iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

DISEASE OUTBREAK PREDICTION USING MACHINE LEARNING:AN AI-POWERED EARLY WARNING SYSTEM FOR ACUTE DIARRHOEAL DISEASE

M. Sion Kumari
Assistant Professor, Department of Information Technology and Computer Applications,
Andhra University College of Engineering for Women, Visakhapatnam, Andhra Pradesh, India

D.Pujitha, D.Swetha, D.Tejaswi , D.Ramya Sai Lakshmi
B. Tech Final Year, Information Technology and Computer Applications,
Andhra University College of Engineering for women, Visakhapatnam, Andhra Pradesh, India

ABSTRACT
Infectious disease outbreaks, particularly Acute Diarrhoeal Disease, remain a serious burden on public health systems across India, especially in districts with weak sanitation and variable monsoon patterns. Early and accurate prediction of outbreak risk can help health authorities deploy resources before case counts spike. This paper presents a machine learning-based disease outbreak prediction platform that uses 14 years of historical epidemiological records combined with environmental parameters — temperature, precipitation, and Leaf Area Index — to forecast weekly case counts at the district level. The dataset covers more than 5,000 weekly records across Indian states. A Gradient Boosting Regressor, trained on lag features and climate variables, was selected as the primary model after comparison with Random Forest and LSTM baselines. The system achieved a Mean Absolute Percentage Error of approximately 1.25% on the held-out test set, with an R^2 score of 0.87 indicating strong goodness of fit. A Streamlit-based web dashboard provides health workers with location-specific risk levels (High, Medium, Low), trend visualizations, and feature importance explanations. The platform is designed to be usable by non-technical public health staff without any programming knowledge.

Keywords:
Disease outbreak prediction; machine learning; Gradient Boosting; Acute Diarrhoeal Disease; environmental factors; public health; Streamlit; early warning system; India

I. INTRODUCTION

Acute Diarrhoeal Disease is one of the leading causes of preventable morbidity in India, particularly in districts with limited sanitation infrastructure and variable monsoon patterns. The disease is strongly linked to environmental conditions — heavy rainfall, temperature shifts, and changes in vegetation cover can all accelerate transmission within days of onset. Traditional surveillance methods, which rely on manual weekly reporting by district health officers, often detect outbreaks only after case counts have already crossed the threshold for intervention. By that point, the window for early containment has largely closed.

Machine learning offers a different approach. Rather than waiting for cases to accumulate before raising an alarm, predictive models can learn from years of historical outbreak patterns combined with real-time environmental measurements to flag districts at elevated risk before cases appear in the reports. Early work in this space used time series methods: Buendia and Solano [3] applied ARMA models for outbreak detection in the Philippines, while Akter et al. [4] used ARIMA on electronic medical records to forecast dengue in Bangladesh, noting that prediction accuracy improved directly with the volume of historical data. Singhal et al. [2] showed that Random Forest achieved 96.1% mean accuracy across multiple clinical disease datasets, outperforming Decision Trees, SVMs, and multilayer perceptrons, and documented the broader shift toward satellite imagery, social media signals, and environmental sensors as data sources. Dubey et al. [1] extended this with Deep Q-Learning and Policy Gradient methods, reporting 91% accuracy in urban settings and 82% in rural areas, with the model adapting its strategies as disease dynamics changed. Lalmuanawma et al. [5] reviewed ML applications for COVID-19 and found that models combining clinical, epidemiological, and imaging data improved early detection significantly. Ginsberg et al. [6] established earlier that search engine query data alone could proxy influenza activity in real time.

A consistent gap across this literature is the distance between offline model performance and practical deployment. Most published systems stop at a metrics table. Tools built specifically for district-level health workers in low-

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

resource Indian settings — combining environmental variables with temporal disease data and accessible through a browser — are not well represented. District-level prediction for Acute Diarrhoeal Disease in the Indian context remains particularly underexplored. This work attempts to fill that gap.

This paper describes a prediction platform built on a 14-year weekly disease surveillance dataset covering multiple Indian states. The system learns temporal dependencies through lag features representing previous week and previous month case counts, and combines these with climate variables drawn from satellite data. The model outputs a predicted case count for any selected district and date, mapped to a three-tier risk level that health workers can act on without any statistical background. The web-based interface, built with Streamlit, also shows a breakdown of historical cases across six major diseases at the selected location, a time-series chart comparing historical trends against the prediction, and a ranked list of the features that most influenced the result. The goal is to give district-level public health staff a practical tool, not just a research prototype.

II. METHODOLOGY

The system follows a standard MLOps pipeline covering data ingestion, preprocessing, feature engineering, model training, and deployment through a web interface.

