Received 20 April 2025

SYMBIOSIS STATISTICAL INSTITUTE

Disease Outbreak Prediction and Geospatial Modeling Using Time Series Analysis

NITIKA SONY, AKSHAT KUMAR, ANKIT YADAV

ABSTRACT Accurate modeling and forecasting of disease outbreaks is crucial for public health planning and policy making. This study aims to perform a time series analysis of seasonal vs non seasonal disease outbreak using statistical models like SARIMA and prophet, the objective to understand how different types of disease propagate and the factors behind their spread and forecast future cases of outbreaks. The dataset used for this project was sourced by compiling weekly disease outbreak reports published by the Ministry of Health and Family Welfare under the Integrated Disease Surveillance Programme. This study also aims to create an early warning system for disease outbreak and perform spatial-temporal analysis of disease spread.

Keywords: time series analysis, SARIMA, prophet, spatial-temporal analysis, weekly disease outbreak report

I. INTRODUCTION

The importance of predicting infectious disease cannot be understated especially for a developing nation like India, forecasting disease outbreaks is crucial for maintaining public health, ensuring timely intervention and availability of resources.This report has classified diseases broadly into two categories; seasonal and nonseasonal, we aims to explore the similarities and differences in the propagation of these different diseases, for achieving such an objective time series analysis can prove to be a very useful tool, traditional models suggest ARIMA and SARIMA Have proven to be highly effective in capturing the behaviour of seasonal diseases.The study also aims to compare the performance of time series models like Sarima with machine learning models like random forest in forecasting disease outbreaks, The project also aims to develop an early warning system for disease outbreak, As well as form a special temporal analysis of disease outbreak to understand the Spread of diseases over across the country. Therefore this research aims to combine statistical and machine learning models to understand and predict seasonal and non seasonal disease outbreaks capturing their behaviour to analyse and create an early warning system for detection as well as understand the spread of these diseases across different states and regions.

By leveraging historical disease data and analyzing temporal patterns, this research provides a foundation for proactive public health planning. Time series models enable the detection of recurring outbreak trends, while machine learning models enhance predictive accuracy through the use of complex, non-linear relationships in the data. The integration of these methodologies allows for a more nuanced understanding of outbreak dynamics, including seasonal fluctuations, regional variations, and lag effects between districts. This multi-faceted approach not only improves forecasting performance but also equips public health authorities with actionable insights for deploying targeted interventions, optimizing resource allocation, and minimizing the socio-economic impact of epidemics.

II. Methodology

1) Data Collection:

The data used for this study was collected by compiling weekly outbreak reports of different diseases published by National Centre for Disease Control under their integrated disease Surveillance programme, the time frame of the dataset used is from 2022 till the 6th week of 2025. The dataset consists of 10 columns starting from

- 'unique id' - This is the unique identifying id for each case.
- 'State' - The state where the outbreak took place.
- 'District' - The district where the outbreak took place.
- 'Diseases' - Particular disease under study
- 'Number of cases' - Number of cases of the particular disease.
- 'Number of deaths - Number of deaths caused by a particular disease.
- 'Date of outbreak' - Date of initial outbreak.
- 'Date of reporting' - Date of when the outbreak was reported.
- 'current status' - Status of the patient being classified as either under observation or under control
- 'Action taken' - The action taken by the local authorities

There are a total of 6200 entries in the dataset

2) Data Preprocessing:

The data set initially obtained was in a pdf format which was then converted into csv ensuring proper data structuring These files were then compiled to create the final data set used in the study the data was pre processed through the move missing values and the Date of outbreak and report were standardized to datetime format.

3) Diseases Categorization and Time Series Aggregation:

Diseases were categorized into:
Seasonal: Dengue, Malaria, Influenza
Non-Seasonal: Chickenpox, Food Poisoning
Other diseases in the dataset were excluded.

Data was arranged monthly using the date of outbreak and number of cases for each disease type was aggregated, the Resulting time series had two columns: Seasonal and Non-Seasonal.

4) Time Series Analysis:

Stationarity test for the time series for both Seasonal and non seasonal diseases was performed using the augmented dickey fuller adf test, and then the time series was decomposed to identify trends, seasonal patterns and residuals for both types of diseases.

The Prophet model was also used for forecasting the disease cases, mainly for seasonal diseases, as it can account for seasonality and holidays, which may impact disease outbreaks.

SARIMAX models were trained separately for Seasonal and Non-Seasonal cases with parameters (1,1,1)(1,1,1,12). Forecasts were compared against actual values on historical data.

SARIMA (Seasonal ARIMA) was chosen for its ability to model univariate time series with and without seasonal components. ARIMA captures linear trends, while SARIMA accounts for seasonal variations, making it ideal for diseases with clear seasonal patterns. These models serve as a baseline for comparison.

Prophet is selected due to its robustness in handling seasonality, holidays, and irregularities in the data. It is particularly effective for diseases with strong seasonal fluctuations or sudden outbreaks that may not fit traditional time series assumptions. Prophet's flexibility in modeling multiple seasonalities and missing data is a key advantage in disease forecasting.

5) Development of Machine Learning Models:

An early warning classification model was developed to predict future outbreaks, a binary outbreak label was created:

Outbreak = 1 if monthly cases ≥ 10

Outbreak = 0 otherwise

The features used were as follows:
Month_num, Year, and Lag_1 (previous month's cases)

Target Definition: A binary target variable Outbreak was created, where values ≥ 500 cases on a day were labeled as 1 (outbreak), else 0.

Upon observation the outbreak class was highly imbalanced, therefore SMOTE (Synthetic Minority Over-sampling Technique) was applied on the training set to balance the positive and negative samples.

Random Forest Classifier:
Stratified 5-fold cross-validation was used to evaluate accuracy and the model was trained/tested with an 80/20 temporal split.
Predictions were visualized and compared with actual outbreaks and performance was assessed using,

- ROC-AUC
- Precision-Recall AUC
- F1-Score
- Confusion Matrix

Random Forest, a machine learning ensemble model, was chosen for its ability to capture nonlinear relationships and interactions between multiple features, such as environmental factors, population density, and previous outbreaks. It is less sensitive to model assumptions, making it suitable for complex disease dynamics that may not follow linear patterns.

Prophet, a time series forecasting tool was also used to predict future outbreaks, for this the target diseases were selected: Dengue, Malaria, Chickenpox, and Food Poisoning. A separate Prophet model is trained for each disease.

Prophet configuration:

- Yearly seasonality enabled us to capture seasonal disease trends.
- Daily seasonality disabled (as the data is not truly daily or has long gaps).
- The model was fitted on historical data and then used to forecast the next 365 days.

Each disease was treated as a univariate time series, where the number of reported cases was tracked over time.The Prophet model was employed to learn patterns of trend, seasonality, and holiday effects from historical data and extrapolate them into the future.Time series decomposition by Prophet helped in understanding underlying trends (e.g., increasing disease burden) and periodicity (e.g., annual peaks)

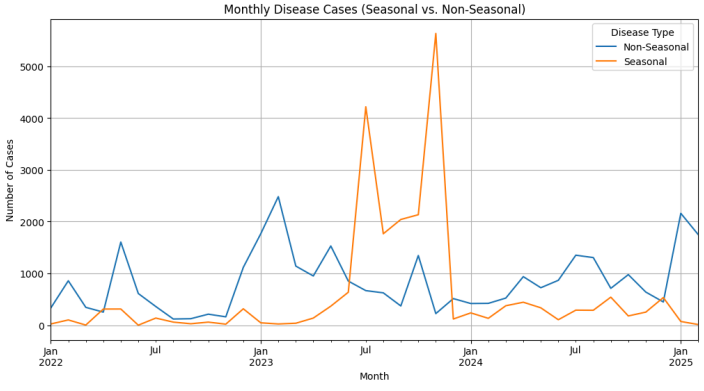
6) **GeoSpatial Analysis:**

A subset of 17 districts across India was mapped using latitude-longitude coordinates.Disease outbreak cases were aggregated yearly by district and visualized using plotly.express on a geo-scatter plot. This provided a dynamic visual summary of spatial hotspots and case intensity over time. Monthly case counts were aggregated by district. A Pearson correlation heatmap was constructed to identify synchronous patterns between districts. Cross-correlation with lag analysis was applied between district pairs (e.g., Malappuram and Kozhikode) to examine delayed relationships, identifying lead-lag disease transmission patterns.

