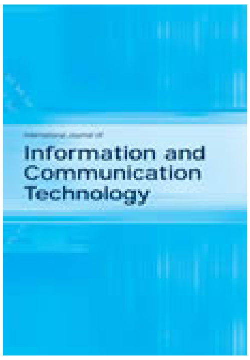


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Forecasting trend of agricultural talents flow by spatio-temporal graph neural network and LightGBM

Jie Liu

DOI: 10.1504/IJICT.2026.10075962

Article History:

Received:	29 August 2025
Last revised:	17 September 2025
Accepted:	18 September 2025
Published online:	09 February 2026

Forecasting trend of agricultural talents flow by spatio-temporal graph neural network and LightGBM

Jie Liu

School of Management,
Xinxiang University,
Xinxiang, 453003, China
Email: jieliuu@163.com

Abstract: Current agricultural talent flow prediction mainly uses single models (e.g., linear regression, ARIMA, LSTM), which fail to capture non-Euclidean spatial-temporal relationships and automatically extract complex spatio-temporal interactions, limiting accuracy and interpretability. This paper proposes a hybrid framework integrating a spatio-temporal graph neural network (STGNN) and LightGBM. Using 2010–2020 data from 17 cities in Henan Province, a spatio-temporal graph is built with city nodes and geographic-threshold edges. STGNN combines graph convolution and temporal convolution (TCN) to automatically learn spatio-temporal features, while LightGBM regresses lag and socio-economic indicators for interpretability. Benchmark comparisons with ARIMA, LSTM, and LightGBM, plus ablation and sensitivity tests, confirm the hybrid model's superiority. It reduces error by 10%–14% versus standalone STGNN/LightGBM, achieving under 12.3% overall error, with significantly improved accuracy and stability.

Keywords: agricultural talent flow; spatio-temporal graph neural network; STGNN; LightGBM; hybrid prediction.

Reference to this paper should be made as follows: Liu, J. (2026) 'Forecasting trend of agricultural talents flow by spatio-temporal graph neural network and LightGBM', *Int. J. Information and Communication Technology*, Vol. 27, No. 8, pp.1–18.

Biographical notes: Jie Liu graduated from Zhejiang University with a Master's degree in 2007. She is currently an Associate Professor at the School of Management, Xinxiang University. Her research interests are mainly in human resource management, business management, and labour economics.

1 Introduction

With the comprehensive advancement of China's rural revitalisation strategy, agricultural talents, as a key factor in promoting agricultural modernisation and the sustainable development of the rural economies, have attracted considerable attention to their flow trends and distribution patterns. Especially under the background of regional coordinated development and optimisation of rural talent structure, accurately predicting the inflow and outflow of agricultural talents will not only help to speed up the allocation efficiency of agricultural production factors, but also provide a scientific basis for policy-making departments, thus helping to balance and rationally allocate talent resources among regions.

The flow of agricultural talents has an obvious temporal and spatial correlation (Geng and Yang, 2017; Qu et al., 2022). On the one hand, there is often a gradient transfer of talents between geographically adjacent or economically and culturally similar regions. On the other hand, factors such as the adjustment of annual planting structures, fluctuations in agricultural prices, and changes in the macroeconomic environment make the talent flow exhibit significant temporal dynamic characteristics. Therefore, how to effectively capture the spatial dependence and historical sequence evolution among regions is the core challenge to realise high-precision agricultural talent flow prediction (Yang et al., 2024).

Traditional talent flow prediction methods are mostly based on time series models and regression analysis. Models such as ARIMA exponential smoothing can describe linear trends and seasonal fluctuations (Rabbani et al., 2021), but it is difficult to take into account the non-Euclidean spatial structure between regions; Regression trees and support vector machines based on machine learning have made breakthroughs in multi-factor fusion (Zhang et al., 2025), but rely on manual feature engineering, and it is difficult to extract complex spatial-temporal interaction information automatically. In recent years, with the rapid development of deep learning technology, long- and short-term memory networks (LSTMs) (Greff et al., 2016) and gated recurrent units (GRUs) (Mim et al., 2023) have performed well in mining time dependencies, but have limited modelling capabilities for spatial topologies.

Graph neural networks (GNNs) (Li et al., 2022) enable efficient information propagation and aggregation in complex networks by embedding entities and their adjacency relationships into graph structures. Spatio-temporal graph neural networks (STGNNs) (Chen et al., 2025b; Wang et al., 2022) combine graph convolution with time series modelling to capture spatial dependencies and temporal dynamics. In this framework, cities are treated as nodes and geographic distances as edges, enabling the analysis of complex migration flows. STGNNs have shown strong performance in areas like traffic and air quality prediction, and are well-suited for modelling agricultural talent migration. By reflecting real-world spatial structures and evolving patterns, they offer valuable insights for regional planning and resource optimisation. However, at present, there are still few studies on the systematic application of the STGNN method in the agricultural field, especially in the prediction of regional agricultural talent flow. Its generalisation ability and practical value in different-scale regions and time granularities need to be further explored.

