Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Trustworthy AI for Dengue Outbreak Forecasting Using a Temporal Fusion Transformer with Explainability and Uncertainty Estimation

Anima Dahal

anima.bit_J2023@padmashreecollege.edu.np

Nilai University https://orcid.org/0009-0000-2718-1524

Bibash Basnet

Nilai University https://orcid.org/0009-0006-9610-9546

Research Article

Keywords: Dengue forecasting, Temporal Fusion Transformer, Explainable AI, Uncertainty estimation, Quantile forecasting, Decision support

Posted Date: September 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10975609/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Accurate and trustworthy dengue case forecasting can support timely public health planning and intervention; however, many forecasting approaches focus primarily on point predictions and provide limited uncertainty information, interpretability, or decision support. This study proposes a trustworthy Artificial Intelligence (AI) framework for dengue case forecasting that integrates a Temporal Fusion Transformer (TFT) with probabilistic forecasting, uncertainty estimation, intrinsic explainability, and rule-based risk stratification. The framework was evaluated using the DengAI: Predicting Disease Spread dataset, containing weekly dengue cases and environmental variables from San Juan, Puerto Rico, and Iquitos, Peru. A chronological forecasting pipeline was employed, including data preprocessing, temporal feature engineering, lag and rolling statistical features, and training-based feature scaling. The TFT used a 24-week encoder and a 4-week prediction horizon and was trained using quantile loss to generate 10th, 50th, and 90th percentile forecasts. Forecasting performance was evaluated using MAE, RMSE, and R^2, while uncertainty was assessed using Prediction Interval Coverage Probability (PICP) and Prediction Interval Normalized Average Width (PINAW). The proposed TFT achieved an MAE of 6.560, RMSE of 10.724, and R^2 of 0.840, outperforming the Naive, Seasonal Naive, and LSTM baselines. The Q10–Q90 prediction interval achieved a PICP of 93.12% with a PINAW of 0.1623. Model interpretability was obtained through TFT variable selection and temporal attention mechanisms, while predicted case counts were further converted into study-specific low, moderate, and high-risk categories with corresponding decision-support recommendations. The proposed framework therefore combines multi-horizon forecasting, uncertainty quantification, intrinsic explainability, and risk-oriented decision support within a unified framework for dengue surveillance and public health planning.

1. Introduction

Infectious disease outbreaks continue to pose major challenges to global public health, causing substantial illness, mortality, and socioeconomic disruption. Among these diseases, dengue fever remains one of the most rapidly spreading mosquito-borne viral infections, particularly in tropical and subtropical regions where climatic factors such as temperature, precipitation, and humidity strongly influence disease transmission [1]. Accurate forecasting of disease outbreaks enables healthcare authorities to allocate medical resources efficiently, implement timely vector-control measures, and support evidence-based public health interventions before outbreaks escalate. Consequently, developing reliable forecasting systems has become an important research topic at the intersection of artificial intelligence and public health.

With machine learning and deep learning methods continuously advancing, the disease breakout prediction was largely enhanced as these methods are able to discover and understand complex temporal patterns in past data. Conventional statistical models like Autoregressive Integrated Moving Average (ARIMA) are generally unable to find nonlinear correlations and also they are limited in their capacities to detect very long-term temporal dependencies beyond a point. Deep learning architectures have demonstrated strong performance for multivariate time-series forecasting tasks with architectures

such as Long Short-Term Memory networks (LSTM), Gated Recurrent Units (GRU), and of course Transformer-based models, which have been very effective for multivariate time-series forecasting. Out of these, the Temporal Fusion Transformer (TFT) is probably the most promising model for forecasting in the way it not only combines the different types of time dependencies (short and long) but also through use of the attention mechanism incorporates both static and dynamic inputs [2] .

There have been quite a few developments yet a large proportion of disease outbreak forecasting systems still mainly focus on the accuracy of prediction while neglecting several other characteristics needed for deployment in real-life situations. Most models provide deterministic point forecasts without any uncertainty quantification which makes it challenging for decision-makers to evaluate forecast reliability. Besides that, numerous deep learning-based forecasting systems operate as black boxes with limited interpretability and make it very hard to figure out the effect of different input variables on their output. Besides that, relatively small number of studies have combined forecasting results with intelligent decision-support systems that would be capable of changing predictions into real public health interventions. These shortcomings make AI-based disease forecasting systems less transparent, less trustworthy, and of less practical help [3].

To address these challenges, this paper proposes a Trustworthy Artificial Intelligence Framework for Disease Outbreak Forecasting using a Temporal Fusion Transformer integrated with probabilistic forecasting, uncertainty estimation, explainability, and intelligent decision support. The framework is developed using the DrivenData DengAI: Predicting Disease Spread dataset and incorporates chronological data preprocessing, temporal feature engineering, feature normalization, and chronological time-based data partitioning. Probabilistic forecasting is performed using quantile loss to generate prediction intervals, whose reliability is evaluated using Prediction Interval Coverage Probability (PICP) and Prediction Interval Normalized Average Width (PINAW). The intrinsic variable selection networks and attention mechanisms of the TFT provide model interpretability, while a rule-based decision support module automatically assigns outbreak risk levels and recommends appropriate public health actions.