Districts with at least six months of data were analyzed for intra-annual variation. Monthly average case counts were computed and plotted across districts to reveal seasonal patterns (e.g., monsoon or post-monsoon peaks).These line plots highlighted districts with recurring seasonal outbreaks and guided understanding of disease cycles.

III. Results and Discussion

This section provides the results and interpretation of the time series and ML models and the geospatial analysis



1) **Time Series Interpretation and SRIMA Model:**

Figure 1 : Line Plot – Monthly Disease Cases (Seasonal vs. Non-Seasonal)

Seasonal diseases (orange line) show sharp peaks at specific times (notably mid-2023 and late 2023), indicating highly concentrated outbreaks. Non-seasonal diseases (blue line) display frequent fluctuations without a consistent pattern, suggesting they are sporadic but persist throughout the year.

Seasonal cases tend to drop to nearly zero during off-peak months, while non-seasonal ones maintain a baseline presence.

```
Augmented Dickey-Fuller Test: Seasonal Diseases
ADF Test Statistic : -3.990011480892016
p-value : 0.0014637267146866797
#Lags Used : 0
Number of Observations Used : 37
Stationary

Augmented Dickey-Fuller Test: Non-Seasonal Diseases
ADF Test Statistic : -3.721081780747065
p-value : 0.0038262046667176695
#Lags Used : 0
Number of Observations Used : 37
Stationary
```

Figure 2: ADF Test Results

ADF Test Statistic for Seasonal Diseases: -3.998

ADF Test Statistic for Non-Seasonal Diseases: -3.721

p-values in both cases are < 0.05 therefore we reject the null hypothesis of a unit root, hence both time series are stationary, a necessary condition for ARIMA modeling.

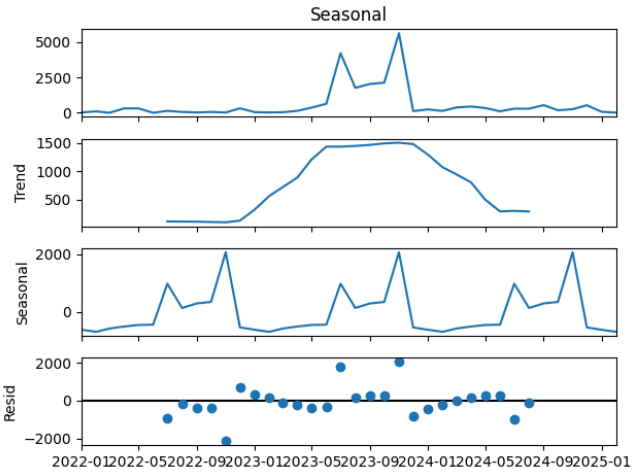


Figure 3: Decomposition of Seasonal Disease Time Series

The decomposed time series for seasonal diseases reflects actual disease counts with clear peaks.There is a gradual rise through 2023 and peak mid-2023, followed by a decline, indicating a medium-term increase and decrease in disease burden.

Seasonality shows repeating spikes and dips roughly every 12 months, consistent with annual seasonality; this confirms strong cyclical behavior, possibly linked to weather or environmental factors.

Residual dots fluctuate around zero, suggesting that most

variation is explained by the trend and seasonality.

A few large residuals might point to unusual outbreaks or reporting anomalies.

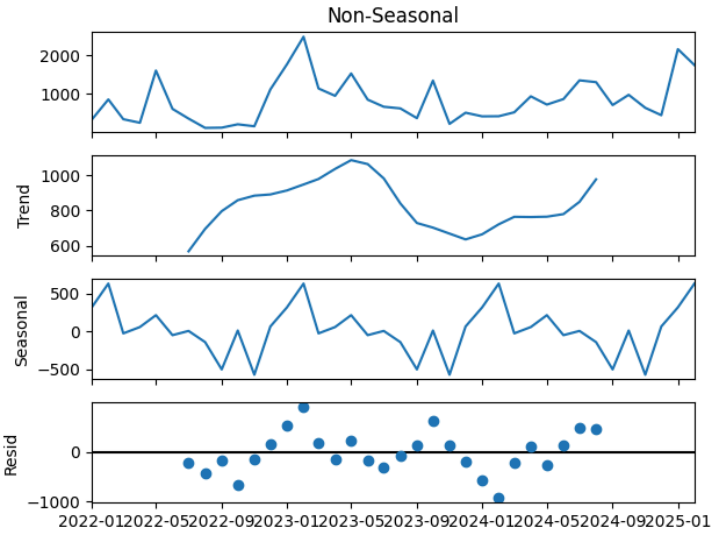


Figure 4: Decomposition of Non - Seasonal Disease Time Series

The observed series reflects the raw number of reported cases each month. The time series displays multiple fluctuations, with notable peaks in early 2023 and late 2024, indicating possible outbreak periods or heightened reporting. The trend line illustrates the long-term movement in disease incidence. There is a clear upward trend starting in early 2022, reaching a peak in mid-2023, followed by a decline and a mild resurgence towards the end of the period. This trend may indicate an overall increase in disease prevalence or improvements in surveillance and reporting systems. Although categorized as non-seasonal, the decomposition reveals a minor seasonal pattern. The repeating wave-like fluctuations suggest that certain non-seasonal diseases may still be influenced by periodic factors such as environmental conditions or social behavior, albeit not strongly enough to be classified as seasonal. The residuals represent irregular variations not explained by trend or seasonality. While most values hover around zero, several outliers are present, indicating months with unexpectedly high or low case numbers.

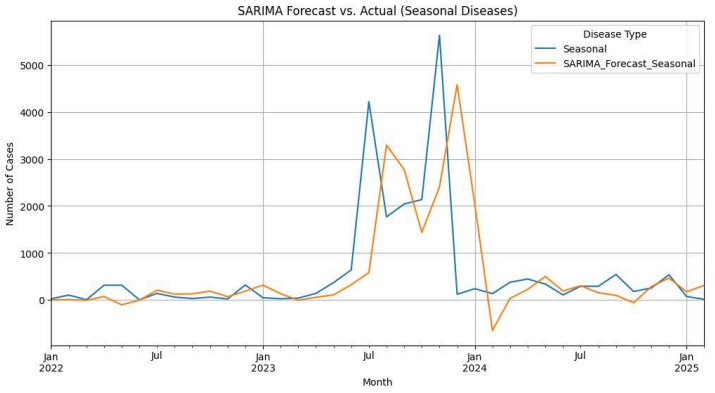


Figure 5 : SRIMA Forecast for Seasonal Disease

The line graph illustrates the comparison between actual seasonal disease cases and the SARIMA model's forecasted cases over a three-year period (January 2022–January 2025). The analysis indicates that the SARIMA model effectively captures the overarching trends and seasonal patterns in disease cases, as evidenced by the alignment of peaks and troughs in the actual and forecasted data. Notable peaks, such as mid-2023 and early 2024, highlight the model's competence in predicting periodic fluctuations. However, deviations observed in specific intervals suggest potential areas for improvement, particularly in addressing abrupt surges or anomalies. These findings reinforce SARIMA's utility as a robust forecasting tool while underscoring the importance of further refinement for greater precision in public health surveillance and intervention planning.

