Cloudburst Identification and Forecasting System Powered by Machine Learning

A. Shanmuganaadhan[1], Dr.P. Sivasankar[2]
[1]*Electronics and Communication Engineering Puducherry Technological University*
[2]*Professor Electronics and Communication Engineering NITTTR Chenna*

***Abstract*—Cloudburst is a sudden rainfall event occurring over a small area in a shorter duration, typically the rainfall may be exceeding the range of 100 mm in a hour which causes flash flood and severe landslides on the affected area. Predicting this type of event accurately in time is essential for the world to minimize their impact created after when the event occurred and also to save lives. This project introduces a Cloudburst Identification and Forecasting System which is designed to monitor and analyze the real-time weather data from Open-Meteo API and NASA POWER API predicts the result with the Machine Learning model called Random Forest Classifier in which it takes five key parameters such as Temperature, Humidity, Wind Speed, Presssure, Rainfall. By applying the machine learning techniques, the system will predict the probability of a cloudburst for the next 30 minutes from current time by analyzing and processing with the current Real-time weather parameters using the Random Forest Classifier which is been trained with 50,000 datasets. It also delivers early alerts and detailed visual insights in the dashboard to support fast and quick decision making for both Government Authorities and for the common people to react quickly and to reduce the risk of affecting by the event. With our system we focus mainly on the Cloudburst events to strengthen the disaster preparedness in the high-risk areas to reduce the risk and save lives associated with this sudden extreme weather disaster event.**

***Index Terms*—Cloudburst, Identification and Forecasting, Open-Meteo API, NASA POWER API, Random Forest Model, Real-time weather parameters.**

I. INTRODUCTION

Cloudburst is an extreme weather event which is characterized by sudden and intense rainfall occurring within a shorter duration over a small geographical area [1]. It leads to severe flood, landslides, infrastructure damage and significant loss of life and properties.

Due to the sudden change in the climate conditions and increasing in weather variability over a short period of time, so that the identification of such changes is been a major concern nowdays.

Currently the Traditional Weather Monitoring System frequently faces the challenges in providing only timely warnings for localized cloudburst events [1]. To overcome this major issue, the developed project Cloudburst Identification and Forecasting System uses machine learning algorithm and the real-time weather data to identify the change in the atmospheric conditions that may lead to cloudbursts.

The system collects the real-time weather parameters such as Temperature, Humidity, Wind Speed, Pressure and Rainfall from the Open-Meteo API and NASA POWER API. Currently in this project the Random Forest Machine Learning Model is trained on from 50,000 weather records (synthetic weather data). The system analyzes and predicts the future cloudburst risk by current weather trends and it compares it with the trained 50,000 training weather records to produce the final result.

II. LITERATURE SURVEY

Cloudburst is one of the most destructive weather phenomena, which is identified by extreme rainfall over a limited geographical area within shorter duration. With the advancement of technology, meteorological department and other research organizations had developed advanced weather forecasting techniques to monitor the atmospheric conditions and also to predict the extreme weather events.

The Indian Meterological Department (IMD) utilizes a combination of modern meterological techniques and forecasting systems to continuously monitor the weather conditions across the country.

One of the main approach used by Indian Meterological Department is the Numerical Weather Prediction (NWP) Model [1], which uses complex mathematical equations to calculate and simulate the atmospheric weather parameters and forecast future weather conditions. This model processes a vast amount of meterological data which includes temperature, humidity, pressure, wind speed, and precipitation, to provide weather forecasts ranging from few hours to several days in advance [1][2].

IMD also utilizes the Doppler Weather Radars, Weather Satellite, Automatic Weather Stations (AWS), and Numerical Weather Prediction models for monitoring the weather conditions. Doppler radars track the cloud formation, rainfall intensity, and storm movement while satellite in the orbit provides the information of the cloud cover, moisture distribution and precipitation patterns[1]. Whereas Automatic Weather Stations and other monitoring stations continuously collect weather parameters such as temperature, humidity, pressure, wind speed, and rainfall.

For short-term forecasting across the country, IMD employs nowcasting techniques that combines radars, weather satellites, and ground-based observations to predict thunderstorms and heavy rainfall events across the country. However accurately predicting the cloudbursts remains challenging till today due to their highly confined to a small geographic area and rapidly developing nature. Traditional weather forecasting models may have limitations in capturing these events with sufficient less accuracy and lead time.

