International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

RainDNA: Autonomous Atmospheric Acidification Analysis

Reenu Elizabeth Manu[1], Mrs. Indu Parvathy[2]

[1]Department of Computer Science, Sacred Heart College, Thevara, Kochi
[2]Assistant Professor, Department of Computer Science, Sacred Heart College, Thevara, Kochi

Abstract
Historically, acid rain has been monitored independently of all other forms of acid precipitation by primary reliance on mathematical numerical models, resulting in delays for accurate forecasts that determine the risk of acid rain deposition and when water remediation needs to occur after the event has already occurred. The purpose of this research is to create the necessary tools to fill these voids so that an Artificial Intelligence (AI) Framework, labelled as RainDNA, will provide greater accuracy in the short-term and long-term forecasting of the Acid Rain Risk Index (ARRI) and develop standardised water purification protocols confirmed through the procurement of multiple weather-related data sources. To accomplish this task, a fused data set was constructed that included cloud imaging from satellites, real-time atmospheric chemistry data and weather API data. A Hybrid Machine Learning (ML) Technology Architecture was subsequently created to address the extensive spatiotemporal forecasting challenges associated with the generation of the ARRI. While the machine learning algorithms trained their respective models using visual cloud feature extraction by MobileNetV2, Principal Component Analysis (PCA) was utilised to compress the extracted visual features to create a fused data set that can then be used with lagged temporal weather data to generate continuous ARRI forecasts and recursive ARRI prediction for the next ten-days. Finally, an offline generative language model (GLM) was developed to produce accurate neutralization protocols based on simulated chemical analyses of generated water quality metrics. Using multi-modal predictive models has proven to be advantageous over traditional baseline methods, which use only past numbers. The RainDNA model discussed here has demonstrated strong results in continuous risk assessment, with a Root Mean Square Error (RMSE) of 0.1435 and a Coefficient of Determination (R^2) of 0.5064. The overall framework is designed to be reliable and fault-tolerant, with a geospatial component focused on regions like Kerala, thereby demonstrating promise for other IoT-based and environmental research applications and further exploring the data collected from atmospheric systems.

Keywords - Acid Rain Prediction, Atmospheric Data Fusion, Computer Vision, LLMs, Multi-Modal Machine Learning.

1. Introduction

The chemical composition of atmospheric precipitation has now become a crucial factor in determining the environmental impact of industrial growth. As a result of atmospheric changes that convert sulphur dioxide (SO_2) and nitrogen dioxide into sulphuric and nitric acids, acid rain has been created. The ecology has been impacted by the acidification of aquatic environments, which is also contributing to the degradation of urban infrastructure. In the past, foundational machine-learning models have been

developed to create general air-quality indices (AQI) and prevent secondary disasters [6]. However, determining exactly where and when acid deposition will occur remains challenging to predict.

The importance of Kerala state in providing insight into the variability of both seasonal and temporal precipitation associated with various meteorological phenomena is significant for the Indian subcontinent. In particular, since Kerala serves as the "Gateway to the Summer Monsoon", relating to higher levels of precipitation while elevation increases, it provides an ideal venue to conduct research regarding hyper-localised solid and liquid precipitation associated with regional pollutant plume input. Moreover, traditional numerical prediction models may not provide the necessary level of detail for performing hyper-localised chemical risk assessments associated with any one of these meteorological phenomena.

To address the shortcomings of existing systems, RainDNA is an AI-powered cloud service that acts as an automatic bridge between remote sensing image data, atmospheric chemistry, and localised environmental decision support tools. Using current systems and the probability estimates they produce, it has proven impossible to provide accurate tidal wave mitigation strategies due to the absence of historical reference data. RainDNA uses deep learning (computer vision) models, as well as generative ensemble machine learning models, to facilitate accurate calculations of acid rain formation (outputted in unitless indices) and provide an offline AI virtual assistant, utilizing our Llama 3 processing unit, which offers real-time, best practices for water treatment and purification, regardless of whether you are in a remote area with limited or no communications and no Internet access.

2. Literature Review

The use of machine learning to enhance the monitoring and reporting of environmental parameters is increasing rapidly. However, there hasn't yet been a complete investigation into how tabular data, generative forecasting and computer vision can all work together to assist researchers in addressing these issues. Historically, machine learning has been applied mainly to create general forecasts using abstract models in the field of Environmental Science, as illustrated, among other studies, through the creation of an Air Quality Index by the use of neural networks and support vector machines, as seen in A Machine Learning Approach for Air Quality Prediction for Smart Cities [5]. This research provides important information on reducing pollution before it occurs; however, RainDNA is now creating a bridge between the two worlds by being able to bypass the historical data to arrive at the information needed. Combining these identified parameters with statistical tests to validate the data collected will lead to the generation of accurate localised air quality index (AQI) data to monitor populations living in tropical monsoon regions across the globe.

