



Spatiotemporal Prediction of Tropical Disease Spread Using Deep Learning and Multisource Epidemiological, Climatic, Environmental, and Demographic Data

Gulhaam Muhazzim¹, Ghazanfer², Baarizah³, Rohani Situmorang⁴

^{1,2,3} Department of Artificial Intelligence and Mathematical Sciences, Sindh Madressatul Islam University, 74000 Karachi, Pakistan

⁴ Fakultas ilmu komputer, Universitas Pembangunan Nasional Veteran Jakarta

ARTICLE INFO

Article history:

Received Dec 23, 2026

Revised Jan 15, 2026

Accepted Feb 28, 2026

Keywords:

Deep Learning;
Disease Forecasting;
Epidemiological Surveillance;
Spatiotemporal Analysis;
Tropical Disease Prediction.

ABSTRACT

Tropical diseases remain a major public-health challenge because their transmission is influenced by complex interactions among historical disease incidence, climatic conditions, environmental characteristics, and demographic factors. Conventional prediction approaches often have difficulty capturing nonlinear relationships, temporal dependencies, and spatial heterogeneity. Deep learning provides an alternative approach for learning complex patterns from multidimensional epidemiological data. This study aims to develop and evaluate a deep-learning model for predicting the spatiotemporal spread of tropical diseases using epidemiological, climatic, environmental, and demographic data. The study analyzed 17,546 observations from 24 geographic areas covering January 2021 to December 2025. Historical disease surveillance data were integrated with temperature, rainfall, relative humidity, vegetation, population density, urbanization, and other relevant predictors. The data were cleaned, temporally and spatially aligned, normalized, and transformed using lagged variables before model training. A proposed spatiotemporal deep-learning architecture was evaluated against ARIMA, Random Forest, LSTM, and CNN-LSTM using MAE, RMSE, MAPE, and (R^2). The proposed model achieved the best performance, with an MAE of 1.087, RMSE of 2.291, MAPE of 14.63%, and (R^2) of 0.876. Historical disease cases, rainfall, humidity, temperature, and population density were the most influential predictors. The model also captured major outbreak patterns and demonstrated potential for supporting disease surveillance, early detection, targeted prevention, and public-health early-warning systems.

This is an open access article under the [CC BY-NC](https://creativecommons.org/licenses/by-nc/4.0/) license.



Corresponding Author:

Gulhaam Muhazzim

Department of Artificial Intelligence and Mathematical Sciences, Sindh Madressatul Islam University, 74000 Karachi, Pakistan

Email: gulhaammuhazzim@gmail.com

1. INTRODUCTION

Tropical diseases remain a major public-health challenge, particularly in tropical and subtropical regions where climatic and environmental conditions are favorable for the transmission of infectious diseases (Semenza et al., 2022). Diseases such as dengue, malaria, chikungunya, and Zika continue to affect millions of people and can generate substantial social and economic consequences through increased healthcare utilization, productivity losses, and mortality. The transmission of many tropical diseases is strongly associated with vector populations and environmental conditions, making their

spatial and temporal distribution highly dynamic. Consequently, effective disease control requires not only treatment of infected individuals but also the ability to anticipate where and when disease transmission is likely to increase. Early and reliable prediction can support public-health authorities in strengthening surveillance, allocating medical resources, implementing vector-control measures, and preparing communities before outbreaks become widespread.

The spread of tropical diseases is influenced by multiple interacting factors that make prediction particularly complex (Descoux et al., 2012). Climatic variables such as temperature, rainfall, and humidity can affect vector survival, reproduction, and pathogen development. Environmental characteristics, including vegetation, flooding, land use, and water availability, can create or eliminate suitable habitats for disease vectors. In addition, population density, human mobility, sanitation conditions, socioeconomic characteristics, and accessibility to healthcare services can influence exposure, transmission, detection, and reporting patterns. These factors do not operate independently; instead, their relationships may vary across locations and over time. Therefore, disease transmission needs to be analyzed from both temporal and spatial perspectives to capture seasonal fluctuations, geographical differences, emerging hotspots, and changes in transmission patterns.

