International Journal of Biology and Life Sciences
ISSN: 2957-9511 | Vol. 14, No. 1, 2026

Research on Tuberculosis Incidence Prediction Based on the ST-GNN Model

Ziheng Gui, Zhuofan Su, Yunjing Liu, Xiaoyu Cheng

North China University of Science and Technology, Tangshan, Hebei, 063200, China

Abstract: Tuberculosis (TB) Is a Chronic Infectious Disease Caused by Mycobacterium Tuberculosis. Its Epidemic Process Is Influenced by Long-Term Socioeconomic Differences, Population Mobility, and Metabolic and Behavioral Risk Factors, Exhibiting Significant Spatiotemporal Heterogeneity in My Country. To More Accurately Characterize the Spatiotemporal Evolution of TB Incidence and Improve Predictive Performance, This Paper Constructs a Spatio-Temporal Graph Neural Network (ST-GNN) Prediction Model Based on Monitoring Data from 31 Provincial-Level Administrative Regions in My Country. This Model Integrates Spatial Correlation, Temporal Dynamics, and Attention Mechanisms, and Incorporates a Comparative Risk Assessment (CRA) Framework to Conduct Multi-Dimensional Risk Attribution Analysis. This Study Treats Provincial-Level Administrative Regions as Graph Nodes, Constructs a Spatial Association Network by Integrating Geographical Adjacency Relationships and Population Flow Information, Uses Graph Convolutional Networks (GCNs) to Learn Inter-Provincial Spatial Dependence Features, Employs Temporal Convolutional Networks (TCNs) to Characterize the Long-Term Dynamic Changes in Tuberculosis Incidence, and Introduces a Spatiotemporal Attention Mechanism to Enhance the Model's Ability to Identify Key Regions and Key Time Periods. Simultaneously, Bayesian Optimization Is Used to Automatically Adjust the Model's Hyperparameters to Improve Its Stability and Generalization Performance. Experimental Results Show That, Compared with Traditional Time Series Models and Various Deep Learning Models, the Constructed ST-GNN Model Exhibits Higher Prediction Accuracy and More Robust Spatiotemporal Feature Capture Capabilities Across Multiple Evaluation Metrics, and Can More Effectively Identify Regional Clustering Trends and Abnormal Fluctuations. Based on the Predictive Analysis, This Paper Uses the CRA Framework Proposed by the Global Burden of Disease Study to Conduct Attribution Analysis on Modifiable Risk Factors Such as High Body Mass Index (BMI), Smoking, and Insufficient Physical Activity. The Results Show That the Attribution Contributions of Different Risk Factors to Tuberculosis Incidence Vary Significantly Across Regions, with Some Behavioral and Metabolic-Related Factors Having a More Prominent Impact in High-Burden Areas. The Relevant Research Results Provide Quantitative Evidence for the Precise Prevention and Control of Tuberculosis, the Formulation of Regionally Differentiated Intervention Strategies, and the Rational Allocation of Public Health Resources.

Keywords: Tuberculosis; Spatiotemporal Graph Neural Network; Incidence Prediction; Risk Attribution; Comparative Risk Assessment.

1. Introduction

1.1. Research Background and Significance

Tuberculosis (TB) is a chronic infectious disease caused by Mycobacterium tuberculosis. It is characterized by a long transmission cycle, an insidious incubation period, and significant challenges in prevention and control, and has long been a major public health concern globally and in my country. Driven by the "Healthy China 2030" strategy and the global framework for ending TB, the overall incidence rate of TB in my country has shown a downward trend. However, due to multiple factors such as uneven regional economic development, increased inter-regional population mobility, population aging, and rising burden of chronic underlying diseases, the prevalence of TB in different regions still shows significant differences, and the prevention and control situation remains severe in some areas.

The epidemic process of TB exhibits significant spatiotemporal coupling characteristics. On the one hand, its incidence rate shows seasonal fluctuations and long-term trends over time; on the other hand, differences in population mobility, economic ties, and medical resource allocation between regions result in significant spatial correlations and spillover effects in disease transmission. Existing studies have mostly used time series models or spatial statistical models to characterize the above characteristics, but these methods usually struggle to simultaneously model temporal dynamics and spatial dependencies within a unified framework. This limits the ability to characterize complex transmission mechanisms and affects prediction accuracy and policy explanatory power.

In recent years, with the rapid development of big data and artificial intelligence technologies in the field of public health, the application of deep learning methods in disease monitoring and prediction has gradually increased. In particular, the Spatio-Temporal Graph Neural Network (ST-GNN) models spatial dependence in a graph convolutional network by treating regions as graph nodes and regional relationships as graph edges, and combines temporal convolution or recurrent structures to characterize temporal dynamics, providing a new technical paradigm for the analysis of high-dimensional complex spatio-temporal data. Related methods have shown good performance in traffic flow prediction, weather forecasting, and some infectious disease research, but their systematic application in the provincial-level prediction and risk assessment of tuberculosis in my country is still relatively limited.

Based on this, this paper takes 31 provincial-level administrative regions in my country as the research object, introduces the ST-GNN method to model and predict the incidence of tuberculosis, and focuses on characterizing the

comprehensive impact of inter-provincial spatial relationships and temporal evolution on disease transmission. Building upon this foundation, and further integrating the Comparative Risk Assessment (CRA) framework, quantitative attribution analysis was conducted on various modifiable socioeconomic and health-related risk factors to identify the main drivers of tuberculosis burden in different regions. The research results contribute to improving the accuracy of tuberculosis incidence prediction and provide a scientific basis for the formulation of regionally differentiated prevention and control strategies, the optimal allocation of intervention resources, and scenario simulation assessment.

1.2. Research Objectives and Content

This study aims to construct a predictive model that comprehensively characterizes the spatiotemporal characteristics of tuberculosis incidence based on multi-source public health and socioeconomic data, and to conduct attribution analysis on major controllable risk factors based on this model. Specific research objectives include:

Analyzing the distribution characteristics of tuberculosis incidence in my country from both temporal and spatial dimensions, exploring its changing patterns in different regions and time periods;

Constructing a tuberculosis incidence prediction model based on a spatiotemporal graph neural network, improving prediction accuracy and model stability by integrating spatial correlation structure and time series information;

Quantitatively analyzing the attribution contributions of various behavioral and metabolic risk factors based on a comparative risk assessment framework, revealing their differences in different regions and populations;

Based on the prediction results and risk attribution analysis, conducting simple scenario simulation analysis to assess the potential impact of different public health interventions on tuberculosis incidence levels.