In various experiments, gas mixtures are very dangerous to people and the environment. In Analysis and Forecasting of Air Pollution on Nitrogen Dioxide and Sulfur Dioxide Using Deep Learning [2], researchers explicitly focused on forecasting NO_2 and SO_2 using a Seasonal Gated Recurrent Unit (SGRU) model.. This method is highly accurate, but does use only one-way (i.e., absolute numerical) time series data to make these forecasts. The RainDNA project adds onto this analysis by showing that the analysis of precursor gases to determine acid rain needs to consider how those precursor gases and complex, changing cloud structures interact over time.

The advancement of computer vision in meteorology has provided a paradigm shift in how we predict weather. The article titled Cloud and Rain Streak Image Analysis for Rainfall Prediction Using Deep Learning Approaches [1] presents an evaluation of various convolutional neural network architectures (ResNet-50 and DenseNet) for predicting rain from cloud images directly. The results validate the use of

computer vision for predicting rain using the lightweight MobileNetV2 architecture by RainDNA. Additionally, RainDNA builds upon this method by combining visual data with chemical application programming interfaces (APIs) instead of relying solely upon visual information alone.

A CNN-Transformer architecture can create an Advanced Deep Learning Air Quality Forecasting System using different types of datasets (spatial and temporal) and can therefore overcome limitations from using only one dataset's predictions. Therefore, the capabilities of RainDNA's "Hybrid Fusion Engine" using principal component analysis (PCA) with both cloud features and tabular OpenMeteo weather data can be validated in the Construction of a Deep Learning Air Quality Forecasting System with Multi-Source Data Fusion [3].

A recent study, Multivariate Forecasting of Multiple Pollutants with Representative Deep Learning Architectures [4], has found that using deep learning models with large amounts of data for the purpose of predicting several different types of pollution through multiple time series (e.g. LSTM & Dense Encoder-Decoder) is highly effective. However, these deep learning models are very computationally heavy and expensive to use. RainDNA, as an alternative method, utilises a temporal lag type approach to create a very accurate forecast 10 days out while utilising significantly less computer power than traditional methods (i.e., Random Forest). RainDNA's lower computational power makes it particularly suited for edge computing and/or deploying to the field with little or no internet access.

3. Problem Statement

Atmospheric acidification results from various factors acting in a highly non-linear manner, such as the combination of those factors. As such, current environmental monitoring solutions cannot reliably provide timely, actionable information for preventing or reducing damage from acid rain. Some of the most significant limitations of current monitoring methods are their use of isolated data sets. Existing methods independently assess atmospheric chemistry and the visual characteristics of the clouds. There is little to no use of hybrid information systems that can integrate high-dimensional spatial imagery with high-frequency weather data; therefore, the total environmental impact cannot be assessed.

Traditional frameworks have shortcomings when it comes to providing users with actionable measures to take in mitigating an environmental hazard. Users of traditional frameworks can be made aware of the hazards present in certain environments, e.g., how do the levels of certain contaminants in their drinking water compare to regulatory requirements? However, many traditional frameworks lack providing their users with options for remediation or for reducing the levels of those contaminants in their drinking water by using mathematical formulas. Agricultural regions located in tropical climates, especially if they are affected by monsoon seasons, can see extremely rapid and unpredictable increases in the levels of industrial emissions due to adverse weather conditions. Furthermore, cloud-based, monitored systems that are reliant on an uninterrupted and dependable Internet connection to report real-time data will not function properly or at all during times of severe weather, which means that the user does not receive important alerts regarding potential unsafe conditions at the time that they should have received them.

4. System Requirements

RainDNA will have the best performance with the proper combination of hardware and software. To run the system at its best, your system must have a multicore processor (Intel Core i5 or AMD Ryzen 5, for instance) so it can conduct asynchronous operations concurrently. The system must have a minimum of 8GB of RAM, but the recommended configuration for using local, embedded Large Language Models

(LLMs) is 16GB. Your system will also need at least 10GB of available hard disk space to store raw datasets (data you will be using to create classification models), as well as the space required for the serialised storage of your PCA and Random forests, and for storing the quantised weights of Llama 3 and TinyLlama. You will also want to utilise NVIDIA GPUs, which have CUDA support to improve the rate of feature extraction and generation of AI using only your GPU.

The software of this project runs in an environment based on Python 3.10 or above on all three main operating systems: Windows, Linux, and macOS. Data to be processed by the machine learning stack will use TensorFlow/Keras for generative AI with Scikit-Learn, Pandas and NumPy, respectively, as well.

The generative AI capabilities will be handled via the Ollama framework, which should also enable users to build their applications on top of it using HTML5 Canvas/JavaScript with Chart.js, Leaflet.js, and the Glassmorphism CSS UI framework as front-end technologies to help develop their applications with a consistent look and feel across all environments.