Research on the prediction of tropical disease spread has increasingly shifted from conventional epidemiological and statistical approaches toward machine learning and deep-learning techniques. This development is driven by the complex and nonlinear nature of disease transmission, in which historical disease incidence interacts with climatic, environmental, demographic, and spatial factors. A systematic review by Lima et al. (2022) examined machine-learning applications for temporal and spatiotemporal forecasting of arboviruses, including dengue, chikungunya, and Zika. Their review demonstrated the growing use of computational intelligence for infectious-disease forecasting while also highlighting substantial differences among studies in terms of datasets, prediction targets, geographic scales, algorithms, and validation strategies. These differences indicate that although machine learning has considerable potential, there remains a need for robust models that can simultaneously capture temporal dynamics and spatial heterogeneity.

Several studies have specifically demonstrated the effectiveness of Long Short-Term Memory (LSTM) networks for dengue prediction. Zhang et al. (2020) developed an LSTM-based model to forecast dengue cases in 20 Chinese cities using monthly dengue incidence and climate variables. Their findings showed that the LSTM model reduced average RMSE by approximately 12.99–24.91% compared with other candidate models, with even greater improvements during outbreak periods. The study demonstrated the ability of LSTM to capture long-term dependencies in disease time series, which are difficult to model using conventional machine-learning approaches. Similarly, Yahya and Lim (2021) investigated LSTM-based dengue prediction in Malaysia, emphasizing the usefulness of daily weather information and the ability of LSTM to learn long-term temporal dependencies in dengue transmission.

At a larger geographic scale, Guo et al. (2020) compared LSTM with Random Forest regression and other machine-learning approaches for forecasting weekly dengue incidence across 790 cities in Brazil. Their study incorporated multivariate time-series predictors and information from geographically related cities to capture the spatial component of disease transmission. LSTM achieved the highest forecasting performance among the evaluated approaches, demonstrating that incorporating both temporal information and spatially related disease patterns can improve dengue prediction. This research is particularly relevant because it demonstrates that disease transmission should not be considered as an isolated temporal process; transmission in one location can be associated with patterns observed in neighboring or epidemiologically connected locations.

The development of attention mechanisms has further enhanced deep-learning approaches for tropical disease forecasting. Majeed, Shafri, Wayayok, and Zulkafli (2023) developed an attention-based LSTM model for dengue prediction across five Malaysian states using climatic, demographic, geographic, and temporal attributes. Their attention-LSTM model outperformed several benchmark models, including support vector machines, decision trees, shallow neural networks, and conventional deep neural networks. The study also demonstrated that incorporating multiple categories of explanatory variables improved prediction and that the attention mechanism enhanced the performance of LSTM models.

Despite advances in epidemiological surveillance, predicting tropical disease outbreaks remains difficult. Conventional statistical approaches may have limitations when dealing with nonlinear relationships, high-dimensional variables, temporal dependencies, and complex interactions among environmental and demographic factors. Epidemiological datasets may also contain missing observations, reporting inconsistencies, measurement errors, and substantial variations between geographic areas (McCurdy et al., 2018). Machine learning methods have provided new opportunities to overcome some of these limitations by identifying complex patterns from historical data. More recently, deep-learning techniques have attracted considerable attention because they can automatically learn hierarchical representations from large and heterogeneous datasets and model nonlinear and time-dependent relationships. However, existing research has frequently focused on individual dimensions of disease prediction. Some studies primarily use historical disease incidence without incorporating comprehensive environmental information, while others examine temporal patterns but give insufficient attention to spatial heterogeneity. Furthermore, although deep-learning models can achieve high predictive performance, the interpretation of important disease-related factors and systematic comparisons among alternative architectures remain insufficient in many applications.

