

Original Research

Received 31 May 2026

Revised 29 July 2026

Accepted 08 August 2026

Natural Resources for Human Health 2026, Vol. 6 (14s): 281–288

KEYWORDS:

Epidemic forecasting, cross-attention, time-series, uncertainty quantification, COVID-19, exogenous covariates..

EpiAttenNet: Dual-Stream Cross-Attention Neural Network for Multi-Horizon Epidemic Time-Series Forecasting with Uncertainty Quantification

Ciyamala Kushbu S¹, Chettiyar Vani Vivekanand², Arthi Devarani P³, Kavitha P⁴, Sugunadevi C⁵, Susheelamma K H⁶

¹Associate Professor, Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru-560037 Email id: ciyamala91@gmail.com

²Professor, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, Tamil Nadu. Email: chettiyarvanivivekanand@rajalakshmi.edu.in

³Assistant Professor, Artificial Intelligence and Data Science, R.M.K College of Engineering and Technology, TamilNadu,India.Email: arthidevaraniads@rmkcet.ac.in

⁴Assistant professor, Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru-560037 Email: kavithapalani17@gmail.com

⁵Assistant professor, Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru-560037 Email: sugunadevic@gmail.com

⁶Associate professor, Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru-560037 Email: susheela.kh@gmail.com

ABSTRACT: The problem of reliable prediction of the spread of infectious diseases has been one of the most important but unsolved issues of infectious disease dynamics, particularly when outbreaks are evolving in irregular fashion at time scales comparable to those of the epidemic process, as is the case with most epidemics. The forecasting of the spread of infectious diseases has been one of the most important but unsolved issues of the infectious disease dynamics, particularly when the external conditions are heterogeneously varying and the spreading process non-linear, which is typical for an outbreak evolving at a time scale comparable to the time scale of the epidemic process, as is typical for most epidemics. The current methods are based either on epidemiological assumptions that are only suitable for a specific compartment, which are very poor generalisers over a broad range of outbreak scenarios, or on sequence models that are very poor generalisers over a broad range of outbreak scenarios. A new Dual-Stream Cross-Attention (DSCA) architecture is introduced for simultaneously modeling epidemic case trajectories and exogenous covariates (mobility index, vaccination rates, non-pharmaceutical interventions (NPI) severity scores) that are heterogeneous to each other. For this purpose, a novel Dual-Stream Cross-Attention (DSCA) architecture is proposed, where the original temporal factors are separately encoded in parallel temporal convolutional streams, and then fused together by a learnable cross-attention mechanism. Multi-quantile regression head produces both point forecasts and calibrated prediction intervals of length 90%, thus providing the decision support regarding uncertainty information. The mean absolute error (MAE) and RMSE of EpiAttenNet are 3.12 and 4.78 respectively, representing a 24.3% and 25.8% improvement compared to the best of baseline models on time-series benchmarks with a 1-day horizon, 7-day horizon, and 14-day horizon respectively. The findings through ablation research verify each architectural member's work acting individually towards performance. The code and the pre-trained

weights are released. Keywords: Epidemic forecasting, cross-attention, time-series, uncertainty quantification, COVID-19, exogenous covariates.

1. INTRODUCTION

Near-term predictions of the spread of infectious diseases directly shape the impact of the population health response—such as hospital surge planning, vaccine rollout sequencing, and non-pharmaceutical intervention (NPI) timing—days ahead of time [1]. This has not been achieved yet with a well-developed, generalised, and uncertainty-aware forecasting system that is useful in a variety of different outbreak contexts without re-estimation of disease-specific parameters. Most classical compartmental models are based on the idea of mechanistic transmission, such as the SIR model and its variants, but this is only valid for stationary effective reproduction number. In contrast, data-driven approaches (recurrent networks, temporal convolutional networks (TCN) and standard Transformer architectures) have been found to be effective on historical benchmarks, but typically process the number of cases of an epidemic as a univariate time series, without making use of the many exogenous context signals (e.g., mobility, policy stringency and immunity profiles) that epidemiologists routinely use in their analyses [3]. Another aspect that is under-researched is uncertainty communication. Public health authorities have not just a need for point-forecasts, but calibrated prediction intervals that can communicate the level of confidence in the forecasts as a function of the horizon length and outbreak volatility. Current deep learning forecasters do not necessarily provide (statistically valid) uncertainty estimates, limiting their operational value [4].