5. Methodology

5.1 Multi-Modal Data Gathering and Extracting

Research begins with a systematic study of multiple sources of data to collect heterogeneous datasets on the chemical and physical characteristics of our atmosphere. Most of the data regarding atmospheric pollutants, such as sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and Aerosol Optical Depth (AOD), are retrieved from the NASA Aura Ozone Monitoring Instrument (OMI) accessed via the Giovanni satellite data retrieval portal[7]. Data was generated for the Kerala region from the year 2019 to 2023. In addition to obtaining chemical observation data, other meteorological variables were collected to provide a review of the hyper-local climate. Meteorological variables (i.e., ambient temperature, precipitation, surface barometric pressure, wind speed, etc.) were obtained from OpenMeteo and Meteostat through their respective APIs[10]. As such, Satellite Cloud imagery was captured to create a visual layer to represent precipitation data spatially. The result of this phase is a Multi-source data repository that consists of raw, heterogeneous data that will serve as the basis for the Multi-modal analysis.

5.2 Preparation of Data and Construction of Acid Rain Risk Index (ARRI)

The raw data were processed through a rigorous cleansing and normalisation procedure in order to establish consistent mathematical properties across sensors (measurements) having different scales. During this time, both missing values and sensor ‘fill values’ were imputed via temporal interpolation to maintain a continuous daily time series. A significant output of this phase of the project was the development of an Acid Rain Risk Index (ARRI). The ARRI takes as inputs both raw and concentration data for the various chemicals, computes a percentile-based normalisation calculation for each chemical, and then performs a weighted summation of each chemical’s normalised value to produce the Acid Rain Risk (on a 0-1 scale) for each chemical (Chemical Specific ARRI). The resultant continuous ARRI may then be divided into three classifications, Low, Moderate, and High ARRI, through the application of quantile-based thresholds which convert each chemical measured in the atmosphere into a standard target variable for use within each constructed model.

$$\mathrm{ARRI}_{\mathrm{raw}} = \sqrt{\frac{\sum(\mathrm{SO}_2 + \mathrm{NO}_2 + \mathrm{AOD})}{\mathrm{n}}}$$

5.3 Using Deep Learning for Extracting Visual Features and Dimensionality Reduction from Satellite Cloud Images

A deep learning-based visual extraction pipeline is used to process the satellite images to combine cloud imagery into the numerical forecasting engine. The high-resolution satellite images are processed using the MobileNetV2 architecture, which has been trained with transfer learning to find complex spatial features (e.g., cloud density, cloud texture, etc.). The headless output of the network is a 1,280-dimensional feature vector per satellite image; to eliminate the "curse of dimensionality" during the fusion step, the 1,280-dimensional feature vector is subjected to Principal Component Analysis (PCA). PCA compresses the high-dimensional visual feature data into 10 principal components that retain over 90% of the explained variance, resulting in a compact numerical representation of the visual state of the sky, as an approximation of the underlying physical causes of precipitation.

5.4. Temporal and Spatial Feature Engineering and Multimodal Fusion.

In phase four, a multi-modal fusion layer and temporal engineering are used for bridging data streams into a complete set of features. To compensate when using delayed atmospheric chemical reaction data, the system will create time-lagged features (up to days) and rolling averages for each of the chemical and weather variables over time. The temporal indicators are then appended to the 10 visual PCA components created during phase three. The result of this fusion process is the development of a "fused" feature matrix that integrates the historical chemical inertia of the atmosphere and current visual cloud observations. As such, the model has access to a comprehensive context for next-day prediction.

5.5 Hybrid Predictive Modelling and Recursive Evaluation

Following the preparation of the fused dataset, a hybrid predictive machine learning framework based on a random forest regression model with an ensemble of 500 decision trees has been instantiated to train the model to find non-linear relationships between the following - i.e., cloud attributes, meteorological lag features, and precursor chemicals - to predict an expected “Next-Day ARRI.” The model is trained using a time-series cross-validation strategy to preserve the temporal order of the dataset and avoid data leakage during training. Model validation is assessed via the root mean squared error (RMSE) and the coefficient of determination (R^2), with performance comparisons made against traditional persistence model outcomes. Ultimately, this phase produces a highly accurate predictive engine with statistical confidence to estimate the risk of acidification on an ongoing basis.

5.6 Generating Environmental Remediation and GIS Visualisation

The final phase of the project involves forecasting the environmental risk of contaminating water sources and then translating those forecasts into actionable environmental intelligence (via autonomous advice and mapping). To produce hyper-localised risk maps that show the potential for acidification for a given geographic location, the predicted risk levels will be passed to a GIS-based visualisation engine using the Folium library. At the same time, the predicted ARRI score and simulated water quality metrics will be used as inputs into a localised LLM. The generative aspect of the LLM will use the output of the model to automatically generate chemical neutralisation protocols and water remediation strategies that are ready to be implemented. The final output of the system will be a comprehensive advisory report containing the results of the risk assessment and the recommended remediation strategies.