At the same time, LightGBM, as an efficient gradient boosting decision tree framework, is widely used in various prediction tasks due to its advantages of fast training speed, support for large-scale data, and the ability to process multiple types of features (Chen et al., 2025a). However, its essence remains a decision tree model based on feature splitting, and it is challenging to directly model the deep dependence of geographical neighbourhood structure and historical sequence information. The goal of this study is to organically combine the advantages of LightGBM and STGNN to construct a hybrid prediction framework that incorporates both spatial sensitivity and feature interpretation.

Talent migration has become an important topic in regional development studies. However, existing models often fail to capture the complexity of agricultural labour flows. In the context of rural revitalisation and agricultural modernisation, accurately forecasting the migration patterns of agricultural talent is essential for optimising resource allocation, improving productivity, and supporting policy-making. Traditional forecasting methods struggle to model the non-Euclidean spatial structures and dynamic temporal dependencies inherent in talent mobility. Traditional forecasting methods, like regression and basic time-series models, often assume linear relationships and overlook the complex spatial dependencies in agricultural talent migration. These models rely on Euclidean assumptions, failing to reflect irregular geographic and economic interactions between rural regions. Talent flows are shaped by diverse factors such as regional policies, infrastructure gaps, and seasonal labour needs, requiring models that can learn dynamic, non-linear, and graph-based patterns. As a result, conventional approaches are inadequate for capturing the spatio-temporal complexities of migration, highlighting the need for more advanced and adaptive modelling frameworks.

Based on the above background, this study is supported by data on agricultural talent flow in Henan Province from 2010 to 2020, and proposes a method to predict the trend of agricultural talent flow that integrates a STGNN and LightGBM. Firstly, the transfer records and socio-economic indicators of 17 prefectures and cities in Henan Province are used to construct a spatio-temporal map, where nodes represent prefectures and cities and edges are determined by geographical distance thresholds; Subsequently, the spatio-temporal features are encoded in parallel by graph convolution and time series convolution (TCN) to extract complex dependencies between cities and cities automatically. At the same time, gradient lifting training is carried out on pre-constructed lag features and economic indicators using LightGBM to supplement the shortcomings of STGNN in terms of characteristic interpretability. Finally, by comparing ARIMA, LSTM, and single LightGBM models, and conducting ablation experiments and parameter sensitivity analysis, the model performance and module contributions are systematically evaluated. The main contributions of this paper are as follows:

1 Propose a prediction scheme that organically fuses a STGNN and a LightGBM gradient lifting tree. The STGNN module extracts spatial dependencies and temporal dynamics between regions in parallel through graph convolution and TCN, while the LightGBM module performs efficient regression on pre-constructed lag features and socio-economic indicators. We reconstruct the phrase so that readers can clearly understand its intention. Here are the specific modifications: The complementary advantages of the two ensure the interpretability of the model while achieving high prediction accuracy.

2 Complete the construction of the spatio-temporal map. Based on the agricultural talent mobilisation records of 17 cities in Henan Province from 2010 to 2020, we use the geographical distance threshold to construct a regional adjacency network and integrate multidimensional socio-economic characteristics, such as annual planting structure, agricultural material prices, and labour costs, to provide high-quality spatio-temporal input for STGNN. The spatio-temporal map construction process is replicable and can be extended to other regions and industries.

3 Multi-benchmark comparison experiments with ARIMA, LSTM and single LightGBM models were designed, and ablation research and parameter sensitivity analysis were carried out. The experimental results show that the mean square error, mean absolute error and goodness of fit of the mixed model are improved by more

than 10% on average, and the stability and generalisation ability of the model are also significantly enhanced.

2 Related work

2.1 Agricultural talent flow and prediction method

In recent years, with the deepening of China's rural revitalisation strategy, the cross-regional flow of agricultural talents has become the core issue of rural modernisation development. Early studies mostly focused on qualitative interviews, focus groups and questionnaires. Through in-depth analysis of factors such as farmers' family background, education level, income level and social capital, the subjective drive of individual mobility willingness was revealed (Wójcik et al., 2019; Valentini et al., 2021). This kind of research provides a valuable perspective on the institutional environment and subjective motivation, but is limited by sample size and investigation depth, making it difficult to quantify the interaction effects between different factors and their temporal and spatial evolution in large-scale and high-dimensional data.

To explain the influence path, scholars have employed classical statistical methods, including multiple linear regression (Hensher and Greene, 2003), the Logit model (Mardani et al., 2017), and structural equation modelling (SEM) (Parzen, 2003), to quantitatively analyse the key driving factors of agricultural talent flow. For example, the Logit model is often expressed as formula (1):

$$P\left(y_i = 1 \middle| x_i\right) = \frac{1}{1 + \exp\left(-x_i^{\top}\beta\right)} \tag{1}$$

where y_i denotes the flow decision for i observations, x_i is the feature vector, and β is the parameter to be estimated.