The main contributions of this work are summarized as follows:

- A trustworthy AI framework that integrates Temporal Fusion Transformer forecasting, uncertainty estimation, explainability, and intelligent decision support within a unified disease outbreak forecasting pipeline.
- A robust temporal data processing pipeline incorporating chronological data partitioning, feature engineering, lag features, rolling statistics, and feature normalization for multivariate disease forecasting.
- A probabilistic forecasting approach based on quantile prediction, evaluated using PICP and PINAW to provide reliable confidence intervals alongside point forecasts.
- An automated decision support module that converts probabilistic forecasts into outbreak risk levels and actionable public health recommendations to enhance practical disease surveillance and

response.

The remainder of this paper is organized as follows. Section II reviews related work on disease outbreak forecasting and trustworthy AI. Section III presents the proposed framework, while Section IV describes the dataset, preprocessing, feature engineering, and model development. Section V reports the experimental results, uncertainty estimation, explainability analysis, and decision support evaluation. Finally, Section VI concludes the paper and discusses future research directions.

2. Related Work and Research Gap

Machine learning and deep learning are increasingly being used to model the factors of disease transmission and temperature and environment over time in the field of dengue forecasting. Recent research has been looking at recurrent networks; attention mechanisms; Transformer-based models; explainable machine learning; and probabilistic forecasting. From a single system perspective, these approaches have improved prediction accuracy and have provided additional insights into dengue patterns, yet, not many of the issues such as reliability of forecasts, understanding of the decision making process, uncertainty quantification, and actionable decision support have been addressed in a comprehensive manner.

Nguyen et al. [4] developed deep learning models for dengue forecasting using climate data from 20 provinces in Vietnam. The study compared CNN, Transformer, LSTM, and attention-enhanced LSTM (LSTM-ATT) models using dengue incidence and meteorological variables, including temperature, humidity, rainfall, evaporation, and sunshine hours. The LSTM-ATT model achieved the best overall performance among the evaluated approaches and was able to forecast dengue incidence and outbreak months up to three months ahead. However, the study mainly evaluated point forecasts using MAE and RMSE and did not explicitly quantify the uncertainty of the predictions [4].

Majeed et al. [5] proposed an LSTM-based dengue forecasting approach incorporating spatial attention in Malaysia. The study evaluated several LSTM variants using climate, geographic, demographic, and land-use information. The results showed that incorporating attention mechanisms improved the ability of the models to capture relevant spatial and temporal information for dengue prediction. However, the approach focused primarily on prediction performance and did not provide probabilistic prediction intervals or a systematic uncertainty evaluation [5].

Pillay et al. [6] applied a Temporal Fusion Transformer (TFT) to long-term climate and disease data for malaria and dengue forecasting in South Africa and Vietnam. The framework used multiple climatic variables and climate indices for multi-horizon forecasting and employed model-derived variable importance to investigate the factors influencing disease incidence. For dengue, the best short-term model achieved an R^2 of 0.90, demonstrating the potential of TFT for modeling complex climate-disease relationships. This study is particularly relevant to the proposed work because it combines TFT, multi-horizon forecasting, and interpretation for dengue prediction. However, its main emphasis was on

predictive performance and identifying climate drivers rather than integrating explicit prediction-interval evaluation with risk-based decision support [6].

Rahman and Shiddik [7] developed an explainable machine learning framework for predicting dengue outbreaks in Bangladesh using eco-climatic, sociodemographic, and landscape variables. XGBoost and LightGBM were combined with SHAP and LIME to identify influential factors affecting dengue outbreaks. XGBoost achieved an AUC of 0.89, while population density, precipitation, temperature, and land-use variables were identified as important predictors. The study also generated future dengue predictions and early-warning trends. Although the framework improved model transparency and supported early warning, it was based on tree-based machine learning models rather than a multi-horizon Transformer architecture and did not explicitly evaluate prediction-interval reliability. [7]

GulMohamed et al. [8] proposed DengueGNN, a dynamic spatio-temporal graph neural network for regional dengue forecasting. The model incorporated environmental information, historical epidemiological characteristics, and regional connectivity to capture spatial and temporal disease patterns. In addition to point prediction, the study evaluated probabilistic forecasts using CRPS, PICP, and mean prediction interval width (MPIW). The proposed model achieved a PICP of 0.95 for one-week forecasts and 0.93 for four-week forecasts, with narrower prediction intervals than the evaluated baseline models. While the study demonstrates the value of uncertainty-aware dengue forecasting, its main focus is spatial disease transmission and graph-based modeling rather than combining TFT-based temporal forecasting, intrinsic variable interpretation, and an integrated risk-based recommendation mechanism [8].

Although these studies demonstrate substantial progress in dengue forecasting, several gaps remain. Deep learning and attention-based models have improved temporal prediction, while recent TFT research has demonstrated strong performance for climate-driven dengue forecasting and provided information about influential variables. Explainable machine learning has further improved transparency, and probabilistic approaches have demonstrated the importance of quantifying forecast uncertainty. However, these capabilities have largely been addressed separately. In particular, the combined use of multi-horizon TFT forecasting, intrinsic interpretability, quantitative prediction-interval evaluation, and risk-based decision support remains limited. This study addresses this gap by integrating a Temporal Fusion Transformer with quantile-based forecasting, model-based variable and temporal interpretation, prediction-interval evaluation using PICP and PINAW, and a risk-based decision-support layer. The proposed framework therefore provides not only dengue case forecasts but also information about influential factors, forecast reliability, and corresponding public health response levels.