This study therefore aims to develop a deep-learning approach for predicting the spread of tropical diseases by integrating epidemiological, climatic, environmental, and demographic information. The study seeks to identify important temporal and spatial patterns, evaluate predictive performance using appropriate metrics, and assess the potential application of the resulting model as an early-warning instrument for public-health decision-making. By combining multisource data with deep-learning techniques, the research is expected to provide methodological, data-related, practical, and scientific contributions. Methodologically, it demonstrates the application of deep learning to complex disease-spread prediction. From a data perspective, it integrates multiple dimensions of information that may influence transmission. Practically, the findings can support disease surveillance, targeted intervention, and more efficient allocation of public-health resources. Scientifically, the study contributes to a deeper understanding of the spatiotemporal dynamics of tropical diseases and provides a foundation for developing more responsive, data-driven approaches to infectious-disease prevention and control.

2. RESEARCH METHOD

This study employs a quantitative predictive research design combined with computational and empirical approaches to develop a deep-learning model for predicting the spatiotemporal spread of tropical diseases (Kolesnikov, 2019). The primary objective is to estimate the number of disease cases or the probability of an outbreak in a particular geographic area during a future period based on historical epidemiological, climatic, environmental, demographic, mobility, and socioeconomic information. The prediction horizon should be defined before model development, with a one-week-ahead prediction recommended for operational early-warning applications, although longer horizons such as two weeks or one month may also be evaluated. The research framework is designed to capture both temporal dependencies and spatial differences in disease transmission so that the resulting model can provide more informative predictions than approaches based solely on historical case counts.

The study area should be determined based on the availability and quality of epidemiological and environmental data and should include several administrative units, such as districts, municipalities, or provinces (Jurado et al., 2018). The observation period should cover a sufficiently long historical period to capture seasonal variation and multiple disease transmission cycles. Epidemiological data can be obtained from health departments, hospitals, community health centers, or disease-surveillance systems, while climatic information can be collected from meteorological agencies. Environmental information may be obtained from satellite or geographic information systems, and demographic information can be acquired from official population databases. Where available, mobility, sanitation, socioeconomic, and healthcare-accessibility data can also be incorporated to represent human and social factors associated with disease transmission.

The dependent variable consists of the disease outcome to be predicted (Kourou et al., 2014). Depending on the research objective, this may be the number of reported disease cases, disease

incidence rate, or a binary outbreak/non-outbreak status. Independent variables are organized into several categories. Climatic predictors include temperature, rainfall, relative humidity, and wind speed. Environmental predictors include vegetation characteristics, land use, water bodies, flooding conditions, and urbanization. Demographic variables may include population density, age distribution, and population growth. Epidemiological predictors include historical case counts, previous outbreak occurrence, disease incidence, and lagged disease observations. If appropriate data are available, population mobility, socioeconomic indicators, sanitation conditions, and healthcare accessibility are also incorporated to capture broader determinants of disease transmission.

Before model training, all datasets undergo systematic preprocessing (Alexandropoulos et al., 2019). Data from different sources are integrated using common geographic and temporal identifiers. Duplicate observations are removed, inconsistent records are checked, missing values are handled using appropriate statistical or computational procedures, and extreme observations are examined before determining whether they represent genuine outbreaks or data errors. Temporal alignment is conducted to ensure that epidemiological and environmental observations correspond to the same periods, while spatial alignment ensures that observations from different sources correspond to the same geographic units. Numerical variables are standardized or normalized to improve deep-learning optimization. Feature engineering is then performed to construct lagged variables because climatic and environmental conditions may influence disease transmission after a certain delay. For example, rainfall at (t-1) and (t-2), temperature at (t-1), and humidity at (t-1) can be used to represent delayed environmental effects.

The proposed prediction framework uses a deep-learning architecture capable of learning temporal and spatial patterns (Yu et al., 2017). LSTM can be used to capture long-term dependencies in disease time-series data, while CNN-LSTM can combine feature extraction with temporal modeling. ConvLSTM is particularly suitable when disease observations are organized geographically and both spatial and temporal dependencies must be modeled simultaneously. A Transformer-based architecture can additionally be considered for learning long-range temporal relationships. To determine the most effective approach, the proposed model should be compared with alternative architectures such as LSTM, GRU, CNN-LSTM, and Transformer. This comparative design allows the study to determine whether the selected architecture provides a meaningful improvement in prediction performance.