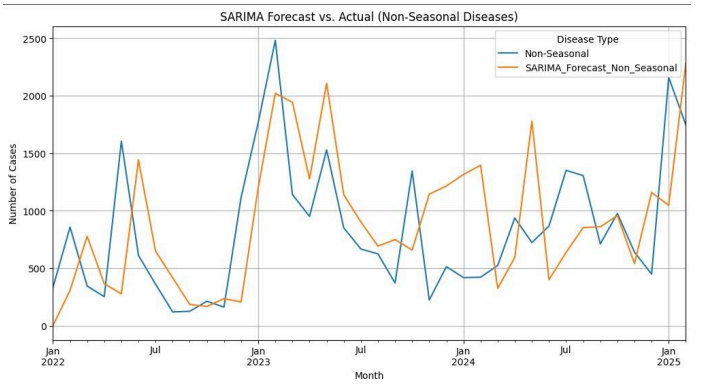


Figure 6 : SRIMA Forecast for Non - Seasonal Disease

The graph demonstrates the forecasting accuracy of the SARIMA model applied to non-seasonal disease cases over a three-year span (January 2022–January 2025). By comparing actual observed cases (blue line) with SARIMA's forecasting cases (orange line), it is evident that the model successfully captures overall trends and periodic fluctuations. While SARIMA proves effective in predicting regular patterns, certain deviations—such as abrupt surges or drops—highlight limitations in accounting for unexpected variations. Peaks and troughs are generally well aligned, suggesting the model's robustness in handling steady behaviors. However, further optimization may enhance the model's sensitivity to sudden changes. These findings underscore the potential and boundaries of SARIMA in public health forecasting, offering insights for researchers focused on improving predictive analytics for disease surveillance.

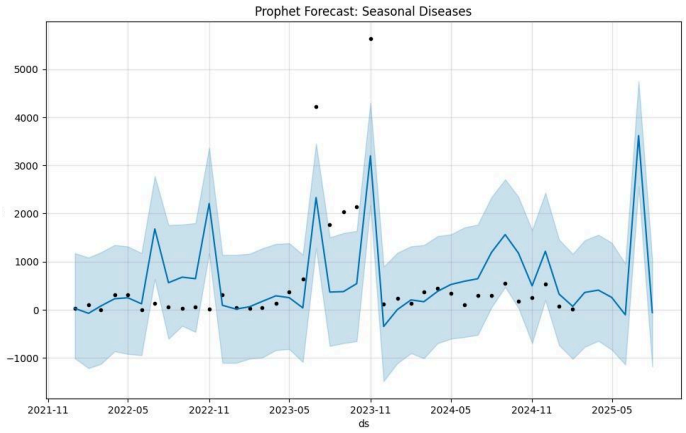


Figure 7: Prophet Forecast for Seasonal Disease

The SARIMA model's forecast, depicted in the image, demonstrates its efficacy in predicting seasonal disease cases over a three-year timeline (January 2022–January 2025). The model successfully aligns with the general seasonal trends observed in the actual data, as highlighted by the congruence of key peaks and troughs. Noteworthy periods, such as mid-2023 and early 2024, underscore the model's ability to capture recurring seasonal variations effectively. However, certain discrepancies between the forecasted and observed cases signal challenges in accounting for abrupt surges or anomalies, emphasizing areas for model refinement. These findings underscore the potential of SARIMA as a powerful tool for seasonal trend analysis in public health forecasting, while also advocating for iterative enhancements to improve predictive precision. Such insights provide valuable direction for optimizing forecasting models to better inform public health resource allocation and intervention strategies.

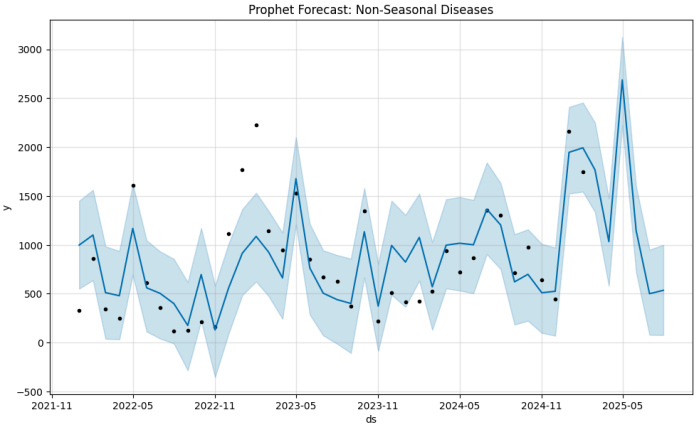


Figure 8: Prophet Forecast for Non - Seasonal Disease

The time series forecast displayed in the image illustrates the prediction of non-seasonal disease cases using the Prophet model for a period extending from November 2021 to May 2025. The model effectively captures underlying trends and provides a reliable forecast, as shown by the alignment of historical data points with the blue forecasted trend line. Peaks and troughs in disease cases are evident, with a prominent peak anticipated around May 2025, which could signify critical periods requiring heightened public health preparedness.

The shaded blue uncertainty intervals surrounding the forecast highlight the expected range within which actual case numbers may fall, thus offering a quantifiable measure of the model's prediction confidence. Despite its overall reliability, occasional deviations from historical data highlight the potential need for model adjustments to better account for unexpected variability.

These findings emphasize the utility of the Prophet model as a robust analytical tool for predicting non-seasonal disease trends. Such insights are vital for policymakers and healthcare professionals to strategically allocate resources and plan interventions to mitigate disease impact. Further exploration and refinement of the model could enhance predictive precision and its applicability in dynamic healthcare environments.

2) **Interpretation Machine Learning Models:**

The original dataset displayed a significant class imbalance, with only 16 outbreak instances compared to 3,711 non-outbreaks:

```
Original Class Distribution:
 Outbreak
0    3711
1      16
Name: count, dtype: int64

Resampled Training Class Distribution:
 Outbreak
0    2968
1    2968
Name: count, dtype: int64
```

Figure 9: Showing Class imbalance

To address this imbalance and prevent biased learning, Synthetic Minority Over-sampling Technique (SMOTE) was applied on the training set. This resulted in a perfectly balanced training distribution:

A Stratified K-Fold Cross-Validation approach was adopted to ensure that both classes were proportionally represented in all folds. The model demonstrated excellent and consistent performance across all folds, as shown below:

- Stratified CV Accuracy Scores:
  [0.9973, 0.9946, 0.9973, 0.9987, 0.9987]
- Mean Accuracy: 0.9973
- Standard Deviation: 0.0015

This high stability across folds suggests that the model generalizes well and is not overly sensitive to the training data. The model's ability to discriminate between classes was evaluated using:

ROC-AUC Score: 0.9996
Almost perfect separability between outbreak and non-outbreak cases.

Precision-Recall AUC Score: 0.9028
High precision and recall trade-off despite the rarity of outbreaks.

These metrics further affirm that the model can effectively identify outbreaks, a critical requirement for real-world early warning systems. By analyzing the F1-score across thresholds, the best balance between precision and recall was achieved at:

- Optimal Threshold: 0.30
- F1 Score: 0.8571

This low threshold was selected to favor sensitivity, ensuring that true outbreaks are not missed even if a few false positives are allowed.

