Predictive Analysis using Machine Learning: Assessing Rainfall Dynamics and Flood Vulnerabilities in Bihar

Guru Dayal Kumar[a], Shekhar Tyagi[b] and Kalandi Charan Pradhan[a]

[a]School of Humanities and Social Sciences; [b]Discipline of Computer Science and Engineering; [a,b]Indian Institute of Technology, Indore_452020, India

ARTICLE HISTORY
Compiled July 21, 2024

ABSTRACT
In recent years, Bihar has faced recurrent flood challenges, with rainfall patterns playing a pivotal role. This research harnesses the power of advanced machine learning techniques to meticulously analyze and predict these patterns, aiming to understand their subsequent impact on floods. Using a comprehensive dataset that captures not only annual rainfall but also flood-induced damages over several years, this study delves deep into district-wise flood severity. It further projects potential future rainfall patterns using state-of-the-art Machine Learning algorithms. The invaluable findings from this research not only shed light on the regions most vulnerable to flood-related damages but also offer a robust predictive framework. This framework stands as a beacon for policymakers and local authorities, guiding mitigation efforts and preparing the state for forthcoming challenges.

KEYWORDS
Bihar floods; Mitigation Efforts; Machine learning; Random forest; Rainfall patterns

1. Introduction

Bihar, situated in the heart of eastern India, has long been a region of historical significance and agricultural prominence. This state, which has seen the rise and fall of ancient empires, thrives on its fertile plains, nourished by the annual monsoon showers[1]. However, the very waters that rejuvenate its fields can also bring widespread devastation in the form of floods. Geographically, Bihar's vulnerability to floods stems from the mighty Ganges River, as well as several other rivers that traverse its landscape. These rivers, originating from the Himalayan catchment areas, carry with them the excess waters from monsoon rains and melting snow, often leading to floods when they overflow[2]. The recurring floods, while natural, have been intensified by factors like deforestation in catchment areas, rapid urbanization, and inadequate flood management infrastructure[3].For a state where a significant portion of the population depends on agriculture, the unpredictability of rains becomes a double-edged sword[4]. While essential for crops, uneven or excessive rainfall can disrupt sowing and harvesting cycles, impacting food security and the state's economy. This research takes a deep dive into the intricate relationship between Bihar's rainfall patterns and the re-

contact:gurudayal1996@gmail.com,shekhartyagirocks007@gmail.com,kcpradhan@iiti.ac.in

sulting floods. Utilizing advanced machine learning techniques, including regression models and time series forecasting, we aim to predict future rainfall trends and analyze their potential impacts on flood severity[5]. Such predictive analytics can provide invaluable insights for policymakers and stakeholders, enabling proactive measures and better resource allocation. In the era of big data, traditional statistical methods often fall short in capturing the nuances and complexities inherent in large datasets. Machine Learning (ML) offers a robust alternative, allowing for pattern recognition, anomaly detection, and predictive modeling on vast scales[6]. In this study, we employ a combination of regression models and time series forecasting, specifically focusing on the Random Forest Regressor and the ARIMA model. These techniques, backed by feature engineering strategies like lagged values and moving averages, provide a holistic approach to understanding and predicting rainfall and flood patterns.