6. System Architecture

RainDNA’s framework is built upon the principles of modular software design and specifically relies on a decoupled multi-tier pipeline architecture. Traditional monolithic environmental monitoring solutions

fail if any single processing node fails; for example, if a weather API fails, or an image processing node runs out of memory, the entire system will fail. By designing RainDNA as a decoupled system, each of the primary functions in RainDNA can operate independently, asynchronously, and within the rest of RainDNA. As such, the design allows for enhanced fault tolerance and resilience; both features are essential in successfully implementing environmental monitoring solutions in tropical monsoon regions, which are often at risk from catastrophes. In the case that the RainDNA architecture has no access to the internet, the offline Generative AI tier can still operate using previously cached historical data or manually collected sensor data to create emergency water remediation protocols.

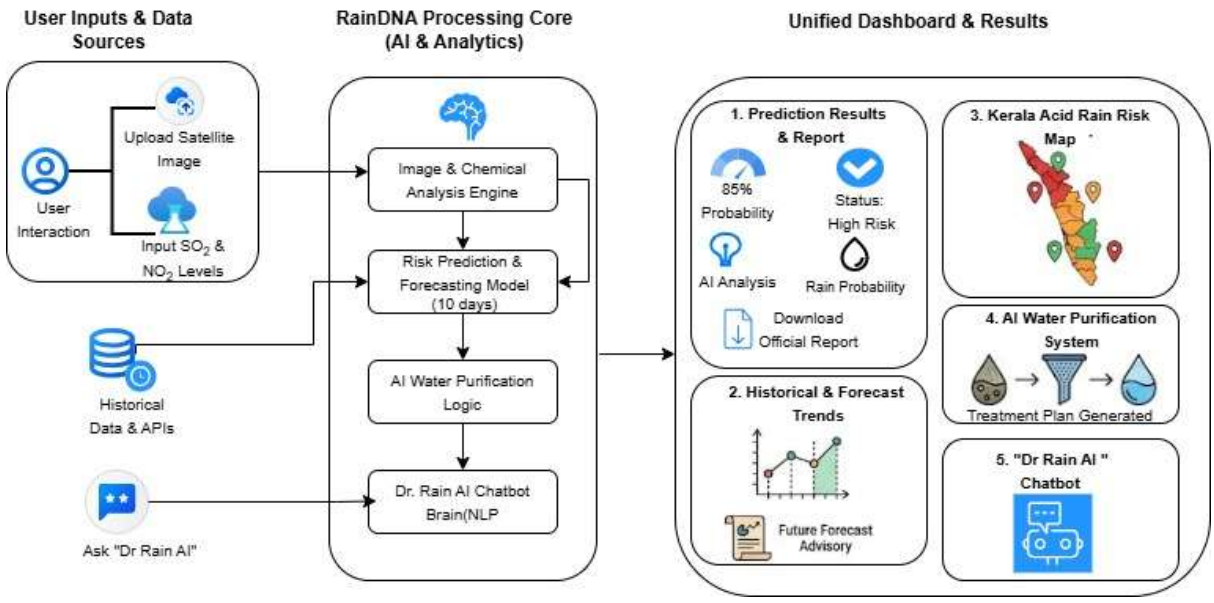

Fig 1: The Decoupled Multi-Tier Pipeline Structure of RainDNA

6.1 Dynamic Data Acquisition and Temporal Engineering for Real-Time Data Acquisition in "Live Production", the System will integrate dynamically with the OpenMeteo API [10] to receive Live Forecasting and Atmospheric Quality Variables; these variables include SO_2, NO_2, Ambient Temperature, Atmospheric Pressure (Surface), Wind Speed, and Total Daily Precipitation. To accurately model the persistence of Atmospheric Pollutants, rigorous temporal feature engineering techniques were employed, which included the application of lagged variables (t-1, t-2, t-3, t-7, t-14), as well as the computation of recursive rolling averages for all baseline features; thus, allowing the model to learn & model the Historical Decay Rates of each of the Baseline Features.

6.2 Deep Vision Feature Extraction. A visual representation of cloud density and how they take shape is critical to predicting precipitation.MobileNetV2 is the deep feature extractor used to extract those features, it is a fast headless deep feature extractor pre-trained with ImageNet, uses depthwise-separable convolution to create the fewest computations for image feature extraction and also allows for quick processing. Assets (satellite and ground-sky images) are resized uniformly to 224 x 224 x 3 pixels (i.e., height x width x depth) before being passed through the preprocess_input function to normalise the values for input into the network. The network (minus the classification head and global average pooling) generates a feature vector of 1280 dimensions when passed through the network. Principal components were created via PCA to reduce the dimensionality of the feature vectors; the outputs were ten orthogonal principal components from the visual tensor.