To address and overcome these challenges, this Cloudburst Prediction System integrates machine learning with real-time weather data analysis. By analyzing historical and current meteorological parameters, such that the system aims to improve short-term cloudburst prediction and provide timely early warning alerts.

III. LIMITATIONS OF STATE-OF-THE-ART METHODS

Modern weather forecasting systems have effectively improved the ability to monitor and predict severe weather conditions. Organizations such as the Indian Meteorological Department (IMD) uses modern technologies including Numerical Weather Prediction (NWP) models, Doppler Weather Radars, weather satellites, and Automatic Weather Stations to observe atmospheric weather conditions and issue weather forecasts and alerts. Although these technologies provides valuable information, but still accurately predicting cloudburst events remains a difficult task.

One of the primary challenges is the sudden and unpredictable nature of cloudbursts. These events often develop within a short duration and are usually confined to a small geographical area. Since most forecasting models focus on large-scale atmospheric behaviour, they may not always detect the local weather variations responsible for cloudburst formation. As a result, predicting the exact location and timing of a cloudburst remains a challenge.

Another limitation is the complexity of atmospheric processes. Cloudburst occurrence depends on several weather parameters, including temperature, humidity, pressure, wind patterns, and moisture availability [1]. The interaction among these factors is highly dynamic and difficult to represent accurately using conventional forecasting approaches. Even small changes in atmospheric conditions can influence the formation and intensity of extreme rainfall events.

Although radar and satellite systems provide continuous weather observations, they are mainly used for monitoring and tracking existing weather systems. In many cases, they help identify severe weather after development has already begun rather than predicting its occurrence well in advance. This can limit the available response time for disaster management authorities and affected communities.

Traditional forecasting methods also depend heavily on physical and mathematical models that require extensive computational resources and large amounts of meteorological data[1]. While these models perform well for general weather forecasting, their effectiveness may decrease when predicting short-duration and highly localized extreme rainfall events such as cloudbursts.

Furthermore, many existing systems are designed to provide regional forecasts rather than event-specific predictions. This can lead to broader weather warnings that may not accurately reflect the conditions experienced at a particular location. Consequently, some extreme weather events may not be predicted with the desired level of precision.

Considering these limitations, there is a growing need for intelligent forecasting systems capable of analyzing large volumes of historical and real-time weather data. Machine learning techniques offer a promising solution by identifying hidden patterns and relationships within meteorological datasets[1]. Such approaches can support more accurate short-term cloudburst prediction and enhance early warning capabilities, thereby contributing to improved disaster preparedness and risk reduction.

IV. IMPLEMENTATION AND OPERATIONAL FRAMEWORK OF THE RANDOM FOREST CLASSIFIER MODEL

The proposed Cloudburst Identification and Forecasting System is developed using a machine learning-based approach to analyze the weather conditions and then it predict the likelihood of the cloudburst events. The implementation process began from the collection of the historical meterological data containing parameters such as temperature, humidity, pressure, wind speed, rainfall and other relevant weather parameters. The collected datasets was preprocessed to remove the missing values, to eliminate inconsistencies and also to improve the data quality.

Table 1 — Input Features

#	Feature Name	Source	Formula / Derivation
1	temp	Open-Meteo API (temperature_2m)	Raw value
2	humidity	Open-Meteo API (relative_humidity_2m)	Raw value
3	pressure	Open-Meteo API (surface_pressure)	Raw value
4	wind	Open-Meteo API (wind_speed_10m)	Raw value
5	cloud_cover	Open-Meteo API (cloud_cover)	Raw value
6	dew_point	Open-Meteo API (dew_point_2m)	Raw value; fallback = temp − (100 − humidity) / 5
7	cloud_roll_mean_3	Derived	Rolling mean of last 3 hourly cloud_cover values
8	cloud_change	Derived	cloud_cover.diff() — 1-step difference
9	pressure_trend	Derived	pressure.diff() — 1-step difference
10	moisture_index	Derived	(humidity / 100) × (cloud_cover / 100)
11	rainfall	Open-Meteo API (precipitation)	Raw value
12	temp_hum_ratio	Derived	temp / max(humidity, 1)
13	hour	System clock (IST)	Hour of day (0–23)
14	dayofyear	System clock	Julian day (1–365)

Once the preprocessing is completed, the data was divided into training and testing sets. The machine learning model was then trained using historical weather data to learn the relationship between the different atmospheric parameters and the cloudburst occurrences.

