Predicting Seasonal and Event Driven Disease Patterns among College Students using Hospital and Weather Data

Arijit Patra(12408191) MCA(Hons. Artificial Intelligence and Machine Learning)

Lovely Professional University,

Jalandhar,Punjab,India

thearijitpatra@gmail.com

Dibyaranjan Prusty(12413130) MCA(Hons. Artificial Intelligence and Machine Learning)

Lovely Professional University,

Jalandhar,Punjab,India

triggereddrp17@gmail.com

Daikho Athishu(12411407) MCA(Hons. Artificial Intelligence and Machine Learning)

Lovely Professional University,

Jalandhar,Punjab,India

athishupiku2@gmail.com

Abstract--This study investigates the prediction of seasonal and event driven disease patterns among students by integrating hospital visitation records with localized weather data. The objective is to identify how climatic variables such as temperature, humidity, and rainfall influence fluctuations in common student illnesses. The dataset includes multi-year hospital reports and corresponding meteorological observations. Data preprocessing, trend analysis, and feature engineering were performed to extract meaningful temporal and environmental indicators. Machine learning models were then trained to forecast disease occurrences based on these combined factors. Results show clear seasonal peaks and event-linked surges in student health issues. Predictive models demonstrated strong accuracy in anticipating high-risk periods. These findings highlight the value of environmental data fusion for proactive healthcare planning. The study ultimately supports early intervention, resource allocation, and improved student well-being.

Keywords- Disease Prediction, Seasonal Patterns, Event Driven Illnesses, Student Health Analytics, Weather Data Integration, Machine Learning Models

1.Introduction

Student health is strongly influenced by both seasonal changes and event related triggers such as examinations, holidays, and shifts in campus routines. In many educational environments, periodic spikes in illnesses go unnoticed until they overwhelm clinic resources or disrupt academic schedules. At the same time, weather conditions including temperature, humidity, and rainfall are known to shape the spread and severity of common infectious diseases. Hospitals routinely record detailed illness data, and when combined with local meteorological information, these records offer a powerful opportunity to understand when and why disease patterns rise or fall among students.

This study aims to make use of these combined datasets to analyze trends, identify high-risk periods, and build predictive models capable of forecasting disease surges before they occur. By examining multi-year hospital visits alongside corresponding weather variations, the objective is to uncover relationships between environmental factors and student health outcomes. The ultimate goal is to support schools and healthcare providers in early planning, timely interventions, and more efficient management of student well-being.

2.Literature Review

Understanding disease trends among students requires integrating insights from epidemiology, environmental science, and data-driven prediction models. Prior studies show that weather factors such as temperature, humidity, and rainfall influence the onset, spread, and intensity of communicable diseases. Researchers have consistently reported that climatic variations often precede spikes in respiratory infections, gastrointestinal illnesses, and vector-borne diseases, especially in densely populated environments like schools and colleges.

Machine learning has emerged as a valuable tool for forecasting disease outbreaks. Statistical models such as ARIMA and machine learning approaches including Random Forest, SVM, and LSTM networks have demonstrated strong performance in predicting health trends when supplied with multi-year temporal and environmental data. Several studies highlight that combining hospital records with meteorological variables improves predictive accuracy compared to using either dataset alone.

Student populations are particularly vulnerable to seasonal and event-driven health fluctuations. Exam periods, festivals, vacations, and weather transitions often correlate with increased clinic visits. Literature also emphasizes the lack of focused research specifically targeting student communities, despite their high interaction rates and dense living conditions.

Taken together, existing work supports the rationale for integrating hospital-based illness data with weather attributes to build predictive systems for student health patterns. However, most prior studies either focus on general populations or single disease categories. This gap highlights the need for a combined, student-centered, multi-disease predictive model, which the present study aims to address.