Through the above research, we hope to provide data support and methodological reference for the regional differentiated prevention and control of tuberculosis, and at the same time verify the feasibility of applying the spatiotemporal graph neural network method in the spatiotemporal data analysis of public health.

2. Current Status of Domestic and International Research

2.1. Research Progress of Traditional Models

To systematically grasp the research trajectory of tuberculosis incidence prediction and burden driver identification, relevant domestic and international work can be broadly summarized into two main lines: one is incidence prediction model research based on statistics, machine learning, and deep learning; the other is risk attribution and intervention priority assessment research based on the Comparative Risk Assessment (CRA) framework proposed by the Global Burden of Disease Study (GBD). Existing research provides an important foundation for tuberculosis surveillance and early warning, but there is still room for improvement in "long-term spatiotemporal coupling modeling" and "integrated analysis of prediction-attribution-scenario simulation."

Early tuberculosis incidence prediction studies mainly rely on traditional statistical models and time series analysis methods, including linear/generalized linear regression, piecewise regression (trend identification), Poisson/negative binomial regression (count data modeling), and autoregressive integral moving average (ARIMA/SARIMA) models. These methods can provide relatively robust trend descriptions under conditions of limited data and few explanatory variables, and can describe certain periodic fluctuations and short-term dependencies through seasonal and lagged terms. However, tuberculosis transmission is driven by multiple factors such as socioeconomic factors, population mobility, accessibility to medical services, and comorbidity burden, and there are potential transmission links and spillover effects between different regions. Models based solely on the time dimension often cannot fully reflect the complex changes caused by inter-regional linkages.

To compensate for the lack of spatial information, researchers have further introduced spatial and spatiotemporal statistical models, such as spatial autoregressive models (SAR), conditional autoregressive (CAR)/Bayesian hierarchical models, spatiotemporal autoregressive (STARIMA), and geographically weighted regression (GWR), to characterize the correlation and spatial heterogeneity between neighboring regions. These methods are advantageous in explaining inter-provincial/municipal differences, but they typically rely on pre-defined adjacency relationships or linear assumptions, limiting their ability to characterize nonlinear and higher-order interactions. Furthermore, when faced with multi-source, high-dimensional covariates, non-stationary changes, and dynamic evolution over long periods, the model structure and parameter updates are often complex, affecting their scalability and generalization ability in practical early warning scenarios.

In general, traditional prediction models have played an important role in the analysis of tuberculosis incidence trends and the explanation of mechanisms, but they still have the following limitations:

(1) Insufficient spatial dependence modeling: Most models based solely on time series assume that samples from different regions are independent of each other, making it difficult to reflect the regional transmission associations and spillover effects caused by population mobility, economic ties, etc.;

(2) Limited ability to characterize nonlinearity and interaction: Tuberculosis transmission is affected by multiple factors such as environment, behavior, socioeconomics, and health services. Traditional linear or shallow models are unable to fully learn complex nonlinear relationships and multi-factor interactions;

(3) Weak dynamic adaptation and multi-source fusion capabilities: When prevention and control measures, social environment, or monitoring systems change, the model needs to be frequently re-evaluated or manually adjusted, and the fusion and utilization of multi-source high-dimensional data is insufficient, affecting the stability and transferability of predictions.

2.2. Introduction and Development of Deep Learning Models

In recent years, deep learning methods have received widespread attention in the field of infectious disease prediction. Recurrent Neural Networks (RNNs) and their variants (LSTM/GRU) can model long-term dependencies in time series, while Temporal Convolutional Networks (TCNs) and attention-based sequence models (such as Transformers) also demonstrate strong capabilities in multi-step prediction

and long-sequence modeling. Although these models can improve the fitting and prediction performance in the time dimension, in inter-provincial scale studies, each region's sequence is often modeled independently, and spatial interactions are usually only reflected by exogenous variables or simple lag terms, making it difficult to explicitly characterize cross-regional propagation links and spatial structural dependencies.

To simultaneously model temporal dynamics and spatial relationships, spatiotemporal deep learning models have gradually become a research hotspot. Among them, the Spatiotemporal Graph Neural Network (ST-GNN) treats regions as graph nodes and inter-regional relationships (adjacency, migration, transportation, economic ties, etc.) as graph edges, learning spatial dependencies in graph convolutional/graph attention networks, and combining temporal convolution, gating mechanisms, or attention structures to capture temporal evolution patterns. This type of method has been validated in traffic flow prediction, spatiotemporal prediction of air pollution, and some infectious disease transmission studies[1][2]. It can learn complex nonlinear spatiotemporal coupling relationships without requiring strong linear assumptions and has good multivariate fusion capabilities and scalability.

The above studies provide important methodological references for the spatiotemporal prediction of tuberculosis, but it is still necessary to fully consider the disease characteristics of tuberculosis, such as "long incubation period, relatively slow epidemic process, and long-term trend": on the one hand, the model needs to take into account both short-term fluctuations and long-term dependencies, capturing cross-seasonal and cross-year evolutionary patterns; on the other hand, significant differences in socioeconomic background and medical accessibility in different regions lead to strong spatial heterogeneity and time-lag and non-stationary characteristics in transmission associations. Therefore, it is necessary to construct a highly robust spatiotemporal model oriented towards the provincial scale, which can improve prediction accuracy while enhancing the ability to identify key regions and key periods, and provide a unified data-driven framework for subsequent risk interpretation and decision support.

2.3. Progress in Risk Attribution Research

In the field of disease risk attribution research, the Comparative Risk Assessment (CRA) framework proposed by the Global Burden of Disease (GBD) study provides a standardized path for quantitatively assessing the contribution of modifiable risk factors to disease burden. This framework typically combines exposure distribution, relative risk (RR), and theoretical minimum risk exposure level (TMREL) to estimate the population attributable fraction (PAF) and its corresponding changes in morbidity, mortality, and disability-adjusted life years (DALYs), thereby achieving a comparable assessment of risk factors. In tuberculosis research, factors such as malnutrition, diabetes, smoking, alcohol consumption, air pollution, and socioeconomic deprivation have been repeatedly proven to be closely related to the occurrence and development of tuberculosis [3]. CRA provides an important basis for identifying high-contribution risk factors and prioritizing interventions (GBD Collaborators, 2020).