- Accuracy: 99.87%
- Macro F1 Score: 0.9282
- Weighted F1 Score: 0.9988

This indicates that the model correctly identified all outbreaks in the test set (Recall = 1.00), albeit with a slight drop in precision — a trade-off well-justified in public health contexts where false negatives are far costlier than false positives.

```
ROC-AUC Score: 0.9996
Precision-Recall AUC Score: 0.9028

Best Threshold based on F1: 0.30 → F1 Score: 0.8571

Classification Report (Threshold = 0.30):
              precision    recall  f1-score   support

 No_Outbreak     1.0000    0.9987    0.9993       743
    Outbreak     0.7500    1.0000    0.8571         3

    accuracy                         0.9987       746
   macro avg     0.8750    0.9993    0.9282       746
weighted avg     0.9990    0.9987    0.9988       746
```

Figure 9: Random Forest Model Performance

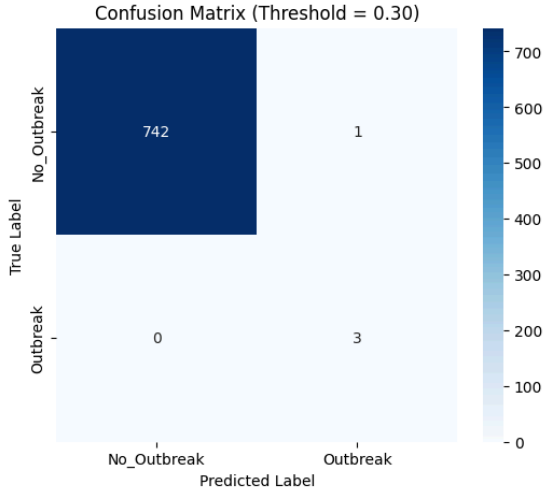


Figure 10: Confusion Matrix

The performance of the outbreak prediction model, evaluated at a decision threshold of 0.30, demonstrates impressive sensitivity and precision; two crucial metrics in any epidemiological early warning system. From the confusion matrix, it is evident that the model correctly identified 742 out of 743 non-outbreak cases and accurately detected all three actual outbreak instances, without a single false negative. The choice of a lower threshold (0.30), guided by F1 score optimization, reflects a well-informed public health strategy. In the context of disease surveillance, especially in time-sensitive scenarios, the cost of missing an actual outbreak can be catastrophic. Therefore, a threshold that slightly increases false positives while capturing all true outbreaks is not just acceptable; it is desirable.

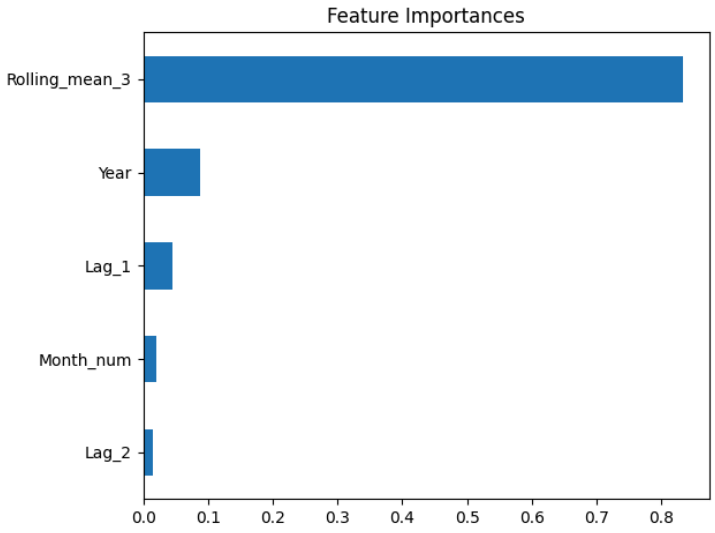


Figure 11: Feature Importance

From a time series perspective, the modeling pipeline exhibits a nuanced understanding of temporal structure in epidemiological data. The introduction of lag features (Lag_1, Lag_2) and a rolling mean (Rolling_mean_3) effectively transforms the raw case counts into temporal signals that encode past dependencies and short-term trends. This is a classic strategy in time series forecasting, often seen in ARIMA, state-space models, and other autoregressive frameworks. Among these engineered features, the 3-period rolling mean emerged as the most critical variable, dominating the feature importance chart with a contribution exceeding 80%. This suggests that the model learns outbreak precursors not from isolated spikes, but from sustained rises in incidence—exactly the kind of pattern time series models are built to capture.

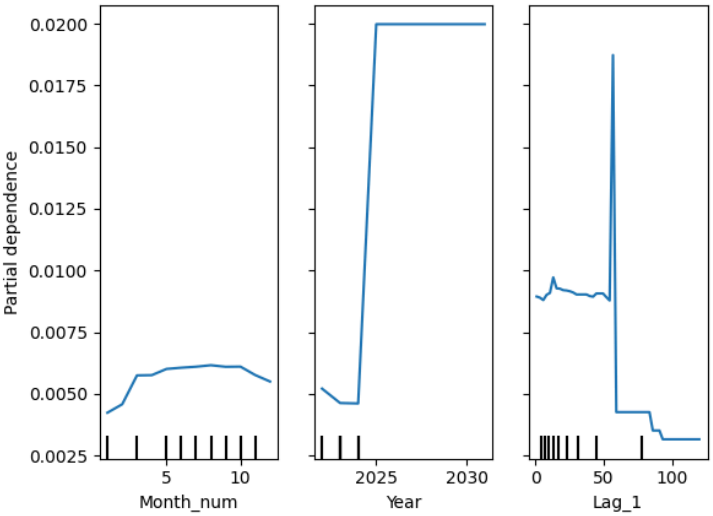


Figure 12: Partial Dependence Plots

Partial dependence plots (PDPs) further reinforce the temporal dynamics at play. For instance, the Month_num variable showed seasonal fluctuation in outbreak probabilities, hinting at a cyclic trend likely tied to climatic or environmental conditions—such as monsoons or seasonal pathogen activity. The Year variable interestingly showed a marked increase in predicted outbreak risk from 2024 onwards, which could point to systemic shifts such as improved reporting, climate changes, or emerging new patterns in disease spread. Meanwhile, the Lag_1 variable showed threshold-like behavior, where outbreak probabilities sharply increase when the previous month's case count crosses a certain threshold (around 50 cases). This nonlinear temporal behavior is emblematic of many infectious disease outbreaks that often accelerate once critical community transmission levels are reached.

These insights confirm that the model is not treating data as static tabular input but is responding to the time-based evolution of the features. This blending of supervised machine learning with time series-informed features is a powerful hybrid approach. It leverages the predictive strength of ensemble classifiers like Random Forests while retaining the interpretability and forecasting intuition inherent to classical time series modeling.

3) **Prophet Model Interpretation:**

Prophet models were developed to forecast disease case counts and identify potential future outbreaks. Below is the summary of forecasts and critical insights for each disease.

The top 10 forecasted dates for Dengue showed moderate expected case counts, all well below outbreak thresholds:

4) **Delay Chain analysis:**

Delay Chain analysis produced the following Steady State Distributions and

	ds	yhat	yhat_lower	yhat_upper
325	2025-02-11	38.502988	-156.152696	218.554737
326	2025-02-12	29.546110	-173.211380	215.844118
327	2025-02-13	35.356284	-148.890990	240.535102
328	2025-02-14	70.040584	-124.847234	260.340969
329	2025-02-15	63.597228	-124.317456	249.336521
330	2025-02-16	59.202481	-126.174626	254.422414
331	2025-02-17	42.296926	-152.893416	234.732441
332	2025-02-18	49.686580	-135.049801	225.255447
333	2025-02-19	38.975302	-148.381419	213.502265
334	2025-02-20	42.878073	-148.606865	244.780311

Figure 13: Forecast Dates for dengue

Although upper bounds occasionally approached ~250, no prediction exceeded 500, indicating no imminent outbreak risk for Dengue in the forecast horizon.

Malaria exhibited high forecast variability, including a notable predicted outbreak:

	ds	yhat	yhat_lower	yhat_upper
140	2025-02-01	304.186780	-146.130577	725.42909
141	2025-02-02	29.209455	-382.161981	472.77824
142	2025-02-03	48.409876	-424.588610	519.44680
143	2025-02-04	41.240855	-407.094571	488.88369
144	2025-02-05	3.546617	-436.965344	430.88285
145	2025-02-06	-4.944002	-438.023876	444.49491
146	2025-02-07	-83.229291	-525.762199	372.61657
147	2025-02-08	127.014380	-295.783246	574.55570
148	2025-02-09	-161.135300	-609.900508	294.08201
149	2025-02-10	-151.230117	-592.204305	289.81708

Figure 14: Forecast Dates for Malaria

A future outbreak is predicted on November 15, 2025, with expected cases exceeding 500. High volatility was observed in earlier forecasts, with yhat values ranging from -161 to 304, emphasizing uncertainty.