By constructing a regression framework including economic income, land scale, social security, public services and other variables, the researchers quantified the marginal contribution of each index to the probability and scale of mobility. Although this kind of method has strong readability in feature interpretation, it relies on linear assumptions and prior variable selection, which cannot fully capture the nonlinear linkage between high-dimensional features, and it is difficult to automatically adapt to the changing macro environment.

With the improvement of big data technology and computing power, time series models such as ARIMA, exponential smoothing and seasonal decomposition models are gradually applied to the short-term prediction of regional agricultural talents (Han et al., 2019; Bashir and Wei, 2018). The general form of the *ARIMA*(p, d, q) model is equation (2):

$$\phi(L)(1-L)^d y_t = \theta(L)\varepsilon_t \tag{2}$$

where L is the lag operator and $\phi(L)$ and $\theta(L)$ are autoregressive and moving average polynomials, respectively. The expressions are shown in equations (3)–(4):

$$\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p \tag{3}$$

$$\theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q \tag{4}$$

Such models can achieve high forecast accuracy on monthly or quarterly scales by stabilising, trending, and seasonally adjusting the historical flow number series. However, the time series method lacks the endogenous characterisation of spatial adjacency effect, and cannot reflect the talent gradient transfer mechanism brought about by geographical proximity, industrial linkage or regional collaboration.

To overcome the above limitations, some studies employ multivariate spatio-temporal analysis methods, such as vector autoregression (VAR) (Millo and Piras, 2012) and panel data models (Yin et al., n.d.), to integrate cross-regional talent flows, economic indicators, and policy variables into a unified framework. The VAR model can be shown by equation (5):

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + \varepsilon_t \tag{5}$$

where y_t is the multivariate series, A_k is the coefficient matrix, and ε_t is the error term. The VAR model can handle the interaction between multivariate sequences, while the panel model enhances the estimation efficiency with the help of cross-regional information. However, this method is highly dependent on the prior setting of the spatial weight matrix and lag order, and is prone to dimensional disaster and overfitting risk in large-scale, high-dimensional, and multi-period data scenarios.

2.2 Application of STGNN in the field of prediction

The STGNN maps regional nodes and their geographical or functional associations into graph structures, and on this basis, introduces graph convolution and time series modelling modules to realise spatial dependence and temporal dynamic collaborative coding (Mao et al., 2024; Zhao et al., 2019). The propagation rule of the standard graph convolutional network (GCN) (Al-Selwi et al., 2024) at the layer is shown in equation (6):

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}\right) \tag{6}$$

where $\tilde{A} = A + I$ is the adjacency matrix of the plus self-loop, $\tilde{D}$ is the degree matrix, W^l is the weight matrix, σ is the activation function. On this basis, STGNN often combines one-dimensional TCN or cyclic unit (RNN/GRU/LSTM) to perform sliding window encoding on node feature sequences.

Compared with the hybrid architecture of traditional GNNs combined with recurrent neural networks (RNN) (Limouni et al., 2023) or time series convolutional networks (TCN) (Assis et al., 2021), STGNN can integrally capture neighbourhood information propagation and node feature evolution, providing a new idea for large-scale spatio-temporal data prediction.

In the field of talent flow prediction, representative models such as STGCN alternately stack graph convolution and TCN, and their core calculation can be expressed as formula (7):

$$H = TCN\left(GCN(X)\right) \tag{7}$$

DCRNN is based on the graph convolution of the diffusion process, combining the network diffusion mechanism of talent flow with the gated cyclic unit (GRU) (Li et al., 2018), and defining k^{th} third diffusion convolution as equation (8):

$$H_t = \sum_{k=0}^{K-1} (D^{-1}A)^k X_{t-k}\theta_k \tag{8}$$

θ_k is the learnable weight of the k^{th} order diffusion kernel. The above methods have achieved excellent results in the fields of integrating multi-source heterogeneous features of talent flow and enhancing the interpretability of results.

2.3 LightGBM and hybrid prediction strategy

Gradient boosting decision tree (GBDT) is widely used in financial risk control, e-commerce demand forecasting, and traditional traffic forecasting due to its powerful regression and classification capabilities (Ju et al., 2019). LightGBM is a high-performance implementation of GBDT. It utilises a histogram-based node splitting algorithm and a leaf-wise growth strategy to achieve fast training and memory optimisation for large-scale datasets. Its prediction model can be expressed as shown in equation (9):

$$\hat{y}_i = \sum_{m=1}^{M} f_m(x_i) \tag{9}$$

where each tree f_m, from the function space F, has an objective function of equation (10):

$$L = \sum_i \ell(y_i, y_i) + \sum_i \Omega(f_m) \tag{10}$$

where ℓ is the loss function and Ω is the regularisation term.