In recent years, domestic research has gradually introduced the GBD/CRA concept to conduct tuberculosis risk attribution at the regional level. Commonly used methods include cross-sectional correlation analysis, panel regression, counterfactual inference, and static PAF estimation, which can provide a reference for identifying major risk factors in specific regions. However, existing work still mainly relies on static statistical attribution, which generally has three shortcomings: First, it lacks deep coupling with spatiotemporal prediction models, making it difficult to characterize the spatiotemporal dynamics and regional heterogeneity of risk factor effects; second, it is difficult to conduct intervention scenario and long-term effect simulations within a unified framework, thus limiting the forward-looking assessment of policy combinations and resource allocation; third, there is still room for improvement in the spatiotemporal alignment, lag effects, and potential confounding control of multi-source data. Therefore, combining high-performance spatiotemporal prediction models with CRA attribution analysis and conducting dynamic attribution and scenario assessment at the provincial scale has clear methodological value and practical significance.

Based on the above research status, the main innovations of this paper are:

(1) Combining spatiotemporal prediction models with risk attribution analysis to achieve dynamic assessment of the tuberculosis burden;

(2) The ST-GNN model is used to perform spatiotemporal weighted estimation of the impact of risk factors, thereby realizing the quantitative analysis of the risk action path;

(3) Conducting long-term effect simulations based on policy intervention scenarios to provide a more forward-looking scientific basis for public health decision-making.

3. Research Methods and Model Design

3.1. Overall Research Framework

This study focuses on the spatiotemporal prediction and risk attribution analysis of tuberculosis incidence, constructing a data-driven analysis process. The overall technical route follows the approach of "data preprocessing—model construction—parameter optimization—result evaluation—risk analysis". This framework aims to improve the interpretability and practical application value of the analysis results while ensuring the predictive performance of the model.

Specifically, firstly, multi-source public health and socioeconomic data are cleaned and standardized to reduce the impact of outliers and missing data on model training; secondly, a spatiotemporal graph structure is constructed using provincial-level administrative regions as the basic analysis unit, incorporating both spatial relationships and time-series characteristics between regions into the modeling process; subsequently, a spatiotemporal graph neural network is used to predict tuberculosis incidence, and key model parameters are adjusted using Bayesian optimization methods to improve model stability; finally, based on the prediction results, risk attribution analysis is conducted using a comparative risk assessment framework, and scenario simulations are performed to evaluate the impact of potential interventions.

This research framework maintains methodological operability while balancing prediction accuracy and analytical interpretability, making it suitable for research on chronic infectious diseases with significant spatiotemporal

characteristics, such as tuberculosis.