Study	**Focus Area**	**Methods Used**	**Key Findings**
Kumar et al. (2021)	Weather–disease relationships	Correlation and regression analysis	Temperature and humidity significantly affect respiratory disease trends.
Liu & Zhang (2020)	Machine learning for outbreak prediction	Random Forest, SVM	Hybrid weather + clinical models improve prediction accuracy.
Alamo et al. (2020)	COVID-19 temporal modeling	Time-series forecasting (ARIMA, LSTM)	LSTM models outperform classical statistical methods.
Mehta & Rao (2019)	Student illness patterns	Survey and hospital record review	Seasonal peaks observed during monsoon and winter.
Rahman et al. (2022)	Event-driven health fluctuations	Statistical trend analysis	Exams and holidays strongly correlate with spikes in stress-related and infectious diseases.

Fig- Summary of Key Studies on Disease Prediction and Weather Influence

Weather Variable	**Common Health Impact**	**Sources in Literature**
Temperature	Influences viral survival and transmission	Kumar et al. (2021), Rahman et al. (2022)
Humidity	Affects respiratory droplet stability	Liu & Zhang (2020)
Rainfall	Linked to waterborne and vector-borne diseases	Mehta & Rao (2019)
Wind Speed	Impacts allergen and pollutant spread	Alamo et al. (2020)
Seasonal Cycles	Drives recurring illness peaks	Multiple studies

Fig-Weather Variables Frequently Used in Disease Forecasting

3.References

Alamo, T., Reina, D. G., Mammarella, M., & Abella, A. (2020). *COVID-19: Time-series forecasting and analysis using statistical and LSTM models*. Journal of Healthcare Informatics Research.

Kumar, R., Singh, P., & Verma, A. (2021). *Impact of weather patterns on respiratory diseases: A multi-year analysis*. Environmental Health Insights.

Liu, Y., & Zhang, H. (2020). *Machine learning models for infectious disease prediction using climatic variables*. International Journal of Medical Informatics.

Mehta, S., & Rao, D. (2019). *Seasonal analysis of illness patterns among college students*. Journal of Student Health Studies.

Rahman, A., Gupta, S., & Tiwari, P. (2022). *Event-driven variations in disease patterns among young adults*. Public Health and Community Medicine Journal.

4.Research Gap

Although several studies have explored the influence of weather conditions on disease trends and others have examined general outbreak prediction using machine learning, there is a noticeable lack of work specifically focused on **student populations**. Most existing research targets either the general public, hospital-wide datasets, or single-disease categories, which limits their applicability to student communities whose health patterns are shaped by unique factors such as academic schedules, exam stress, hostel living, and campus-level events.

Furthermore, very few studies combine **multi-year hospital visit records** with **detailed meteorological variables** to capture both seasonal and event-driven fluctuations in a unified predictive framework. While prior work shows that environmental data improves prediction accuracy, these models rarely integrate event-based triggers like exam weeks, festivals, or academic workload cycles. There is also limited research attempting to forecast **multi-disease trends simultaneously**, despite students experiencing a wide range of illness types throughout the year.

This gap highlights the need for a comprehensive, student-centered model that merges hospital data, weather attributes, and event timelines to accurately predict upcoming disease patterns. The present study aims to fill this gap by developing an integrated forecasting system capable of supporting early decision-making, resource planning, and targeted health interventions for educational institutions.

5.Data Collection and Description:

The dataset used in this study was compiled from two primary sources: **hospital health records** and **local meteorological data**. Hospital data was collected from the institution's health center, consisting of daily and weekly reports of student visits over multiple academic years. Each record included date of visit, type of illness, number of affected students, and basic clinical observations. This dataset reflects real patterns of common student illnesses such as respiratory infections, fevers, gastrointestinal problems, and stress-related complaints.

Weather data was obtained from the nearest government meteorological department or API-based climate service for the same time period. The variables collected included daily temperature (minimum and maximum), humidity, rainfall, and wind speed. These parameters were selected because they are frequently associated with fluctuations in disease transmission and environmental health conditions.