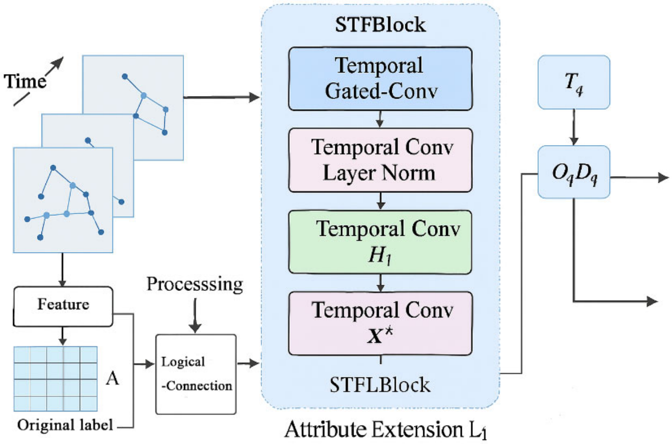
LightGBM has natural limitations in modelling endogenous spatial dependence and temporal dynamics (Simaiya et al., 2024; Beg, Pateriya and Tomar, 2024), and thus relies on hand-constructed time series lag features, spatial neighbourhood indicators, or network centrality measures to indirectly reflect spatio-temporal connections. To compensate for this deficiency, this study employs a range of hybrid and integration strategies that combine neural networks and GBDT. Model-level fusion first uses a deep neural network to extract high-dimensional features, and then submits them to GBDT for final prediction; Feature-level fusion takes the key split features generated by GBDT as neural network input to enhance interpretability and stability; Decision-level fusion fuses the prediction results of multiple models through weighted averaging, stacking or meta-learning to balance the bias and variance of different models.

3 Models and methods

3.1 Design of STGNN model

In this study, a STGNN is designed based on the parallel coding of graph convolution and temporal convolution. As shown in Figure 1, it is divided into four stages from left to right: input, expansion, extraction and mapping.

Figure 1 Architecture diagram of STGNN model (see online version for colours)



The input layer receives graph data at different times. The graph is composed of nodes and connected edges. The nodes carry dynamic features and spatial distance information, as well as attribute vectors and location features of location nodes. After stacking multiple time graphs, a timing diagram tensor is formed to provide raw samples for subsequent processing. The attribute expansion layer includes two parallel fusion units, which process the source-end and position-end map signals, respectively. Each unit extracts features, expands dimensions and aggregates information at the previous query time and the current time, and outputs the extended feature representation at the corresponding time. The spatio-temporal feature extraction layer integrates time-sequential gated convolution and spatial attention convolution submodules. The time-series gating module captures the multi-scale temporal evolution characteristics of the graph structure, and the spatial attention module measures the spatial dependency between nodes and focuses on key regions. The outputs of the two sub-modules are fused in the channel dimension to generate a unified spatio-temporal representation vector. The prediction mapping layer calculates the query code, observation vector, and duration vector based on the spatio-temporal representation vector and inputs them into a multi-layer fully connected network, together with the spatio-temporal representation. After linear transformation and nonlinear activation, the final prediction result is output.

3.2 LightGBM module and fusion training strategy

To enhance the interpretability of the model and address the shortcomings of STGNN in feature importance analysis, we developed a LightGBM regression model in parallel. The model takes standardised lag flow characteristics, socio-economic indicators, and quarterly proportions as inputs, and adjusts hyperparameters such as tree depth, leaf number, and learning rate through a histogram-based leaf growth strategy and five-fold cross-validation. It then realises the single-step regression prediction of talent flow for the next year. LightGBM supports the natural processing of category variables and provides an output feature importance ranking, offering an intuitive driver reference for policy formulation.

Figure 2 LightGBM model training and fusion structure diagram (see online version for colours)

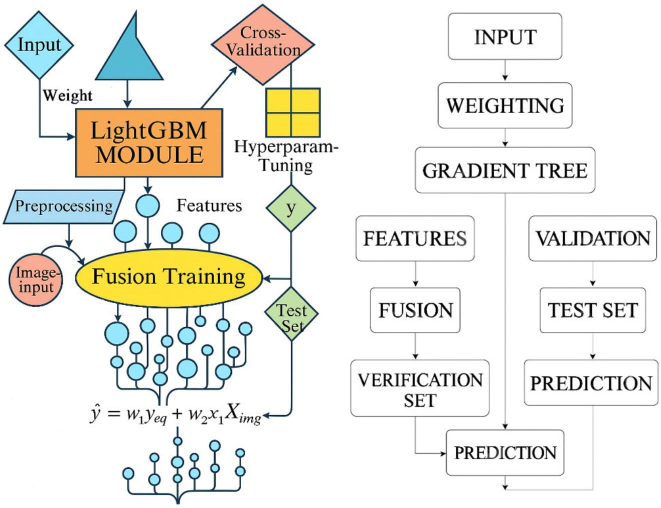


Figure 2 shows the overall training and prediction process of the proposed method. In this process, the 'input' module first receives the raw data, and then assigns learnable weights to different data channels in the 'weight' module to enhance the model's responsiveness to important information. The weighted data is sent to the 'gradient lifting tree module' for preliminary training, which continuously improves the prediction performance through iterative fitting of residuals. The training results enter the 'cross-validation' stage, where the model's generalisation ability is evaluated through multi-fold partitioning. The optimal model configuration is then determined through a systematic search in the 'hyperparameter tuning' link.