3. Methodolgy

3.1 Dataset

The proposed framework was evaluated using the DengAI: Predicting Disease Spread dataset from DrivenData [9] .It includes weekly reported cases of dengue as well as environmental data from San Juan, Puerto Rico and Iquitos, Peru. The input variables include weather conditions, rainfall, humidity, temperature, vegetation-related indicators, and other environmental observations, while the prediction target is the weekly number of reported dengue cases. The training features and corresponding dengue case labels were merged using city, year, and week of the year as common identifiers. Exploratory analysis was conducted to examine the data structure, missing values, feature characteristics, and chronological consistency before subsequent preprocessing and model development.

3.2 Data Preprocessing

Before model development, the dataset was cleaned and organized separately for San Juan and Iquitos to preserve the temporal structure of dengue observations. The records were sorted chronologically according to the week start date, and duplicate observations were identified and removed. Missing numerical values in the environmental variables were handled within each city to maintain temporal continuity. The data were then divided chronologically into training, validation, and testing sets rather than using random sampling, with 70% of the observations allocated for training, 15% for validation, and 15% for testing. This temporal partitioning ensured that the model was evaluated on observations occurring after the training period and reduced the risk of future information being used during model development. The resulting partitions were checked for missing values, duplicate timestamps, and chronological consistency before feature engineering and model training.

3.3 Feature Engineering

Feature engineering was performed to capture temporal, seasonal, historical, and environmental patterns associated with dengue transmission. Feature engineering was performed to capture temporal, seasonal, historical, and environmental patterns associated with dengue transmission. Previous dengue forecasting studies have demonstrated the usefulness of lagged dengue incidence and meteorological variables, including temperature, humidity, and rainfall, for forecasting dengue incidence [10]. Calendar-based variables, including year, month, epidemiological week, and quarter, were extracted from the observation date. Monthly and weekly seasonality were further represented using sine and cosine transformations to provide continuous representations of periodic patterns. City information was retained as a categorical variable to distinguish between the two study locations.

Historical dengue case information was incorporated through lag features representing the previous 1, 2, 4, and 8 weeks. Rolling statistical features, including rolling mean, standard deviation, maximum, minimum, and median, were generated to summarize recent case patterns. Additional temporal and environmental features were constructed to represent rainfall accumulation, temperature and humidity trends, historical dengue-case patterns, and changes in selected climatic variables. All target-derived temporal features were calculated using only previously observed dengue cases to prevent future target information from entering the forecasting inputs. Records without sufficient historical observations for the required lag and rolling-window calculations were removed.

3.4 Feature Scaling

Continuous numerical features were standardized using the StandardScaler from Scikit-learn to improve numerical stability and optimization during model training [11]. The scaler was fitted exclusively on the training data, and the learned transformation parameters were subsequently applied to the validation and testing data. This ensured that information from the validation and testing periods was not used when estimating the scaling parameters. Categorical variables, including city, were not standardized because they were handled separately by the Temporal Fusion Transformer through categorical embeddings.

3.5 Time-Series Dataset Construction

Redundant variables, including raw year, weekofyear, and numerical city encoding, were removed where their information was already represented by engineered features or categorical embeddings. The processed data were converted into sequential forecasting samples using the TimeSeriesDataSet class from PyTorch Forecasting. City was treated as a static categorical variable, while temporal and environmental variables were provided as time-varying inputs.

After evaluating different configurations, an encoder length of 24 weeks and a prediction horizon of 4 weeks were selected. Thus, the model used the previous 24 weeks of observations to forecast dengue cases for the next four weeks. Training, validation, and testing dataloaders were then created with a batch size of 64 for mini-batch training.

3.6 Temporal Fusion Transformer Architecture

The Temporal Fusion Transformer (TFT) was employed for multi-horizon dengue forecasting. The architecture combines LSTM layers, variable selection networks, gating mechanisms, and multi-head attention to model temporal and environmental relationships [2]. The model used a 24-week encoder and 4-week prediction horizon, with a hidden size of 64, hidden continuous size of 16, two LSTM layers, four attention heads, and 0.20 dropout. The learning rate was set to (1\times10^{-3}). Quantile Loss was used to predict the 10th, 50th, and 90th percentiles. Early stopping with a patience of 10 epochs and gradient clipping with a maximum norm of 0.1 were applied during training.

3.7 Model Training

The TFT was trained using PyTorch Forecasting and PyTorch Lightning with the Adam optimizer and Quantile Loss for probabilistic multi-horizon forecasting [12]. The training set was used for model learning, while the validation set was used to monitor performance and select the best model checkpoint. Early stopping was applied with a patience of 10 epochs, and gradient clipping was used to improve training stability. The model was trained for a maximum of 50 epochs and stopped at epoch 17, with the best checkpoint obtained at epoch 7. Final performance was evaluated on the unseen test set.

3.8 Baseline Models and Comparative Evaluation

Three baseline models were used to evaluate the proposed TFT: Naive, Seasonal Naive, and LSTM. The Naive model used the most recently observed dengue case count as the forecast, while the Seasonal Naive model used the corresponding value from the previous seasonal cycle. The LSTM model used a 24-week historical window to forecast the following four weeks. All models were evaluated on the same unseen testing period using MAE, RMSE, and R^2, enabling a consistent comparison with the proposed TFT.

3.9 Performance Evaluation

Forecasting performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2) as the primary metrics. MAE measures the average absolute prediction error, RMSE gives greater weight to larger errors, and R^2 measures the proportion of target variation explained by the model. Performance was evaluated over the four-week forecast horizon and separately for each prediction week.