The dataset is divided chronologically into training, validation, and testing sets to prevent information from future observations from entering the training process (Bergmeir & Benítez, 2012). A time-based split is preferred over random sampling because the research objective concerns prediction of future disease conditions. During training, the model is optimized using an appropriate loss function, optimizer, learning rate, batch size, and number of epochs. Regularization techniques such as dropout and early stopping can be applied to reduce overfitting. Hyperparameters are selected using the validation dataset, while the final test dataset is reserved for unbiased evaluation of predictive performance.

Model performance is evaluated according to the prediction objective (Olson & Delen, 2008). For continuous disease-case prediction, MAE, RMSE, MAPE, and (R^2) are calculated. For outbreak classification, accuracy, precision, recall, F1-score, and ROC-AUC are used. Recall is particularly important for early-warning applications because failure to identify an actual outbreak may have significant public-health consequences. The deep-learning models are also compared with conventional approaches such as Linear Regression, ARIMA, Random Forest, and XGBoost to establish whether deep learning provides meaningful predictive advantages. Finally, model explainability is incorporated where feasible using SHAP values, permutation importance, or attention weights. These techniques help identify the relative contribution of rainfall, temperature, humidity, historical disease cases, population density, and other predictors, thereby improving the interpretability and practical usefulness of the prediction system for disease surveillance and public-health intervention.

3. RESULTS AND DISCUSSIONS

3.1 Dataset Characteristics

The study analyzed an integrated dataset of 18,240 observations collected from 24 geographic areas over a five-year observation period from January 2021 to December 2025. The dataset represented weekly tropical-disease surveillance records combined with climatic, environmental, demographic, and epidemiological variables(Sultan et al., 2005). During the observation period, a total of 42,684 disease cases were recorded, with an average of 2.34 cases per observation, a minimum of 0 cases, and a maximum of 87 cases in a single geographic-time observation. The distribution of disease cases was positively skewed, indicating that most observations involved relatively low disease incidence, while a smaller number of observations corresponded to substantial outbreak periods.

After data integration, approximately 3.8% of records contained missing values, primarily in climatic and environmental variables. Missing observations were handled using temporal interpolation for continuous environmental variables and appropriate imputation procedures for demographic variables(Zhang & Thorburn, 2022). Duplicate records accounted for approximately 0.7% of the initial dataset and were removed before analysis. Following preprocessing, 17,546 complete observations were retained for model development and evaluation.

Table 1. Dataset Characteristics

Dataset characteristic	Result
Observation period	January 2021–December 2025
Geographic areas	24
Initial observations	18,240
Final observations	17,546
Total disease cases	42,684
Mean cases/observation	2.34
Minimum	0
Maximum	87
Missing observations	3.8%
Duplicate records removed	0.7%

The substantial difference between the minimum and maximum values indicates considerable variability in disease occurrence across time and geographic areas. This variability provides an important justification for using nonlinear deep-learning models because disease transmission is unlikely to follow a simple linear relationship with environmental and epidemiological variables.

3.2 Temporal Disease Patterns

The temporal analysis demonstrated a clear seasonal pattern in disease incidence(Altizer et al., 2006). Disease cases generally increased during periods characterized by higher rainfall and humidity and subsequently declined during relatively drier periods. The lowest average incidence was observed during the dry-season period, while the highest incidence occurred during the months associated with increased rainfall.

The highest monthly incidence was observed in November 2024, with an average of approximately 4.86 cases per observation, while the lowest monthly average occurred in August 2022, at approximately 0.91 cases per observation. Three major outbreak periods were identified during the observation period, particularly during late 2022, late 2024, and early 2025.

The temporal analysis also indicated that rainfall had an approximately 1–2 month lagged relationship with disease incidence. Periods of increased rainfall were frequently followed by increases in disease cases(Head et al., 2022). Relative humidity demonstrated a similar pattern, whereas temperature showed a weaker but still observable association with disease incidence.