To address the complex challenges of agricultural talent migration forecasting, this study integrates STGNN and LightGBM based on their complementary strengths. STGNN effectively captures spatial-temporal dependencies by modelling cities or regions as graph nodes and using adjacency relationships to represent geographic and economic links. Through graph convolutions and temporal attention mechanisms, it learns dynamic migration patterns. However, STGNN lacks interpretability and requires high computational resources. LightGBM compensates with efficient training on structured data, low memory usage, and strong performance on small to medium datasets, while providing feature importance scores for transparency. In the fusion training module, features from the original data are combined with STGNN outputs, and both validation and test sets are introduced to ensure robustness and generalisation across diverse data distributions.

4 Data and experimental component

4.1 Experimental preparation

This study examines the flow of agricultural talent in Henan Province from 2010 to 2020. The core data originate from the transfer filing system of the Henan Provincial Department of Human Resources and Social Security, including prefecture-level administrative division codes, locations where talents are transferred in/out, the year of transfer, and the number of people. The auxiliary socio-economic indicators are sourced from the Statistical Yearbook of Henan Province, which covers annual GDP, total population, crop planting area, and per capita disposable income for various cities. To prevent bias toward data-rich cities, we applied a data balancing strategy before training. Underrepresented regions were upsampled using temporal interpolation, while overrepresented areas were downsampled to ensure uniform learning. This approach helped the STGNN-LightGBM model maintain low error rates across both urban hubs and remote regions, enhancing its generalisation and fairness.

Table 1 Core characteristics of agricultural talent flow in Henan Province from 2010 to 2020

Trait	*Data type*	*Description*
Year	int	Year of transfer
City_code	string	Administrative division code of prefecture and city
inflow	float	Number of inflows in the year (Z-score)
outflow	float	Number of outflows for the year (Z-score)
GDP	float	Gross regional product of the year (Z-score)

Table 1 summarises the key fields of agricultural talent flow in 17 cities in Henan Province from 2010 to 2020, including the mobilisation year (year), city code (city code), inflow and outflow (both standardised by Z-score), and regional gross domestic product (GDP, as standardised). The data undergoes preprocessing processes, including missing value interpolation, outlier elimination, and single-hot coding, to form a unified numerical feature input. This approach not only retains the time series characteristics but also takes into account regional economic attributes, meeting the requirements of STGNN and LightGBM models for structured data.

Table 2 Configuration table of experimental environment

Link	*Specific configuration*
Computing hardware	CPU: IntelXeonE5-2630v4 GPU: NVIDIA RTX2080Ti (11 GB), RAM: 32GB
Operating system	Windows10
Programming language	Python3.8. 10
Deep Learning Framework	PyTorch1.10. 0
Scientific computing library	NumPy1.20. 1, pandas1.2. 3
CUDA Environment	CUDA11.1
Development tools	Jupyter Notebook6.4. 3

Table 2 lists the software and hardware environments used for model training and evaluation, which ensure experimental reproducibility and stability.

During the experiment, agricultural talent flow data from Henan Province (2010–2020) were divided into three stages: 2010–2017 for training, 2018 for validation, and 2019–2020 for testing. This division ensures temporal independence and supports hyperparameter tuning while enabling rolling window prediction. The three-stage setup reflects real-world forecasting conditions, allowing the model to learn long-term migration patterns and evaluate its adaptability to recent shifts. It also captures seasonal cycles and policy-driven changes, enhancing the model's spatio-temporal generalisation and practical relevance. A three-year rolling window was adopted for temporal modelling, balancing the need to capture meaningful historical patterns with maintaining forecasting relevance. This window length effectively reflects seasonal labour cycles, policy changes, and economic fluctuations, while avoiding outdated or irrelevant information.