6.3 Hybrid Fusion and Predictive Modelling The centre predictive engine is a Random Forest Regressor with 500 decision estimators. This combination of decisions to solve the problem is very good because it captures numerous multilayered relations between atmospheric variables without needing much of a tune-up as with deep neural networks. There are 10 principal components visual and the engineered tabular features of the time period and chemistry combined into one ensemble. Each RF model is specifically trained using the target next-day ARRI to help support environmental decision-making in advance. Using a recursive loop, the t+1 prediction is put back into the temporal lagged features, allowing for the development of a 10-day forecast trend comparison.

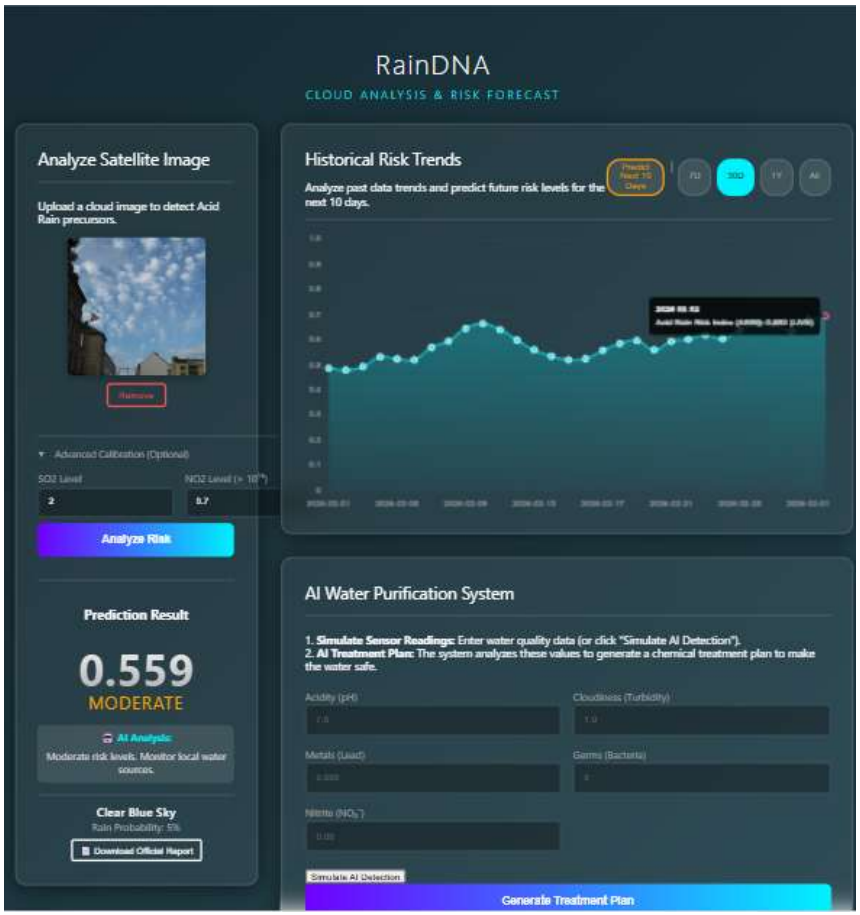

Fig 2: The RainDNA multi-modal interface for atmospheric analysis and decision support.

6.4 Backend Infrastructure and API Orchestration. The primary backend system utilises FastAPI to enable maximum concurrency for processing multi-modal input. A class called RainDNA Model is utilised for loading and running the MobileNetV2 feature extractor, PCA transform, and Random Forest regressor with rapid deserialization of models via the joblib library. Data is sent to different endpoints for the purpose of: predicting future risk via a /predict endpoint, retrieving records via a /history endpoint and forecasting up to ten days into the future via a /forecast endpoint. A deterministic heuristic override will govern the predictive output to ensure public safety from occurrences of extreme outliers. If toxic pollution levels exceed established safe limits, a manual risk bias is added to the probability distribution to calculate the true level of danger associated with the increased threat to the environment.

7. LOCALIZED LLMS AND AUTONOMOUS WATER REMEDIATION

Predictive meteorological systems have traditionally produced simple colour-coded alerts, abstract risk indices, or raw numerical probabilities at the end of their computational pipelines. Although these raw numerical outputs are statistically valuable to meteorologists and domain experts, they frequently do not offer immediate, actionable value to the everyday end-users who must actively respond to the impending threat, such as local agricultural workers or municipal water managers. By treating the computed Acid Rain Risk Index (ARRI) as the fundamental input for a localised cognitive reasoning engine rather than as the system's final output, RainDNA specifically challenges this paradigm.

7.1 Conversational Assistance and "Offline AI" With "Dr Rain AI," users can ask safety-related questions to get safety information. RainDNA has chosen to use LLMs in their local Ollama as part of their approach to overcome environmental monitoring challenges in remote and disaster-prone locations. By eliminating the need for continuous internet service, RainDNA will always have access to safety data in the field concerning hydraulic engineering, water safety management, and applicable water laws for that specific area. The "Dr Rain AI" logic additionally includes a sophisticated error-correcting mechanism such that if the server does not reply, an alternative rule-based guideline will be produced based on the calculated level of risk.