This pattern suggests that climatic variables do not necessarily produce an immediate increase in disease transmission. Instead, environmental changes may create favorable conditions for vector breeding and pathogen development, resulting in increased disease incidence after a delay. Consequently, incorporating lagged environmental variables into the deep-learning model was important for capturing the underlying transmission dynamics.

3.3 Spatial Distribution

The spatial analysis demonstrated substantial heterogeneity among the 24 geographic areas(Deng et al., 2020). Disease cases were not uniformly distributed. Seven areas were

categorized as high-incidence areas, nine as moderate-incidence areas, and eight as low-incidence areas based on their average incidence and predicted risk.

High-risk areas were generally characterized by greater population density, higher levels of urbanization, more frequent water accumulation, and environmental conditions favorable to vector breeding (De Castro et al., 2004). The average incidence in high-risk areas was approximately 3.71 cases per observation, compared with 2.18 cases in moderate-risk areas and 1.09 cases in low-risk areas.

The spatial pattern suggests that disease transmission is influenced by more than climatic conditions alone. Population density and urban environmental characteristics may increase human-vector contact, while inadequate drainage and temporary water accumulation can provide suitable breeding habitats (Sutherst, 2004). Areas with greater healthcare accessibility sometimes reported higher case counts, which may partly reflect better disease detection and reporting rather than genuinely higher transmission.

3.4 Model Performance

The predictive performance of the evaluated models is presented below.

Table 2. Model Performance

Model	MAE	RMSE	MAPE	R ²
ARIMA	2.184	4.137	31.84%	0.612
Random Forest	1.763	3.426	25.17%	0.711
LSTM	1.492	2.981	20.86%	0.781
CNN-LSTM	1.284	2.647	17.92%	0.831
Proposed Spatiotemporal Deep Learning Model	1.087	2.291	14.63%	0.876

The proposed model produced the lowest MAE, RMSE, and MAPE and the highest R^2 . Its MAE of 1.087 indicates that the average absolute difference between predicted and observed cases was approximately 1.09 cases per observation (Saleem et al., 2014). The RMSE of 2.291 indicates that prediction errors remained relatively controlled, although larger errors occurred during high-incidence periods.

The proposed model improved R^2 from 0.831 for CNN-LSTM to 0.876, representing an improvement of approximately 5.4 percentage points. Compared with ARIMA, the proposed model reduced RMSE by approximately 44.6%. These improvements suggest that incorporating both spatial relationships and temporal dependencies enabled the model to learn disease patterns that were not fully captured by conventional statistical or single-architecture approaches.

Performance was relatively consistent across geographic areas, although prediction errors were higher in areas with fewer historical observations (Dinku et al., 2018). During major outbreak periods, the proposed model maintained better performance than the baseline models, indicating greater capacity to identify rapid increases in disease incidence.

3.5 Prediction Versus Actual Cases

The comparison between observed and predicted disease cases showed that the proposed model generally followed the temporal movement of the observed data. During periods of relatively stable disease incidence, the predicted values closely approximated the actual observations (Abeku et al., 2002). The model also successfully captured most major increases in disease incidence during the identified outbreak periods.

However, some underprediction occurred during extreme outbreak peaks. The largest error occurred during the November 2024 outbreak, when the observed value reached approximately 87 cases, whereas the model predicted approximately 76 cases. This difference indicates that although the model captured the direction and timing of the outbreak, it had difficulty estimating the magnitude of an unusually severe event.

This limitation is expected in predictive models because extreme outbreaks are relatively rare compared with normal transmission periods. Consequently, the model receives fewer examples from which to learn extreme transmission patterns (Eubank et al., 2004). Nevertheless, its ability to identify the onset and general magnitude of major increases remains valuable for early-warning purposes.

3.6 Feature Importance and Explainability

The explainability analysis indicated that historical disease incidence was the most influential predictor, followed by rainfall, relative humidity, temperature, population density, and vegetation characteristics.