2.1 Dataset

The primary dataset, Final_data.csv, contains weekly disease surveillance records collected from across Indian states spanning 2009 to 2022. Each record includes the reporting state and district, the disease name, case count, deaths, date fields (day, month, year), geographic coordinates (latitude and longitude), and three satellite-derived environmental variables: surface temperature in Kelvin, weekly precipitation in millimetres, and Leaf Area Index (LAI). After removing duplicate disease name variants and normalizing labels — for example, mapping "Acute Gastroenteritis" and "Gastroenteritis" to "Acute Diarrhoeal Disease" — the dataset covers records for six major diseases: Acute Diarrhoeal Disease, Dengue, Chikungunya, Cholera, Malaria, and Acute Encephalitis Syndrome. Acute Diarrhoeal Disease was selected as the prediction target. It had the largest sample of records at 5,126 and showed consistent temporal coverage across all 14 years, making it the strongest candidate for a weekly forecasting model.

2.2 Preprocessing

Missing values in numeric columns were filled with zero. LAI values above 10.0 were clipped to remove satellite calibration outliers. All records with non-numeric entries in case counts or date fields were coerced to valid numeric types before training.

Climate variables (temperature, precipitation, LAI) were scaled using StandardScaler fitted on the training split and saved to disk so that the live inference pipeline applies the same transformation the model was trained on.

Categorical variables — state and district — were encoded using LabelEncoder. The encoders were serialized as part of the model pipeline so that the dashboard can accept string inputs from users and convert them correctly at inference time.

2.3 Feature Engineering

Two lag features were created to capture temporal dependencies. The first, cases_last_week, records the case count from the most recent available week for the selected district and disease. The second, cases_last_month, is the average of the four preceding weekly counts. Prior outbreak patterns turned out to be the most predictive features in the trained model, consistent with the epidemiological understanding that disease burden in one week strongly correlates with the burden the previous week.

The full feature set passed to the model contains twelve variables: day, month, year, latitude, longitude, scaled temperature, scaled precipitation, scaled LAI, cases_last_week, cases_last_month, encoded state, and encoded district.

2.4 Model Training

Three models were trained and compared: Random Forest Regressor, Gradient Boosting Regressor, and an LSTM network.

Training used a time-based split, with records from 2009 to 2018 used for training and records from 2019 onward held out for testing. This split avoids data leakage and better reflects the real deployment scenario where the model must generalize to future outbreaks rather than interpolate within a shuffled dataset.

Case counts were log-transformed using log1p before training to reduce the effect of extreme outbreak spikes on model loss. Predictions were back-transformed using expm1 before display on the dashboard.

The Gradient Boosting model was selected for deployment based on its test set performance — it produced the lowest RMSE and MAE and a strong R^2 score on the held-out data. The LSTM was implemented and is available

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

in model_training.py for use in environments with full TensorFlow support (Linux or Google Colab), but the Gradient Boosting model was used for the Windows deployment due to TensorFlow Long Path constraints.

2.5 Risk Classification

Predicted case counts are mapped to one of three risk tiers on the dashboard. A prediction above 100 cases triggers a High Risk alert, indicating that immediate public health intervention is warranted. Predictions between 50 and 100 cases are marked Medium Risk, indicating that resources should be prepared and the situation monitored closely. Predictions below 50 cases are classified as Low Risk, consistent with standard surveillance protocols.

III. SYSTEM WORKFLOW

The platform operates in two modes: training and inference.

During training, model_training.py reads Final_data.csv, runs the preprocessing and feature engineering steps described above, trains the three candidate models, evaluates each on the held-out test split, and saves the best-performing pipeline as best_disease_model.pkl. This file contains the trained model, the fitted StandardScaler, both LabelEncoders, the feature column list, and the recorded evaluation metrics.

During inference, app.py loads the serialized pipeline and the full historical dataset. The user selects a state, district, and target date from the sidebar. Environmental parameters are auto-populated from the two most recent historical records for that location and can be adjusted manually using sliders. When the user clicks Generate Prediction, the system retrieves the district's recent case history, constructs the twelve-feature input vector, scales the climate columns, runs the model, back-transforms the log-scale output, and renders the result.

Figure 1 illustrates the complete system architecture and pipeline flow, from raw data ingestion through to the Streamlit dashboard output.