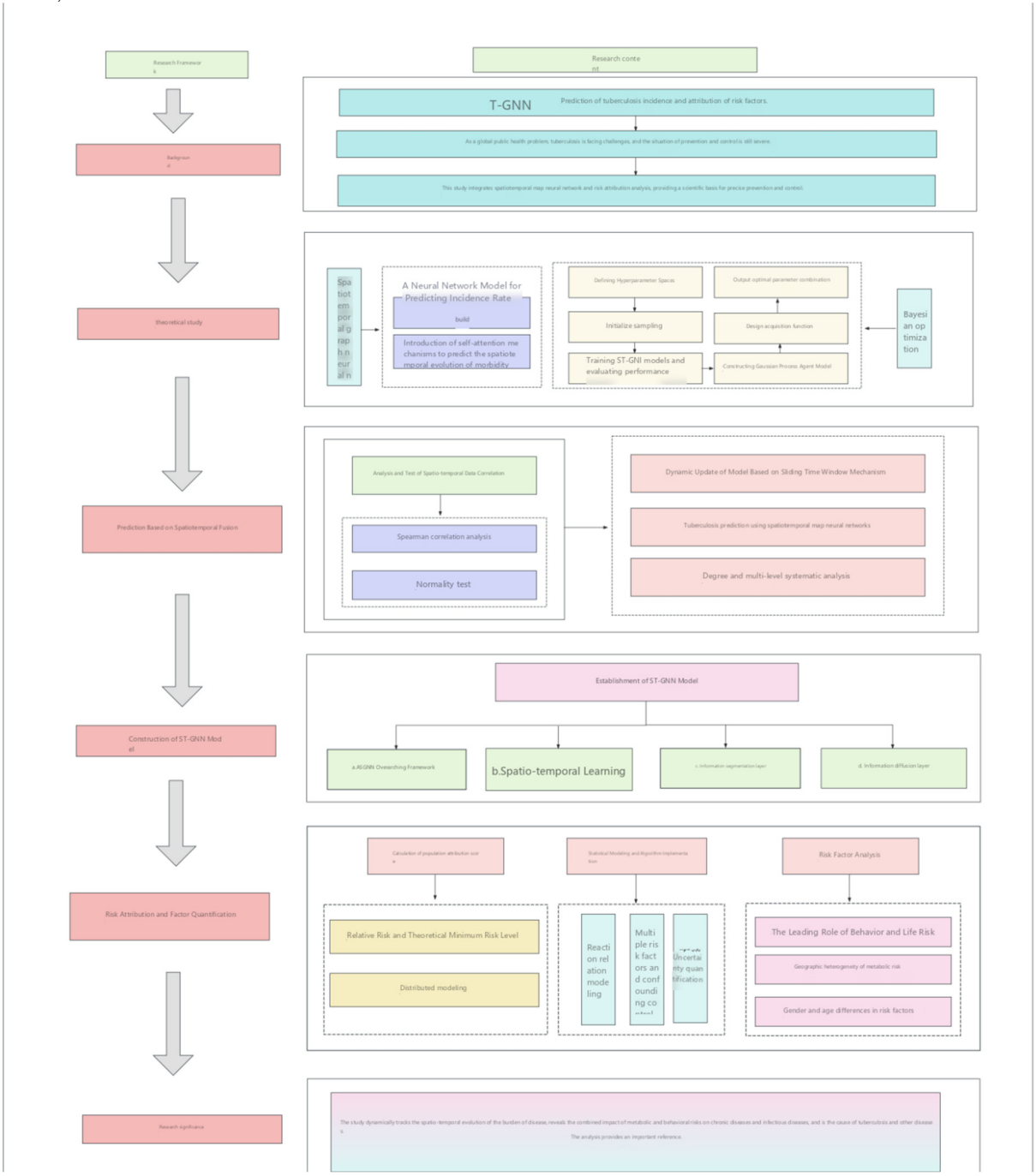


Figure 1. Research Technology Roadmap

3.2. Data Preprocessing and Feature Construction

1. Outlier Identification and Handling

To improve the reliability of the model input data, this paper performs outlier detection and correction on the incidence of tuberculosis and related indicators. Outlier identification is carried out by combining the box plot method with the Z-score statistic. First, the possible outlier observations are initially identified by the box plot method, and their upper and lower limits are defined as:

Upper Limit $=Q_3+1.5\times$IQR, lower limit $=Q_1-1.5\times$IQR

Where,$Q_1$and$Q_3$where IQR represents the first and third quartiles, respectively, and IQR is the interquartile range. For observations that exceed the above range, their Z-score is further calculated:

$$Z_I=\frac{X_i-u}{\sigma}$$

When $|Z_i|>3$, it is identified as an outlier. For confirmed outliers, this paper does not directly remove them, but uses the method of interpolation between adjacent time points to correct them, so as to preserve the overall trend characteristics of the original time series as much as possible.

2. Missing Value Imputation

In view of the data missing problem in some years and regions, this paper uses the spatio-temporal kriging method for estimation. This method considers the influence of spatial distance and time lag on the correlation of variables, and is suitable for data with obvious spatiotemporal dependence characteristics.

Its covariance structure is defined as:

$$C(h_s,h_t)=p^2\exp(-\frac{h_s}{\phi_s}-\frac{h_t}{\phi_t})$$

Where,$h_s$and$h_t$representing spatial distance and temporal

distance respectively, $\emptyset_s$and$\emptyset_t$are spatial and temporal scale parameters. The relevant parameters are determined by the maximum likelihood estimation method, so that the imputation results are consistent with the existing observation data as a whole, thereby reducing the potential impact of missing values on model training.

3. Spatiotemporal Feature Matrix Construction

In the feature construction stage, this paper considers the 31 provincial-level administrative regions in China as nodes in a graph structure, with each node corresponding to the observation data of a province in different years. The node feature vector mainly includes the incidence rate of tuberculosis, mortality rate, socioeconomic indicators, and exposure levels of major risk factors.

In terms of time dimension, a time series is constructed using annual data, with each node containing continuous data T Characteristic information of a time step. The final spatiotemporal feature tensor is represented as:

$X \in R^{N \times T \times F}$

Where N represents the number of nodes, T represents the number of time steps, and F represents the feature dimension.

The spatial adjacency matrix is used to describe the correlation between different provinces. This paper considers both geographical adjacency and population mobility intensity to construct the adjacency matrix, which is defined as:

$A_{ij}=\alpha G_{ij}+(1-\alpha)M_{IJ}$

Where, G_{ij} representing geographical adjacency, M_{ij} representing population mobility intensity index, αbeing the balance coefficient. This construction method can reflect the potential transmission links between regions to a certain extent.

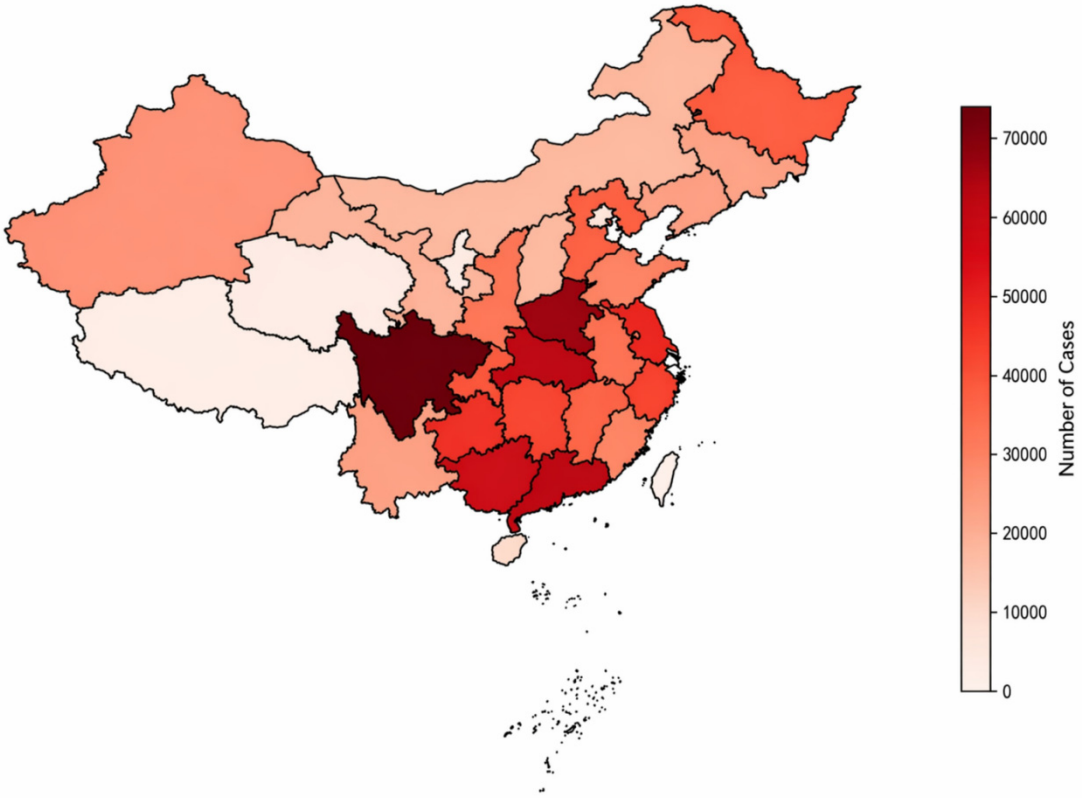


Figure 2. Heat map of the number of cases in each province in 2020

3.3. Construction of the Spatiotemporal Graph Neural Network Model

1. Overall structure of the model

The spatiotemporal graph neural network model constructed in this paper mainly consists of three modules: spatial feature extraction module, temporal feature extraction module, and spatiotemporal attention module. The basic idea of the model is to learn the variation patterns of tuberculosis incidence in the spatial and temporal dimensions, and to assign higher weights to key regions and key time periods through an attention mechanism.

