Problem Statement

Cloudbursts are intense, short-duration rainfall events—typically exceeding 100 mm of rain within an hour over a localized area. Meteorologically, they are often associated with convective thunderstorms, rapid updrafts, and saturated atmospheric conditions, but their exact precursors remain difficult to isolate due to their sudden onset and small spatial scale. In urban contexts, cloudbursts can cause severe flash flooding, disrupt transportation and utilities, damage infrastructure, and endanger lives, particularly in densely populated and poorly drained areas. Socioeconomically, such events strain emergency response systems and disproportionately impact vulnerable communities, making timely and accurate early warnings crucial.

Despite advances in meteorological modeling, predicting cloudbursts remains a major technical challenge. These events are rare and spatially concentrated, leading to class imbalance in datasets. Additionally, weather station networks often have limited spatial coverage, reducing data granularity. Models trained in one region (e.g., the U.S.) may fail to generalize to other geographies with differing climatic patterns. Finally, accurate predictions may require explicit modeling of dynamic atmospheric processes, which are computationally intensive and data-dependent.

To address these challenges, the INDRA project aims to develop a machine learning–powered predictive system that can: (1) detect cloudbursts within a 3-hour window; (2) surpass the current 50% accuracy benchmark in target areas; and (3) integrate seamlessly with early warning systems to improve disaster preparedness for residents, businesses, and city planners. By combining public climate data, performance analysis, and real-world validation, the project seeks to deliver an effective and scalable cloudburst forecasting solution.

Literature Review

Recent advances in cloudburst prediction have leveraged a range of machine learning (ML) and deep learning (DL) approaches, each with distinct methodologies, data sources, and forecasting capabilities. Murakami et al. (2022) utilized a deep learning autoencoder to detect precipitation anomalies in Japan based on long-term climate data and simulations. Their approach, while effective in identifying broad anomaly trends linked to climate change, employed unsupervised learning without precision or recall metrics, limiting its applicability for short-term forecasting such as the 3–6 hour window targeted by the INDRA project.

In contrast, Patil and Kulkarni (2023) combined Global Forecast System (GFS) atmospheric data with supervised ML models, including XGBoost and neural networks, to predict cloudburst occurrences in India. Their models demonstrated strong binary classification performance (Accuracy = 88%, F1 = 0.84, Precision = 0.85, Recall = 0.83) and maintained high multiclass performance as well, indicating strong potential for short-term operational use. Their work is

particularly relevant to INDRA's goal of surpassing the current 50% accuracy benchmark in target locations.

Shani and Nagappan (2024) proposed a CNN-based classifier using Gramian Angular Field (GAF) transformations of weather vectors to detect burst events. Although their method achieved high precision (0.78), the relatively low recall (0.58) suggests it may miss a significant portion of actual cloudburst events, making it less reliable as a standalone solution. However, it could serve as a useful architectural baseline or component within a hybrid ensemble system.

Together, these studies highlight the promise of data-driven approaches in extreme weather forecasting and underscore the importance of model interpretability, data resolution, and performance trade-offs. For the INDRA project, the integration of real-time forecast data with supervised learning techniques, as demonstrated by Patil and Kulkarni, offers a particularly valuable foundation.

Data Processing and Feature Engineering

The primary dataset used in this project is the NOAA Local Climatological Data (LCD), which provides high-resolution meteorological observations from stations such as Miami International Airport (Station ID: 72202012839). Spanning from January 1, 2015 to December 31, 2024, the dataset includes over 100,000 hourly and daily records capturing variables such as temperature, dew point, wind speed and direction, atmospheric pressure, humidity, visibility, sky condition, and precipitation. In total, the dataset comprises over 120 columns, including both raw measurements and derived variables (e.g., heating/cooling degree days).

To prepare this data for modeling, several preprocessing steps were implemented. First, only the most relevant hourly features and metadata were retained. When multiple records existed for a given timestamp, entries were filtered based on a predefined priority hierarchy (FM-15 > FM-12 > FM-16 > SOD > SOM). All timestamps were converted into a uniform datetime format to ensure consistency. Missing values were inspected visually, and both duplicates and extreme outliers (beyond ±3 standard deviations) were removed to improve data quality.

Feature engineering was critical for enhancing the predictive signal. Temporal features such as hour of day, month, and weekend indicators were extracted, with cyclic patterns encoded using sine and cosine transformations. Derived meteorological variables were calculated, including vapor pressure, dew point depression, and wind vector components (U/V). Rainfall-specific features included cumulative precipitation over multiple time intervals (1h, 3h, 6h, 12h, 24h), maximum values, and binary rain event flags. Additionally, trend-based features capturing rate-of-change and lagged variables (e.g., 1-hour, 3-hour delay) were generated to reflect short-term weather dynamics preceding potential cloudbursts.

For model readiness, the dataset was scaled and split into time-aware training, validation, and test sets. Class balancing techniques were applied to mitigate the inherent rarity of cloudburst events and address the resulting class imbalance problem. Together, these steps formed a comprehensive preprocessing pipeline optimized for high-frequency, event-level precipitation forecasting.

Model Implementation and Evaluation

Given the rarity and unpredictability of cloudburst events, our modeling strategy focused on balancing recall and precision under extreme class imbalance conditions. We implemented and evaluated three primary approaches: a baseline Random Forest classifier, a Random Forest with SMOTEENN (a hybrid oversampling and cleaning method), and a custom Two-Stage Model. Among these, the Random Forest with SMOTEENN emerged as the most balanced and practically viable solution.