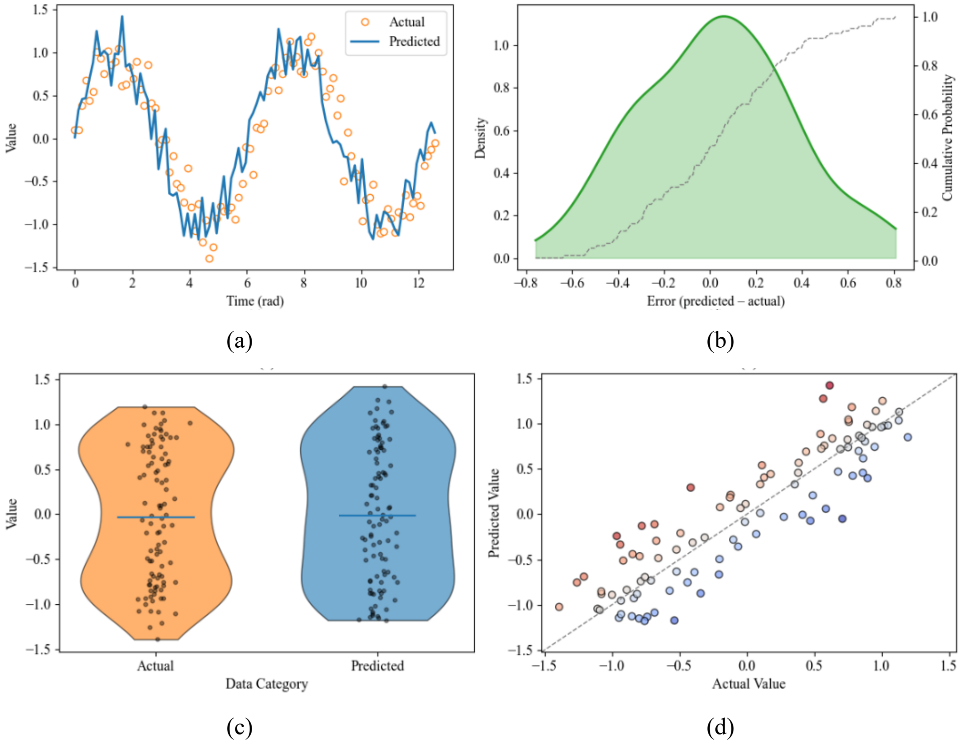
To ensure consistency across features and improve model convergence, all input variables – including migration counts, economic indicators, and regional attributes – were standardised using Z-score normalisation. This approach transforms each feature to have a mean of zero and a standard deviation of one, effectively eliminating scale disparities and preventing dominant features from biasing the learning process. Z-score normalisation was applied prior to graph construction and model training, ensuring that both the STGNN and LightGBM components received uniformly scaled inputs. This preprocessing step is particularly important for models sensitive to feature magnitude, such as gradient-based learners and GCNs.

4.2 Comparative test

In order to evaluate the advantages of the proposed STGNN in the task of forecasting agricultural talent flow, we selected three representative baseline models for comparison: the traditional linear time series model ARIMA, the univariate RNN LSTM, and the gradient lifting tree LightGBM based on feature engineering. All models were trained using data from 2010 to 2017 in Henan Province, tested with data from 2019 to 2020, and hyperparameter-tuned using 2018 data on the training set. A five-fold cross-validation strategy was employed during LightGBM training to enhance model stability, reduce overfitting, and ensure that the model generalises well across different subsets of the data by repeatedly validating performance on unseen samples.

As shown in Figure 3, this figure evaluates the performance of the prediction model from different angles with four subgraphs. Figure 3(a) compares the changing trends of real and predicted sequences with time. The solid polyline and hollow point represent the predicted value and the real value, respectively, which can intuitively reflect the timing dynamics and phase error. Figure 3(b) superimposes the kernel density estimation curve and the cumulative distribution curve of errors in the same coordinate system. The former illustrates the degree of concentration and fluctuation range of errors, while the latter reveals the cumulative probability distribution of prediction bias. Figure 3(c) displays the density visualisation and sample dispersion of the real and predicted numerical distributions, respectively, which helps distinguish the distribution patterns and median trends of the two sets of data. Figure 3(d) presents the corresponding relationship between the real value and the predicted value with a scatter plot colored with error, and draws a y = x reference line to reflect the systematic deviation and outlier of the model intuitively.

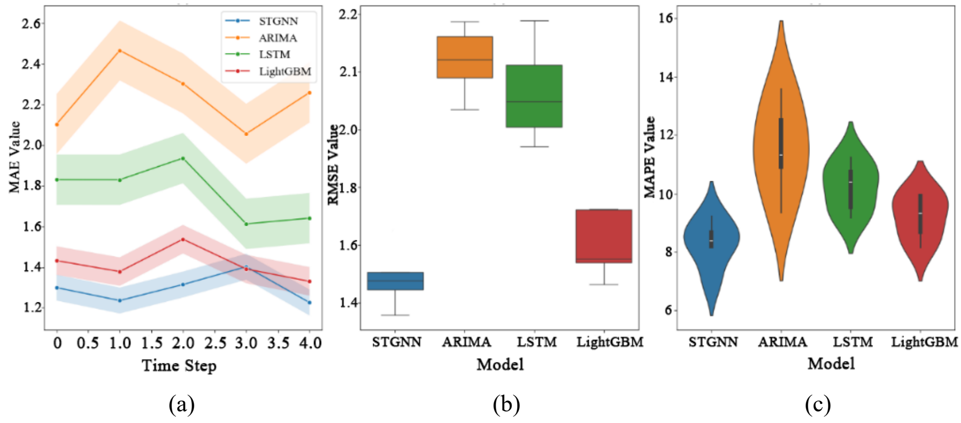
Figure 3 Multi-faceted evaluation of time series prediction performance (see online version for colours)



(a) (b)

(c) (d)

