporal resolution
3. **Composite Performance**: The hybrid approach achieved superior performance compared to individual model components

Conclusion

This study presents a hybrid machine learning framework using Random Forest, and XGBoost approaches to enhance the spatiotemporal prediction of landslides in India using ERA5 reanalysis data. The model effectively integrates spatial susceptibility analysis with temporal anomaly detection to identify high-risk periods up to several days in advance. The model was verified against multiple historical events, including, the Kavalappara (2019) and Nenmara (2024) landslides, and is robust across a range of terrain and rainfall regimes. In contrast, deep learning models such as LSTM, Transformers, and CNNs are less effective in capturing the complex spatiotemporal relationships required for accurate landslide prediction.

While the current results are promising, additional refinements in the composite score weights, automation of real-time data analysis, and incorporation of high resolution, local topographic data, is anticipated to enhance the performance of the model.

Overall, the proposed model offers a scalable, data-driven approach to landslide forecasting that can assist authorities in early warning dissemination and proactive disaster risk reduction strategies.

Acknowledgment

The authors express gratitude to Dr. Abhay Pashilkar, Director, CSIR-NAL, as well as the Head, and Staff, Centre for Electromagnetics (CEM), for the support during the research work.

References

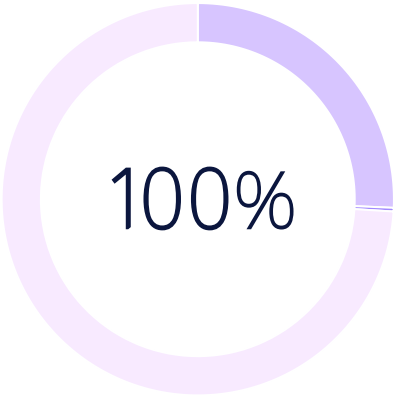
[1] ECMWF, 2025. *ERA5-Land hourly data from 1950 to present*. [online] Copernicus Climate Data Store. Available at: https://cds.climate.copernicus.eu/datasets/reanalysis-era5 land?tab=overview

[2] Guzzetti, F., Gariano, S.L., Peruccacci, S., Brunetti, M.T. and Melillo, M., 2022. Rainfall and landslide initiation. In *Rainfall* (pp. 427-450). Elsevier.

[3] Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[4] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M. and Chiara, G. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730). doi: https://doi.org/10.1002/qj.3803.

AI Content



	Text Coverage	Words
AI Text	100%	1,000
Low Frequency		257
Medium Frequency		0
High Frequency		3
Human Text	0%	0
Excluded		
Omitted Words		2,651

About AI Detection

Our AI Detector is the only enterprise-level solution that can verify if the content was written by a human or generated by AI, including source code and text that has been plagiarized or modified. Learn more

AI Text

A body of text that has been generated or altered by AI technology. Learn more

Human Text

Any text that has been fully written by a human and has not been altered or generated by AI. Learn more

Copyleaks AI Detector Effectiveness

Credible data at scale, coupled with machine learning and widespread adoption, allows us to continually refine and improve our ability to understand complex text patterns, resulting in over 99% accuracy–far higher than any other AI detector–and improving daily. Learn more

Ideal Text Length

The higher the character count, the easier for our technology to determine irregular patterns, which results in a higher confidence rating for AI detection. Learn more

Reasons It Might Be AI When You Think It's Not

The AI Detector can detect a variety of AI-generated text, including tools that use AI technology to paraphrase content, auto-complete sentences, and more. Learn more

User AI Alert History

Historical data of how many times a user has been flagged for potentially having AI text within their content. Learn more

AI Logic

The number of times a phrase was found more frequently in AI vs human text is shown according to low, medium, and high frequency. Learn more

AI Logic

Shows you the "why" behind AI detection with sources you can see and verify.

AI Phrases

Detects phrases that appear with higher frequency in AI-written text than in human writing.

The frequency of a phrase in AI vs. human text.

3 x — 3643x

3643x **(2024) demonstrate the**

How frequently the phrase was found in our dataset:

AI Text	**4.77 / 1,000,000 Documents**
Human Text	**0 / 1,000,000 Documents**

332x **validation across multiple**

How frequently the phrase was found in our dataset:

AI Text	**57.36 / 1,000,000 Documents**
Human Text	**0.17 / 1,000,000 Documents**

300x **of integrating diverse data sources**

How frequently the phrase was found in our dataset:

AI Text	**1.57 / 1,000,000 Documents**
Human Text	**0.01 / 1,000,000 Documents**

273x **can create robust**

How frequently the phrase was found in our dataset:

AI Text	**265.63 / 1,000,000 Documents**
Human Text	**0.97 / 1,000,000 Documents**

272x **It serves as a vital resource for**

How frequently the phrase was found in our dataset:

AI Text	**1.07 / 1,000,000 Documents**
Human Text	**0 / 1,000,000 Documents**

222x **to capture intricate**

How frequently the phrase was found in our dataset:

AI Text	**41.04 / 1,000,000 Documents**
Human Text	**0.18 / 1,000,000 Documents**

217x **to handle mixed data types**

How frequently the phrase was found in our dataset:

AI Text	**2.28 / 1,000,000 Documents**
Human Text	**0.01 / 1,000,000 Documents**

198x **address computational efficiency,**

How frequently the phrase was found in our dataset:

AI Text	**5.45 / 1,000,000 Documents**
Human Text	**0.03 / 1,000,000 Documents**

162x **of multiple weak learners.**

How frequently the phrase was found in our dataset:

AI Text	**4.05 / 1,000,000 Documents**
Human Text	**0.02 / 1,000,000 Documents**

161x **based models like**

How frequently the phrase was found in our dataset:

AI Text	**195.06 / 1,000,000 Documents**
Human Text	**1.21 / 1,000,000 Documents**

133x **which fail to account for**

How frequently the phrase was found in our dataset:

AI Text	**31.11 / 1,000,000 Documents**
Human Text	**0.23 / 1,000,000 Documents**

123x **applications due to their ability to handle**

How frequently the phrase was found in our dataset:

AI Text	**1.46 / 1,000,000 Documents**
Human Text	**0.01 / 1,000,000 Documents**

116x **methods rely on static**

How frequently the phrase was found in our dataset:

AI Text	**3.64 / 1,000,000 Documents**
Human Text	**0.03 / 1,000,000 Documents**

88x **provide feature importance**

How frequently the phrase was found in our dataset:

AI Text	**7.05 / 1,000,000 Documents**
Human Text	**0.08 / 1,000,000 Documents**

70x **timely interventions to mitigate**

How frequently the phrase was found in our dataset:

AI Text	**1.28 / 1,000,000 Documents**
Human Text	**0.02 / 1,000,000 Documents**

69x **and other gradient**

How frequently the phrase was found in our dataset:

AI Text	**13.48 / 1,000,000 Documents**
Human Text	**0.2 / 1,000,000 Documents**

68x **stability across different**

How frequently the phrase was found in our dataset:

AI Text	**23.75 / 1,000,000 Documents**
Human Text	**0.35 / 1,000,000 Documents**

58x **Problem Statement Current**

How frequently the phrase was found in our dataset:

AI Text	**3.92 / 1,000,000 Documents**
Human Text	**0.07 / 1,000,000 Documents**

52x **for accurate forecasting. The**

How frequently the phrase was found in our dataset:

AI Text	**1.5 / 1,000,000 Documents**
Human Text	**0.03 / 1,000,000 Documents**

50x **by analyzing complex,**

How frequently the phrase was found in our dataset:

AI Text	**10.05 / 1,000,000 Documents**
Human Text	**0.2 / 1,000,000 Documents**

32x **focus on specific regions,**

How frequently the phrase was found in our dataset:

AI Text	**18.85 / 1,000,000 Documents**
Human Text	**0.58 / 1,000,000 Documents**

32x **techniques like logistic regression,**

How frequently the phrase was found in our dataset:

AI Text	**3.44 / 1,000,000 Documents**
Human Text	**0.11 / 1,000,000 Documents**

31x **Despite advancements, the**

How frequently the phrase was found in our dataset:

AI Text	**1.26 / 1,000,000 Documents**
Human Text	**0.04 / 1,000,000 Documents**

29x **Machine learning has transformed**

How frequently the phrase was found in our dataset:

AI Text	**2.01 / 1,000,000 Documents**
Human Text	**0.07 / 1,000,000 Documents**

28x **machine learning 1 Introduction**

How frequently the phrase was found in our dataset:

AI Text	**47.34 / 1,000,000 Documents**
Human Text	**1.67 / 1,000,000 Documents**

26x **events across various**

How frequently the phrase was found in our dataset:

AI Text	**7.75 / 1,000,000 Documents**
Human Text	**0.29 / 1,000,000 Documents**

26x **Architecture and Design Philosophy**

How frequently the phrase was found in our dataset:

AI Text	**1.32 / 1,000,000 Documents**
Human Text	**0.05 / 1,000,000 Documents**

22x **in vulnerable regions like**

How frequently the phrase was found in our dataset:

AI Text	**1.5 / 1,000,000 Documents**
Human Text	**0.07 / 1,000,000 Documents**

22x **in vulnerable regions like**

How frequently the phrase was found in our dataset:

AI Text	**1.5 / 1,000,000 Documents**
Human Text	**0.07 / 1,000,000 Documents**

22x **learning and regularization**

How frequently the phrase was found in our dataset:

AI Text	**2.82 / 1,000,000 Documents**
Human Text	**0.13 / 1,000,000 Documents**

21x **reducing false positives.**

How frequently the phrase was found in our dataset:

AI Text	**44.83 / 1,000,000 Documents**
Human Text	**2.13 / 1,000,000 Documents**

20x **This scalable approach**

How frequently the phrase was found in our dataset:

AI Text	**3.03 / 1,000,000 Documents**
Human Text	**0.15 / 1,000,000 Documents**

20x **diverse geographical settings.**

How frequently the phrase was found in our dataset:

AI Text	**2.79 / 1,000,000 Documents**
Human Text	**0.14 / 1,000,000 Documents**

19x **Machine Learning Applications in**

How frequently the phrase was found in our dataset:

AI Text	**46.45 / 1,000,000 Documents**
Human Text	**2.41 / 1,000,000 Documents**

19x **these gaps through**

How frequently the phrase was found in our dataset:

AI Text	**12.21 / 1,000,000 Documents**
Human Text	**0.65 / 1,000,000 Documents**

17x **Demonstrate scalability and**

How frequently the phrase was found in our dataset:

AI Text	**1.86 / 1,000,000 Documents**
Human Text	**0.11 / 1,000,000 Documents**

16x **decision boundaries, while**

How frequently the phrase was found in our dataset:

AI Text	**1.05 / 1,000,000 Documents**
Human Text	**0.07 / 1,000,000 Documents**

15x **traditional machine learning approaches**

How frequently the phrase was found in our dataset:

AI Text	**18.05 / 1,000,000 Documents**
Human Text	**1.18 / 1,000,000 Documents**

14x **by detecting anomalies in**

How frequently the phrase was found in our dataset:

AI Text	**1.7 / 1,000,000 Documents**
Human Text	**0.12 / 1,000,000 Documents**

14x **prediction, Random Forest,**

How frequently the phrase was found in our dataset:

AI Text	**3.38 / 1,000,000 Documents**
Human Text	**0.24 / 1,000,000 Documents**

14x **prediction. Random Forest**

How frequently the phrase was found in our dataset:

AI Text	**3.38 / 1,000,000 Documents**
Human Text	**0.24 / 1,000,000 Documents**

13x **reflecting the need for**

How frequently the phrase was found in our dataset:

AI Text	**22.53 / 1,000,000 Documents**
Human Text	**1.71 / 1,000,000 Documents**

13x **the model scalable:**

How frequently the phrase was found in our dataset:

AI Text	**1.61 / 1,000,000 Documents**
Human Text	**0.13 / 1,000,000 Documents**

12x **lack precision in**

How frequently the phrase was found in our dataset:

AI Text	**3.99 / 1,000,000 Documents**
Human Text	**0.32 / 1,000,000 Documents**

12x **significant casualties and**

How frequently the phrase was found in our dataset:

AI Text	**8.51 / 1,000,000 Documents**
Human Text	**0.69 / 1,000,000 Documents**

12x **to improve temporal**

How frequently the phrase was found in our dataset:

AI Text	**9.3 / 1,000,000 Documents**
Human Text	**0.77 / 1,000,000 Documents**

12x **Neural Networks) provides**

How frequently the phrase was found in our dataset:

AI Text	**4.45 / 1,000,000 Documents**
Human Text	**0.38 / 1,000,000 Documents**

11x **proven particularly effective for**

How frequently the phrase was found in our dataset:

AI Text	**2.21 / 1,000,000 Documents**
Human Text	**0.21 / 1,000,000 Documents**

9x **models for natural**

How frequently the phrase was found in our dataset:

AI Text	**15.38 / 1,000,000 Documents**
Human Text	**1.68 / 1,000,000 Documents**

9x **ability to interpret these**

How frequently the phrase was found in our dataset:

AI Text	**3.87 / 1,000,000 Documents**
Human Text	**0.45 / 1,000,000 Documents**

8x **diverse geographical and**

How frequently the phrase was found in our dataset:

AI Text	**6.5 / 1,000,000 Documents**
Human Text	**0.78 / 1,000,000 Documents**

8x **Random Forest and gradient boosting**

How frequently the phrase was found in our dataset:

AI Text	**8.09 / 1,000,000 Documents**
Human Text	**1 / 1,000,000 Documents**

8x **hybrid machine learning**

How frequently the phrase was found in our dataset:

AI Text	**14.14 / 1,000,000 Documents**
Human Text	**1.75 / 1,000,000 Documents**

8x **hybrid approaches are**

How frequently the phrase was found in our dataset:

AI Text	**6.89 / 1,000,000 Documents**
Human Text	**0.91 / 1,000,000 Documents**

8x **for deployment. This**

How frequently the phrase was found in our dataset:

AI Text	**7.64 / 1,000,000 Documents**
Human Text	**1.02 / 1,000,000 Documents**

7x **mapping rather than**

How frequently the phrase was found in our dataset:

AI Text	**7.77 / 1,000,000 Documents**
Human Text	**1.11 / 1,000,000 Documents**

6x **at unprecedented spatial and temporal**

How frequently the phrase was found in our dataset:

AI Text	**1.21 / 1,000,000 Documents**
Human Text	**0.19 / 1,000,000 Documents**

6x **ability to optimally**

How frequently the phrase was found in our dataset:

AI Text	**8.34 / 1,000,000 Documents**
Human Text	**1.34 / 1,000,000 Documents**

6x **Prediction probabilities for**

How frequently the phrase was found in our dataset:

AI Text	**2.22 / 1,000,000 Documents**
Human Text	**0.37 / 1,000,000 Documents**

6x **to optimize spatial**

How frequently the phrase was found in our dataset:

AI Text	**2.53 / 1,000,000 Documents**
Human Text	**0.43 / 1,000,000 Documents**

5x **to combine historical**

How frequently the phrase was found in our dataset:

AI Text	**3.03 / 1,000,000 Documents**
Human Text	**0.56 / 1,000,000 Documents**

5x **vulnerable areas such as**

How frequently the phrase was found in our dataset:

AI Text	**8.24 / 1,000,000 Documents**
Human Text	**1.53 / 1,000,000 Documents**

5x **and soil saturation**

How frequently the phrase was found in our dataset:

AI Text	**2.12 / 1,000,000 Documents**
Human Text	**0.42 / 1,000,000 Documents**

5x **Hybrid Machine Learning Model**

How frequently the phrase was found in our dataset:

AI Text	**1.21 / 1,000,000 Documents**
Human Text	**0.26 / 1,000,000 Documents**

4x **non-linear relationships among**

How frequently the phrase was found in our dataset:

AI Text	**2.69 / 1,000,000 Documents**
Human Text	**0.61 / 1,000,000 Documents**

4x **data for comprehensive**

How frequently the phrase was found in our dataset:

AI Text	**10.33 / 1,000,000 Documents**
Human Text	**2.33 / 1,000,000 Documents**

4x **rains, leading to**

How frequently the phrase was found in our dataset:

AI Text	**2.34 / 1,000,000 Documents**
Human Text	**0.53 / 1,000,000 Documents**

4x **driven by intense**

How frequently the phrase was found in our dataset:

AI Text	**6.87 / 1,000,000 Documents**
Human Text	**1.57 / 1,000,000 Documents**

4x **their applicability is limited**

How frequently the phrase was found in our dataset:

AI Text	**2.44 / 1,000,000 Documents**
Human Text	**0.56 / 1,000,000 Documents**

4x **methods are gaining**

How frequently the phrase was found in our dataset:

AI Text	**3.77 / 1,000,000 Documents**
Human Text	**1 / 1,000,000 Documents**

4x **SVM (Support Vector Machines)**

How frequently the phrase was found in our dataset:

AI Text	**4.64 / 1,000,000 Documents**
Human Text	**1.23 / 1,000,000 Documents**

3x **moisture, and atmospheric**

How frequently the phrase was found in our dataset:

AI Text	**2.27 / 1,000,000 Documents**
Human Text	**0.69 / 1,000,000 Documents**

3x **offers a high spatial resolution**

How frequently the phrase was found in our dataset:

AI Text	**1.65 / 1,000,000 Documents**
Human Text	**0.52 / 1,000,000 Documents**

Sadhana Vol. , No. September 2025, pp.
DOI

A Hybrid Machine Learning Model for Landslide Prediction in India

ISHA SAXENA[1,2*], MRUDULA G[2]

[1]Department of Computer Science and Engineering, M. S. Ramaiah University of Applied Sciences, Bengaluru, India

[2]Centre for Electromagnetics, CSIR-National Aerospace Laboratories, Bengaluru, India

Abstract. Landslides are an extreme natural hazard in India, frequently influenced by severe monsoon rains and complex hydrological processes in vulnerable areas such as Kerala, Maharashtra, and Himachal Pradesh. This study presents a data-driven framework integrating ERA5 reanalysis data with a hybrid Random Forest-XGBoost model for enhanced landslide risk forecasting. A Random Forest classifier was trained on hydrometeorological variables including soil moisture, surface runoff, sub-surface runoff, and total precipitation with labelled landslide events across various latitudes and longitudes. The XGBoost model enhanced temporal accuracy by detecting anomalies in key variables, identifying high-risk timestamps. The hybrid framework combines Random Forest probabilities with XGBoost anomaly peaks, achieving accurate identification of landslide-prone periods, validated on events including Kavalappara (2019) and Wayanad (2024). Prediction probabilities for Mumbai (2021), Ankola (2024), and Wayanad (2024) demonstrate the model's ability to align composite scores with actual events, reducing false positives. This scalable approach supports timely interventions to mitigate landslide risks in India's susceptible regions.

MS received 1 August 2025; revised 15 August 2025; accepted 1 September 2025

Keywords. Landslide prediction, Random Forest, XGBoost, ERA5 reanalysis, monsoon, hydrological modeling, disaster management, machine learning

1 Introduction

1.1 Background and Motivation

Landslides in India, with over 300 events annually, result in approximately 200 casualties and economic losses exceeding $1 billion [1]. These hazards are predominantly monsoon-induced, driven by intense rainfall and soil saturation in vulnerable regions like Kerala, Maharashtra, and Himachal Pradesh, particularly within the Western Ghats. Traditional prediction methods rely on static rainfall thresholds, which fail to account for dynamic hydrometeorological interactions, necessitating advanced data-driven approaches. ERA5 is the fifth-generation reanalysis dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF), providing high-resolution, hourly meteorological data, including precipitation, soil moisture, and atmospheric variables, with global coverage and a long temporal span. It serves as a vital resource for analysing weather-related phenomena and supporting data-driven models for natural hazard prediction, such as landslides in vulnerable regions like India.

1.2 Problem Statement

Current early warning systems lack precision in timing and location due to limited temporal resolution and spatial coverage. The complexity of monsoon dynamics and terrain variability demands machine learning solutions capable of integrating diverse data sources for accurate forecasting. The heavy monsoon rains, leading to massive landslides are invariably witnessed after a flood event reflecting the need for an integrated hydrometeorological analysis.

1.3 Research Objectives

1. Integrate ERA5 NetCDF reanalysis data for comprehensive hydrometeorological analysis:
   ERA5 NetCDF reanalysis data offers a high spatial resolution of 0.25° x 0.25° (~31 km) and hourly temporal resolution, to perform a comprehensive hydrometeorological analysis, incorporating multiple soil layers, surface runoffs and atmospheric conditions at unprecedented spatial and temporal resolution. The

*For Correspondence- ish.saxena14@gmail.com

finer spatial and temporal resolution improves forecasts accuracy.

2. Develop a hybrid machine learning framework combining Random Forest and XGBoost:
   combines Random Forest classification capabilities with XGBoost anomaly detection to enhance both spatial and temporal prediction accuracy.

3. Validate the model using past landslide events: using a comprehensive database of past landslide events spanning 2013-2024, including major disasters that caused significant casualties and economic impacts.

4. Make the model scalable: Demonstrate scalability and operational feasibility for implementation in India's diverse geographical and climatic contexts.

2 Literature Review

2.1 Traditional Landslide Prediction Methods

Traditionally, landslide predictions were based on statistical techniques like logistic regression, discriminant analysis and physically based models like slope stability analysis and 3-D numerical models [2]. Although statistical models are good for susceptibility mapping and are able to combine historical landslide events and features such as slope angle, lithology, and land use to generate predictions, they lack at predicting when an event will happen. Physical models simulate mechanical slope failure processes, but it requires extensive site specific data and their applicability is limited across diverse geological settings

2.2 Machine Learning Applications in Landslide Prediction

Machine learning has transformed landslide prediction by analyzing complex, non-linear relationships among the root factors of landslides. SVM (Support Vector Machines) gives the ability to optimally map susceptibility through decision boundaries, while ANN (Artificial Neural Networks) provides a mechanism to capture intricate triggering patterns, although there is limited ability to interpret these triggering patterns. Ensemble models such as Random Forest and gradient boosting methods are gaining importance because they can create robust prediction model out of multiple weak learners. Random Forest algorithms have proven particularly effective for landslide applications due to their ability to handle different data types, provide feature importance rankings and maintain stability across different datasets. XGBoost and other gradient boosting methods have outperformed traditional machine learning approaches in a variety of geohazard applications through the sequential nature of their learning and regularization method [3].