2. Details of our prepared Dataset: Rainfall and Flood Damages in Bihar

Our dataset provides a comprehensive record of rainfall taken from the India-WRIS, Government of India and its associated flood damages in the state of Bihar from the Disaster Management department, Goverment of Bihar reference period 1991 to 2022. Bihar is situated in eastern India. Bihar is historically known to grapple with the challenges of heavy monsoon rains, which, while essential for agriculture, can also lead to devastating floods that impact lives and livelihoods.
Features of the Dataset:
1. Year: This column represents the specific year for which the data has been recorded. It serves as a temporal identifier, essential for any time series analysis. Given the dataset's scope, it allows for an annual analysis of rainfall and flood damage trends.
2. District: Bihar is administratively divided into several districts, each with its unique geographical and topographical characteristics. This feature provides district-wise data, enabling a granular analysis of rainfall and flood patterns.
3. Annual Rainfall (mm): Representing the total rainfall in millimeters received in a district for a particular year, this feature is central to the dataset. It provides insights into the variability and intensity of monsoon rains.
4. Flood Affected Area (hectares): This column denotes the total area (in hectares) impacted by floods in a specific district for a given year. It's a direct measure of the flood's spatial impact and can be used to gauge the severity of the flooding.
5. Population Affected: Representing the number of individuals affected by floods in a district for a given year, this feature provides a human dimension to the data. It reflects the societal impact of floods and can be correlated with rainfall to understand human vulnerabilities.
6. Crops Damaged (hectares): Agriculture is a significant livelihood in Bihar. This feature provides data on the total crop area (in hectares) damaged due to floods in a particular year for each district. It offers insights into the economic impact of floods on agriculture.
7. Houses Damaged: Denoting the number of houses damaged due to floods for a specific district in a given year, this feature provides insights into the infrastructural impact of floods.
8. Human Lives Lost: A grim reminder of the destructive power of nature, this column provides the number of lives lost due to floods in each district for a particular year. It underscores the urgency of predictive analyses and early warning systems.

This dataset, with its range of features, offers a holistic view of the interplay between rainfall, floods, and their impacts in Bihar. Analyzing this data can yield insights that are valuable for policymakers, disaster management authorities, and researchers alike.

3. Methodology

Understanding the intricate relationship between rainfall and floods in Bihar. It demands a comprehensive analytical approach. The methodology section elucidates the step-by-step procedures followed in this research, encompassing data preprocessing, exploratory analysis, feature engineering, and predictive modeling.

3.1. Data Acquisition and Preprocessing

The foundational element of this study is the dataset, which encompasses annual rainfall figures and associated flood damages across various districts in Bihar. Initial steps involved data cleaning and handling missing values. Given the temporal nature of the dataset, interpolation methods were employed to fill gaps, ensuring continuity and maintaining the time series' integrity[7].

3.2. Exploratory Data Analysis (EDA)

Before delving into machine learning models, a thorough EDA was conducted to understand the dataset's underlying patterns and anomalies[8]. This step was crucial in identifying trends, seasonality, and potential outliers. Visualizations, including time series plots and heat maps, were instrumental in revealing district-wise disparities in flood damages and variations in rainfall over the years.

3.3. Feature Engineering

Machine learning models thrive on informative features. Given the time series nature of the data, several derived features were introduced to capture temporal dependencies and patterns[9]: Lagged Values: These represent previous years' data points. Given that rainfall and flood patterns often exhibit temporal correlations, introducing lags(e.g., Lag_1, Lag_2) helped the model recognize such dependencies. Moving Averages: A moving average smoothens the time series, capturing the underlying trend by averaging values over a predefined window. For instance, a 3-year moving average was used to ascertain the general direction in which rainfall or flood damages were moving. Rolling Variances: Variance provides a measure of data dispersion. In a time series, rolling variance indicates how much the data fluctuates around a mean over a specific window. This feature was vital in recognizing periods of high variability, potentially corresponding to anomalous events or severe floods.

3.4. Predictive Modeling with Machine Learning

Given the study's objective of forecasting rainfall and understanding its relationship with flood damages, regression models were deemed most suitable[10]: Random Forest Regressor: A popular ensemble learning method, the Random Forest Regressor, constructs multiple decision trees during training and outputs the average prediction

of individual trees for regression tasks[11]. Its ability to handle non-linear relationships and its inherent feature importance metric made it a prime choice for this study. Time Series Forecasting Models: Given the dataset's temporal nature, specific models tailored for time series forecasting, such as ARIMA, were considered. However, based on preliminary tests and the complexity introduced by multiple districts, the study primarily relied on the Random Forest Regressor, which effectively captured the time dependencies via feature engineering.