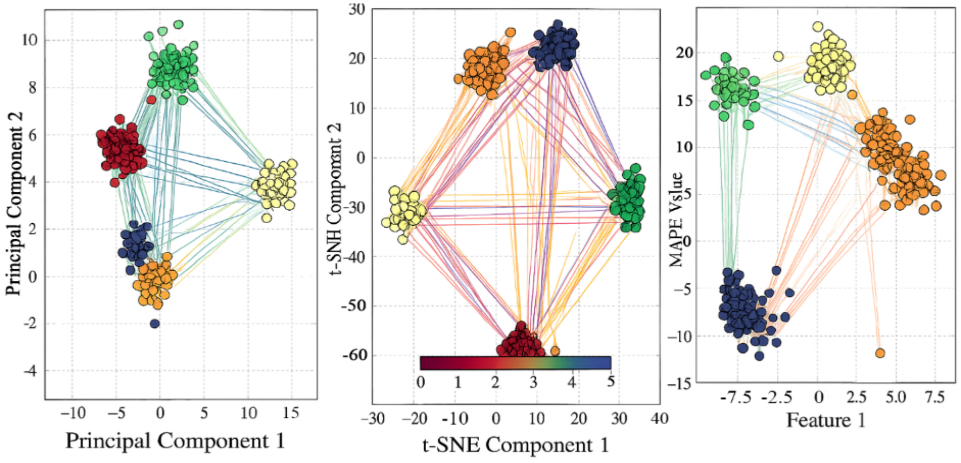
Figure 4 Performance comparison of four models in agricultural talent flow prediction (see online version for colours)



(a) (b) (c)

Figure 4 presents the performance of four models in forecasting agricultural talent flow, evaluated using MSE, RMSE, and MAPE. These metrics were chosen for their complementary strengths: MSE emphasises large errors, RMSE provides interpretable error magnitude, and MAPE expresses errors as percentages, allowing fair comparison across regions with varying migration scales. Together, they offer a balanced assessment of both absolute and relative prediction accuracy, making them well-suited for evaluating spatio-temporal models in heterogeneous agricultural contexts. The three subgraphs in the figure show the performance of each model in the task. Figure 4(a) shows the error change trend of the four models in different experiments. It can be seen that the STGNN model exhibits the most stable performance in all tests, with minimal error variation, while the ARIMA model displays significant error fluctuation, indicating its poor stability when dealing with complex spatio-temporal data. LSTM and LightGBM performed relatively moderately but remained below ARIMA. Figure 4(b) shows the error distribution of each model over multiple trials. The error distributions of STGNN and LightGBM are more concentrated, showing their high stability and consistency. In contrast, ARIMA and LSTM exhibit a wide range of errors. Figure 4(c) provides more detailed error distribution information. STGNN and LightGBM are more concentrated in areas with lower errors, while ARIMA and LSTM present a larger error range, indicating that these two models may have large deviations when predicting.

Figure 5 Comparison of data dimensionality reduction and clustering effects: PCA, t-SNE and OPTICS (see online version for colours)



(a) (b) (c)

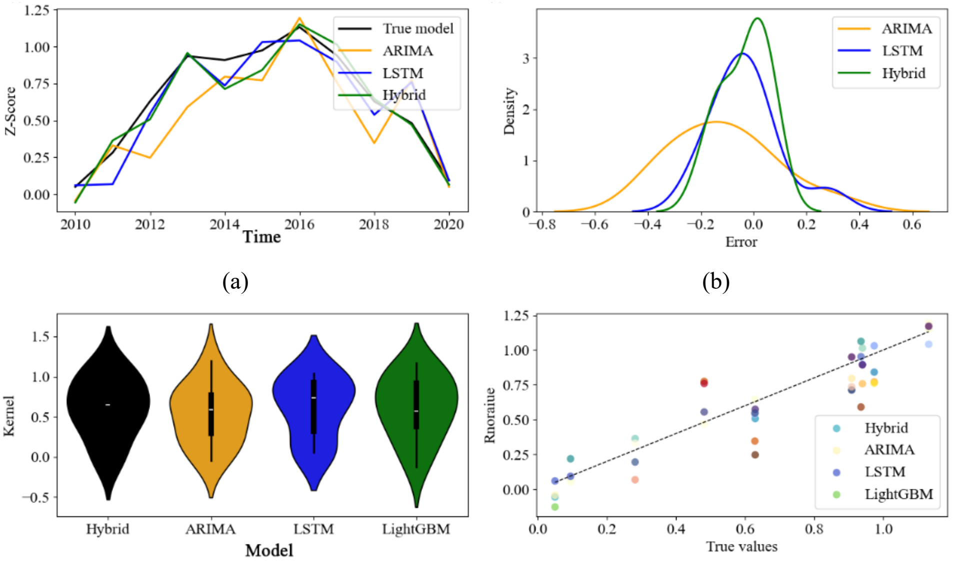
Figure 5 illustrates three commonly used dimensionality reduction and clustering methods: principal component analysis (PCA), t-distribution random neighbourhood embedding (t-SNE), and ordered point identification clustering structure (OPTICS). It performs a visual comparison through three subgraphs. Figure 5(a) displays the PCA dimensionality reduction results, where the projection of the data into a two-dimensional space is used to represent different clusters. The points of each cluster are coloured according to their original labels, and the structural association of data points is illustrated by the connecting lines between every 5 points. Figure 5(b) shows the distribution of data points after dimensionality reduction by t-SNE. Compared to PCA, t-SNE can better preserve the local structure of the data, and the distance between each cluster is visible, further strengthening the dense connections between points. The polyline element enhances the readability of the data and helps illustrate the hierarchical relationships between clusters. Finally, Figure 5(c) displays the data distribution after clustering using OPTICS, with the clustering results represented by colour coding. This method can dynamically identify the structure of clusters and show the changing trend of data points in feature space through the line connecting every 5 points.