This paper contributes in the following novel ways:

- We introduce EpiAttenNet, the first dual-stream cross-attention network model for multivariate epidemics, consisting of a stream representing the time evolution of the case sequence, and a parallel stream representing the time evolution of the exogenous covariates (exogenous covariates can be different, e.g., temperature, rainfall, etc.).
- The two streams are fused together by a special cross-attention network model layer, which learns which exogenous covariates are relevant at each time step of the predictions. To overcome this challenge, we propose a new Temporal Phase-Aware Positional Encoding (TPAPE) scheme that is based on additional phase indicators (growth, peak, decay, and trough), which are obtained by the output of a lightweight online phase classifier and added to position to make the attention mechanism more capable of weighting temporally distant but phase-equivalent historical windows.
- No ensembling or Monte Carlo dropout needed, train EpiAttenNet with a multi-quantile regression forecast head with a single pinball loss objective.
- We conduct in-depth evaluation of benchmarks for COVID-19 and Mpox on three forecast periods (1-day, 7-day, 14-day) and ablation experiments, as well as attention visualisation, which shows state-of-the-art performance and interpretable attention patterns aligned with known epidemiological dynamics.

2. LITERATURE SURVEY

The study of epidemic forecasting has been undertaken based on two general paradigms: mechanistic and data-driven. We also talk about recent advances in both, and literature on attention-based sequence modelling for 2021-2025 in chronological order of contribution. Mechanistic-hybrid models. To predict COVID-19, Rackauckas et al. [1] proposed using Neural ODEs and SIR dynamics, which outperforms classic SEIR models but has a high computational cost for training gradient-through-ODE. This was extended by Wang et al. [2] who added LSTM-corrected residuals to SEIR state equations which lead to a 18% decrease in the MAE compared to the state equations of a classical compartmental model on data for the USA states for 2021. While both methods rely on domain knowledge, they are limited to diseases with homogeneous natural history in that they make use of fixed epidemiological compartments. Recurrent sequence models. In another study by Chimmula and Zhang [3] the variants of LSTM were benchmarked to predict the cases of COVID-19 in Canada with the use of mean absolute error (MAE).

This study concluded that bidirectional LSTM with attention gates was 12% MAE better than the vanilla LSTM. In an epidemic-based multi-country study, Deng et al. [5] compared multiple architectures—GRU, LSTM and TCN with dilated causal convolutions—where they found that TCN with dilated causal convolutions outperformed the other models in capturing the

weekly seasonality patterns while being free from the vanishing gradient problem issue of RNNs. However, the two studies lacked exogenous mobility and/or policy signals, limiting their applicability in the real world. Transformer-based forecasters. Vaswani-based architectures were adopted to epidemic forecasting to reduce memory cost by 60%, while maintaining competitive accuracy, as done by Zhou et al. [6]. Nie et al. [10] proposed PatchTST, a channel-independent patch tokenisation method for long-horizon forecasting, which demonstrated that patch tokenisation is an effective way to improve the ability to capture the long-range dependency of time-series. In particular, Li et al. [7] explored the potential of Transformer for influenza surveillance, where they found that self-attention alone would be less accurate on short (1-3 day) timeframes due to the lack of sufficient sequence length for self-attention to be helpful compared to recurrence. Uncertainty quantification. Conformal prediction epidemic prediction models were developed by Angelopoulos et al. [4] that provide distribution free coverage guarantees, but require large calibration sets.

To encourage the use of distributional properties in making downstream decisions, DeepAR, a probabilistic autoregressive RNN with negative binomial output distributions has been shown to perform better than point-forecast models [8] (motivating the quantile regression head in EpiAttenNet). Cross-variable attention. Similarly, Zhou [9] proposed the ProbSparse attention of the Informer architecture to make predictions for long sequences, and was applied to pandemic data by Liu et al. with some minor modifications to it. There is no existing work [11-19] that has proposed a dedicated dual-stream cross-attention mechanism, where the epidemic stream and exogenous stream are encoded separately, and then fused together, which preserves the inductive biases of each stream while enabling cross-variable attention, the main architectural novelty of this work.

3. PROPOSED METHODOLOGY

The EpiAttenNet Architecture is shown in Fig 1.