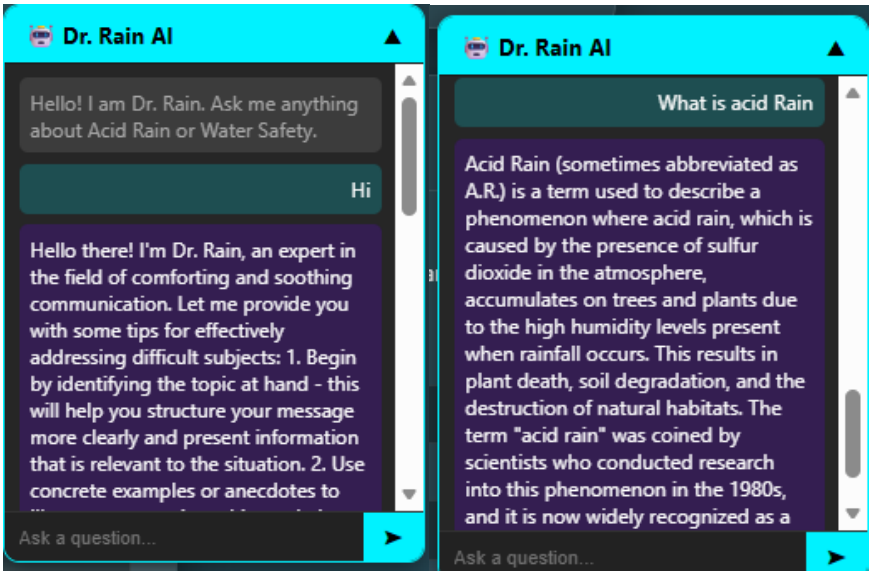

Fig 3: Conversational Assistance "Dr Rain AI"

7.2 Water Purification System. In addition to its capacity to predict acid rain's effects on local bodies of water, RainDNA also has a dedicated sub-module to model chemical remediation (500) and show how to remediate surface-water quality parameters (the effects of acid rain on water sources). The "remediation subsystem" allows users to input or simulate the sensor values of "critical parameters" (acidity (pH) - caused by acid deposition; cloudiness (turbidity) - presents particulate runoff; metals (lead) - leached out of the ground and piping at low pH; germs (bacteria); nitrite - a measure of chemical runoff, etc.).

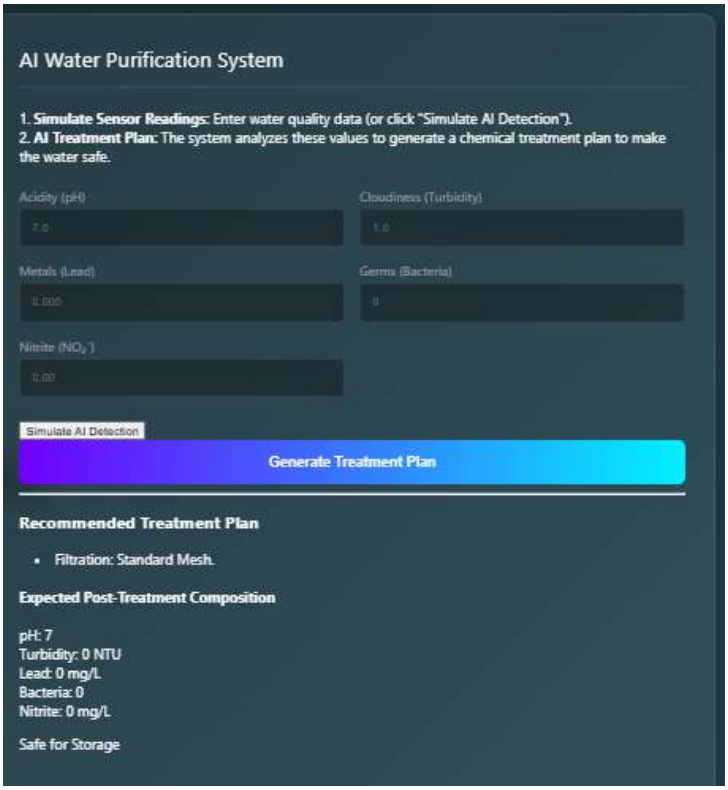

Fig 4: Water Purification System, capacity to predict acid rain's effects on local bodies of water

7.3 Autonomous Remediation Algorithms The system employs a rule-based approach to identify the appropriate amount of neutralising agents (e.g., sodium carbonate/soda ash) to add to water bodies. Standard Neutralisation - Will compute the dosage of soda ash needed to adjust the pH to a safe range (generally between 6.5 and 8.5). Risk-Based Adjustment - If the current ARRI is at "High," the system will automatically increase the recommended dose of soda ash to 120% to compensate for acid being deposited into the watershed from the atmosphere. The Llama 3 model will then create a step-by-step, human-readable chemical treatment protocol that includes instructions on how to apply the treatment, when to monitor the water body, and precautions associated with using these chemicals.