From an intuitive perspective, this model is equivalent to learning the pattern of tuberculosis evolution over time on an "inter-provincial network," allowing nodes with closer spatial connections or more significant temporal changes to play a greater role in the prediction process.

2. Spatial Feature Extraction Module

Spatial feature extraction is implemented using Graph Convolutional Network (GCN). Its core idea is to learn the spatial dependencies between nodes by aggregating information from neighboring nodes. This paper adopts the normalized adjacency matrix form with self-loops, and its calculation process can be expressed as:

$H^{(k+1)}=\sigma(\tilde{D}-\frac{1}{2}\tilde{A}\tilde{D}-\frac{1}{2}H^{(k)}W^{(k)}+b^{(k)})$

Where$\tilde{A}=A+I$represents the adjacency matrix after adding self-loops, $\tilde{D}$is the corresponding degree matrix, $H^{(k)}$is the node feature representation of the k-th layer, $\sigma(\cdot)$is the activation function.

This module can reflect the potential correlation characteristics between provinces in the process of tuberculosis transmission to a certain extent.

3. Temporal Feature Extraction Module

In the temporal dimension, this paper uses a temporal convolutional network (TCN) to model the TB incidence sequence. TCN expands the receptive field without

significantly increasing the number of parameters through dilated convolutions, making it suitable for capturing long-term time-dependent features.

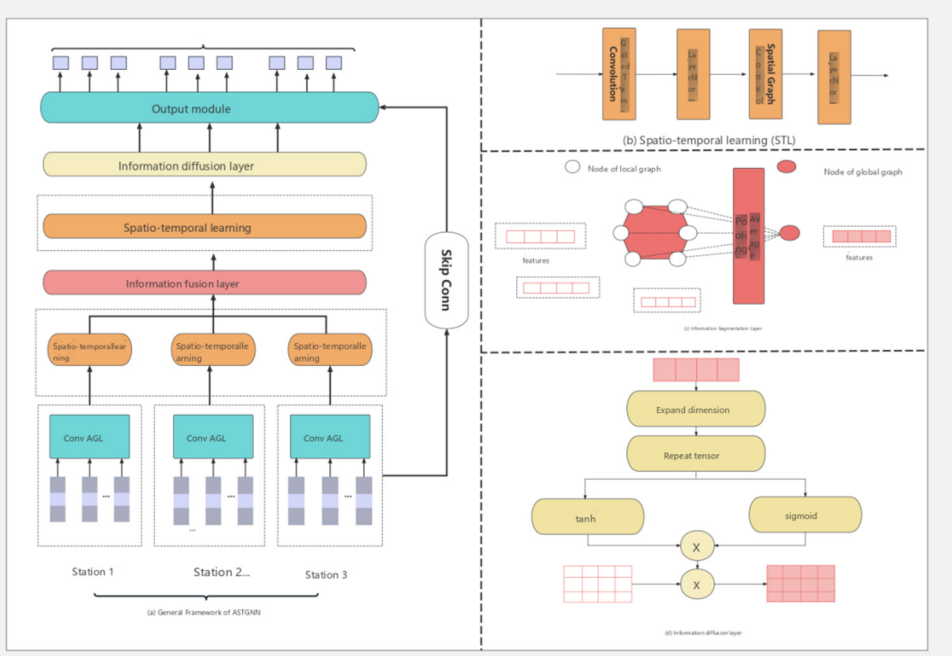


Figure 3. Flowchart of ST-GNN model

Its basic calculation form is:

$H_t^{(l)}=\mathrm{ReLU}(W_t^{(l)}*X_t^{(l)}+b_t^{(l)})$

By setting convolutional layers with different dilation rates, the model can simultaneously learn short-term fluctuations and long-term trend changes, thereby more comprehensively depicting the temporal dynamic characteristics of tuberculosis incidence.

4. Spatiotemporal Attention Mechanism

To enhance the model's ability to express heterogeneous spatiotemporal features, this paper introduces a spatiotemporal attention mechanism to adaptively weight the importance of different nodes and different time periods. Spatial attention is used to measure the relative importance of different provinces in the propagation network, while temporal attention is used to characterize the contribution of different time steps to the prediction results.

By introducing this mechanism, the model can identify regions and key time periods that have a significant impact on the prediction results to a certain extent, thereby improving the stability of the prediction performance.

5. Output Layer and Loss Function

The output layer of the model adopts a fully connected structure, mapping the extracted spatiotemporal features to the predicted value of tuberculosis incidence:

$\hat{y}_t=W_O\mathrm{h}_t+b_o$

Mean squared error (MSE) is used as the loss function during model training:

$$L=\frac{1}{NT}\sum_{i,t}(Y_{it}-\hat{Y}_{it})^2$$

L2 regularization and Dropout strategies are introduced to reduce the risk of model overfitting.

3.4. Application of Bayesian Optimization in Model Parameter Adjustment

Considering that there are multiple hyperparameters in the model that have an important impact on the prediction performance, this paper uses the Bayesian optimization method to automatically adjust the model parameters. This method constructs a probabilistic mapping relationship between hyperparameters and model performance, finding optimal parameter combinations within a limited number of search iterations. It is suitable for model training tasks with high computational costs.

In the specific implementation, this paper uses the prediction error on the validation set as the optimization objective, uses Gaussian process to model the objective function, and uses the Expected Improvement (EPI) acquisition function to determine the next set of parameters to be evaluated. This process helps to improve the overall stability of the model while ensuring search efficiency.

It should be noted that Bayesian optimization in this study is mainly used as an auxiliary parameter tuning method. Its purpose is to reduce the uncertainty caused by manual empirical parameter setting, rather than to improve the optimization algorithm itself.