Food poisoning forecasts were slightly higher in magnitude but still well below outbreak levels:

	ds	yhat	yhat_lower	yhat_upper
520	2025-02-16	85.310784	-12.196686	182.218556
521	2025-02-17	64.066291	-30.414402	158.402324
522	2025-02-18	62.192098	-29.879191	169.429035
523	2025-02-19	80.214670	-20.172913	178.677256
524	2025-02-20	66.071124	-30.665761	165.193094
525	2025-02-21	57.482493	-48.444868	155.974452
526	2025-02-22	90.496016	-9.565016	192.023890
527	2025-02-23	82.851736	-15.337458	182.144289
528	2025-02-24	62.280050	-35.172405	164.156824
529	2025-02-25	61.112430	-46.620734	157.827009

Figure 15: Forecast Dates for Food Poisoning

Forecasts reflect **periodic spikes**, possibly linked to seasonal or event-driven factors, but none are high enough to trigger outbreak warnings.

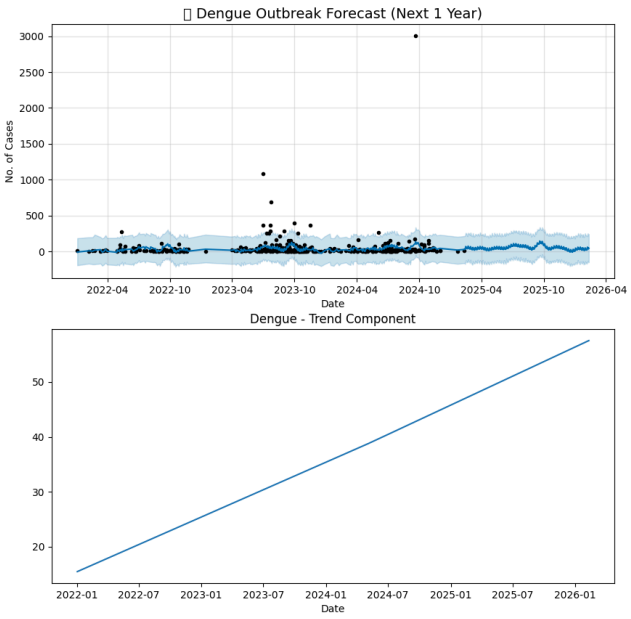


Figure 16: Dengue Outbreak Forecast and Trend line

The upper plot displays predicted Dengue cases for the upcoming year, with actual past cases marked as black dots. The model's forecast (blue line) and its confidence interval (shaded region) indicate expected trends and variability. Significant spikes in past cases highlight outbreak events that the model learns from to anticipate future surges. The trend component in the lower plot shows a steady upward trajectory, signaling a gradual long-term increase in reported cases. This rising trend, combined with forecasted case estimates, can inform health authorities in advance and support timely preventive action. The visualization demonstrates the system's ability to not only capture existing patterns but also provide reliable early warnings, which are crucial for effective disease surveillance and public health preparedness.

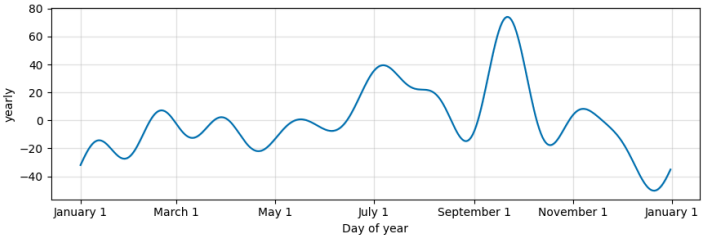


Figure 17: Dengue Outbreak Yearly Pattern

The yearly pattern (bottom plot) highlights strong seasonal effects, with a pronounced peak around September, suggesting that dengue outbreaks are most likely during late monsoon and early post-monsoon months. This decomposition enables us to anticipate high-risk periods and improve outbreak preparedness using time-driven insights.

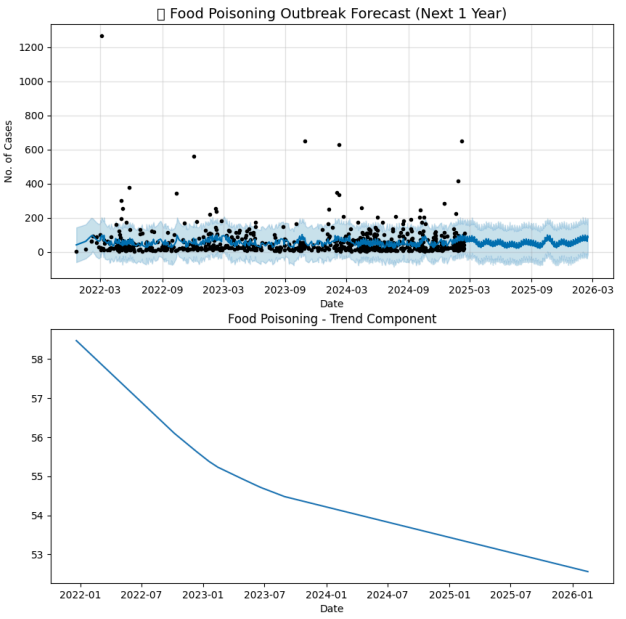


Figure 18: Food Poisoning Outbreak Forecast and Trend line

The forecast plot presents projected case counts for the upcoming year, with uncertainty bands indicating the confidence range of predictions. Despite observed fluctuations, the model effectively captures the underlying temporal dynamics, enabling timely identification of potential future surges. The trend component, extracted using additive decomposition, reveals a gradual decline over time—an encouraging sign for long-term public health outcomes. This forecasting pipeline enhances situational awareness and empowers authorities to act preemptively based on data-driven insights.

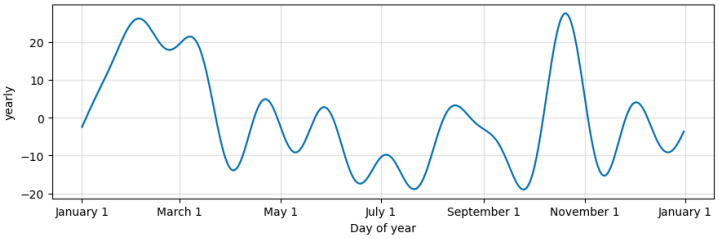


Figure 19: Food Poisoning Outbreak Yearly Pattern

The yearly seasonality plot for Food Poisoning reveals pronounced cyclical trends, indicating that outbreaks tend to follow a recurring annual pattern. Peaks in cases are observed around late February to early March and again in October to November. The spike during October and November aligns with the festival season in India, a time marked by increased food preparation, gatherings, and consumption of street or perishable foods, which could contribute to higher incidences of food poisoning. Conversely, the mid-year months like June to August and

the end of December show dips in cases, indicating relatively lower incidence during those periods. These insights can help guide food safety campaigns and health interventions to target high-risk months more effectively.

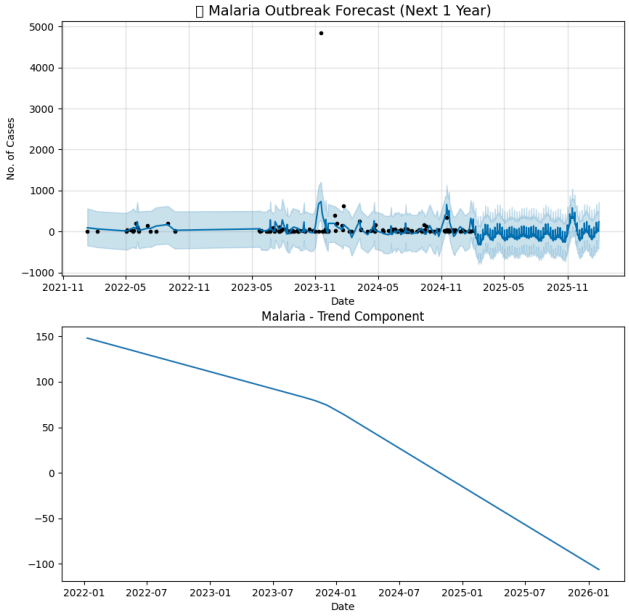


Figure 20: Malaria Outbreak Outbreak Forecast and Trend line

The upper plot presents the malaria outbreak forecast for the next one year using a time series model. The black dots represent historical case counts, while the solid blue line indicates the predicted number of future cases. The surrounding shaded blue region is the confidence interval, which accounts for uncertainty in the forecast — wider bands reflect greater uncertainty. The model anticipates that malaria cases will remain relatively stable, with no strong upward spikes, though some degree of fluctuation is expected. A few historical outliers (notably high spikes) are visible, but the model does not project similar spikes in the near future. This suggests that, if current conditions hold, the likelihood of a severe outbreak is low, making this forecast a useful tool for early warning and preventive planning.The extracted trend component shows a clear decline in malaria incidence over time, suggesting a positive shift in public health conditions. By integrating such forecasts into real-time systems, authorities can make informed, proactive decisions to mitigate potential outbreaks effectively.