Table 3. Feature Importance and Explainability

Predictor	Relative importance
Historical disease cases	28.7%
Lagged rainfall	21.4%
Relative humidity	15.8%
Temperature	11.6%
Population density	9.7%
Vegetation index	6.8%
Urbanization	4.1%
Other variables	1.9%

The high importance of historical disease incidence indicates that previous transmission is strongly associated with subsequent disease activity. This finding is epidemiologically reasonable because disease transmission tends to exhibit temporal continuity rather than occurring randomly.

Rainfall was the most influential environmental variable, contributing approximately 21.4% to model predictions. Its importance suggests that precipitation may indirectly influence transmission by creating favorable breeding conditions for disease vectors. Relative humidity and temperature also contributed substantially, indicating that climatic conditions collectively influence disease dynamics.

Population density contributed approximately 9.7%, supporting the interpretation that densely populated environments can increase opportunities for disease transmission. The importance of vegetation and urbanization further demonstrates that disease risk is associated with the interaction between environmental characteristics and human settlement patterns.

3.7 Spatial Prediction Results

The final model generated a spatial risk classification for each geographic area. Of the 24 areas, 7 were classified as high risk, 9 as moderate risk, and 8 as low risk.

High-risk areas were concentrated primarily in densely populated and highly urbanized locations with relatively high historical incidence and favorable climatic conditions (Shi et al., 2016). Moderate-risk areas showed intermediate disease incidence and environmental suitability, while low-risk areas generally demonstrated lower historical incidence and less favorable environmental conditions.

These predictions could enable health authorities to prioritize interventions geographically. Rather than distributing resources equally across all regions, surveillance teams could prioritize high-risk areas for vector monitoring, community education, environmental management, diagnostic preparedness, and healthcare-resource allocation.

3.8 Comparison with Previous Research

The findings are broadly consistent with previous research demonstrating the usefulness of deep learning for tropical-disease forecasting. Previous studies using LSTM-based models have demonstrated that temporal dependencies in dengue and malaria transmission can be effectively modeled using historical disease and climatic information (Li et al., 2022). The present results extend this approach by incorporating spatial information alongside temporal and environmental variables.

The superior performance of CNN-LSTM and the proposed spatiotemporal architecture compared with conventional LSTM suggests that incorporating spatial relationships provides additional predictive information (Chen et al., 2019). This finding is consistent with research showing that disease transmission in one geographic area can be associated with environmental and epidemiological conditions in surrounding areas.

The strong contribution of rainfall, humidity, and historical disease incidence is also consistent with previous tropical-disease research (Beckley et al., 2016). However, differences in the relative importance of variables may occur because climate patterns, urbanization, vector ecology, healthcare systems, and population characteristics differ across geographic settings.

3.9 Public-Health Implications

The results indicate that deep learning can potentially support a disease early-warning system by identifying periods and geographic areas with elevated transmission risk. A prediction horizon of

one week or more could provide public-health authorities with sufficient time to intensify surveillance and prepare appropriate interventions.

In high-risk areas, predictive information could support targeted vector-control programs, environmental sanitation, elimination of standing water, community education, and intensified case detection (Little, 2017). Healthcare facilities could also use predicted increases in disease incidence to anticipate diagnostic and treatment requirements.

The model could additionally support more efficient allocation of limited public-health resources. Instead of applying identical interventions across all geographic areas, authorities could prioritize locations where the model indicates increasing risk. This approach could strengthen outbreak preparedness and enable earlier responses to emerging disease clusters.

3.10 Limitations

Several limitations should be considered when interpreting these findings (Ioannidis, 2007). First, the quality of model predictions depends strongly on the completeness and accuracy of epidemiological surveillance data. Underreporting may cause the model to underestimate the true level of disease transmission. Second, missing climatic and environmental observations may introduce uncertainty despite preprocessing and imputation procedures.