3.10 Uncertainty Estimation

Forecast uncertainty was estimated using quantile regression. The Temporal Fusion Transformer predicted the 10th, 50th, and 90th percentiles for each forecast period. The intervals were evaluated using Prediction Interval Coverage Probability (PICP) and Prediction Interval Normalized Average Width (PINAW). PICP measures the proportion of observed dengue cases falling within the predicted interval, while PINAW measures its normalized average width [13]. Together, these metrics assess the coverage and sharpness of the prediction intervals.

3.11 Explainable Artificial Intelligence

The TFT's built-in explainability mechanisms were used to interpret the forecasting process [2]. Variable selection weights were analyzed to identify the relative importance of input features, while temporal attention weights were examined to determine the importance of historical time steps within the 24-week encoder window. These outputs provided insights into the environmental and historical factors influencing dengue forecasts.

3.12 Risk Stratification and Decision Support

A rule-based risk stratification layer was developed to translate predicted dengue cases into low, moderate, and high-risk categories. Thresholds were derived from the training-data distribution using the 33rd percentile (7 cases) and 66th percentile (24 cases). Forecasts were then mapped to the corresponding risk levels and linked to predefined recommendations for surveillance, vector control, preparedness, and medical-resource allocation. Prediction intervals were also considered to provide uncertainty information alongside the risk classifications.

4. Result and Discussion

4.1 Experimental Setup

The proposed framework was implemented in Python using Google Colaboratory. Model development employed the PyTorch, PyTorch Forecasting, PyTorch Lightning, Scikit-learn, NumPy, and Pandas libraries. The Temporal Fusion Transformer (TFT) was trained using the Adam optimizer with Quantile Loss to generate probabilistic forecasts. The final model configuration consisted of a 24-week encoder length, 4-week prediction horizon, hidden size of 64, hidden continuous size of 16, two LSTM layers, four attention heads, dropout rate of 0.20, and a learning rate of 1×10^{-3}. Performance was evaluated on the unseen testing dataset using MAE, RMSE, R^2, PICP, and PINAW.

4.2 Forecasting Performance

The forecasting performance of the proposed Temporal Fusion Transformer (TFT) was compared with Naive, Seasonal Naive, and Long Short-Term Memory (LSTM) models on the same unseen testing dataset. MAE, RMSE, and R^2 were used as the primary evaluation metrics.

Table 1
Comparative forecasting performance of the evaluated models

Model	MAE	RMSE	R^2
Naive	7.601	12.948	0.391
Seasonal Naive	18.049	29.350	−2.091
LSTM	7.430	12.536	0.458
Proposed TFT	**6.560**	**10.724**	**0.840**

The proposed TFT achieved the best performance across all three primary metrics, with an MAE of 6.560, RMSE of 10.724, and R^2 of 0.840. Compared with the Naive model, the TFT reduced MAE by 13.70% and RMSE by 17.18%. It also improved upon the LSTM model by 11.71% in MAE and 14.47% in RMSE. The results indicate that the TFT was better able to capture the temporal and environmental patterns associated with dengue case variation than the evaluated baseline models.

Figure 1 presents the actual versus predicted dengue cases for the proposed TFT over the testing period, illustrating the model's ability to track both the overall trend and major outbreak fluctuations.

The Seasonal Naive model performed considerably worse, with an MAE of 18.049, RMSE of 29.350, and R^2 of − 2.091, indicating that fixed seasonal repetition was insufficient to represent the observed dengue dynamics. Overall, the comparative results demonstrate the effectiveness of the proposed TFT for dengue case forecasting.

4.3 Uncertainty Estimation

Table 2
Prediction-interval performance of the proposed TFT

Metric	PICP	PINAW
Value	93.12%	0.1623

The proposed framework generated probabilistic prediction intervals using quantile regression. The Q10–Q90 prediction interval achieved a Prediction Interval Coverage Probability (PICP) of 93.12%, indicating that the actual weekly dengue case counts were contained within the predicted bounds for the large majority of test forecasts. Specifically, 555 of 596 forecast samples were covered by the prediction interval. The Prediction Interval Normalized Average Width (PINAW) was 0.1623, indicating a relatively limited normalized interval width alongside the observed coverage. Because prediction intervals involve a trade-off between coverage and sharpness, both measures should be considered when assessing their usefulness for forecasting applications. The relatively high empirical coverage observed in this study suggests that the proposed TFT produced conservative prediction intervals, as the observed coverage exceeded the nominal 80% coverage of the Q10–Q90 interval. Unlike the Naive, Seasonal Naive, and LSTM baseline models evaluated in Section 4.2, which were used only to produce point forecasts, the proposed TFT directly generated quantile-based probabilistic forecasts. This capability can be valuable in public health forecasting because it provides not only an expected case count but also a plausible range of future outcomes. Figure 2 illustrates the prediction intervals alongside the actual and median predicted dengue case counts over the testing period.

The shaded region represents the Q10–Q90 quantile interval, the black line shows actual dengue cases, and the dashed blue line shows the median forecast (Q50). The model achieved a PICP of 93.12% (555 of 596 forecast samples covered) with a PINAW of 0.1623.