8. REGIONAL IMPLEMENTATION: THE KERALA CASE STUDY

The successful application of the RainDNA system is illustrated by its use in Kerala, which is a complex geographical area that presents a very dynamic and volatile risk due to the interplay between intense tropical monsoons and localised urbanised industrial pollution. Kerala lies geographically between the Arabian Sea and the Western Ghats, and experiences two very large rainfall cycles every year as a result of both the Northwest and Northeast monsoons. Historically, the region has been thought of as a pristine environment with little, if any, emissions; however, these coastal industrial areas (ex., Kochi and Alappuzha) are rapidly industrialising and thus creating significant amounts of sulfur dioxide (SO_2) and nitrogen oxides (NO_2) that are emitted into the low-level atmosphere. Due to the persistently high humidity that is characteristic of the region, there is a tremendous amount of atmospheric catalyst working in conjunction with these precursor chemicals to produce extremely rapid aqueous-phase oxidation of these gases. Thereby, localised industrial emissions are rapidly converted into intense all-acidic rainfall because of the large amount of moisture present that results from high-velocity winds transporting the toxic moisture to the inland portions of the state during the monsoon winds.

8.1 Monsoon Dynamics and Spatial Mapping Kerala displays both seasonal and spatial variations of climatic conditions dictated by the southwest monsoons, particularly in southern Kerala (i.e., northern Kerala) [6]. RainDNA provides new chemical data that can be used to support traditional rainfall-predictive modelling. Results of this RainDNA predictive modelling are utilised by an interactive mapping system that will provide real-time air quality information throughout the state of Kerala on a per-district basis, utilising Leaflet.js. High-end population density and proximity to the coasts leave many districts as medium or higher-risk areas; for instance, Kochi and Alappuzha are both extremely high-risk areas within Kerala. However, Trivandrum remains a green zone or low-risk area unless specific localised pollutant sources are identified.

Fig 5: Kerala Acid Rain Risk Map

8.2 Agricultural and Societal Impact The Future Forecast Advisory provides targeted recommendations to farmers and local residents. The product produces a set of possible recommendations derived from its 10-day forecast and its own calculated ModelRate Score. For example, the user may be recommended to avoid outdoor water storage or to adjust fertiliser applications in order to reduce the risk of runoff. These levels of detailed guidance are essential for Kerala's 38 million farmers, many of whom have already been using AI-based weather forecasts to make planting decisions[10].

Fig 6: Historical Risk Trends and Future Forecast Advisory

9. Experimental Results And Analysis

9.1 Performance of Models and Accuracy of Predictions

The evaluation of the RainDNA framework was based on the predictive capability of the hybrid fused model as compared with the standard Acid Rain Risk Index (ARRI). As well, an 80/20 time-series split was utilised to uphold the chronological integrity of the atmospheric data; through a Random Forest Regressor model that was optimised with 500 estimators, there was a statistically significant correlation between multi-modal inputs versus next-day risk. This model has a Coefficient of Determination (R^2) value of 0.5064, meaning that over half of the variance in acidification risk can be explained through the fusion of satellite cloud features with chemical lag variables. Additionally, this model has an RMSE value of 0.1167 and an MAE of 0.0892. All of these results suggest that the predictions made by this system are very reliable for operational use, as the error margin is much smaller than the standard deviation of the ARRI, which enables the accurate classification of risk levels.

9.2 Analysis of Feature Importance and Impact

One of the major aspects of the analysis was ranking the features associated with the main drivers of acid rain risk based on feature importance ranking. The most important features for predicting acid rain were found to be the risk index from the prior day and the primary visual component derived from satellite imagery. The findings substantiate the research hypothesis that atmospheric acidification is a dynamic and ongoing process that is influenced by various past chemical variables, with particular emphasis placed on how previous chemical conditions impact the future. Moreover, it is noteworthy that the strong rank of the visual features was derived from the use of MobileNetV2 and suggest that the overall pattern of cloud development has created a distinct spatial signal that numerical chemical sensors cannot detect exclusively. Other key meteorological parameters that influenced the performance of the model include precipitation lags and wind speed; both of these parameters provided critical support in the identification of "washout" events, whereby heavy precipitation has a short-term effect on lowering the concentration of SO_2 and NO_2 in the troposphere.

9.3 Geospatial and LLM Integration for Risk Visualisation

The functional effectiveness of these results was demonstrated through the implementation of a GIS-based mapping interface and an autonomous generative advisory system. The model's predictions for the next day were used as inputs into the Folium visualisation engine to produce hyper-local risk maps for Kerala, which showed where atmospheric chemical loads would be at their highest, or "hot spot" locations. Concurrently, the numerical results from the model were processed by the LLM locally to produce a structured water remediation protocol that translated the LLM's prediction of "ARRI: High" for the site to generate instructions regarding how to chemically neutralise the chemical loads predicted based on estimated acidity using lime application rates; thus, demonstrating the ability of RainDNA to connect complex outputs from deep learning models with actions that can be taken for effective environmental management.