4. Experimental Results and Analysis

4.1. Experimental Design and Evaluation Indicators

To evaluate the performance of the constructed spatiotemporal graph neural network model in predicting tuberculosis incidence, this paper selects tuberculosis monitoring data from 31 provincial-level administrative regions in my country from 2008 to 2020 for experimental analysis. According to the time sequence, the data is divided into a training set (2008-2017), a validation set (2018-2019), and a test set (2020) to avoid the impact of future information leakage on the prediction results.

The model is implemented in the PyTorch Geometric framework and uses the Adam optimizer for parameter updates. During training, the model convergence is monitored using validation set error, and key hyperparameter combinations are determined using Bayesian optimization methods to reduce the uncertainty of manual parameter

setting.

To evaluate the model's predictive performance from different perspectives, this paper selects mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2), and mean relative error (MAPE) as evaluation metrics. These metrics reflect the model's performance in terms of error level, stability, and overall fit.

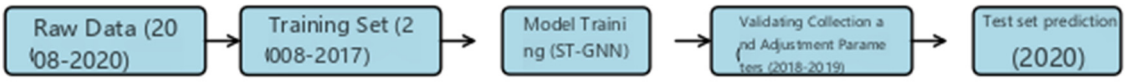


Figure 4. Schematic diagram of model training and evaluation process

4.2. Comparison of Predictive Performance of Different Models

To verify the predictive effect of the constructed model, this paper conducts comparative experiments with several commonly used prediction methods, including the Seasonal Autoregressive Moving Average (SARIMA) model, Long Short-Term Memory (LSTM) network, Gated Recurrent Unit (GRU) network, and XGBoost model. All comparative models are trained and tested under the same data partitioning and evaluation metrics to ensure the fairness of the comparison results.

The predictive performance of each model on the test set is shown in Table 1.

Table 1. The predictive performance of each model on the test set

Model	MAE	RMSE	R^2
SARIMA	1.98	2.531	0.862
LSTM	1.62	2.004	0.889
GRU	1.45	1.865	0.901
XGBoost	1.38	1.781	0.914
ST-GNN (the model presented in this paper)	**1.21**	**1.531**	**0.932**

The results show that the ST-GNN model outperforms the comparison models in terms of MAE, RMSE,$R^2$and MAPE. Among these, deep learning models have a certain advantage in prediction accuracy compared to traditional time series models, and the ST-GNN model further improves error control and fitting ability after introducing spatial correlation modeling.

It should be noted that the performance differences between different models are not solely determined by model complexity. The ST-GNN model, by explicitly introducing inter-provincial spatial correlation structures, allows the potential propagation connections between regions to be reflected, thereby improving the prediction effect to a certain extent.

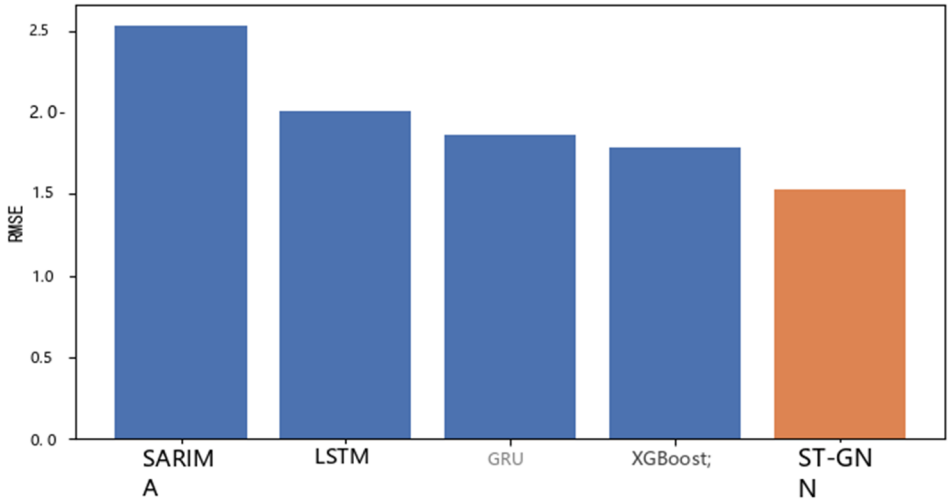


Figure 5. Comparison of prediction errors between models (RMSE)

4.3. Ablation Experiment Analysis

To further analyze the impact of each component module in the model on the prediction performance, this paper designed an ablation experiment. By gradually removing key components in the model, the changes in prediction error were observed. The specific experimental settings and results are shown in Table 2.

Experimental results show that when the spatial feature modeling module is removed and only the time series modeling is retained, the model prediction error increases significantly, indicating that spatial correlation information plays an important role in the prediction of tuberculosis incidence. When only spatial modeling is retained without considering temporal dynamic features, the model performance also decreases, indicating that time series information is indispensable for characterizing long-term trends.

Table 2. The specific experimental settings and results

Model Variants	MAE	RMSE
Removal of Spatial Modeling	1.68	1.974
Removal of Temporal Modeling	1.57	1.853
Removal of Attention Mechanism	1.41	1.648
Removal of Bayesian Optimization	1.36	1.592
Complete Model (ST-GNN)	**1.21**	**1.531**

In addition, after removing the spatiotemporal attention mechanism and Bayesian optimization module, although the model performance did not decline drastically, the prediction error still increased, indicating that these modules played an auxiliary role in improving model stability and parameter configuration rationality.

Overall, the modules in the model have a certain degree of complementarity, jointly promoting the improvement of prediction performance.