4.4 Multi-Horizon Forecasting Performance

Table 3
Forecasting performance by prediction horizon

Horizon	MAE	RMSE	R^2
Week + 1	6.168	10.593	0.843
Week + 2	6.144	10.330	0.851
Week + 3	6.706	10.527	0.846
Week + 4	7.221	11.415	0.819
Overall	**6.560**	**10.724**	**0.840**

Table III presents the forecasting performance broken down by individual prediction horizon. The model maintained strong and consistent performance across all four forecast weeks, with R^2 values ranging from 0.819 at Week + 4 to 0.851 at Week + 2. MAE and RMSE increased modestly at Week + 3 and Week + 4, which is expected behavior in multi-horizon forecasting since predictions further into the future naturally carry greater uncertainty. The fact that R^2 remained above 0.819 even at the 4-week horizon confirms that the TFT retained strong predictive capability across the full forecasting window. This consistent multi-horizon performance is particularly important for public health planning. Such multi-week forecasts can provide additional lead time for surveillance planning, resource preparation, and vector-control activities.

4.5 Explainability Analysis

The proposed framework incorporated explainability through the TFT's built-in variable selection networks and temporal attention mechanism, providing interpretability at both the feature and time-step levels without requiring an external explainability tool.

The variable selection network identified the relative importance of input features for dengue forecasting. As shown in Fig. 3, *case_increase* was the most influential feature (11.23%), followed by humidity and temperature-related variables, including *reanalysis_relative_humidity_percent_change* (7.23%) and *reanalysis_avg_temp_k* (6.97%). Recent case-history features, including *lag_1* (5.42%) and *lag_2* (4.19%), also showed substantial importance, indicating that both recent dengue trends and environmental conditions contributed to the forecasts.

The temporal attention mechanism in Fig. 4 showed that greater attention was generally assigned to recent observations within the 24-week encoder window. This indicates that recent case history contributed strongly to short-term dengue forecasting.

4.6 Risk Stratification and Decision Support

The proposed framework included a rule-based risk stratification layer to translate TFT forecasts into actionable public health guidance. Risk thresholds were derived from the training-data distribution using the 33rd percentile (7.0 cases) and 66th percentile (24.0 cases), defining low, moderate, and high risk categories.

Across the 596 test forecast samples, the model classified 231 weeks (38.8%) as low risk, 252 weeks (42.3%) as moderate risk, and 113 weeks (19.0%) as high risk. The average predicted weekly case count was 18.03, corresponding to the moderate-risk category. Accordingly, the framework recommended routine surveillance and vector-control maintenance during low-risk periods, enhanced surveillance and health-system preparedness during moderate-risk periods, and intensified vector control and medical-resource mobilization during high-risk periods.

The prediction intervals further provided uncertainty information alongside the risk classifications, allowing the decision-support layer to consider forecast reliability when supporting public health

resource allocation.

4.7 Discussion

The proposed TFT framework outperformed all evaluated baseline models across all three primary metrics. Compared with the Naive baseline, MAE was reduced by 13.70% and RMSE by 17.18%. The improvement over LSTM was also meaningful at 11.71% in MAE and 14.47% in RMSE, demonstrating that TFT's combination of variable selection, gating mechanisms, and multi-head attention provided genuine advantages over a conventional recurrent architecture. A PICP of 93.12% with a PINAW of 0.1623 confirmed that prediction intervals reliably captured actual dengue case counts while remaining practically narrow. None of the evaluated baselines provided any uncertainty quantification, highlighting a fundamental gap that the proposed framework directly addresses for public health decision-making.

The explainability analysis identified case_increase as the most influential predictor with an importance score of 11.23%, followed by humidity and temperature reanalysis variables. These findings are consistent with known dengue transmission dynamics where week-on-week case trends and climatic conditions jointly drive outbreak trajectories. Temporal attention weights concentrated on the most recent encoder weeks, confirming the model correctly prioritized recent transmission patterns over older historical observations. The main limitation of this study is the use of only two cities from the DengAI dataset. Broader validation across additional locations and disease contexts would strengthen the generalizability of the findings.

5. Conclusion

This study presented a Trustworthy Artificial Intelligence framework for multi-horizon dengue outbreak forecasting built around the Temporal Fusion Transformer. The framework brings together data preprocessing, temporal feature engineering, probabilistic forecasting, uncertainty estimation, explainability, and decision support into a single unified pipeline addressing a gap that most existing forecasting approaches leave open by focusing only on prediction accuracy. Evaluated on the DrivenData DengAI dataset covering weekly dengue case records from San Juan, Puerto Rico, and Iquitos, Peru, the proposed TFT outperformed Naive, Seasonal Naive, and LSTM baselines across all primary metrics, achieving an MAE of 6.560, RMSE of 10.724, and R^2 of 0.840. Uncertainty estimates produced a PICP of 93.12% and PINAW of 0.1623, confirming that prediction intervals were both reliable and practically narrow. The explainability analysis identified case_increase and humidity-related variables as the strongest predictors, consistent with known dengue transmission dynamics. The decision support layer translated forecasts into risk levels and actionable public health recommendations, making the framework directly usable by providing a foundation for decision-support applications in dengue surveillance. The results confirm that combining strong forecasting performance with transparency and uncertainty quantification produces a more complete and trustworthy tool for disease surveillance than prediction accuracy alone can offer.

6. Future Work

Although the proposed framework performed well on the DengAI dataset, several directions remain open. Future work could extend the framework to other infectious diseases including malaria, influenza, and COVID-19 to evaluate generalizability beyond dengue. Incorporating real-time surveillance feeds, mobility data, and socio-economic indicators such as population density and sanitation access could improve responsiveness to rapidly changing outbreak conditions. City-specific adaptive risk thresholds derived from local outbreak history would make the decision support layer more accurate than the current percentile-based approach. Better uncertainty calibration techniques such as conformal prediction could bring observed PICP closer to the nominal coverage level. Validation in real public health settings alongside epidemiologists and health authorities remains the most critical next step toward practical deployment during actual outbreaks.

