

Integrating Clustering and Multivariate Models for Enhanced Infectious Disease Forecasting in India

Sangita Bose^{1,*}

^{1,*}Swami Vivekananda University, Barrackpore, India

sangitab@svu.ac.in

Abstract

Infectious disease forecasting is of paramount importance in managing public health in India, given its vast population and diverse environmental conditions. This paper introduces a clustering-based multivariate prediction model that forecasts new infectious disease cases by capturing the interaction between various factors such as environmental variables, demographic characteristics, and historical trends. The model aggregates extensive datasets and employs clustering algorithms to segment data, enabling localized and accurate disease forecasting. Advanced multivariate prediction techniques, including machine learning algorithms and regression analysis, enhance the model's forecasting accuracy. This approach offers proactive public health strategies, aiding in identifying high-risk areas and optimizing resource allocation. The proposed framework demonstrates improved predictive accuracy and management of infectious disease threats in India, paving the way for future integration with real-time data for greater effectiveness.

1. Introduction

Infectious diseases represent a significant public health challenge worldwide, accounting for substantial morbidity and mortality. In India, this challenge is intensified due to its large population and diverse environmental conditions, necessitating effective forecasting mechanisms to mitigate outbreaks. Recent advancements in data analytics and machine learning offer promising avenues for developing robust predictive models that address the limitations of traditional univariate approaches.

This paper presents a clustering-based multivariate prediction model tailored to India's unique demographic and environmental landscape. By integrating multiple datasets, such as climatic variables (temperature, humidity, and rainfall) and demographic factors (population density, mobility patterns), the model segments regions into epidemiological clusters, enhancing the accuracy and relevance of forecasts. The proposed approach supports targeted public health interventions, optimizing resource allocation and reducing the healthcare burden.

Keywords: Clustering, Multivariate Prediction, Infectious Disease Forecasting, Data Analytics, Machine Learning, Epidemiology, Public Health Management

2. Literature Review

Traditional infectious disease forecasting methods have primarily relied on univariate models, which fail to capture the intricate interactions between multiple variables influencing disease spread. Recent studies highlight the benefits of integrating clustering techniques with predictive modeling.

- **Clustering and Forecasting:** Researchers have demonstrated the utility of clustering algorithms in epidemiology, enabling the identification of regional patterns in disease incidence. This segmentation enhances the accuracy and contextual relevance of predictive models (Khan et al., 2020).

- **Machine Learning:** Advanced machine learning techniques, such as Long Short-Term Memory (LSTM) networks, are instrumental in analyzing sequential data and capturing complex temporal dependencies.

- **Environmental and Demographic Factors:** Studies underscore the role of climatic and demographic factors in disease transmission, emphasizing the need for multivariate models that integrate these dimensions (Bhattacharya et al., 2021).

These insights form the foundation for developing a hybrid model that combines clustering and multivariate prediction techniques.

3. Methodology

3.1 Data Collection

A robust dataset was compiled from multiple reliable sources covering 2010–2020. Key variables included:

- **Disease Incidence:** Monthly cases of dengue, malaria, and typhoid (public health records).
- **Climatic Variables:** Temperature, humidity, and rainfall data (Indian Meteorological Department).
- **Demographic Data:** Population density, age distribution, and healthcare accessibility (National Sample Survey).

3.2 Data Preprocessing

1. **Missing Data Imputation:** Linear interpolation for continuous variables and mode-based imputation for categorical variables.
2. **Normalization:** Min-max scaling normalized all numeric variables to a $[0, 1]$ range.
3. **Feature Selection:** Correlation-based Feature Selection (CFS) reduced dimensionality by retaining impactful features.

3.3 Clustering Analysis

1. **Algorithm Selection:** K-means clustering was employed to group regions with similar disease and climatic patterns. The optimal number of clusters ($K=5$) was determined using the elbow method.
2. **Cluster Validation:** Silhouette score and Davies-Bouldin index assessed cluster cohesiveness and separation, yielding satisfactory results (silhouette score = 0.67; Davies-Bouldin index = 0.82).

3.4 Multivariate Prediction Model

1. **Model Selection:** An LSTM network was chosen for its ability to capture long-term dependencies in sequential data.

2. Input Variables: Disease incidence, climatic variables, and demographic factors served as input features.

3. Training and Evaluation: The LSTM model was trained on 70% of the dataset, with 15% used for validation and 15% for testing. Performance metrics included RMSE, MAE, and MAPE.

3.5 Nationwide Forecasting System Cluster-level models were integrated into a national forecasting framework, enabling real-time predictions across India and supporting public health decision-making.

4. Results

4.1 Clustering Outcomes

K-means clustering identified five distinct clusters:

- **Cluster 1:** Cold, mountainous regions with low disease incidence.
- **Cluster 2:** Temperate central India with moderate urbanization and disease spread.
- **Cluster 3:** Humid southern coastal areas with high mosquito-borne disease prevalence.
- **Cluster 4:** Urbanized regions with robust healthcare systems and lower disease rates.
- **Cluster 5:** Rural regions with poor healthcare access and high disease incidence.

4.2 Model Performance

The LSTM model outperformed baseline models (ARIMA and Random Forest) across all clusters:

Metric	LSTM	ARIMA	Random Forest
RMSE	18.08	23.46	21.87
MAE	12.33	16.24	15.12
MAPE	10.29%	14.72%	13.56%

4.3 Cluster-Specific Insights

- Urban clusters (Cluster 4) exhibited the highest forecast accuracy due to consistent data and robust healthcare infrastructure.
- Rural clusters (Cluster 5) faced higher prediction errors, highlighting challenges in data quality and access.

5. Discussion

The integration of clustering with multivariate forecasting significantly improved predictive accuracy and localized decision-making. The LSTM model's ability to handle temporal dependencies proved crucial in understanding complex epidemiological dynamics. However, rural regions require enhanced data collection mechanisms and alternative modelling approaches, such as satellite or mobile health data integration.

6. Future enhancements include:

- Real-time data integration from diverse sources, including social media and IoT.
- Exploration of advanced architectures, such as hybrid CNN-LSTM models, to capture spatial-temporal patterns.
- Transfer learning techniques to predict outbreaks of novel diseases.

Conclusion

This study demonstrates the efficacy of a clustering-based multivariate prediction model for infectious disease forecasting in India. By addressing regional diversity and integrating climatic, demographic, and historical data, the model provides accurate and localized forecasts. Future research should focus on real-time data integration and advanced modeling techniques to further enhance predictive capabilities and public health responses.

References

Bhattacharya, S., Dey, A., & Mukherjee, A. (2021). Predicting Infectious Disease Outbreaks.

Khan, A., Gupta, S., & Sharma, R. (2020). Data-Driven Approaches for Infectious Disease Forecasting.

Patel, R., Kumar, A., & Jain, S. (2020). Climate Change and Infectious Diseases.

Singh, R., Gupta, S., & Prakash, S. (2019). Machine Learning for Infectious Disease Prediction.

Sinha, R., & Das, A. (2018). Multivariate Disease Prediction Using Climatic and Demographic Data.