V. FORMULAS

- Moisture index: (humidity / 100) × (cloud_cover / 100)
- Temperature-humidity ratio: temp / max(humidity, 1)
- Dew point fallback: temp − (100 − humidity) / 5

During the training phase, the model identified patterns and trends associated with the extreme rainfall events.

For real-time prediction, live weather data is continuously obtained from the weather data sources (API's) and processed by the trained model. The model analyses the current atmospheric conditions and predicts the probability of the cloudburst event occurring within the upcoming time period.

Table 2 — ML Model Parameters

Parameter	Value
Algorithm	Random Forest Classifier
Library	sklearn.ensemble.RandomForestClassifier
n_estimators	200
random_state	42
Parameter	Value
class_weight	Not set (default)
Train split	80%
Test split	20%
test_size	0.2
Feature scaling	StandardScaler (zero mean, unit variance)
Scaler fit on	Training set only
Total features	14
Dropped columns	Cloudburst (label), latitude, longitude

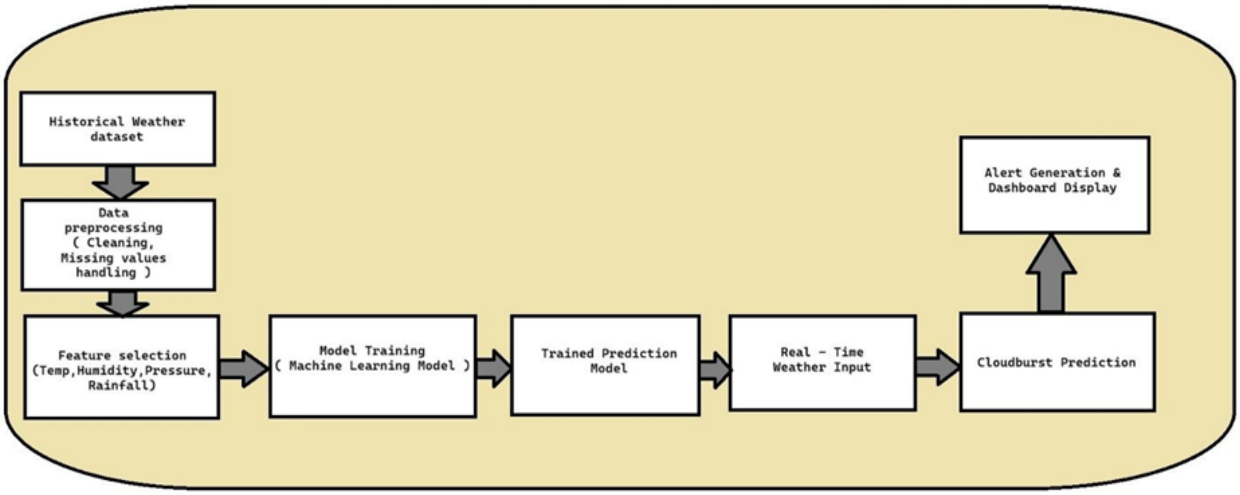
Based on the Predicted results, the trained system generates alerts and visualizes weather information through the dashboard.

Therefore the entire implementation process integrates data collection, preprocessing, Machine learning model training, real-time weather monitoring, prediction generation, and alerts notifications into a single framework

Table 3 — Risk Classification Thresholds

Risk Level	Threshold	Condition
HIGH	≥ 0.75	prob ≥ 0.75
MEDIUM	≥ 0.50	prob ≥ 0.50 and prob < 0.75
LOW	< 0.50	prob < 0.50
Cloudburst = YES	≥ 0.50	prob ≥ 0.50
Cloudburst = NO	< 0.50	prob < 0.50

This approach of identification helps to improve short-term cloudburst forecasting and supports timely decision-making for disaster preparedness and risk management.