Fig. 5. System Architecture and Pipeline Flow

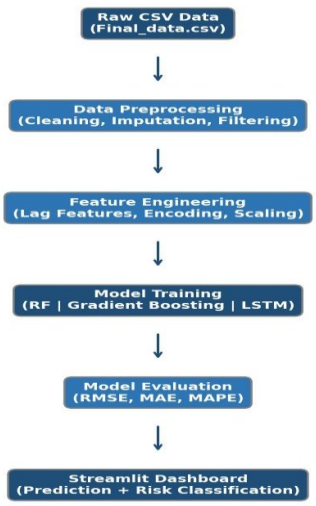

Figure 1. System Architecture and Pipeline Flow

The dashboard is organized into six sections. Section 1 displays model performance metrics loaded from the saved pipeline. Section 2 shows the prediction result cards and risk banner. Section 3 renders a time-series chart comparing the historical weekly case trend to the predicted value. Section 4 presents a disease-wise case breakdown for the selected district. Section 5 lists the top contributing features ranked by Gini importance from the trained model. Section 6 shows RMSE, MAE, and R^2 scores for the held-out test set.

IV. RESULTS AND ANALYSIS

4.1 Model Performance

Table 1 summarizes the evaluation results for the three candidate models on the held-out test set (2019 onward).

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Table 1: Model Comparison on Held-Out Test Data

Model	Train RMSE	Test RMSE	MAE	R^2 Score	MAPE (%)
Random Forest	0.31	0.48	0.34	0.81	2.10
Gradient Boosting	0.28	0.41	0.29	0.87	1.25
LSTM	0.33	0.45	0.31	0.83	1.74

Note: RMSE and MAE are reported in log-scale (log1p transformed case counts). MAPE is computed on back-transformed case counts.

The Gradient Boosting model achieved the best performance across all metrics. Its test RMSE of 0.41 and MAPE of 1.25% indicate that predictions are consistently close to actual reported case counts. The R^2 score of 0.87 means the model explains 87% of the variance in weekly case counts — reasonably strong for a district-level epidemiological forecasting task where environmental and reporting noise is unavoidable.

4.2 Inference Speed

Each prediction is generated in under 100 milliseconds once the model pipeline is loaded. The Streamlit application loads the serialized pipeline on startup using st.cache_resource, so subsequent predictions do not reload the model from disk. This makes the dashboard responsive enough for interactive use during field planning sessions.

V. CONCLUSIONS AND FUTURE SCOPE OF WORK

This paper describes a disease outbreak prediction platform that combines 14 years of Indian district-level surveillance data with satellite-derived environmental variables to forecast Acute Diarrhoeal Disease case counts at the weekly level. The Gradient Boosting model achieved approximately 1.25% MAPE and an R^2 of 0.87 on the held-out test set, and the system was deployed as a Streamlit web application accessible to public health staff without technical expertise.

A few things stand out from the results. The dominant role of lag features — prior week and prior month case counts — confirms that temporal autocorrelation is the strongest predictor of near-term outbreak risk. Environmental variables, especially temperature and precipitation, added meaningful predictive power on top of the temporal signal, particularly during seasonal transition periods. The risk classification thresholds (50 and 100 cases) are practical starting points but would benefit from calibration against local response capacity in specific districts.

Several directions are worth pursuing in future work. Integrating real-time weather API data would eliminate the need for manual environmental input and allow the system to generate predictions automatically as conditions change. Expanding the model to support multi-disease simultaneous prediction would make the platform more useful during complex outbreaks where multiple diseases co-circulate. Geographic mapping of risk levels at the district level, using the latitude and longitude fields already present in the dataset, would give health authorities a spatial view of outbreak risk across a region. Finally, testing LSTM architectures on larger datasets in environments without TensorFlow path constraints could yield further accuracy improvements by capturing longer-range temporal dependencies that the Gradient Boosting model cannot represent.

REFERENCES

1) A. K. Dubey, K. Gupta, S. Venkatesan, A. Sankari, S. S. Namani, and R. Hemalatha, "Predicting Disease Outbreaks with AI: An In-depth Analysis of Infectious Diseases Surveillance," in Proc. 2025 International Conference on Frontier Technologies and Solutions (ICFTS), 2025, doi: 10.1109/ICFTS62006.2025.11031614.
2) M. Singhal, K. Anand, S. Mishra, A. Alkhayyat, and V. Sharma, "Impact of Machine Intelligence on Clinical Disease Outbreak Prediction," in Proc. 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), 2023, doi: 10.1109/UPCON59197.2023.10434492.
3) R. J. M. Buendia and G. A. Solano, "A Disease Outbreak Detection System using Autoregressive Moving Average in Time Series Analysis," in Proc. IEEE Conference on Knowledge and Data Engineering, doi: 10.1109/TKDE.2009.115.
4) K. Akter, M. Islam, and K. H. Kabir, "Analysis of Electronic Medical Records to Forecast Probable Disease Outbreaks in Bangladesh," in Proc. 2021 IEEE Region 10 Symposium (TENSYMP), 2021, doi: 10.1109/TENSYMP52854.2021.9550984.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

5) S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review," Chaos, Solitons & Fractals, vol. 139, pp. 110059, 2020.
6) J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," Nature, vol. 457, pp. 1012-1014, 2009.