4. Algorithms for Predicting Rainfall and Flood Damages in Bihar using Machine Learning

Input:

- Dataset containing annual rainfall figures and associated flood damages across various districts in Bihar.

Output:

- Predicted rainfall and flood damages for future years.

Algorithm 1 Data Acquisition and Preprocessing

1: Load the dataset into a suitable data structure.
2: **for** each district in the dataset **do**
3: Identify and handle missing values using interpolation methods.
4: **end for**

Algorithm 2 Exploratory Data Analysis (EDA)

1: Visualize the dataset to identify trends, seasonality, and potential outliers.
2: Generate summary statistics for district-wise disparities in flood damages and variations in rainfall.

Algorithm 3 Feature Engineering

1: **for** each year in the dataset **do**
2: Calculate `Lag_1` as the rainfall from the previous year.
3: Calculate `Lag_2` as the rainfall from two years ago.
4: Calculate `Moving_Avg_3` as the average rainfall over the current and two preceding years.
5: Calculate `Rolling_Variance_3` as the variance in rainfall over the current and two preceding years.
6: **end for**

5. Results

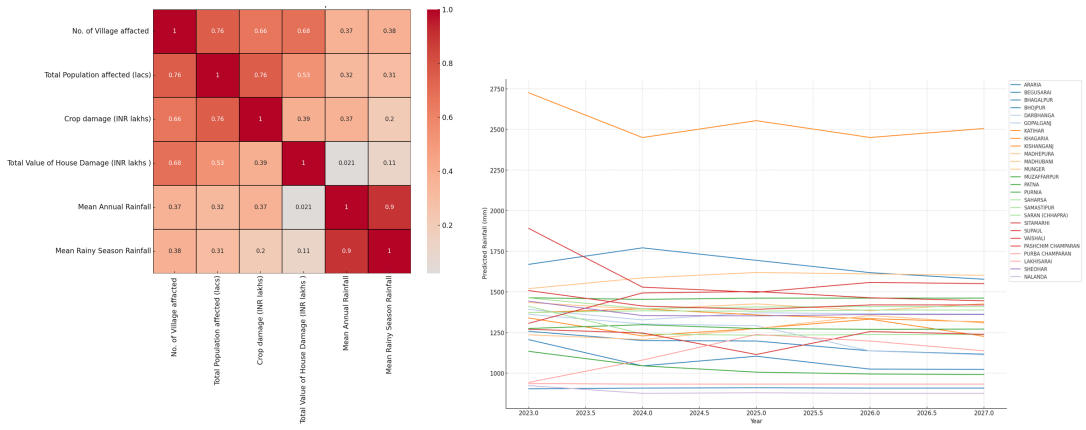
Our analyses revealed significant disparities in flood damages across districts, with some districts exhibiting consistently high flood-induced damages over the years. Predictions indicate fluctuating rainfall patterns for the coming years, with potential implications for flood management strategies.

Algorithm 4 Predictive Modeling using Random Forest Regression

Initialize a Random Forest Regressor with a specified number of decision trees.
for each tree in the Random Forest **do**
    Draw a bootstrap sample from the dataset.
    Build a decision tree using the sample, selecting a subset of features at each node without replacement.
    Split the node using the feature that provides the best split among the selected subset.
    Repeat the splitting until a specified stopping criterion is met, such as maximum tree depth or minimum samples at a leaf.
end for

Algorithm 5 Making Predictions

for each new data point or year to predict **do**
    Initialize a variable `sum` to zero.
    for each tree in the Random Forest **do**
        Calculate the prediction for the new data point using the tree.
        Add the prediction to `sum`.
    end for
    Calculate the final prediction as `sum` divided by the total number of trees in the Random Forest.
end for