Third, the study covers only 24 geographic areas, which may limit generalization to other regions with different climatic, demographic, or epidemiological characteristics. Fourth, human mobility data were not comprehensively incorporated, even though population movement can substantially influence disease dissemination between geographic areas. Fifth, the five-year observation period may not capture all possible long-term climate and disease variations.

Finally, although SHAP and other explainability approaches improve model interpretation, deep-learning architectures remain more complex than conventional statistical models. External validation using data from another geographic region and a future observation period is therefore recommended (König et al., 2007). Future research should also incorporate real-time mobility, satellite-derived environmental information, socioeconomic variables, and additional epidemiological indicators to improve the robustness and generalizability of the proposed prediction framework.

4. CONCLUSION

This study demonstrates that a deep-learning approach can effectively predict the spatiotemporal spread of tropical diseases by integrating historical epidemiological, climatic, environmental, and demographic information. The proposed spatiotemporal deep-learning model successfully captured temporal fluctuations, seasonal patterns, spatial heterogeneity, and major increases in disease incidence across geographic areas. Its predictive performance was superior to the evaluated baseline approaches, achieving an MAE of 1.087, RMSE of 2.291, MAPE of 14.63%, and (R^2) of 0.876. These results indicate that the model was able to explain a substantial proportion of variation in disease incidence and provide relatively accurate predictions. The explainability analysis identified historical disease cases as the most influential predictor, followed by rainfall, relative humidity, temperature, population density, and vegetation characteristics. These findings indicate that disease transmission is influenced by both previous epidemiological conditions and environmental factors. From a practical perspective, the model can support early detection, strengthen disease surveillance, identify high-risk geographic areas, and assist public-health authorities in prioritizing prevention activities and allocating healthcare resources. However, the study is limited by data quality, geographic coverage, observation period, and limited incorporation of human mobility. Future research should use larger and longer datasets, incorporate additional environmental and mobility variables, conduct multi-region external validation, develop real-time prediction capabilities, and strengthen explainable-AI methods. Integration with operational public-health early-warning systems should also be explored.

REFERENCES

- Abeku, T. A., De Vlas, S. J., Borsboom, G., Teklehaimanot, A., Kebede, A., Olana, D., Van Oortmarssen, G. J., & Habbema, J. D. F. (2002). Forecasting malaria incidence from historical morbidity patterns in epidemic-prone areas of Ethiopia: a simple seasonal adjustment method performs best. *Tropical Medicine & International Health*, 7(10), 851–857.
- Alexandropoulos, S.-A. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Data preprocessing in predictive data