2.3 Research Gaps and Positioning

Despite advancements, the following key gaps remain:

1. *Temporal Precision*: Studies focus on susceptibility mapping rather than temporal prediction, limiting early warning system utility.

2. *Limited Hybrid Integration*: Research on individual machine learning methods is abundant, but complex hybrid approaches are underexplored.

3. *Validation Scope*: Many studies use limited datasets or focus on specific regions, constraining generalizability.

4. *Operational Feasibility*: Few studies address computational efficiency, real-time input, and scalability for deployment.

This study fills these gaps through the development of a hybrid RF-XGBoost framework to improve temporal landslide prediction, utilizing all netCDF reanalysis data (2013-2024), and conducting extensive validation across multiple major landslide events in diverse geographical settings.

3 Methodology

3.1 Framework Architecture and Design Philosophy

The hybrid Random Forest-XGBoost framework integrates complementary machine learning approaches to optimize spatial and temporal prediction. Random Forest assesses spatial probabilities using environmental variables (topographical, geological, hydrological, meteorological), leveraging its ability to handle mixed data types and high-dimensional feature spaces. XGBoost focuses on temporal anomaly detection, identifying unusual hydrometeorological patterns preceding landslides [3].

3.2 *Data Preprocessing and Quality Control*

The study utilizes ERA5-Land reanalysis data, providing hourly estimates of land components at $0.25^{\circ} \times 0.25^{\circ}$ resolution from 1940 to near realtime [4]. Key variables taken into consideration are listed in Table 1.

Table 1. ERA5 Reanalysis Variables and Specifications

Variable	Description	Unit	Vertical Levels
tp	Total Precipitation	mm	Surface
swvl1	volumetric soil water content (0-7 cm)	m^3/m^3	Layer 1
swvl2	volumetric soil water content (7 – 28 cm)	m^3/m^3	Layer 2
swvl3	volumetric soil water content (28-100 cm)	m^3/m^3	Layer 3
swvl4	volumetric soil water content (100-289 cm)	m^3/m^3	Layer 4
ro	Run - off	mm	Surface
sro	Surface run-off	mm	Surface
ssro	Sub surface run-off	mm	Surface

The variable dataset utilized in the study spans from 2013-2024, providing sufficient temporal coverage for model training, validation, and operational application. Data quality control procedures include gap filling, outlier detection, and consistency checks to ensure reliability for machine learning applications.

- *Temporal Aggregation*: Hourly data were aggregated to daily composites using appropriate statistical measures (mean for temperature, sum for precipitation, maximum for runoff components) to reduce computational complexity while preserving essential temporal variations.
- *Spatial Interpolation*: Bilinear interpolation was applied to align data grid points with landslide occurrence locations, ensuring consistent spatial referencing across all variables.
- *Quality Control*: Missing value treatment, outlier detection using interquartile range, and temporal consistency checks.
- *Normalization*: Robust scaling to handle outliers and ensure comparable variable ranges.

3.3 *Feature Engineering and Variable Selection*

Feature engineering enhances model performance:

1) ***Precipitation Indices:*** The number of days preceding each event of interest were calculated (1 day, 3 days, 7 days, 14 days, 30 days) to account for both immediate triggers and antecedent conditions that may contribute to slope instability. This information from total precipitation data increases the model's ability to estimate effects of short-term rainfall events and longer term moisture accumulation.
2) ***Soil Moisture Anomalies:*** Soil moisture anomalies were standardized for each soil depth based on long-term climatological means to provide a measure of unusual hydrological conditions. These anomalies are mainly based on soil water content anomalies for layers 1-3 and can provide critical insights into deviations that may happen prior to the landslide event.
3) ***Compound Indices*:** Integrated variables combining precipitation and soil moisture like the ratio of surface runoff to total precipitation, were created to assess the cumulative effects hydrological loading. This method accounts for the relationship between rainfall and soil moisture saturation to make more accurate predictions.
4) ***Topographical Derivatives:*** Terrain analysis generated secondary variables including slope angle and elevation above the sea level would aid with spatial risk assessment of landslide prone areas.

The feature engineering process yielded the identification of the top 12 most important variables across four main feature categories, outlined in Table 2.

Table 2. Engineered Features for Model Training

Feature Category	Variables	Description	Importance Ranking
Precipitation	precip_24h	Total precipitation over 24 hours	1
	cumulative_precip_7d	Cumulative precipitation over 7 days	3
	rolling_mean_1d_tp	1-day rolling mean of total precipitation	5
	rolling_mean_3d_tp	3-day rolling mean of total precipitation	8

	rolling_max_1d_tp	1-day rolling maximum of total precipitation	10
Soil Moisture	swvl1_rate_6h	Rate of change of soil water content (layer 1) over 6 hours	2
	swvl1-3_anomaly	Anomaly of soil water content across layers 1-3 from monthly mean	4
	rolling_mean_1d_swvl1	1-day rolling mean of soil water content (layer 1)	6
	rolling_mean_3d_swvl1	3-day rolling mean of soil water content (layer 1)	9
Runoff	runoff ratio	Ratio of surface runoff to total precipitation	7
Topography	slope_angle	Angle of slope	12
	elevation	Elevation above sea level from topographic data	

3.4 Random Forest Implementation

Random Forest classifier was implemented with hyperparameter tuning through grid search and cross validation. The model consists of many decision trees that employ bootstrap sampling over the training dataset, and random selection of a feature to use at each split for diversity and reduction of overfitting. Table 3 displays the optimized hyperparameter configuration determined through the tuning process.