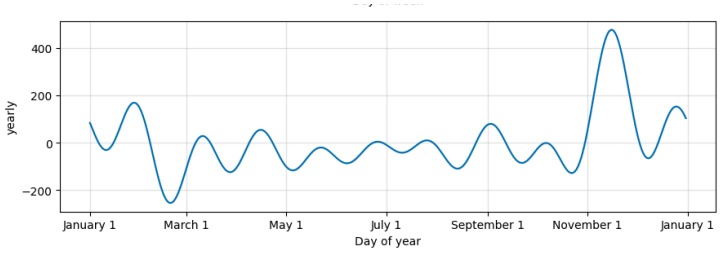


Figure 21: Malaria Outbreak Yearly Pattern

The yearly seasonality) reveals a pronounced peak around November, highlighting this period as a high-risk time for malaria outbreaks. These patterns allow for targeted forecasting and timely intervention, especially in late-year periods, aiding early warning efforts.

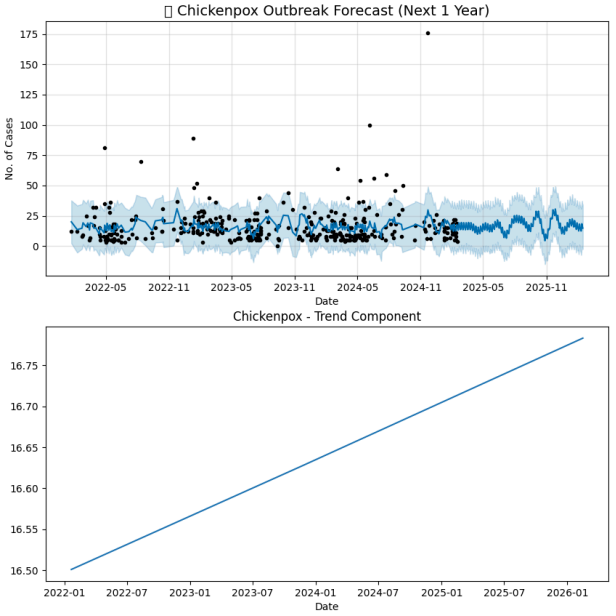


Figure 22: ChickenPox Outbreak Outbreak Forecast and Trend line

The Chickenpox outbreak forecast for the next year indicates a relatively stable pattern, with the predicted number of weekly cases generally fluctuating around 20 to 30. While the majority of historical data points fall within this range, occasional spikes highlight the potential for localized outbreaks. The forecast includes a 95% confidence interval, which broadens slightly over time, reflecting growing uncertainty but remaining within moderate bounds. The underlying trend component shows a subtle but steady increase from early 2022 through 2026, suggesting a marginal rise in baseline transmission, possibly due to shifts in immunity levels or reporting practices. Overall, the projection implies low to moderate risk with no major surges expected, but continued monitoring is essential to manage sporadic upticks effectively.

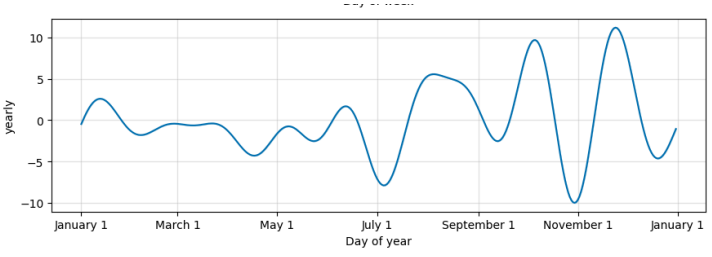
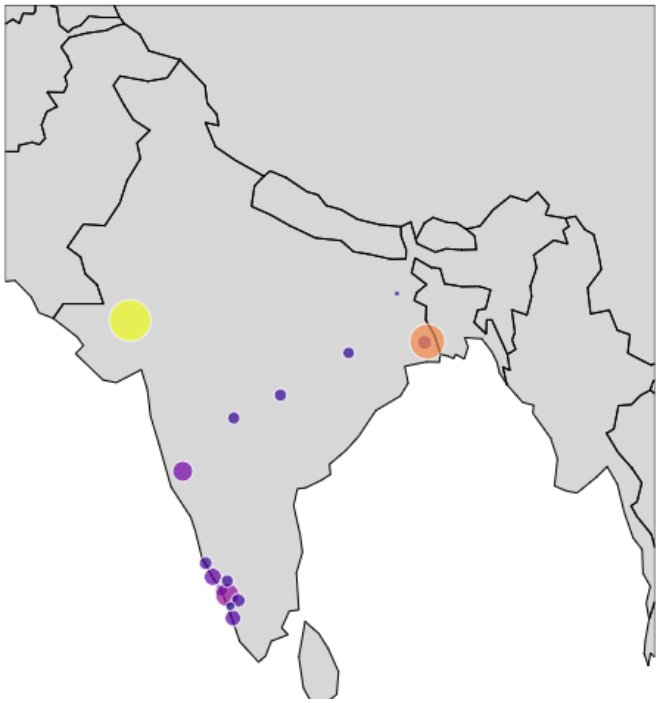


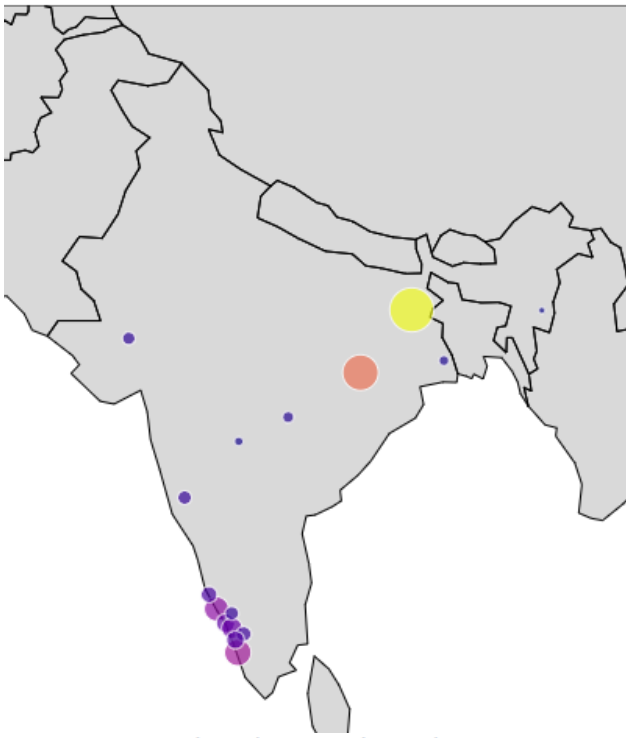
Figure 23: ChickenPox Outbreak Yearly Pattern

The yearly seasonal pattern, although modest in amplitude, highlights periodic fluctuations. Local peaks appear around September and December, and dips around November and July. This suggests potential seasonal influences, such as school terms or changes in indoor congregation, that can subtly affect transmission.