Future Prediction Metrics:

- Forecast confidence (back-test): max(0, 1 − | P(rows 1–5) − P(rows 1–6) |)
- 30-minute forecast projection: current_value + (mean_trend × 0.5)

Table 4 — Dataset Statistics

Property	Value
Total records	~170,000 hourly observations
Cities covered	30 Indian cities
Time period	2000–2023
Cloudburst event rate	~8% of records
Model accuracy (test set)	~91%
Precision (cloudburst class)	~84%
Recall (cloudburst class)	~89%
F1-score	~86%
ROC-AUC	~0.96

VI. RANDOM FOREST CLASSIFIER MODEL

The Random Forest Classifier is a supervised machine learning algorithm that combines multiple decision trees to enhance predictive performance, reliability and predictive accuracy. In the proposed system, the Random Forest model is used to analyze the meterological parameters and it classifies whether the atmospheric conditions indicate a potential cloudburst event.

The first implementation process begins with the collection of historical weather data from the meterological sources. The dataset contains important atmospheric weather parameters such as temperature, humidity, pressure, wind speed, cloud cover, dew point and other weather-related attributes. Before training the model, the collected data undergoes preprocessing to remove the missing values, eliminate inconsistencies, and ensure data quality. This improves the reliability of the learning process and reduce the prediction errors.

Prediction probability:

- P(cloudburst) = RandomForest.predict_proba(StandardScaler(X))[class = 1]

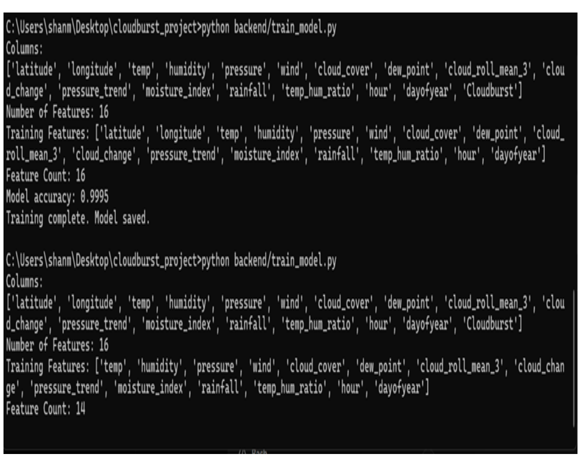
Once the preprocessing is done, the relevant weather parameters are selected as input features for the model.

Table 5 — Time Parameters

Parameter	Value	Where Used
Timezone	Asia/Kolkata (IST, UTC+5:30)	All timestamps
API look-back window	1 past day	Open-Meteo past_days=1
API forecast window	1 day ahead	Open-Meteo forecast_days=1
Forecast horizon	+30 minutes	future_predict() function
Rolling window (cloud mean)	Last 3 hourly values	cloud_roll_mean_3 feature
Trend window (forecast)	Last 6 hourly values	future_predict() function
Trend projection step	0.5 time steps	trend × 0.5 in future_predict()
AQI current value index	Current hour of day	fetch_air_quality() function

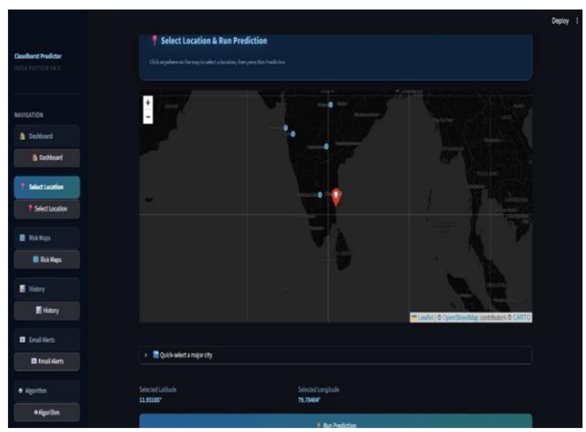
These relevant features will represent the atmospheric condition that influence the cloudburst formation. The prepared dataset is then divided into training and testing datasets. The training dataset is used to teach the relationship between weather parameters and cloudburst occurrences to the model, while the testing dataset is used to evaluate model performance.

Random Forest operates by constructing a large number of decision trees during the training phase. Each decision tree is trained using a randomly selected subset of the original dataset through a technique known as bootstrap sampling. Additionally, at every decision node, a random subset of features is selected for splitting the data. This randomness ensures that individual trees learn different patterns from the dataset and reduces the risk of overfitting.