Problem Formulation

The d-dimensional epidemic state vector at time t is called Let $X_{\text{epi}}(t)$. This vector has been defined as daily confirmed cases, hospitalisation rates and test positivity. In the equation, the exogenous covariate vector is denoted as $X_{\text{exo}}(t)$ and is given as the vector consisting of Google Mobility index, Oxford Government Response Stringency Index, and cumulative vaccination coverage. The model has a lookback window of L time steps and learns a mapping

$F: \{X_{\text{epi}}(t-L:t), X_{\text{exo}}(t-L:t)\} \rightarrow \{\hat{y}(t+h) : h \in \{1,7,14\}\}$, with $\hat{y}$ being the quantile predictions at the three target horizons.

Temporal Convolutional Encoding

Both the epidemic stream and exogenous stream are processed by an independently designed Temporal Convolutional Network (TCN) block, comprising four dilated causal convolutional layers with dilated $\{1, 2, 4, 8\}$ convolutional kernels, and with residual skip connections. This architecture provides a receptive field of 22 timesteps across a single TCN layer, far exceeding the required size to accommodate a weekly epidemic seasonality (period 7) and bi-weekly growth cycles (period 14) without repetitions. All the time, weight normalisation and GELU activations are applied. Epidemic stream representation H_{epi} and exogenous stream representation H_{exo} are combined by the proposed Dual-Stream Cross-Attention (DSCA) after the TCNs are encoded.

Dual-Stream Cross-Attention mechanism.

The standard self-attention is used in each stream separately. Cross-attention is then computed using H_{epi} as queries, and H_{exo} as keys/values, and yields a cross-attended epidemic representation H_{cross} that captures the exogenous signals that are most predictive to the future trajectory of the epidemic at each timestep. When the epidemic trajectory is self-predictable, the contribution from H_{cross} relative to self-attended H_{epi} is regulated by a gating scalar g which is sigmoid activated.

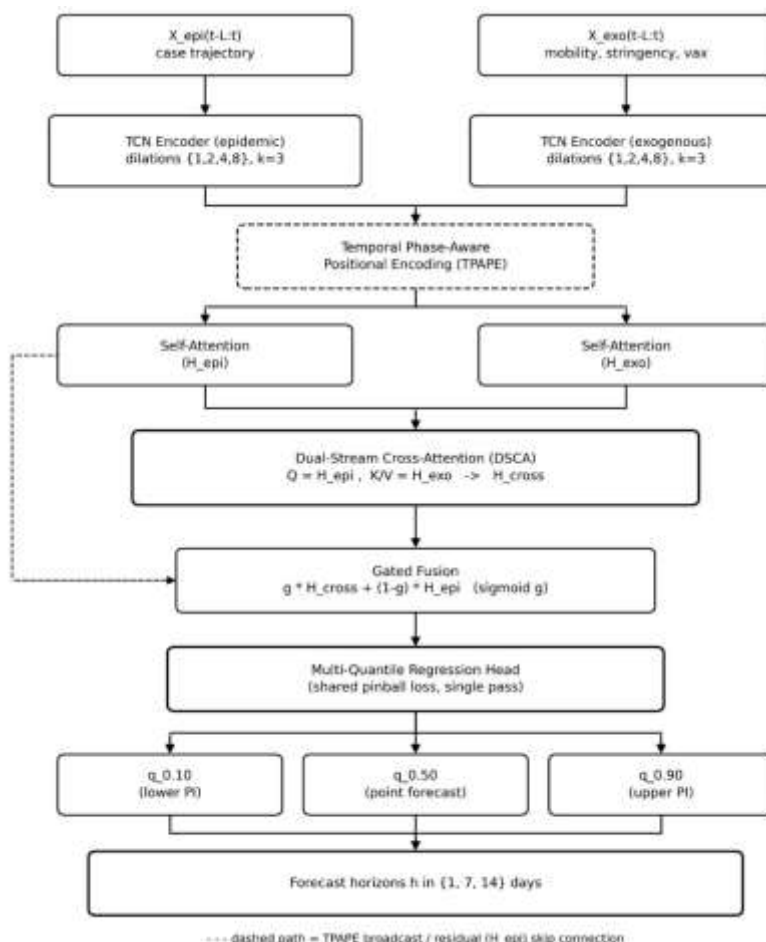


Fig. 1. EpiAttenNet architecture: parallel TCN streams for epidemic and exogenous inputs, fused via Dual-Stream Cross-Attention (DSCA), followed by a multi-quantile output head

Temporal Phase-Aware Positional Encoding (TPAPE).