Declarations

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable. This study used publicly available secondary data and did not involve human participants or animals.

Consent for publication

Not applicable.

Funding

No funding was received for this research.

Authors' contributions

Anima Dahal and Bibash Basnet contributed equally to the conceptualization, methodology, data preprocessing, feature engineering, model development, experimental evaluation, interpretation of results, and preparation of the manuscript. Both authors contributed to reviewing and revising the manuscript and approved the final version.

Data availability

The dataset used in this study is publicly available through the DrivenData DengAI: Predicting Disease Spread competition.

Code availability

The code used in this study is available from the corresponding author upon reasonable request.

References

1. Messina JP, Brady OJ, Golding N, Kraemer MUG, Wint GRW, Ray SE, Pigott DM, Shearer FM, Johnson K, Earl L, Marczak LB, Shirude S, Weaver ND, Gilbert M (2019) The current and future global distribution and population at risk of dengue. Nat Microbiol 4:1508–1515
2. Lim B, Arık SÖ, Loeff N, Pfister T (2021) Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. Int J Forecast 37(4):1748–1764
3. Chaddad A, Peng J, Xu J, Bouridane A (2023) Survey of Explainable AI Techniques in Healthcare, *Sensors*, vol. 23, no. 2, p. 634
4. Nguyen V-H, Tuyet-Hanh TT, Mulhall J, Minh HV, Duong TQ, Chien NV (2022) Deep learning models for forecasting dengue fever based on climate data in Vietnam. PLoS Negl Trop Dis 16(6):e0010509
5. Majeed MA, Shafri HZM, Zulkafli Z, Wayayok A (2023) A Deep Learning Approach for Dengue Fever Prediction in Malaysia Using LSTM with Spatial Attention. Int J Environ Res Public Health 20(5):4130
6. Pillay MT, Le MTQ, Takamatsu Y, Phong TV, Kgalane N, Minakawa N (2026) Application of a Temporal Fusion Transformer and Long-Term Climate and Disease Data to Assess the Predictive Power and Understand the Drivers for Malaria and Dengue. Int J Environ Res Public Health 23(1):75
7. Rahman MS, Shiddik MAB (2025) Explainable artificial intelligence for predicting dengue outbreaks in Bangladesh using eco-climatic triggers. Global Epidemiol 10:100210
8. GulMohamed RB, Hetany W, Almaimani HA, Khiery FAS (2026) DengueGNN: Graph-based deep learning for modeling disease spread dynamics and prediction. Sci Rep 16:10584
9. DrivenData DAI Predicting Disease Spread, [Online]. Available: https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/. [Accessed August 2026]
10. Nguyen V-H, Tuyet-Hanh TT, Mulhall J, Minh HV, Duong TQ, Chien NV, Nhung NTT, Lan VH, Minh HB, Cuong D, Bich NN, Quyen NH, Linh NT (2022) T. N. Q. Tho and Nghia, Deep learning models for forecasting dengue fever based on climate data in Vietnam. PLoS Negl Trop Dis, 16, 6
11. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M (2011) and P, Scikit-learn: Machine Learning in Python. J Mach Learn Res, 12, 85
12. Wen R, Torkkola K, Narayanaswamy B, Madeka D (2017) A multi-horizon quantile recurrent forecaster, in *NeurIPS*

13. Zhang J, Liao Z, Shu J, Yue J, Liu Z, Tao R (2024) Interval prediction of short-term photovoltaic power based on an improved GRU model. Energy Sci Eng 12(7):3142–3156