SMOTEENN first synthetically oversampled minority (cloudburst) instances and then applied Edited Nearest Neighbors (ENN) to clean noise and borderline majority instances. The feature set included lagged precipitation values, temporal encodings (hour, month, weekend), weather-derived metrics (vapor pressure, wind U/V), and short-term rainfall trends.

The final model was evaluated at a 0.50 threshold and produced the following confusion matrix:

- Precision: 0.1510
- Recall: 0.2261
- F1 Score: 0.1811
- ROC-AUC: 0.8753
- PR-AUC: 0.0901

These results highlight a critical tradeoff: while the ROC-AUC score of 0.8753 suggests that the model performs well in distinguishing between classes globally, the low PR-AUC (0.0901) reflects the inherent difficulty of correctly identifying rare cloudburst events in practice. Compared to our baseline and Two-Stage approaches, the SMOTEENN-enhanced model showed improved F1 and precision without drastically sacrificing recall, making it the most appropriate candidate for real-world deployment in early warning systems.

In summary, our model demonstrates that with careful preprocessing and class balancing, machine learning models can provide actionable insights even in highly imbalanced, rare-event

prediction contexts. However, ongoing refinement—particularly in reducing false positives and improving lead time reliability—remains necessary for operational readiness.

Ethical Considerations

Deploying a machine learning–based cloudburst prediction system involves several important ethical considerations, particularly in contexts where public safety is at stake. One of the primary concerns lies in the tradeoff between false positives and false negatives. A false negative—failing to predict an actual cloudburst—could result in insufficient warnings, leading to loss of life, property damage, and inadequate emergency preparedness. On the other hand, false positives—predicting an event that does not occur—could cause unnecessary panic, economic disruption, and erosion of public trust in early warning systems. Striking a balance between sensitivity and specificity is therefore critical, especially in high-stakes urban environments.

Although the current system uses only publicly available meteorological data, privacy and data security must still be considered if the system is expanded to integrate citizen-reported observations, geolocation data, or private infrastructure sensors in the future. Ensuring transparency around data use, secure storage, and compliance with data protection regulations will be essential to maintain public confidence.

In addition, the use of machine learning in safety-critical domains raises concerns about algorithmic bias, model interpretability, and the risk of over-reliance on automated systems. ML models may fail to generalize to new climate patterns or underrepresented regions, potentially reinforcing geographic inequalities in warning accuracy.

AI Assistants Use and Process Reflection

Throughout this project, I made extensive use of AI tools such as ChatGPT and Claude to support various stages of research, modeling, and analysis. These tools played a crucial role in helping me navigate a new and technically demanding domain.

During the research phase, I relied on AI to rapidly acquire foundational knowledge about cloudbursts and meteorological concepts. Instead of spending hours combing through academic papers and technical articles, I used AI to ask targeted questions and receive clear, concise explanations. This interactive learning approach helped me quickly understand the definition, causes, precursors, and broader impacts of cloudbursts, allowing me to construct a solid conceptual framework for the rest of the project.

In the modeling phase, AI tools were even more central to my workflow. I used them to generate base code for machine learning models, refine data preprocessing strategies, and brainstorm feature engineering ideas. For example, when working with time-series weather data, I asked AI

to suggest ways to extract lag features, moving averages, and cyclical temporal encodings. AI also helped me design class balancing strategies for rare event prediction, such as combining SMOTE and ENN. These suggestions saved time and exposed me to a wide range of techniques that I may not have considered independently.

That said, there were limitations in using AI tools. Not all suggestions were directly applicable to my dataset or model pipeline. Sometimes, AI-generated code would make assumptions about data structure that didn't hold in practice. To address this, I regularly verified AI outputs by running the code, inspecting errors, and cross-referencing with official documentation. I also learned to be cautious of edge cases and to critically evaluate the logic of AI-generated solutions. This process improved my debugging skills and taught me how to better balance automation with manual oversight.

Working on this project helped me grow in multiple areas. Tackling a novel topic like cloudburst prediction forced me to think critically and seek help efficiently. Using AI as a "co-pilot" made the learning curve less steep and encouraged me to experiment more confidently with new methods. At the same time, I developed transferable data science skills, including handling imbalanced datasets, designing time-aware features, evaluating models with appropriate metrics (e.g., PR-AUC), and optimizing classification performance in rare-event contexts.

Conclusion and Future Directions

Overall, this project was a highly valuable learning experience that taught me how to approach and execute a complete data science project from start to finish. I gained hands-on experience in defining a problem, conducting background research, collecting and processing complex datasets, building and evaluating predictive models, and reflecting on ethical implications. More importantly, I learned how to think systematically and iteratively—breaking down the problem into stages and addressing each one with a combination of technical tools, critical thinking, and structured experimentation.

Looking ahead, I see clear opportunities for technical improvement, particularly in modeling strategies. One key insight from this project was the importance of temporal features, which consistently emerged as significant across multiple machine learning models. In future iterations, I plan to explore time series–based models such as LSTM, GRU, or Temporal Fusion Transformers, which are more naturally suited to sequential weather data. These models may offer improved performance by capturing longer-term dependencies and subtle temporal patterns that traditional ensemble models might overlook.

In addition, future work could focus on expanding the dataset to include satellite data, citizen reports, or multi-source sensor streams, as well as improving interpretability and uncertainty quantification in model outputs. These enhancements would support more robust and trustworthy deployment in real-world early warning systems.