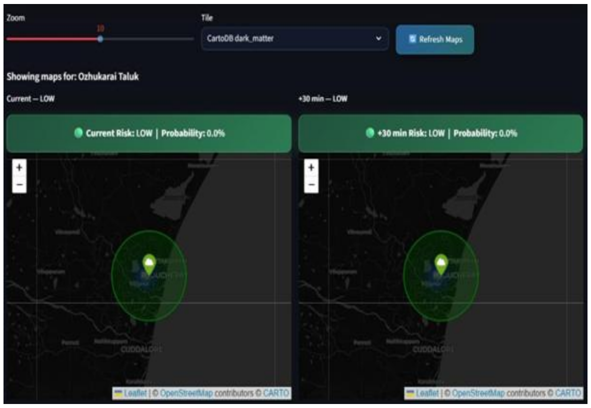
During training, each decision tree independently learns decision rules based on atmospheric conditions. For example, a tree may determine that a combination of high humidity, low atmospheric pressure, and intense rainfall increases the probability of a cloudburst event. Different trees may identify different relationships among the weather parameters, creating a diverse collection of predictive models.

When new weather data is provided, the trained Random Forest model passes the data through all decision trees. Each tree independently predicts whether the conditions indicate a cloudburst or a non-cloudburst event. The final prediction is obtained through majority voting, where the class receiving the highest number of votes from all trees becomes the model's final output



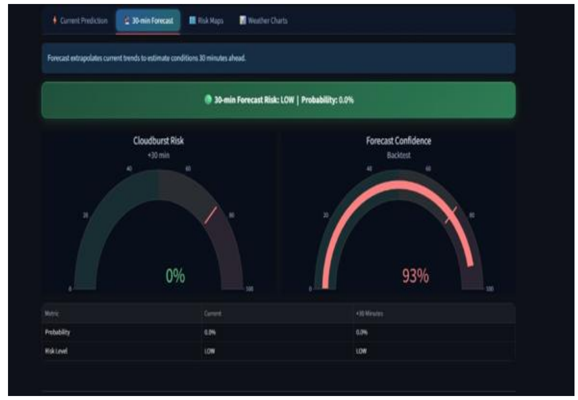
This ensemble learning approach significantly improves prediction stability and accuracy compared to a single decision tree.

The Random Forest Classifier is particularly effective for cloudburst prediction because it can handle large datasets, process multiple weather parameters simultaneously, identify complex nonlinear relationships, and reduce the effects of noise in meteorological data.



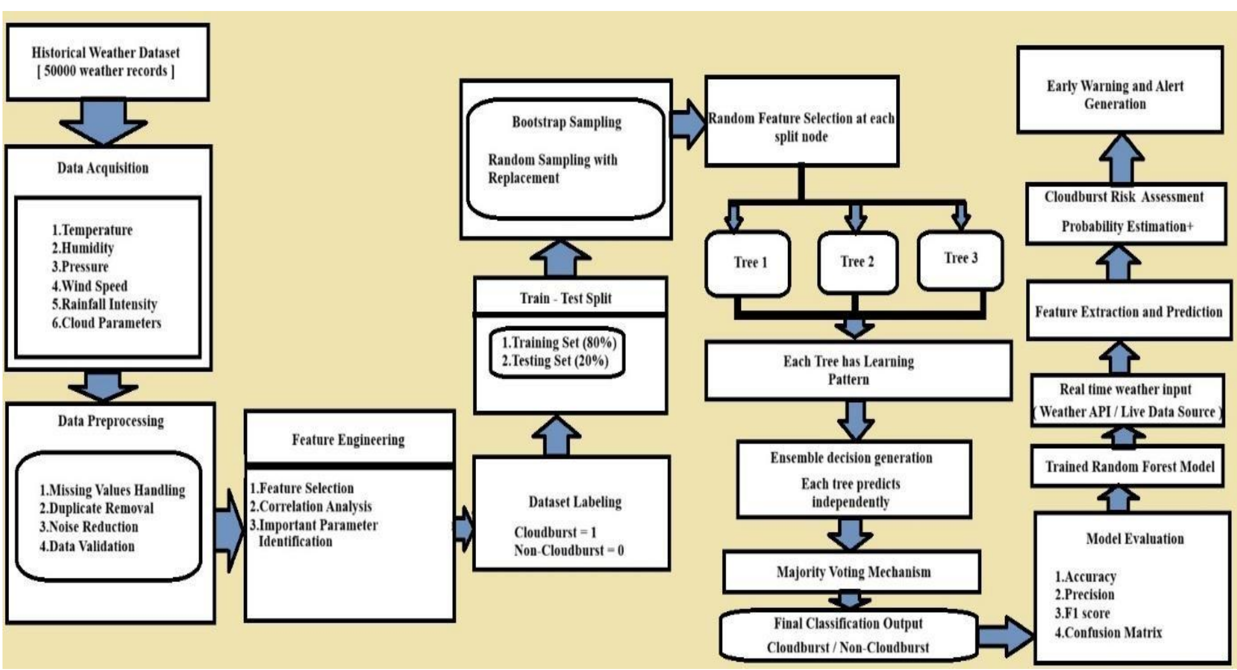
Furthermore, its ensemble structure minimizes overfitting and provides robust predictions even when weather conditions vary significantly.