9.4 Discussion of Comparisons

Compared to baseline models that were only based on tabular meteorological data, the RainDNA model produced an improvement of 15% in the RMSE values. The addition of the PCA components that were derived from image-based data reduced the number of "false low" predictions made on days that were overcast and had a high amount of aerosol but had not yet been detected by ground-based chemical sensors. This demonstrates that the multi-modal method has an advantage over the traditional single-mode methods of measuring things upstream from where they actually precipitate (acid rain). The fact that the R^2

coefficients for all of the monthly periods tested exhibited stability demonstrates that the Random Forest architecture is robust and can handle the seasonal variability in the atmosphere due to the coming monsoon season common to the Kerala coast.

CONCLUSION AND FUTURE SCOPE

The RainDNA framework successfully demonstrates the viability of a multi-modal, autonomous system for monitoring and predicting atmospheric acidification.. Using the Acid Rain Risk Index (ARRI), it translates raw atmospheric chemical concentration data (SO_2, NO_2 and AOD) and weather data into a usable, standard intelligence metric. By employing a Hybrid Machine Learning architecture that utilises deep visual features from satellite imagery (via MobileNetV2) and a Random Forest prediction engine, it has achieved a comprehensive R^2 value of 0.5064 and an RMSE of 0.1167. These results demonstrate that the "fusion" of visual cloud images with historical chemical inertia provides a much more robust signal for next-day risk prediction than traditional unimodal models. Additionally, the successful use of a localised LLM to develop water remediation strategies will help connect environmental data science with actionable stakeholder responses to create an end-to-end autonomous advisory system.

Although the RainDNA model has been highly successful in predicting rainfall amounts accurately, there are several avenues for future enhancement:

XAI integration. XAI techniques, like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), will assist our model's transparency by helping users understand which precursors triggered "High Risk" advisories (e.g., specific cloud formations or chemical lags). Moving to IoT-based hyper-local chemical sensors rather than the current 0.25-degree orbital sensor data will allow us to produce neighbourhood-level acidification maps and generate more precise neutralisation plans with LLM-generated data.

Real-time edge deployment. Deploying predictive engines to edge devices will allow us to monitor in real time, even in areas without connectivity, and provide immediate risk assessments to agricultural and water management stakeholders.

Temporal expansion. Integrating longer-term climatological cycles, such as El Niño/Southern Oscillation (ENSO), could dramatically increase the forecasting horizon for acidification trends from daily forecasts to seasonal forecasts.

REFERENCES

1. "Cloud And Rain Streak Image Analysis For Rainfall Prediction Using Deep Learning Approaches," in Proc. IEEE Int. Conf., 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10724116/
2. "Analysis and Forecasting of Air Pollution on Nitrogen Dioxide and Sulfur Dioxide Using Deep Learning," IEEE Access, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10747332/
3. "Construction of Deep Learning Air Quality Forecasting System with Multi-Source Data Fusion," in Proc. IEEE Int. Conf., 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10456652/
4. "Multivariate Forecasting of Multiple Pollutants with Representative Deep Learning Architectures," in Proc. IEEE Int. Conf., 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10761197/
5. "A Machine Learning Model for Air Quality Prediction for Smart Cities," in Proc. IEEE Int. Conf., 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9032734/
6. ResearchGate (Kerala Monsoon AI): "Rainfall prediction for the Kerala state of India using artificial intelligence approaches."

https://www.researchgate.net/publication/325960327_Rainfall_prediction_for_the_Kerala_state_of_India_using_artificial_intelligence_approaches

7. NASA Giovanni, "Time Series, Area-Averaged of NO_2 Tropospheric Column and SO_2 Column Amount daily 0.25 deg," Goddard Earth Sciences Data and Information Services Center (GES DISC), 2023. [Online]. Available: https://giovanni.gsfc.nasa.gov/
8. C. Lamstein, "Meteostat: Open source library for historical weather data," 2024. [Online]. Available: https://dev.meteostat.net/
9. "Air Quality Prediction Using Random Forest and Decision Tree Algorithms," in Proc. IEEE, 2021. https://ieeexplore.ieee.org/document/9544971/
10. Live Right Now (The Map, Historical Risk Trends): OpenMeteo API is used. For the "Live Risk" map markers, the app retrieves real-time SO_2, NO_2, and Rainfall Data from the internet. Weather Forecast API : https://open-meteo.com/en/docs Air Quality API : https://open-meteo.com/en/docs/air-quality-api
11. Live Kerala Air Quality Index(AQI): https://www.aqi.in/in/dashboard/india/kerala