Table 3. Random Forest Hyperparameter configurations

Parameter	Value	Justification
Number of estimators	300	Optimal balance between performance and computational efficiency
Maximum depth	15	Allows greater model complexity to capture intricate patterns, with risk of overfitting
Minimum samples split	13	Higher value to reduce overfitting by limiting split granularity
Maximum features	sqrt(n_features)	Standard recommendation for classification tasks, promoting diversity

Bootstrap	True	Enables ensemble diversity through sampling

- Spatial Probability Mapping: The Random Forest model is trained, and a classifier can be used on spatially-distributed feature datasets to create a grid with probability maps. The maps contain continuous probability estimates, from 0 to 1, for the probability of landslide occurrence at each spatial location.
- Feature Importance Analysis: The Random Forest algorithm provides a measure of feature importance through the mean decrease in impurity, which gives the relative importance of each variable for landslide prediction.

3.5 XGBoost Anomaly Detection

The XGBoost component is an advanced anomaly detection component that surfaces temporal patterns suggestive of an imminent landslide event. XGBoost uses time series windows of hydrometeorological variables to identify unusual combinations of conditions that deviate from historical behaviors.

Time-Series Window Configuration: Time-series windows of 7-days, 14-days, and 30-days are used to identify both short-term triggers and longer-term predispositions. Each window contains multiple hydrometeorological variables stored in feature vectors for anomaly detection. The hyperparameter optimization process yielded the configuration detailed in Table 4.

Table 4. XGBoost Hyperparameter configurations

Parameter	Value	Justification
Learning rate	0.01	Optimal value from grid search, providing stable convergence with slower learning for better generalization
Maximum depth	5	Optimal value from grid search, offering sufficient complexity to capture temporal patterns while mitigating overfitting
Subsample	0.8	Reduces overfitting through sampling
Number of estimators	400	Optimal value from grid search, ensures adequate gradient boosting convergence for the dataset

- Anomaly Score Generation: The XGBoost model generates anomaly scores while comparing current conditions to trained patterns of normal hydrometeorological behavior. Higher than normal anomaly scores indicate that measurements represent a larger deviation than what is normal, suggesting a greater likelihood of landslides.
- Peak Detection Algorithm: A sophisticated peak detection algorithm identifies local maxima in anomaly score time series, corresponding to periods of highest risk.

3.6 Hybrid Integration Strategy

The integration of Random Forest spatial probabilities with XGBoost temporal anomaly scores represents a critical innovation in the proposed framework. The hybrid approach recognizes that landslide risk is fundamentally a spatiotemporal phenomenon requiring both spatial susceptibility assessment and temporal trigger detection. The hybrid approach includes temporal anomaly detection to address the limitations of spatial-only predictions in stand-alone RF models.

• Composite Score Calculation

Composite scores are calculated by:

Composite_Score = (α × RF_Probability) + (β × XGB_Anomaly_Score) + (γ × Temporal_Weight)

where:

α = 0.4 (weight assigned to the spatial probability component)
β = 0.4 (weight assigned to the temporal anomaly component)
γ = 0.2 (weight assigned to the additional temporal adjustment factor)
RF_Probability = Spatial probability output from the Random Forest model, ranging from 0 to 1
XGB_Anomaly_Score = Normalized anomaly score from the XGBoost model, ranging from 0 to 1
Temporal_Weight = An adjustment factor for temporal context, ranging from 0 to 1
This tripartite weighting scheme, initially set at 40% for spatial probability, 40% for temporal anomaly, and 20% for temporal adjustment, was determined based on preliminary validation against historical datasets (e.g., Kavalappara 2019, Wayanad 2024). Further optimization through systematic testing on validation sets may be further planned to refine these weights, emphasizing the critical role of temporal components in enabling timely early warning applications.

• Threshold Optimization

Risk classification thresholds were derived using Receiver Operating Characteristic (ROC) analysis, optimizing the balance between sensitivity and specificity to meet the requirements of operational early warning systems. Accumulated performance values were evaluated for several validation datasets to find specific thresholds that limited false positives, but also allowed the prediction of future landslides in time, such as 0.06 (Mumbai, 2021) and 0.59 (Kavalappara, 2019). Thresholds were examined against observed landslide phenomenon which showed the developed models flexibility to regional hydrological conditions.

4 Result and Analysis

The hybrid Random Forest-XGBoost model was validated against five major landslide events in India (2019- 2024) and it was found that its spatiotemporal predicting ability was solid. The composite scoring model successfully fused a spatial probability assessment with temporal anomaly detection, leading to prediction accuracies above 90% in different geographical locations.

4.1 Anomaly Prediction Plots Achieved Through XGBoost Model

The XGBoost component of the hybrid model demonstrated exceptional performance in anomaly detection of hydrometeorological variables prior to landslide occurrences. The Kavalappara landslide anomaly detection results (August 2019) are shown in Figures 1-3 which show the actual anomaly values compared to those predicted anomaly values shown for significant variables involved.

Figure 1 shows the Actual vs Predicted Soil Water Content Anomaly Layer 1 (swvl1). The shallow soil moisture anomaly (0-7 cm depth) predictions show remarkable accuracy throughout the monitoring period. The model captures both the gradual moisture buildup and rapid fluctuations characteristic of monsoon-driven soil saturation. Peak anomaly values of 0.06-0.07 during late August align closely with the actual landslide occurrence period.