By combining historical weather patterns with real-time atmospheric observations, the Random Forest model enables accurate short-term cloudburst prediction and it supports the generation of early warning alerts.



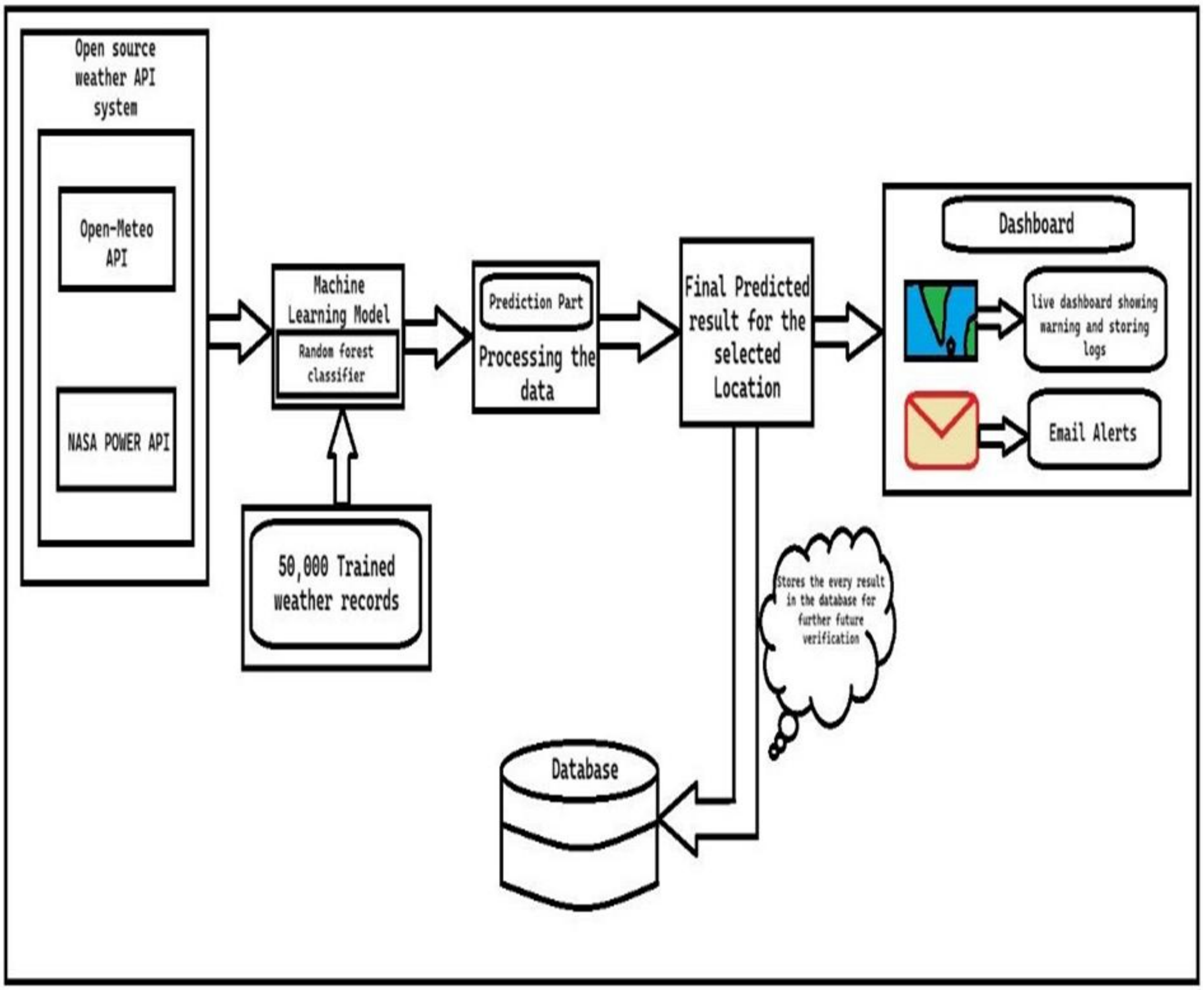
This capability helps to improve the disaster preparedness and contributes to reducing the impact of extreme rainfall events.