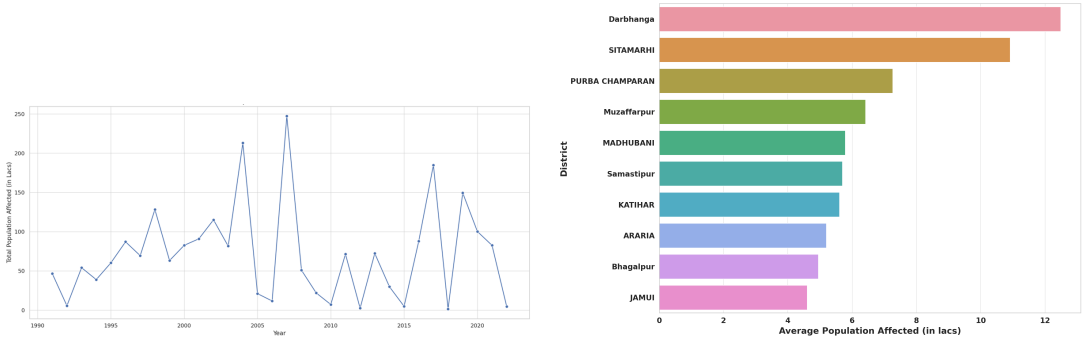
(a) Correlation heatmap.

(b) District wise rainfall for 2023-2027

Figure 1. Example of a two-part figure with individual sub-captions showing that captions are correlation heatmap and district wise rainfall for 2023-2027.

6. Discussion

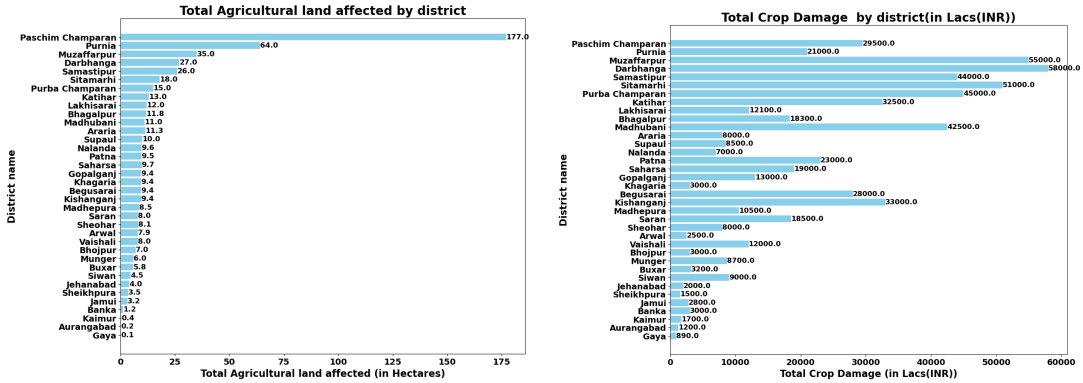
The results underscore the need for targeted flood management strategies. Districts with high flood severity indices should be prioritized for resource allocation and mitigation efforts. Stern's perspective highlights the importance of recognizing the diversity of vulnerabilities and adaptation needs across different regions and countries. A 'one-size-fits-all' approach to climate change adaptation may not be effective or eq-



(a) Total population affected over the years.

(b) Top 10 districts with highest average population affected across all years

Figure 2. Sub-captions showing that captions are the Total population affected over the years and the Top 10 districts with the highest average population affected across all years.



(a) Total agriculture land affected by district.

(b) Total crop damage by district

Figure 3. Diagram showing the total agriculture land affected by district and total crop damage by district.