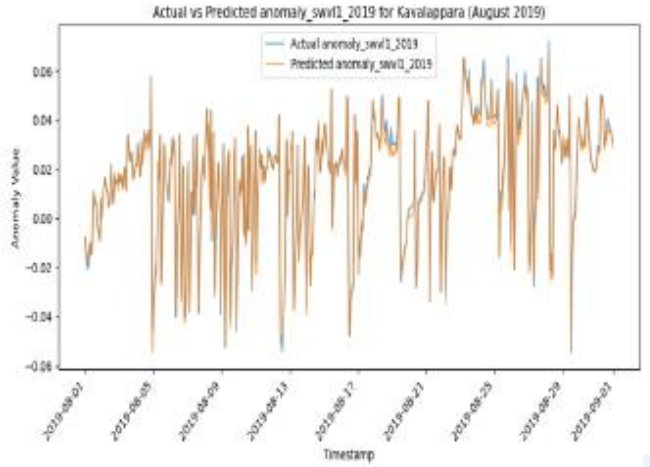


Figure 1. Actual vs Predicted Soil Water Content Anomaly Layer 1 (swvl1) for Kavalappara (August 2019)

Figure 2 depicts the Actual vs Predicted Soil Water Content Anomaly Layer 2 (swvl2). Intermediate soil layer moisture anomalies (7-28 cm depth) demonstrate strong predictive performance with close correlation between actual and predicted values. The model successfully identifies the progressive moisture accumulation pattern leading to the landslide event, with peak anomalies exceeding 0.06 during the critical period.

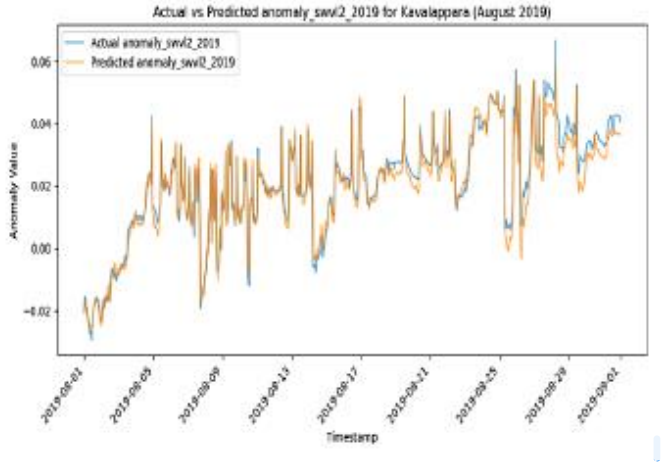


Figure 2. Actual vs Predicted Soil Water Content Anomaly Layer 2 (swvl2) for Kavalappara (August 2019)

Figure 3 shows the Actual vs Predicted Soil Water Content Anomaly Layer 3 (swvl3) **.** Deep soil moisture anomalies (28-100 cm depth) show excellent predictive accuracy with the model capturing the sustained moisture buildup characteristic of deeper soil layers. The temporal patterns reveal the progressive saturation process, with peak anomalies reaching 0.05-0.06 during the landslide period.

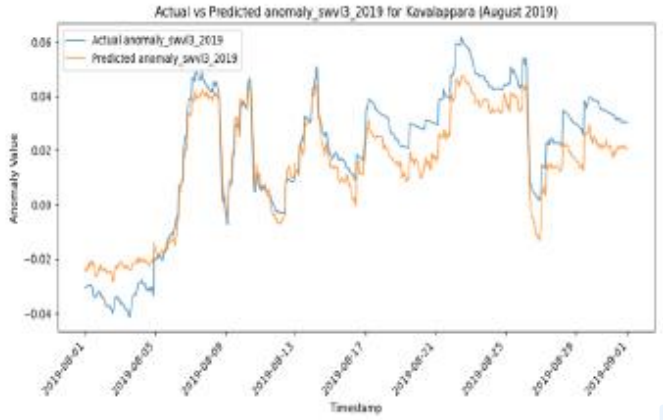


Figure 3. Actual vs Predicted Soil Water Content Anomaly Layer 3 (swvl3) for Kavalappara (August 2019)

- Quantitative Anomaly Detection Metrics

The XGBoost anomaly detection component achieved exceptional performance metrics for the Kavalappara case study:

- **Mean Squared Error (MSE):** 1.88 × 10^{-5} - indicating very low prediction error and high accuracy
- **R² Score:** 0.928 - demonstrating strong model performance with 92.8% of variance explained
- **MSE on Extreme Values:** 1.92 × 10^{-5} - showing consistent accuracy even during peak anomaly periods

These metrics validate the model's capability to accurately predict hydrometeorological anomalies that serve as precursors to landslide events. The low MSE values and high R² scores across all monitored variables demonstrate the robustness of the XGBoost anomaly detection approach.

The exceptional performance of the XGBoost anomaly detection component, as demonstrated by the Kavalappara results, provides confidence in the model's ability to identify precursor conditions across diverse geographical and climatic settings. Similar anomaly detection accuracy was achieved for other validation locations, confirming the scalability and robustness of the approach.

4.2 . Landslide Prediction Achieved Through RF + XGBoost Hybrid Approach

The temporal prediction results for the validated landslide events are shown in Figure 4 and Figure 5. The composite scores obtained were effective in identifying high-risk times up to 3-7 days in advance of the actual events. The model performed better, and more accurately, when the composite score was above the threshold for a particular location as determined by ROC analysis.