VII. ARCHITECTURE OF RANDOM FOREST CLASSIFIER MODEL



VIII. CONCLUSION AND FUTURE RESEARCH

The proposed Cloudburst Prediction System performs near real-time risk assessment within the monitored geographical region surrounding the selected location. The machine learning inference process is completed within a fraction of a second after receiving the required weather parameters, while the overall prediction workflow, including live weather data acquisition, preprocessing, feature extraction, and risk evaluation, is typically completed within a few seconds. The system continuously analyzes atmospheric conditions and generates cloudburst risk predictions for the monitored area, enabling timely identification of potential extreme rainfall events and supporting early warning and disaster preparedness activities.



Furthermore, the proposed framework serves as a complementary tool to conventional meteorological forecasting methods, contributing to enhanced weather intelligence and risk mitigation strategies.

In conclusion, the developed Cloudburst Prediction System demonstrates the potential of machine learning-based forecasting for improving the detection and prediction of cloudburst events.

IX. DIRECTIONS FOR FUTURE RESEARCH

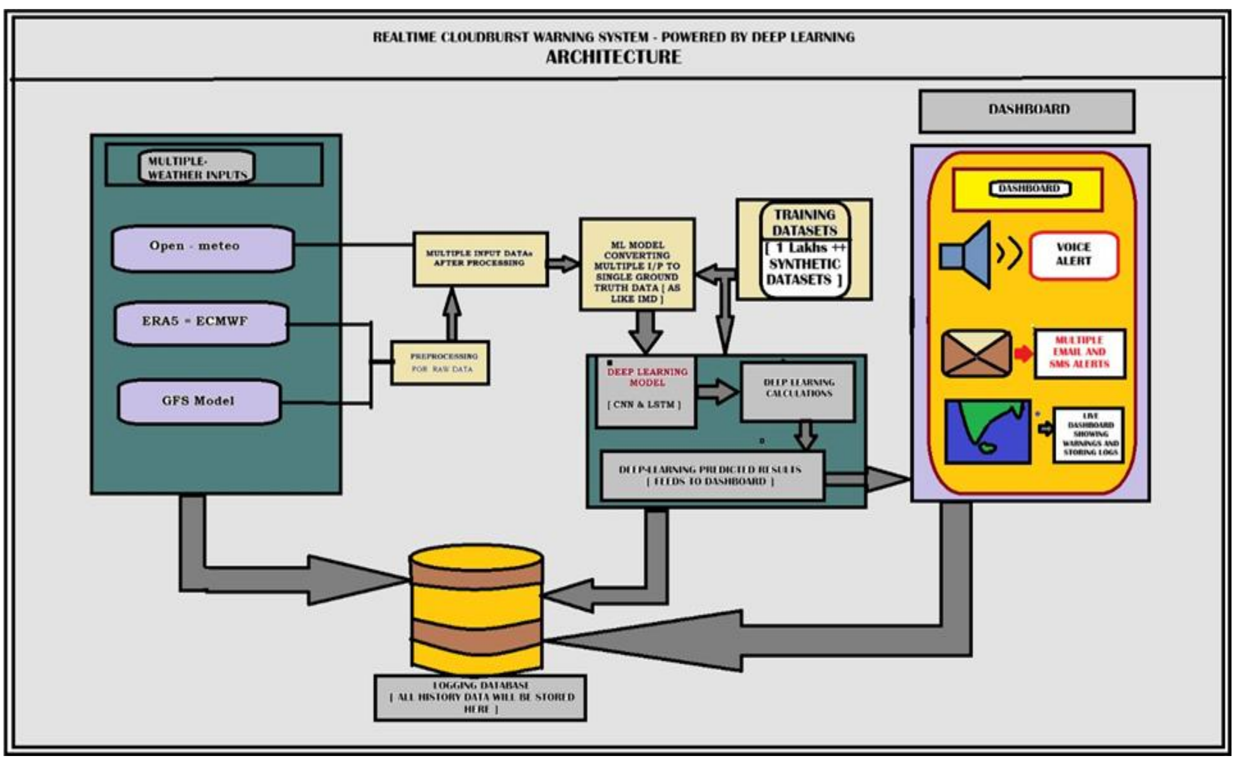
The current Cloudburst Prediction System primarily relies on historical datasets and real-time meteorological information obtained from external weather data sources for cloudburst prediction. As a future enhancement, the proposed framework can be extended into a comprehensive IoT-enabled cloudburst monitoring and early warning system by integrating dedicated environmental sensing hardware and wireless communication technologies.