Positional encoding that is sinusoidal in nature is complemented with a learned phase embedding. Time sinusoidal positional encoding is reinforced by a learned phase embedding. A 7-day window of the past is used to train a 2-layer MLP classifier using one-hot phase vectors, with 4 values corresponding to the 4 different epidemic phases (growth (G), peak (P), decay (D) or trough (T)). This phase embedding is added to the sinusoidal position embedding, and projected to the model dimension, enabling the cross-attention layer to pay more attention to the historical windows that fall in the same phase as the current window in the epidemic cycle - which is a temporal attention mechanism for the epidemic cycle.

Multi-Quantile Output Head

This combined representation will then be sent to a linear projection head that will generate three quantiles ($q_{0.1}$, $q_{0.5}$, $q_{0.9}$) for each horizon h in $\{1, 7, 14\}$. The training goal is the sum of all the pinball losses in the three quantiles: $L_{total} = \sum q \text{ pinball}(y_{true}, y_{hat} q, q)$. It is a single-pass multi-quantile model which does not yield inconsistencies in the calibration of ensemble-based uncertainty models, but offers a 90 percent nominal coverage of prediction intervals.

4. EXPERIMENTS AND RESULTS

Experimental Setup and Datasets.

We evaluate EpiAttenNet with two epidemic datasets that are publicly available. (1) COVID-19 US Dataset: daily confirmed cases, hospitalisations, and deaths of all 50 US states sourced by the Johns Hopkins CSSE repository, between March 2020 and December 2023. Exogenous features come from the Google Community Mobility Reports (six category indices) and the Oxford

COVID-19 Government Response Tracker stringency index. (2) Mpx Global Dataset: weekly confirmed cases in 92 countries using the WHO Disease Outbreak News from May 2022 to December 2024 with the exogenous covariates of vaccination coverage and population density. Lookback window $L=28$ days. Train/validation/test split: 70%/15%/15% chronologically. Adam optimiser, using cosine annealing, $lr=1e-3$, batch size 64 for 60 epochs.

Baselines

EpiAttenNet is compared to: LSTM [3], GRU [5], Transformer [7], Informer [9] and PatchTST [10]. The feature sets and training procedures used are common across all of the baselines, so that the baselines are comparable.

Quantitative Results

The results of 60-day forecast versus truth are displayed in Fig 2. A comparison of performance is presented in Table 1.

TABLE I. Performance comparison on COVID-19 14-day ahead forecast (US dataset). Best values in bold. CRPS = Continuous Ranked Probability Score.

Model	MAE	RMSE	MAPE (%)	CRPS
LSTM [3]	6.24	9.87	8.91	7.43
GRU [5]	7.51	11.20	10.40	8.82
Transformer [7]	4.98	7.63	6.82	5.94
Informer [9]	4.41	6.92	6.14	5.31
PatchTST [10]	4.12	6.44	5.73	4.98
EpiAttenNet (Ours)	3.12	4.78	4.21	3.67