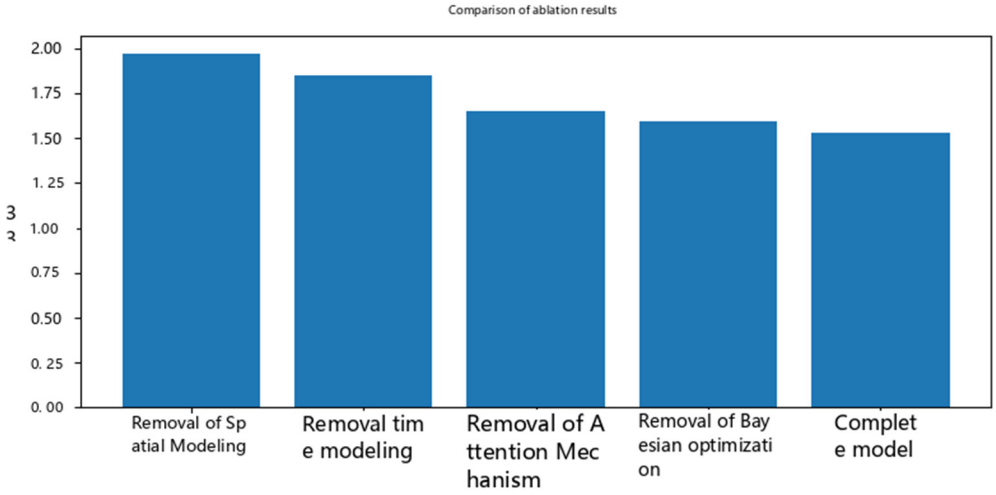


Figure 6. Comparison of ablation results

4.4. Visualization of Prediction Results and Error Distribution Analysis

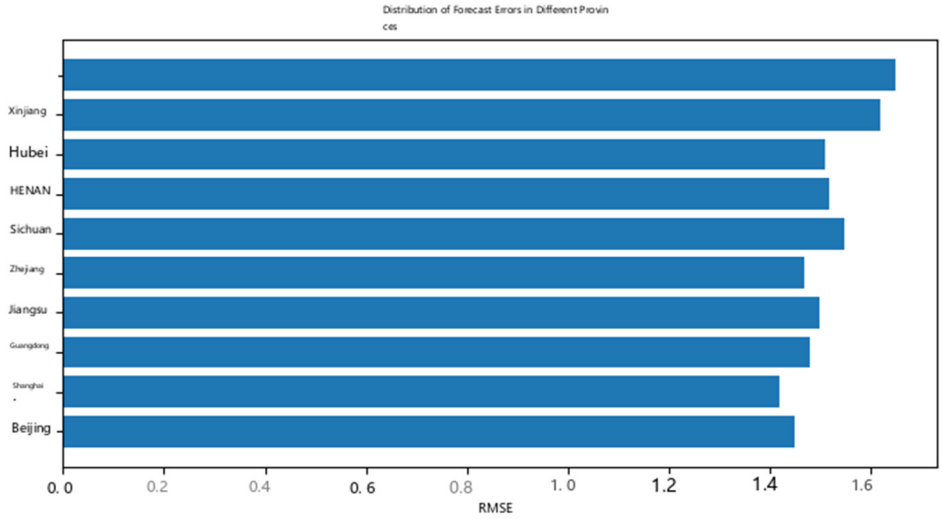


Figure 7. Distribution of forecast errors in different provinces

To intuitively demonstrate the model's prediction effect, this paper selects the national scope and some representative provinces in 2020 to compare and analyze the predicted values with the actual observed values. The results show that the model can generally fit the changing trend of tuberculosis incidence well, and the predicted values and actual values maintain a high degree of consistency in most provinces.

At the spatial level, the error distribution results show that the overall prediction error in the eastern and central regions is relatively low, while the prediction error in some western regions is slightly higher. This phenomenon may be related to factors such as the low density of monitoring data and

complex population flow patterns in the relevant regions. Overall, the model shows relatively stable prediction ability in different regions.

Error distribution analysis shows that the model did not exhibit systematic overestimation or underestimation in high-incidence areas, indicating its robustness across different incidence levels.

4.5. Summary of Results

The experimental results show that the constructed spatiotemporal graph neural network model demonstrates good overall performance in predicting tuberculosis incidence. Compared with models that do not explicitly consider spatial correlations or temporal dynamics, its prediction error is generally lower, and it exhibits relatively stable performance across different regions and time scales.

Experimental results indicate that combining spatial correlation modeling with time series analysis helps to more comprehensively characterize the spatiotemporal evolution of tuberculosis incidence, providing a reliable predictive basis for subsequent risk attribution analysis and scenario simulation.

5. Discussion and Conclusion

5.1. Discussion of Main Research Results

Based on tuberculosis incidence and related covariate data from 31 provincial-level administrative regions in my country from 2008 to 2020, this study constructed a spatiotemporal graph neural network (ST-GNN) model that integrates spatial correlation and temporal dynamic features. The model was compared with baseline models such as SARIMA, LSTM, GRU, and XGBoost under a unified data partitioning and evaluation index system. The results show that the constructed model achieves better predictive performance and more stable generalization on the test set (MAE=1.21, RMSE=1.531, R^2=0.932), indicating that in provincial-level tuberculosis surveillance scenarios, explicitly modeling the coupling structure of "inter-regional correlation-temporal evolution" can effectively improve prediction accuracy and robustness.

From the perspective of methodology, the performance improvement mainly comes from two aspects: First, the graph convolution module uses graph structure to propagate and aggregate information on inter-provincial connections, enabling the model to learn the spatial dependence implied by factors such as population flow, economic ties, and geographical proximity, thereby capturing cross-regional transmission connections and spillover effects; Second, the temporal convolution module extracts local fluctuations and long-term trend features of sequences under multi-scale windows, which helps to simultaneously characterize seasonal changes and cross-year evolution patterns. The combination of the two enables the model to learn more complex nonlinear spatiotemporal relationships without relying on strong linear assumptions when facing the epidemic characteristics of tuberculosis, which are characterized by "low fluctuations, long cycles, and being driven by multiple factors."

Ablation experiments further validated the key contributions of spatial and temporal modeling to performance: removing any module led to an increase in prediction error. Meanwhile, the spatiotemporal attention mechanism adaptively highlighted key regions and key time steps, improving the model's ability to identify "high-risk nodes / high-risk periods"; Bayesian optimization automatically optimized within the hyperparameter search space, helping to reduce the uncertainty of manual parameter tuning and improve training stability. Overall, this indicates that tuberculosis prediction tasks need to strike a balance between model expressive power and generalization ability to adapt to the non-stationary changes in data quality and epidemiological characteristics across different regions.