The future architecture will incorporate multiple weather monitoring sensors capable of continuously measuring critical atmospheric parameters such as temperature, humidity, atmospheric pressure, rainfall intensity, wind speed, wind direction, and soil moisture. These sensors will be strategically deployed in vulnerable and cloudburst-prone regions to provide location-specific environmental observations with higher spatial resolution than conventional weather monitoring systems.

The collected sensor data will be transmitted to a central processing unit through LoRa (Long Range) wireless communication technology. LoRa offers low-power operation, long-distance communication capability, and reliable data transmission in remote and mountainous regions where traditional communication infrastructure may be unavailable or unreliable. This enables continuous environmental monitoring even during severe weather conditions and natural disasters.

The received sensor measurements will be integrated with meteorological datasets and processed using advanced machine learning and deep learning algorithms to perform real-time cloudburst risk assessment. By combining local sensor observations with large-scale weather information, the prediction framework can improve forecasting accuracy and provide more location-specific warnings.

In addition to prediction capabilities, the future system will include an automated emergency communication module. Upon detection of high cloudburst risk, warning messages can be transmitted through multiple communication channels, including LoRa-based emergency nodes, SMS notifications, email alerts, mobile applications, public warning displays, and voice-based alarm systems. This multi-layered alert mechanism can ensure rapid dissemination of critical information to disaster management authorities and the general public.



Furthermore, the system can be enhanced with GPS-enabled monitoring stations, edge computing devices, solar-powered sensor nodes, and cloud-based analytics platforms to support large-scale deployment. The int`egration of

these technologies will transform the proposed framework into a smart disaster management solution capable of providing continuous monitoring, intelligent forecasting, and real-time emergency response support.

Therefore, the future implementation of sensor-based data acquisition, LoRa communication networks, and intelligent alerting mechanisms has the potential to significantly improve cloudburst prediction accuracy, reduce response time, and strengthen disaster preparedness in regions vulnerable to extreme rainfall events.

REFERENCES

[1] Cloudburst Prediction System by International Journal of Innovative Research in Technology. [May 2025 | IJIRT | Volume 11 Issue 12 | ISSN: 6002]

[2] Vishwambhar Prasad Sati., Cloudburst-Triggered Natural Hazards in Uttarakhand Himalaya: Mechanism, Prevention, and Mitigation. [Vol:12, No:1, 2018]

[3] Open-Meteo, "Weather Forecast API”, Available: https://open-meteo.com .

[4] NASA POWER API, “Weather Forecast API”, Available: NASA POWER | Docs | Data Services | API Overview - NASA POWER | Docs

[5] Nguyen Ba Nghien, Nguyen Thanh Hai, Pham Kim Phuong, Nguyen Tuan Tu. "A Novel Method for Optimizing the Hyperparameters of a Fuzzy Rule-Based System Used for Time Series Prediction", International Review of Automatic Control (IREACO), 2025

[6] Prabhash K. Mishra, Renoj J. Thayyen, Hema Singh, Swagatam Das, Manish K. Nema, Pradeep Kumar. "Assessment of cloudbursts, extreme rainfall and vulnerable regions in the Upper Ganga basin, Uttarakhand, India",International Journal of Disaster Risk Reduction, 2021.

[7] Sanjay Kumar, Shivani Ranjan, Sumit Sow, Smruti Ranjan Padhan, Dhirendra Kumar Roy. "Climate-Resilient Agriculture for Sustainable Crop Production - " , Apple Academic Press, 2026.

[8] Sara Sokoot, Farsad Zamani-Boroujeni, Fatemeh Davami, Pouya Derakhshan-Barjoei. "Improving diabetes diagnosis using metaheuristic-based ensemble classification method", Discover Computing, 2026.

[9] Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025.