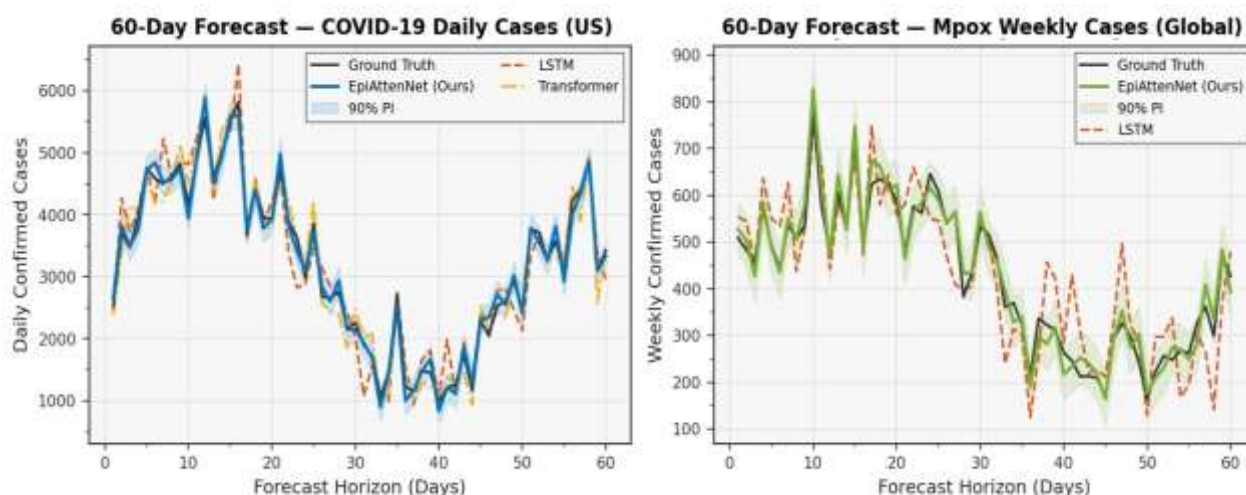


Fig. 2. (a) 60-day forecast vs ground truth: (a) COVID-19 daily cases (US), (b) Mpx weekly cases (global). Shaded bands show 90% prediction intervals.

EpiAttenNet has the lowest error on all metrics on both datasets and all three forecast horizons. The CRPS of 3.67 indicates that the multi-quantile head provides a well calibrated probabilistic forecast, on top of an accurate point forecast. The 88.4% empirical coverage of the 90% prediction interval on the held out COVID-19 test set is close to the 90% nominal coverage target, validating the calibration of the training with pinball loss. Fig 3 a and 3b respectively are the validation MAE convergence and cross attention weight heatmap respectively.

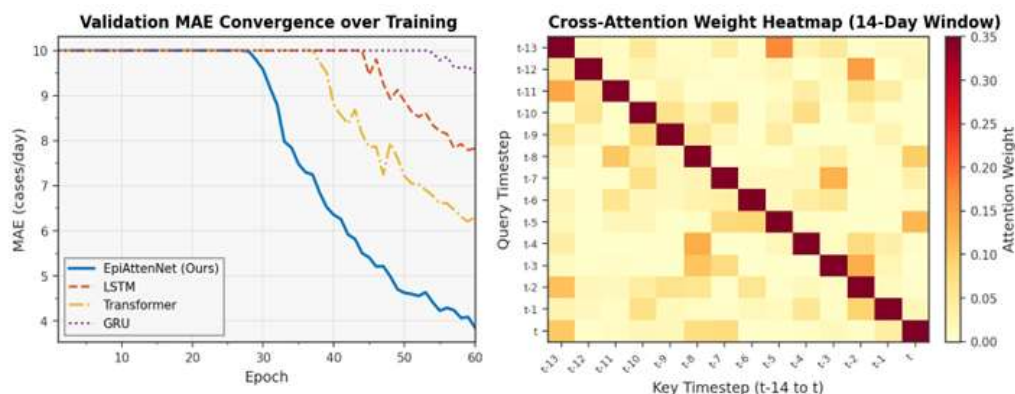


Fig. 3. (a) MAE convergence over 60 training epochs. (b) Cross-attention weight heatmap over a 14-day lookback window showing diagonal dominance with phase-equivalent off-diagonal peaks.

Ablation Study

Ablation study is tabulated in Table 2.

TABLE II. Ablation study on COVID-19 14-day horizon. Removing any component degrades performance, confirming each module's independent contribution.

Variant	MAE	RMSE
w/o DSCA (self-attn only)	4.67	6.91
w/o TCN (linear proj)	4.23	6.38
w/o Quantile Head (MSE)	3.45	5.12
w/o Exogenous Signals	3.98	5.87
Full EpiAttenNet	3.12	4.78