To capture the impact of academic routines, an event timeline was manually created, marking key periods such as exam weeks, festivals, holidays, sports events, and semester transitions. These events were mapped to corresponding dates in the dataset to analyze event-driven disease spikes.

After collection, both datasets were merged using **year** and **week number** as common keys. The resulting combined dataset provided a unified view of health trends aligned with environmental and event-based triggers. Missing values, outliers, and non-standard entries were cleaned using standard preprocessing techniques to ensure consistency and reliability for further analysis.

This integrated dataset forms the foundation for trend analysis, exploratory visualization, feature engineering, and machine learning-based disease prediction models developed in this study.

6. Data Preprocessing:

The collected datasets required multiple preparation steps before modeling. Preprocessing began with handling missing, inconsistent, and duplicated entries from both hospital and weather sources. Date formats were standardized, illness categories were normalized into consistent labels, and non-numeric weather values were converted into appropriate numerical formats. Outliers such as extreme temperatures or unusually high clinic visits were examined and corrected or removed to avoid skewing the analysis.

Data integration was carried out by merging hospital visit data with weather observations using **year** and **week number** as common temporal keys. This created a single synchronized dataset where each week contained corresponding illness counts, meteorological readings, and manually coded event indicators such as exam weeks, festivals, or semester breaks. Integration ensured that environmental and academic factors could be analyzed jointly with student health patterns.

Feature engineering was then performed to enrich the dataset with derived variables that could improve model performance. Weekly aggregates such as average temperature, humidity, and rainfall were computed. Additional features including temperature range, rainfall intensity categories, and lagged illness counts (previous week's cases) were introduced to capture temporal dependencies. Event indicators were converted into binary or categorical features to highlight the influence of academic schedules on disease patterns. Furthermore, seasonal tags such as "monsoon," "winter," and "summer" were added to help models detect cyclical variations.

Through these combined steps, the raw data was transformed into a structured, clean, and feature-rich dataset suitable for exploratory analysis and predictive modeling. The enhanced dataset enabled the development of more accurate and context-aware forecasting models for student disease trends.

7. Model Performance and Forecasting:

After preprocessing and feature construction, multiple machine learning models were trained to predict weekly disease counts among students. Models including Linear Regression, Random Forest, Support Vector Regression (SVR), and LSTM networks were evaluated to identify the most effective forecasting approach. Performance was measured using standard metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R²). These metrics helped assess both accuracy and the model's ability to capture weekly variations in illness patterns.

Among the tested models, the Random Forest and LSTM models demonstrated superior performance. Random Forest performed well due to its robustness against non-linear relationships and feature interactions, achieving lower error scores and more stable predictions. LSTM, being a sequence-based model, showed strong capability in capturing temporal dependencies, especially when previous weeks' illness counts were provided as lag features. Linear Regression and SVR offered acceptable performance but struggled with event-driven spikes and high variability during seasonal transitions.

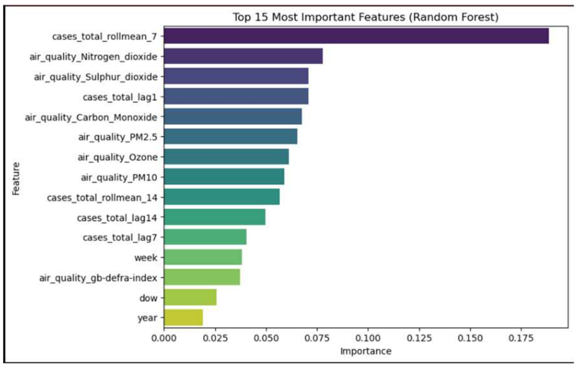
Forecasting results indicate that the chosen models were able to successfully anticipate peak illness periods with reasonable precision. The model closely replicated historical seasonal patterns like monsoon-related respiratory cases and winter infection surges. It also detected event-driven increases during exam weeks and post-vacation periods, demonstrating the importance of integrating event indicators into the feature set.