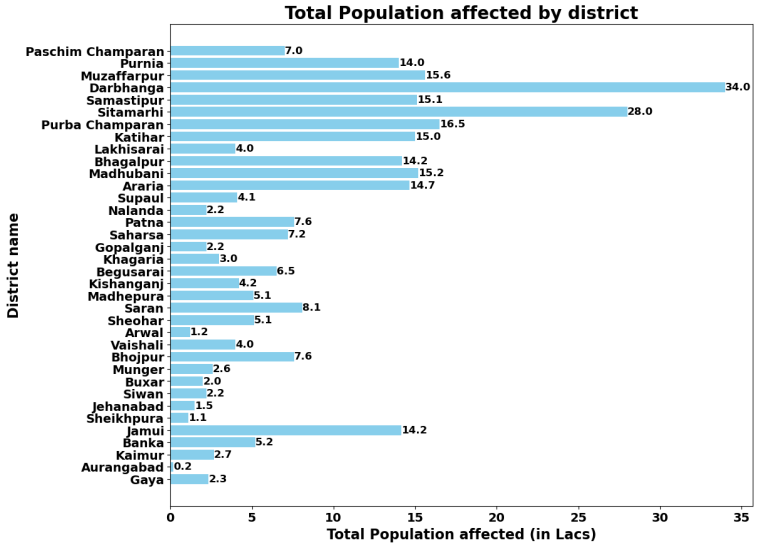
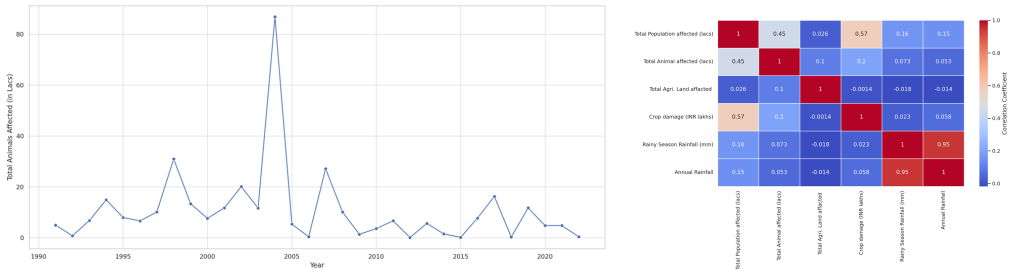


Figure 4. Distribution of affected population by districts



(a) Total animal affected over the years. (b) Correlation between various factors

Figure 5. Example of a two-part figure with individual sub-captions showing that captions are Total animals affected over the years and Correlation between various factors.

uitable, as the impacts and responses to climate change will vary depending on local circumstances. The predictive insights on rainfall provide an opportunity for proactive planning, allowing authorities to brace for potential flood scenarios in advance.

7. Conclusion

Harnessing the power of machine learning, this research illuminates the intricate relationship between rainfall and floods in Bihar. Identifying patterns and forecasting future scenarios paves the way for informed, data-driven decision-making in flood management.The Sustainable Development Goal 11 (SDG-11), which aspires to create inclusive, resilient, and sustainable cities and human settlements, is one area in which the study's conclusions are relevant. Specifically, these findings align with Target 11.5, which strives to reduce deaths and the impact of water-related disasters, while prioritizing the safeguarding of vulnerable and marginalized segments of society.

References

Smith, J. (2005). The Monsoons of India: Historical and Cultural Significance.*Indian Journal of Meteorology*, 58(2), 123-132.

Kumar, R. & Singh, P. (2010). River Dynamics and Flood History of the Ganges River.*Geographical Review*,100(3), 354-373.

Gupta, S. (2015). Urbanization and its Impact on Floods: A Case Study of Bihar. *Urban Studies*, 52(14), 2674-2689.

Verma, M. (2018). Agrarian Economy of Bihar: Challenges and Opportunities.*Economic Affairs*, 63(1), 77-85.

Brown, K. & Sharma, A. (2019). Machine Learning in Hydrology: Applications and Future Directions.*Journal of Hydro informatics*, 21(5), 673-688.

Mitchell, T. (1997). Machine Learning. McGraw Hill.

McKinney, W. (2010). *Data Structures for Statistical Computing in Python.* Proceedings of the 9th Python in Science Conference.

Tukey, J.W. (1977). *Exploratory Data Analysis.* Addison-Wesley.

Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice.* OTexts.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* Springer.

Breiman, L. (2001).*Random Forests Machine Learning*, 45(1), 5-32.

Stern, N. (2007). *The Economics of climate change: the stern review.* Cambridge University Press, Cambridge, UK.