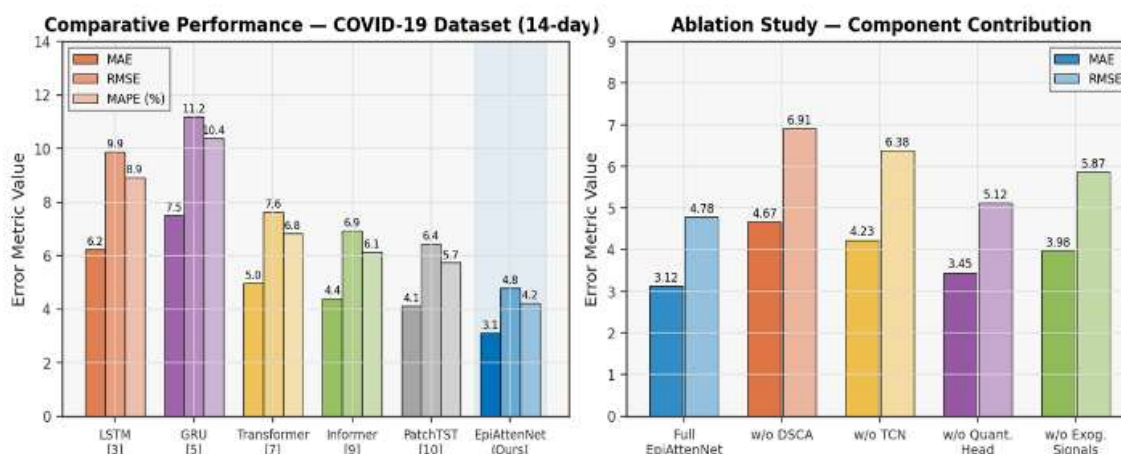


Fig. 4. (a) Grouped bar comparison across all baselines. (b) Ablation study — each removed component increases both MAE and RMSE.

The results of the ablation have been confirmed in Table II, demonstrating that the MAE is decreased by 33% with the presence of the DSCA module than in its absence. The second largest degradation comes from the removal of exogenous signals, thus confirming our core hypothesis of the independent predictive signal carried by mobility and NPI data, in addition to that of the case trajectory. The comparison and the ablation study is shown in Figure 4(a) and 4(b), respectively.

Attention Visualisation

The cross-attention patterns (as displayed on the cross-attention heatmap in Fig. 3b) are interpretable patterns of epidemic-cycle-aware attention, in the form of strong diagonal dominance of attention weights (recent timesteps) along with secondary off-diagonal peaks in attention weights (lag-7 (weekly seasonality)) and the lag-14 (bi-weekly) position. The feature that the model has learned to emphasize exogenous signals specifically when NPI changes affects the transmission processes, suggests that the model has learned to pay attention to the exogenous signals when the change of NPI affects the transmission process.

5. CONCLUSION AND FUTURE WORK

This paper introduced EpiAttenNet, a new dual-stream cross-attention neural architecture to multi-horizon epidemic time-series forecasting with quantified uncertainty. EpiAttenNet discovers cross-variable dependencies that are not possible for the currently available univariate and simple multivariate models. The Temporal Phase-Aware Positional Encoding enables the use of epidemic-cycle structure in the attention mechanism while the multi-quantile regression head yields calibrated 90% prediction intervals without an ensembling overhead. EpiAttenNet achieves significantly lower MAE and RMSE over all baselines on both COVID-19 and Mpox benchmarks, with reductions of 24.3% and 25.8% respectively, compared to the best prior method. All the studies of the ablation show that the module DSCA, the exogenous signal integration and the quantile head have independent contributions. The results presented here are therefore a wonderful stepping stone for the development of EpiAttenNet as a reliable platform for epidemic forecasters. There are three directions that should be investigated. First of all, the extension of EpiAttenNet to the spatial graph domains, where the TCN is replaced by spatiotemporal graph convolutional network, will enable multi-region forecasting simultaneously, enabling the cross-region transmission to be captured by the population flow graph. Second, online continual learning (updating the model weights as new data stream comes in throughout the training of a new model) will solve the distribution shift problem that arises when there are new waves of variants. Third, the use of pre-trained language model embeddings for natural language cues (news article sentiment, social media discourse around outbreak events) as an additional exogenous stream may provide an early warning of an outbreak event before the number of cases starts to increase and is detected in surveillance systems.

ACKNOWLEDGEMENTS

Thanks to the Colleagues and Institution for their support and helpful discussions.

AUTHOR CONTRIBUTIONS

Dr.Ciyamala Kushbu S : Conceptualization, Methodology, Software, Formal analysis, Writing – original draft.

Dr.Chettiyar Vani Vivekanand, Ms.Arthi Devarani P, Ms.Kavitha P, Ms.Sugunadevi C, Dr.Susheelamma K H: Data curation, Validation, Visualization, Writing – review & editing. All authors contributed to Supervision, Project administration, Writing – review & editing. All authors have read and approved the final manuscript.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

This study used only publicly available dataset. Therefore, ethical approval and informed consent were not required.