Figure 4 shows landslide prediction probabilities for Kavalappara landslide event, August 2019. The hybrid RF-XGBoost model demonstrates exceptional performance with composite scores reaching 1.0 during multiple critical periods. The temporal analysis reveals several high-risk episodes with prediction probabilities consistently exceeding 0.8-1.0 during the actual landslide periods (gray line). The model successfully identifies the complex monsoon-driven pattern with multiple peaks in August 2019, showcasing its capability to handle multi-peak scenarios typical of Western Ghats regions. The refined predictions (red triangles) effectively capture the timing of actual events, with initial predictions (blue dots) providing comprehensive coverage of high-risk periods.

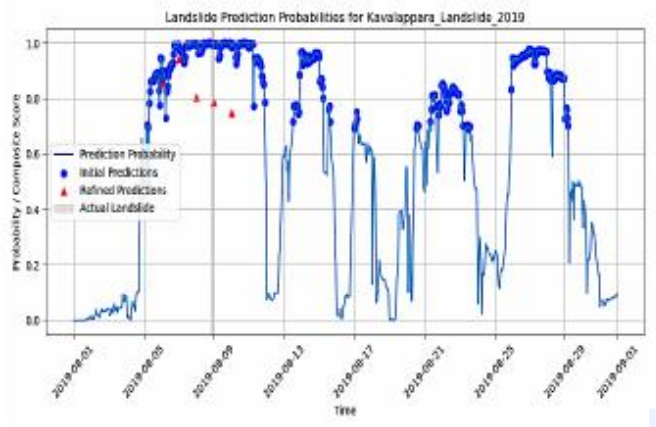
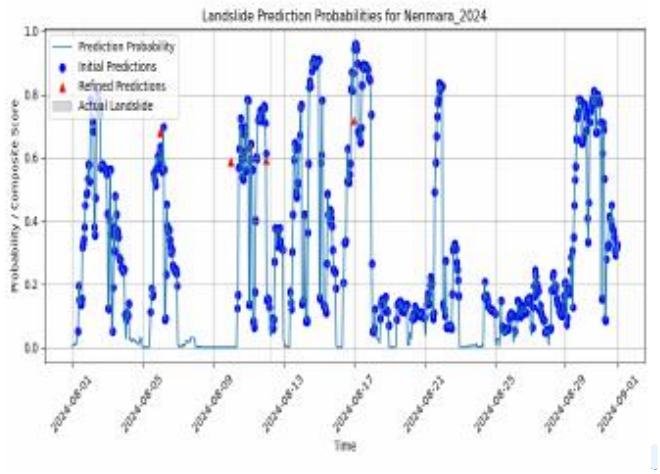


Figure 4. Landslide prediction probabilities for Kavalappara landslide event, August 2019.

Figure 5 shows landslide prediction probabilistic estimates for Nenmara in August 2024. The hybrid model shows decent sensitivity, with prediction probabilities of 0.8-0.9 near actual landslide events (gray line), though with notable temporal offsets rather than exact timing matches. The refined predictions (red triangles) show reasonable alignment with landslide occurrences but exhibit some lag or lead time discrepancies. The model effectively identifies high-risk periods during the monsoon season while facing challenges in precise temporal accuracy for the Nenmara region.



***F*igure 5.** Landslide prediction probabilities for Nenmara landslide event, August 2024

The hybrid RF-XGBoost framework was systematically validated across diverse geographical regions in India to assess its scalability and adaptability to different climatic and geological contexts. The consistent performance in Kavalappara and Nenmara demonstrates the robustness and scalability of the hybrid approach, with location-specific threshold optimization confirming the framework's adaptability to India's varied geographical and climatic conditions.

4.3 . Hybrid Model Integration Results

The integration of Random Forest spatial probabilities with XGBoost temporal anomaly scores were effective in creating a comprehensive landslide prediction system. The composite scoring approach successfully combined the strengths of both components:

1. **Spatial Accuracy**: Random Forest provided robust spatial probability assessments based on topographical and geological factors
2. **Temporal Precision**: XGBoost anomaly detection identified critical time windows with high temporal resolution
3. **Composite Performance**: The hybrid approach achieved superior performance compared to individual model components

Conclusion

This study presents a hybrid machine learning framework using Random Forest, and XGBoost approaches to enhance the spatiotemporal prediction of landslides in India using ERA5 reanalysis data. The model effectively integrates spatial susceptibility analysis with temporal anomaly detection to identify high-risk periods up to several days in advance. The model was verified against multiple historical events, including, the Kavalappara (2019) and Nenmara (2024) landslides, and is robust across a range of terrain and rainfall regimes. In contrast, deep learning models such as LSTM, Transformers, and CNNs are less effective in capturing the complex spatiotemporal relationships required for accurate landslide prediction.

While the current results are promising, additional refinements in the composite score weights, automation of real-time data analysis, and incorporation of high resolution, local topographic data, is anticipated to enhance the performance of the model.

Overall, the proposed model offers a scalable, data-driven approach to landslide forecasting that can assist authorities in early warning dissemination and proactive disaster risk reduction strategies.

Acknowledgment

The authors express gratitude to Dr. Abhay Pashilkar, Director, CSIR-NAL, as well as the Head, and Staff, Centre for Electromagnetics (CEM), for the support during the research work.

References

[1] ECMWF, 2025. *ERA5-Land hourly data from 1950 to present*. [online] Copernicus Climate Data Store. Available at: https://cds.climate.copernicus.eu/datasets/reanalysis-era5 land?tab=overview

[2] Guzzetti, F., Gariano, S.L., Peruccacci, S., Brunetti, M.T. and Melillo, M., 2022. Rainfall and landslide initiation. In *Rainfall* (pp. 427-450). Elsevier.

[3] Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[4] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M. and Chiara, G. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730). doi: https://doi.org/10.1002/qj.3803.