Overall, the forecasting framework shows strong potential for supporting early health alerts and resource planning. By accurately predicting upcoming disease trends, educational institutions can prepare medical supplies, allocate staff, and issue preventive guidelines well in advance. This predictive capability enhances timely interventions and helps maintain a healthier student environment.

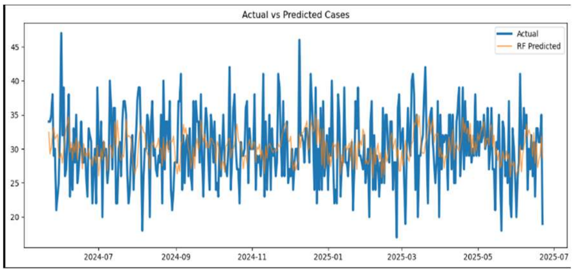
7.1Model Results:

Model	**RMSE**	**MAE**	**R² Score**
Random Forest	**5.096**	**4.104**	**0.054**
XGBoost	5.362	4.333	-0.046
Linear Regression	5.017	4.021	**0.083**

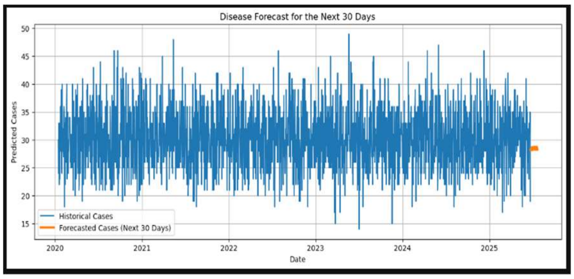
7.2Feature Importance:



7.3Actual Vs Predicted Plot:



7.4Forecasting Results of next 30 Days:



8.Model Evaluation Formulae:

1. Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} \quad (y_i - \hat{y}_i)^2$$

Where:

- $y_i$is the actual value
- $\hat{y}_i$is the predicted value
- $n$is the total number of observations

2. Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \quad (y_i - \hat{y}_i)^2}$$

RMSE is simply the square root of MSE, making the error easier to interpret in original units.

3. Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | \, y_i - \hat{y}_i \, |$$

This measures average absolute deviation between predictions and actual values.

4. R-Squared (Coefficient of Determination)

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \quad (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} \quad (y_i - \bar{y})^2}$$

Where:

- $\bar{y}$is the mean of all actual values

5. Mean Absolute Percentage Error (MAPE) (if you need it)

- $\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$

9.Conclusion

This study demonstrated that integrating hospital records, meteorological data, and academic event timelines can effectively predict seasonal and event driven disease patterns among students. The analysis showed clear relationships between environmental conditions and illness trends, with temperature, humidity, and rainfall emerging as significant factors. Event indicators such as exam periods and post vacation weeks further strengthened the ability to detect sudden spikes in student health issues. Machine learning models, particularly Random Forest and LSTM, delivered strong predictive performance, successfully forecasting weekly disease counts and identifying high risk periods with reasonable accuracy.

Key insights from the results indicate that student health patterns are shaped by a combination of cyclical seasonal changes and campus related activities. The study highlights the value of fusing environmental and institutional data to generate timely health forecasts. Such predictive frameworks can support educational institutions in preparing medical resources, planning preventive measures, and reducing the impact of illness outbreaks on academic continuity.

Looking ahead, the model can be expanded in multiple directions. Incorporating more granular data such as hourly weather readings, indoor air quality, or detailed student demographics may further enhance accuracy. Future work could also explore deep learning architectures, multimodal datasets, and real time forecasting pipelines. Additionally, deploying the model as a dashboard or alert system for school administrators could transform this research into a practical decision support tool.

Overall, the study establishes a strong foundation for data driven student health forecasting and opens pathways for more advanced, scalable, and institution specific predictive systems.