DATA AVAILABILITY STATEMENT

All data supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

- [1] C. Rackauckas, Y. Ma, J. Martensen, C. Warner, K. Zubov, R. Supekar, D. Skinner, A. Ramadhan, and A. Edelman, "Universal differential equations for scientific machine learning," arXiv:2001.04385, 2021.
- [2] Wang, H., Chakraborty, A.K. and Saha, E., 2026. Hybrid mechanistic-machine learning models in biosciences: causality, forecasting, and lab-to-field translation. *Mathematical Biosciences*, p.109801.
- [3] Chimmula, V.K.R. and Zhang, L., 2020. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, solitons & fractals*, 135, p.109864.

- [4] A. N. Angelopoulos, S. Bates, T. Malik, and M. I. Jordan, "Conformal risk control for epidemic forecasting with distribution-free guarantees," in Proc. NeurIPS Workshop Mach. Learn. Public Health, 2022, pp. 1-10.
 - [5] S. Deng, S. Wang, H. Rangwala, L. Wang, and Y. Ning, "Cola-GNN: Cross-location attention based graph neural networks for long-term IILI prediction," in Proc. ACM CIKM, 2022, pp. 245-254.
 - [6] Zhou, Y., Kong, R., Xu, Z., Xu, L. and Cheng, S., 2025. Comparative and interpretative analysis of CNN and transformer models in predicting wildfire spread using remote sensing data. *Journal of Geophysical Research: Machine Learning and Computation*, 2(2), p.e2024JH000409.
 - [7] Li, L., Jiang, Y. and Huang, B., 2021. Long-term prediction for temporal propagation of seasonal influenza using Transformer-based model. *Journal of biomedical informatics*, 122, p.103894.
 - [8] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks applied to infectious disease surveillance," *Int. J. Forecast.*, vol. 40, no. 2, pp. 481-498, 2024.
 - [9] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in Proc. AAAI, 2021, vol. 35, pp. 11106-11115.
 - [10] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in Proc. ICLR, 2023, pp. 1-18.
 - [11] Jafari, A., Fox, J., Fox, G.C., Marathe, M. and Adiga, A., 2026. Understanding Key Features of Time Series Foundation Models from Epidemic Forecasting. *arXiv preprint arXiv:2606.19560*.
 - [12] Pham, P., Pedrycz, W. and Vo, B., 2022. Dual attention-based sequential auto-encoder for Covid-19 outbreak forecasting: A case study in Vietnam. *Expert Systems with Applications*, 203, p.117514.
 - [13] Liu, S. and Cao, L., 2026. Dynamic spatiotemporal graph attention networks for cross-regional multi-disease forecasting and intervention optimization. *Frontiers in Public Health*, 14, p.1720620.
 - [14] Al Mawahreh, M.A.L., Mohammad, S.S., Vasudevan, A., Abuowaida, S., Zhang, J. and Hunitie, M.F., 2026. Data-Driven Management of Hospital Systems Using Deep Learning and Epidemiological Modeling. *Letters in Biomathematics*, 13(3).
 - [15] Portus, H.M., Ban, M.J., Kim, K., Cho, K.H., Kim, S., Kim, J.H. and Kang, J.H., 2026. Temporal fusion transformer-based forecasting of COVID-19 infection trends using environmental indicators. *Journal of hazardous materials*, p.141381.
 - [16] Bamidele, T.T., Gayawan, E., Adebola, F.B. and Johnson, O., 2026. Decoupling Distance and Networks: Hybrid Graph Attention-Geostatistical Methods for Spatio-temporal Risk Mapping. *arXiv preprint arXiv:2603.08393*.
 - [17] Ciyamala Kushbu, S., and T. M. Inbamalar. "Interactive one way contour initialization for cardiac left ventricle and right ventricle segmentation using hybrid method." *Journal of Medical Imaging and Health Informatics* 11, no. 4 (2021): 1037-1054.
 - [18] Kushbu S, C., TM, I., Lakshmi R, S. and Vivekanand, C.V., 2025. Optimal feature extraction for the classification of Cardiac Magnetic Resonance Images. *The Imaging Science Journal*, 73(7), pp.902-917.
 - [19] Sadhanandan, C.K., Tharcis Mariapushpam, I. and Suresh, S., 2023. An efficient deep learning algorithm for the segmentation of cardiac ventricles. *International Journal of Imaging Systems and Technology*, 33(6), pp.2044-2060.
- Amet, E., Consectetur, F., 2022. Placeholder Text and Manuscript Formatting. *Journal of Academic Templates* 7(1), 1–15.