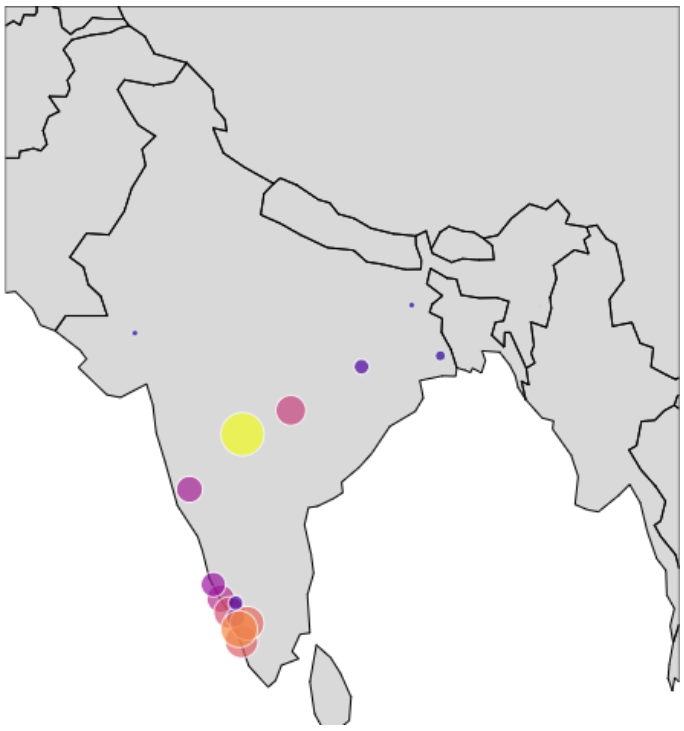
4) Geo-Spatial Analysis [Figues 24 to 27]:



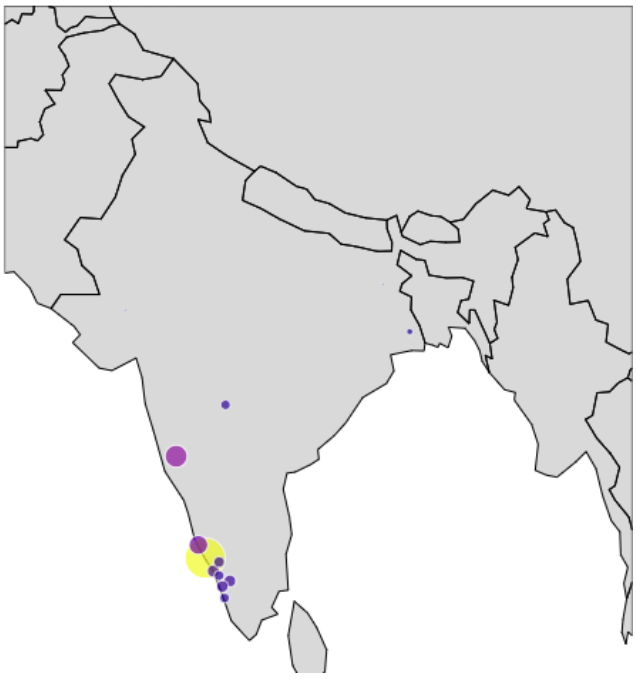
Disease Outbreaks in Selected Districts - 2022



Disease Outbreaks in Selected Districts - 2023



Disease Outbreaks in Selected Districts - 2024



Disease Outbreaks in Selected Districts - 2025

The interactive dashboards (refer to the google colab link in references) present a geographic and temporal visualization of disease outbreaks across selected Indian districts from 2022 to 2024, with each map corresponding to a specific year. The color intensity and size of the plotted points represent the number of reported cases, helping identify hotspots and trends over time. In 2022, districts like Ernakulam, Godda, and Malappuram showed moderate case counts. Moving to 2023, there's a noticeable rise in outbreaks in Kolhapur, Kasaragod, and Thrissur, indicating a potential spread or reporting improvement in western coastal regions. By 2024, districts such as Wayanad, Palakkad, and Kozhikode reflect continued high case numbers, hinting at either recurring outbreaks or inadequate containment efforts. The consistent presence of certain districts across all years may suggest persistent vulnerabilities, poor sanitation, or environmental factors conducive to disease spread. From a time series perspective, these dashboards act as an early-warning layer for public health decision-making by highlighting spatial-temporal clusters and helping prioritize resource allocation in frequently affected areas.

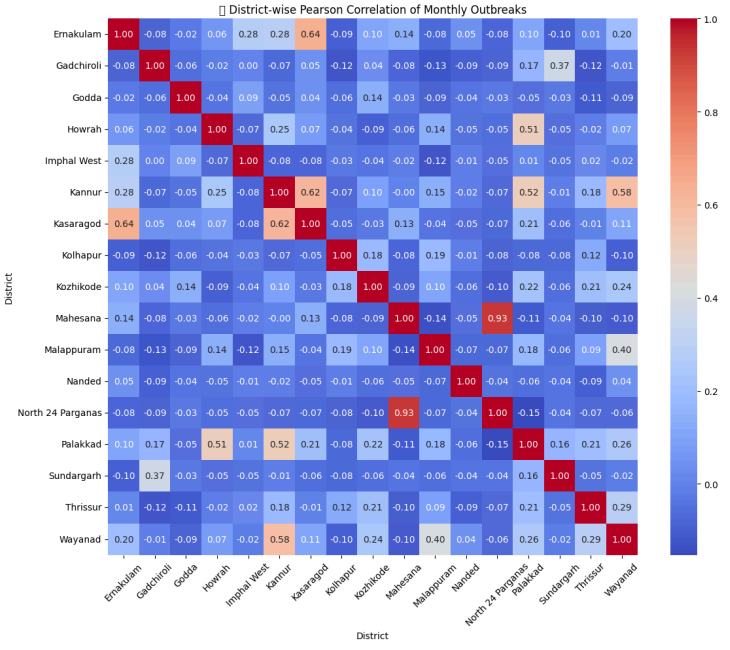


Figure 28: District wise correlation of monthly outbreaks

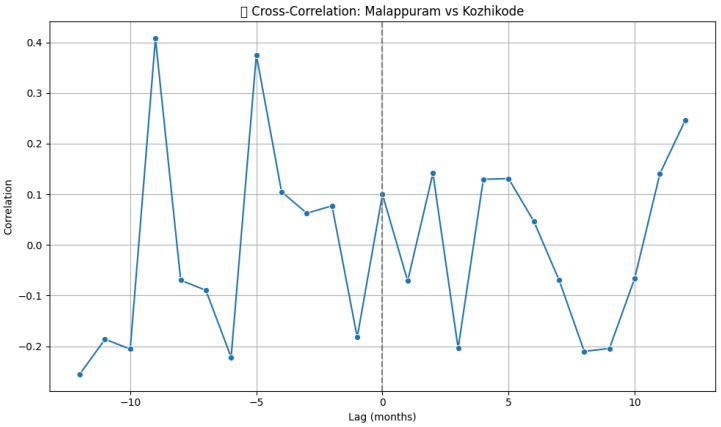
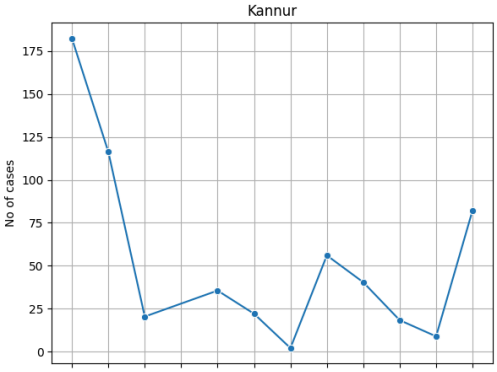