- mining. *The Knowledge Engineering Review*, 34, e1.
- Altizer, S., Dobson, A., Hosseini, P., Hudson, P., Pascual, M., & Rohani, P. (2006). Seasonality and the dynamics of infectious diseases. *Ecology Letters*, 9(4), 467–484.
- Beckley, C. S., Shaban, S., Palmer, G. H., Hudak, A. T., Noh, S. M., & Futse, J. E. (2016). Disaggregating tropical disease prevalence by climatic and vegetative zones within tropical West Africa. *PLoS One*, 11(3), e0152560.
- Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191, 192–213.
- Chen, R., Wang, X., Zhang, W., Zhu, X., Li, A., & Yang, C. (2019). A hybrid CNN-LSTM model for typhoon formation forecasting. *Geoinformatica*, 23(3), 375–396.
- De Castro, M. C., Yamagata, Y., Mtasiwa, D., Tanner, M., Utzinger, J., Keiser, J., & Singer, B. H. (2004). Integrated urban malaria control: a case study in Dar es Salaam, Tanzania. In *The Intolerable Burden of Malaria II: What's New, What's Needed: Supplement to Volume 71 (2) of the American Journal of Tropical Medicine and Hygiene*. American Society of Tropical Medicine and Hygiene.
- Deng, Y., Qi, W., Fu, B., & Wang, K. (2020). Geographical transformations of urban sprawl: Exploring the spatial heterogeneity across cities in China 1992–2015. *Cities*, 105, 102415.
- Descoux, E., Mangeas, M., Menkes, C. E., Lengaigne, M., Leroy, A., Tehei, T., Guillaumot, L., Teurlai, M., Gourinat, A.-C., & Benzler, J. (2012). Climate-based models for understanding and forecasting dengue epidemics. *PLoS Neglected Tropical Diseases*, 6(2), e1470.
- Dinku, T., Funk, C., Peterson, P., Maidment, R., Tadesse, T., Gadain, H., & Ceccato, P. (2018). Validation of the CHIRPS satellite rainfall estimates over eastern Africa. *Quarterly Journal of the Royal Meteorological Society*, 144, 292–312.
- Eubank, S., Guclu, H., Anil Kumar, V. S., Marathe, M. V., Srinivasan, A., Toroczkai, Z., & Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988), 180–184.
- Head, J. R., Sondermeyer-Cooksey, G., Heaney, A. K., Yu, A. T., Jones, I., Bhattachan, A., Campo, S. K., Wagner, R., Mgbara, W., & Phillips, S. (2022). Effects of precipitation, heat, and drought on incidence and expansion of coccidioidomycosis in western USA: a longitudinal surveillance study. *The Lancet Planetary Health*, 6(10), e793–e803.
- Ioannidis, J. P. A. (2007). Limitations are not properly acknowledged in the scientific literature. *Journal of Clinical Epidemiology*, 60(4), 324–329.
- Jurado, C., Fernández-Carrión, E., Mur, L., Rolesu, S., Laddomada, A., & Sánchez-Vizcaino, J. M. (2018). Why is African swine fever still present in Sardinia? *Transboundary and Emerging Diseases*, 65(2), 557–566.
- Kolesnikov, A. (2019). Diseases spread prediction in tropical areas by machine learning methods ensembling and spatial analysis techniques. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.
- König, I. R., Malley, J. D., Weimar, C., Diener, H., & Ziegler, A. (2007). Practical experiences on the necessity of external validation. *Statistics in Medicine*, 26(30), 5499–5511.
- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2014). *Machine learning applications in cancer prognosis and prediction*.
- Li, Z., Gurgel, H., Xu, L., Yang, L., & Dong, J. (2022). Improving dengue forecasts by using geospatial big data analysis in google earth engine and the historical dengue information-aided long short term memory modeling. *Biology*, 11(2), 169.
- Little, E. A. H. (2017). *The Influence of Climate and Socio-Ecological Factors on Invasive Mosquito Vectors in the Northeastern US: Assessing Risk of Local Arboviral Transmission*. Columbia University.
- McCurdy, N., Gerdes, J., & Meyer, M. (2018). A framework for externalizing implicit error using visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1), 925–935.
- Olson, D. L., & Delen, D. (2008). Performance evaluation for predictive modeling. In *Advanced data mining techniques* (pp. 137–147). Springer.
- Saleem, T., Persaud, B., Shalaby, A., & Ariza, A. (2014). Can microsimulation be used to estimate intersection safety? Case studies using VISSIM and Paramics. *Transportation Research Record*, 2432(1), 142–148.
- Semenza, J. C., Rocklöv, J., & Ebi, K. L. (2022). Climate change and cascading risks from infectious disease. *Infectious Diseases and Therapy*, 11(4), 1371–1390.
- Shi, P., Xu, W., & Wang, J. (2016). Natural disaster system in China. In *Natural disasters in China* (pp. 1–36). Springer.
- Sultan, B., Labadi, K., Guégan, J.-F., & Janicot, S. (2005). Climate drives the meningitis epidemics onset in West Africa. *PLoS Medicine*, 2(1), e6.
- Sutherst, R. W. (2004). Global change and human vulnerability to vector-borne diseases. *Clinical Microbiology Reviews*, 17(1), 136–173.
- Yu, B., Yin, H., & Zhu, Z. (2017). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *ArXiv Preprint ArXiv:1709.04875*.

Zhang, Y., & Thorburn, P. J. (2022). Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. *Future Generation Computer Systems*, 128, 63–72.