5.2. Discussion of Risk Attribution Analysis Results

Based on predictive analysis, this paper introduces the Comparative Risk Assessment (CRA) framework proposed by the GBD study to quantify the relative contributions of various modifiable risk factors. The analysis results suggest that behavioral and metabolic-related factors (such as smoking, high BMI, and insufficient physical activity) have a high attributable contribution to the tuberculosis burden, exhibiting significant regional and population differences. This finding reflects the intertwining of tuberculosis and chronic disease risk factors in the real world: in some regions, metabolic abnormalities and unhealthy behaviors may amplify the risk of tuberculosis or affect disease outcomes through pathways such as impaired immune function, increased comorbidities, and changes in healthcare utilization.

It should be emphasized that the risk attribution in this paper belongs to "comparative risk assessment at the population level." Its core lies in estimating the relative contribution of different risk factors to the overall burden based on existing relative risk evidence and exposure distribution, used to identify potential intervention priorities and compare intervention directions, rather than making a direct judgment on causal effects at the individual level. Therefore, the relevant conclusions are more appropriately interpreted as "quantitative hints" in public health strategy formulation, and should be comprehensively evaluated in conjunction with the specific epidemiological background of the region, policy feasibility, and resource constraints.

From the perspective of prevention and control practice, the attribution results suggest that tuberculosis prevention and control can be more closely integrated with chronic disease management and health promotion strategies: for example, strengthening tobacco control, improving nutritional and metabolic health, and increasing physical activity levels in high-burden areas, combined with screening, standardized treatment, and follow-up management of key populations, is expected to create a synergistic effect in reducing the incidence and recurrence risk of tuberculosis. At the same time, the existence of regional differences also means that intervention strategies should not be "one-size-fits-all," but should be based on the risk factor spectrum and health resource endowment, and implement precise intervention and resource allocation in different regions and levels.

5.3. Practical Significance of Scenario Simulation Results

Based on the prediction model output and CRA attribution results, this paper further conducts scenario simulations to compare the relative impacts of different intervention directions on tuberculosis incidence under certain assumptions. Simulation results show that achieving a decrease in smoking rates, an increase in physical activity levels, or an improvement in metabolic health in the medium

to long term could all lead to a downward trend in tuberculosis incidence, although the magnitude and pathway of these effects vary across different regions. These results indicate that incorporating risk factor interventions into a comprehensive tuberculosis prevention and control system, combined with regionally differentiated management, has potential real-world benefits.

It should be noted that the output of scenario simulation is highly dependent on the model settings, relative risk parameters, exposure change magnitude and time lag assumptions, etc. Therefore, its conclusions should not be regarded as accurate predictions of the future, but should be used as a "directional and comparative" quantitative tool to support the discussion of intervention priority ranking and resource allocation. In the future, if more granular exposure data, regionally specific risk parameters, and policy implementation cost and accessibility constraints can be introduced, it will help improve the realism and decision-making value of scenario simulation.

5.4. Research Limitations

Although this paper explores the integrated analysis of spatiotemporal prediction and risk attribution of tuberculosis at the provincial scale, the following limitations still exist:

(1) Data granularity and heterogeneity: This paper uses provincial aggregated data, which makes it difficult to characterize the exposure differences and transmission chains at the individual level, thus limiting the identification of micro-mechanisms and high-risk subgroups; (2) Data quality and non-stationarity: Problems such as monitoring caliber, underreporting and reporting delays in some regions in earlier years may affect model learning, and changes in prevention and control measures may introduce structural mutations, increasing prediction uncertainty; (3) Spatial association setting: The graph structure is mainly based on adjacency or similarity, and has not fully integrated the dynamic connections of migration flow, transportation network and other factors that can better reflect the real transmission path; (4) Externality of attribution parameters: The relative risk used by CRA is mostly from existing research or aggregated evidence, and regional differences and potential confounding may bring uncertainty; (5) Limitations of causal explanation: This paper focuses on association assessment and has not yet introduced a strict causal inference framework to identify policy effects.

The above limitations suggest that the conclusions of this paper need to be calibrated and verified in combination with local data conditions and prevention and control background when they are promoted and applied. At the same time, they also provide clear directions for improvement for subsequent research.

5.5. Conclusion and Outlook

In summary, this paper focuses on the spatiotemporal prediction and risk attribution of tuberculosis incidence. The main conclusions are as follows: First, the constructed ST-GNN model can effectively characterize the spatiotemporal coupling characteristics of tuberculosis incidence at the provincial level, and achieves better predictive performance and stability in comparison with various baseline models; Second, attribution analysis combined with the CRA framework can be used to identify key modifiable risk factors of tuberculosis burden in different regions, providing a quantitative basis for the selection of intervention priorities and regional and stratified prevention and control; Third, scenario simulation based on prediction-attribution results can provide a reference for comparing the relative effects of different intervention directions, and support discussions on public health resource allocation and strategy priorities.

Looking to the future, we can deepen our understanding in the following aspects: First, we can introduce more granular (city/county/district and even individual level) monitoring and exposure data, and integrate dynamic information such as migration flows, transportation networks, and medical service utilization to enhance the characterization of real transmission paths and spatial heterogeneity; Second, we can combine causal inference methods (such as difference in difference, instrumental variables or structural causal models) to assess the causal effects of key policies/interventions and improve the explanatory power of risk analysis mechanisms; Third, we can promote research on model interpretability and deployability, link the explanatory results such as attention weights and feature contributions with business indicators, and build a decision support system that can be used for routine monitoring and early warning; Fourth, we can expand to multi-disease joint prediction and comorbidity risk assessment scenarios to provide more comprehensive data-driven support for the comprehensive prevention and control of chronic infectious diseases.

Acknowledgments

2025 University-level Student Innovation and Entrepreneurship Training Program; Research on Tuberculosis Incidence Prediction Based on the ST-GNN Model. Project Number:X2025068.

References

[1] Cheng Yi Yang, Feng Liu, Jia Yin Qi, et al. Graph convolutional neural network spatiotemporal data learning for COVID-19 epidemic prediction [J]. Chinese Journal of Image and Graphics, 2021, 26 (5): 1128-1137.

[2] M Barman, M Panja, N Mishra, et al. Epidemic-guided deep learning for spatiotemporal forecasting of Tuberculosis outbreak [J]. arXiv preprint, 2025.

[3] Harztek Salavati, Bahani Mailiman, Yanwu Nie, et al. Tuberculosis incidence trend and age-period-cohort analysis in China from 1990 to 2019 [J]. China Preventive Medicine, 2022, 23 (12): 881-887.