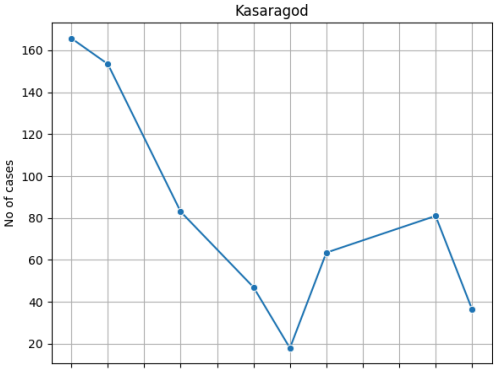
Figure 29: cross-correlation

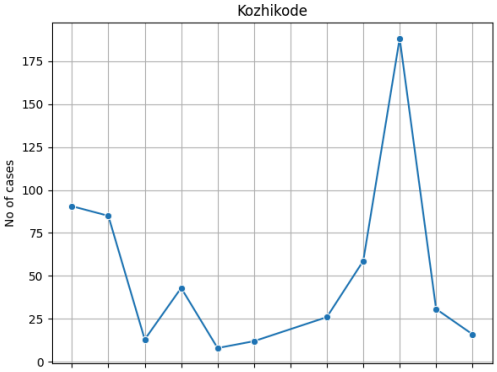
Figure 28 presents a district-wise Pearson correlation heatmap, based on monthly aggregated time series data of disease outbreak cases from 2022 to 2025. In this analysis, each district's monthly case counts form a separate time series. The heatmap reflects how closely the temporal patterns of these districts align. A value close to 1.00 indicates a strong positive correlation, meaning the outbreaks in both districts tend to rise and fall together over time. For example, Malappuram and Kozhikode show a moderately strong correlation, suggesting similar seasonal or environmental trends driving outbreaks. Extremely high correlations like between Mahesana and North 24 Parganas (~0.93) suggest closely synchronized outbreak behaviors. On the other hand, districts like Kolhapur and Ernakulam have near-zero or negative correlations, indicating little to no similarity in their outbreak trends.

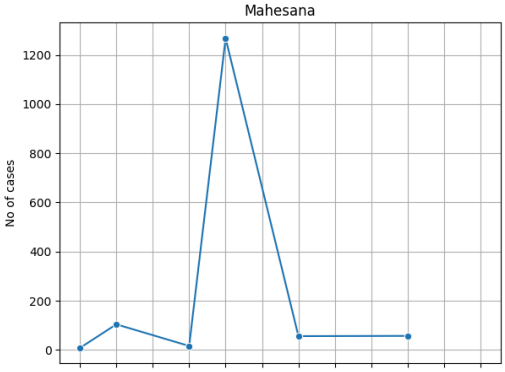
Figure 29 applies cross-correlation analysis, a powerful time series technique, to explore lagged relationships between two districts: Malappuram and Kozhikode. Here, the aim is to determine whether the outbreak in one district can predict future outbreaks in the other. The graph shows correlations at various time lags (in months). A peak correlation at lag -9 indicates that Kozhikode's outbreak pattern leads Malappuram's by 9 months—suggesting that Kozhikode can act as a temporal indicator or early warning signal for Malappuram. This is especially valuable in proactive outbreak response and forecasting.

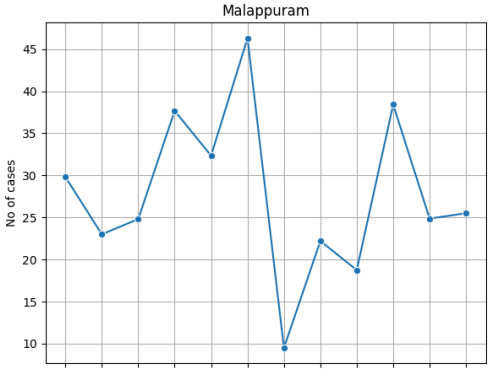
📊 Monthly Average Disease Outbreaks – Selected Districts

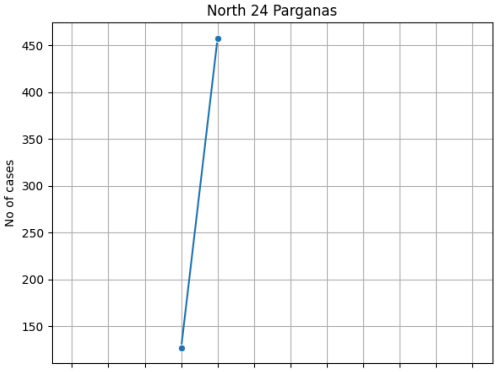


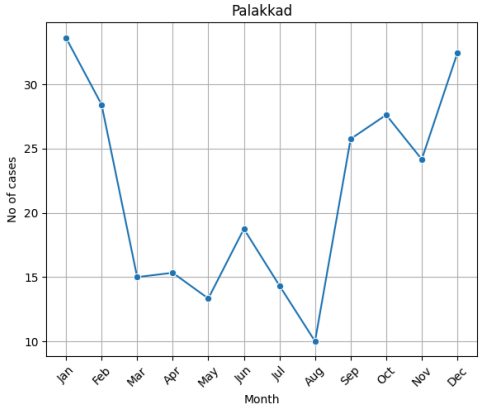












District	Observed Pattern	Interpretation
Malappuram	Spike from **June to August**	Monsoon-related outbreaks; likely water- or vector-borne diseases
Kozhikode	Mid-year peak, especially **around July**	High correlation with Malappuram; suggests regional monsoon seasonality
Kasaragod	Peak during **June–July**, tapering afterward	Consistent with monsoon-linked vector-borne outbreaks
Kannur	Strong spike in **January**, smaller rise in mid-year	Possible dual outbreak seasons (winter viral + monsoon-related)
North 24 Parganas	Sudden rise in **August–September**	Likely post-monsoon/flood-related diseases; strong seasonal outbreak pattern
Palakkad	Small rise in **June–August**, but less consistent	Suggests weak seasonality; may be influenced by outbreaks in neighboring districts
Mahesana	Sharp peak in **May**	Unusual dry-season spike; could point to respiratory or heat-related illnesses

The above charts show the monthly average outbreak plots for all districts using the provided dataset (as shown in the code). However, many districts showed either insufficient data (fewer than 6 months of data) or no discernible seasonal pattern. Therefore, for clarity and insight, we focused the interpretation on seven districts that demonstrated meaningful and consistent trends.

IV. Conclusion

This project presents a comprehensive framework for disease outbreak forecasting using time series analysis combined with machine learning approaches. By leveraging both classical time series models and data-driven classification methods, the project demonstrates how patterns in disease incidence over time can be modeled, understood, and used for predictive insights. The integration of models such as SARIMA and Prophet enables robust forecasting of future case counts while accounting for seasonality and trend structures inherent in health surveillance data.

From a time series standpoint, the project captures the temporal evolution of disease outbreaks at a fine-grained, district level. This includes identifying recurring seasonal peaks, detecting delayed relationships between regions through cross-correlation analysis, and examining how historical patterns inform future risks. These insights are crucial for understanding the cyclical nature of disease spread and for designing proactive public health responses.

For stakeholders such as public health officials, epidemiologists, and policy planners this framework offers a valuable tool for early warning and resource allocation. It supports timely decision-making by forecasting potential outbreaks before they occur, allowing for pre-emptive actions such as awareness campaigns, vaccine distribution, and medical preparedness. The inclusion of district-specific temporal patterns also enables localized intervention strategies rather than generalized national-level policies.

Ultimately, this project contributes to building a predictive, time-aware public health surveillance system. It emphasizes the importance of temporal dynamics in disease control and demonstrates how time series analysis can be practically applied to support data-driven public health planning and response.

References

[1] https://colab.research.google.com/drive/118vIfpIQhX8eI7MJSXp66koLxrnzCV_R?usp=sharing

[2] https://docs.google.com/spreadsheets/d/18r0qhO-xoaxOIuJgBc_qc2oTroU1L9aN0renjkRN5y0/edit?usp=sharing

[3] https://idsp.nic.in/index4.php?lang=1&level=0&linkid=406&lid=3689

[4] https://pedro.unifei.edu.br/download/Montgomery.pdf