Figures

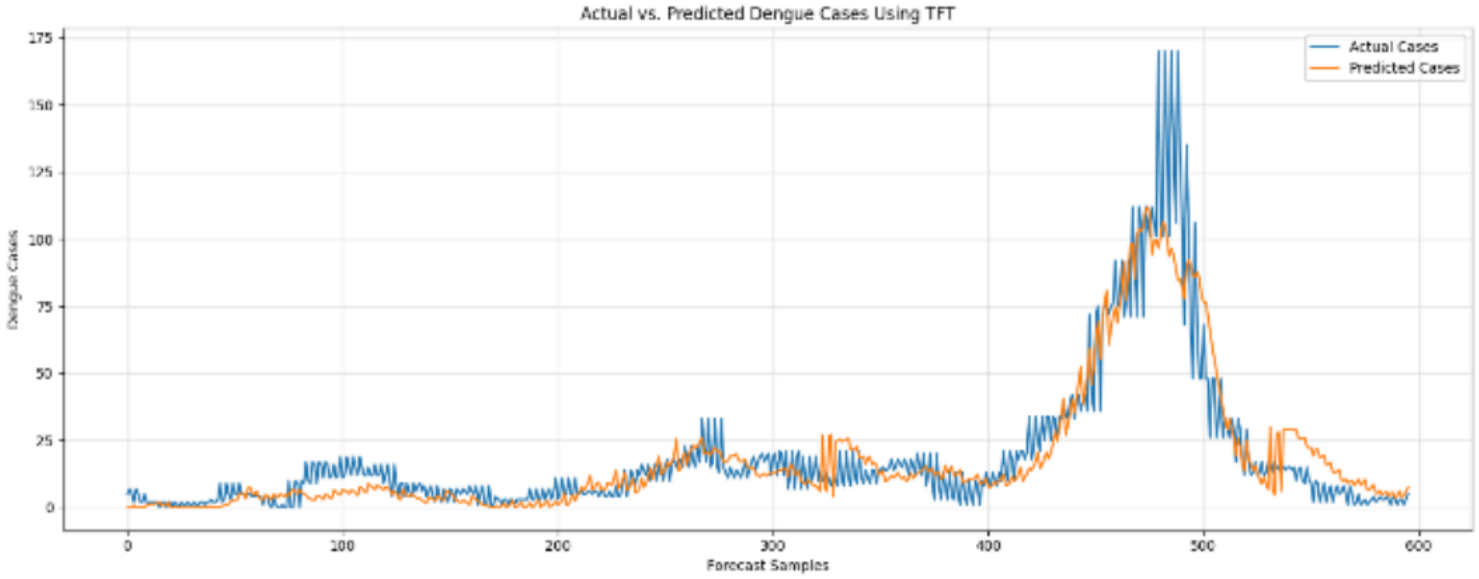


Figure 1

Actual vs. Predicted Dengue Cases

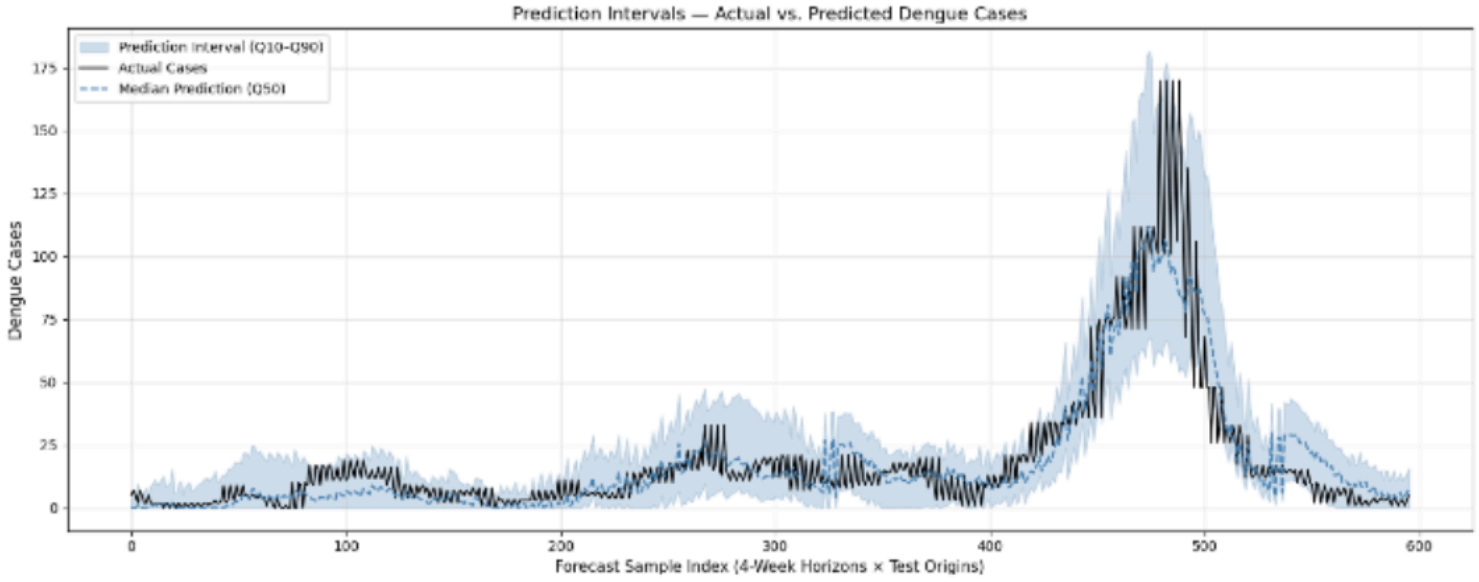


Figure 2

Prediction intervals generated by the proposed TFT framework over the testing period.

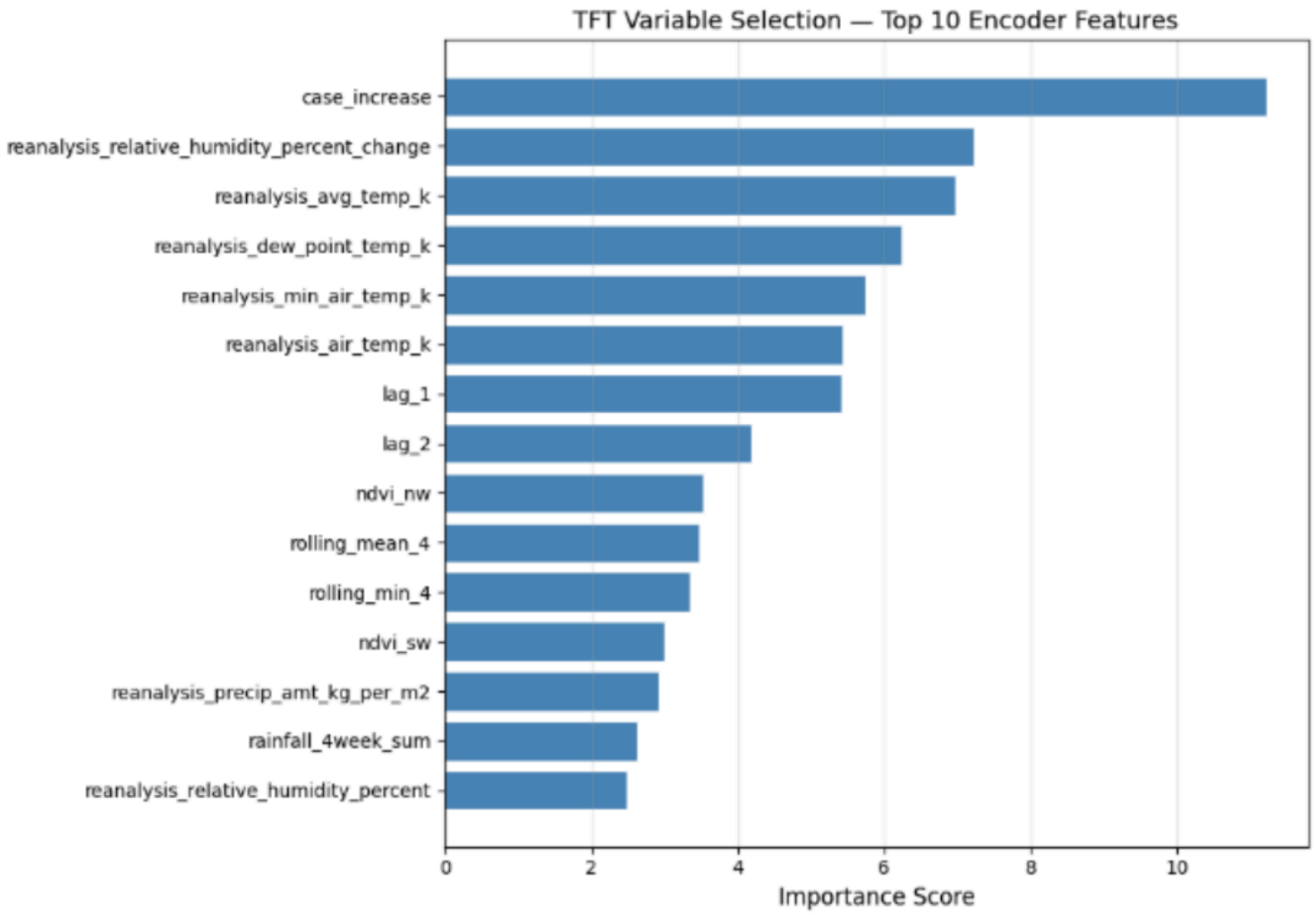


Figure 3

TFT Variable Selection: Top 10 Encoder Feature Importance

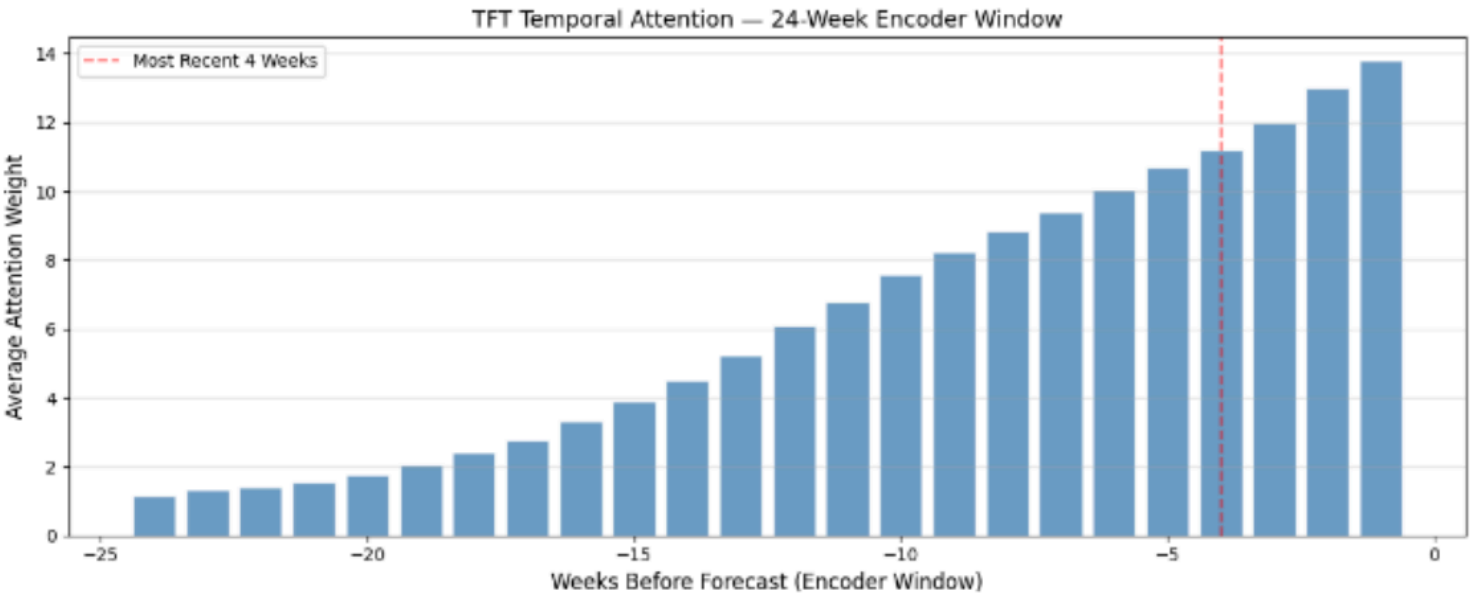


Figure 4

TFT temporal attention weights across the 24-week encoder window