Table 3 Performance comparison of CNN, LightGBM and fusion models in classification tasks

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*	*AUC*
CNN Only	0.88	0.87	0.89	0.88	0.93
LightGBM Only	0.85	0.84	0.86	0.85	0.90
CNN + LightGBM Fusion	0.91	0.90	0.92	0.91	0.95

Table 3 compares the key performance indicators of the three models on the test set, including five commonly used measures: accuracy, precision, recall, F1-score and area under the curve (AUC). With its deep convolutional network architecture, the CNN model exhibits outstanding performance in image feature extraction, achieving an accuracy rate of 0.88 and an AUC of 0.93. The LightGBM model, based on the gradient boosting decision tree algorithm, achieves an accuracy of 0.85 and an AUC of 0.90. To fully leverage the advantages of both, this paper incorporates the high-dimensional depth features extracted by CNN into LightGBM and employs a weighted fusion strategy to optimise the output of each model comprehensively. After verification, the fusion model has been significantly improved in all indicators-the accuracy rate reaches 0.91, the AUC reaches 0.95, and the F1-score and recall rate increase to 0.91 and 0.92, respectively, which fully proves that the fusion strategy can effectively improve the overall performance of the model.

Figure 6 Multi-dimensional performance assessment of regional talent flow prediction model (see online version for colours)



(a)

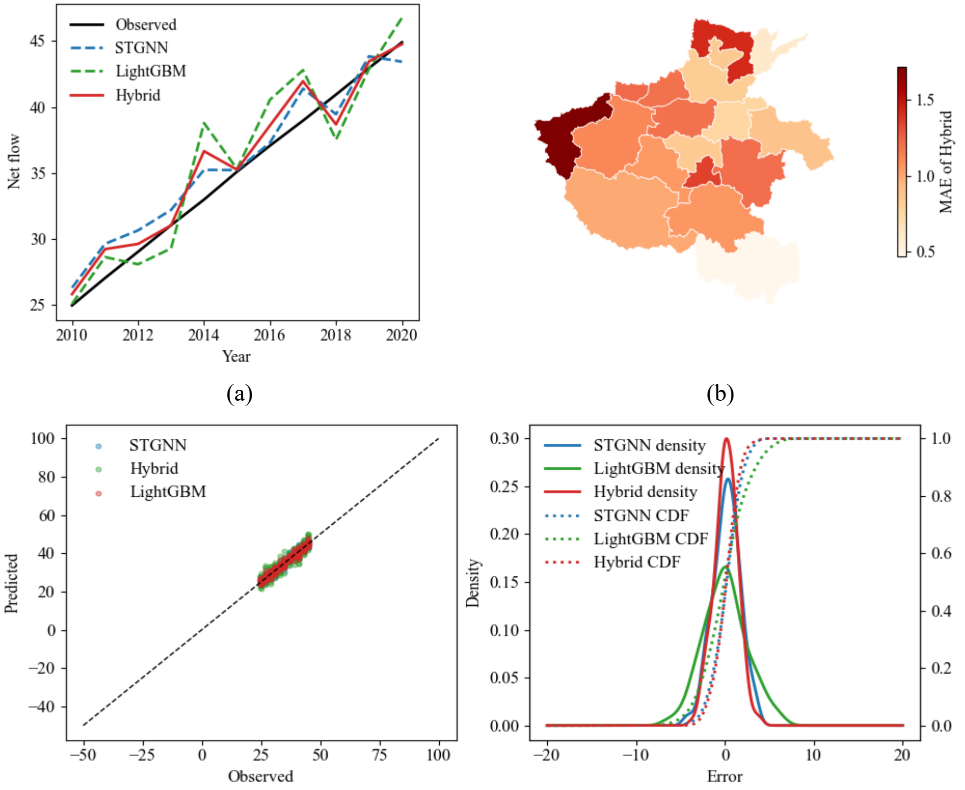
(b)

(c)

(d)

Figure 6 shows the performance comparison of the STGNN-LightGBM hybrid model and three baseline models (ARIMA, LSTM, LightGBM) in the agricultural talent flow prediction task in Henan Province from 2010 to 2020. Figure 6(a) presents the time series change of the real Z-score and the predicted value of each model, and the mixed model curve has the highest coincidence with the observed value; Figure 6(b) uses error kernel density estimation and cumulative distribution function, which shows that the error distribution of the mixed model is the most concentrated and the deviation is the smallest; Figure 6(c) visualises the distribution pattern and dispersion degree of the predicted values of the four models, and the mixed model has the narrowest distribution and the strongest stability; Figure 6(d) shows the real value and the predicted value correspondingly in the form of scatter points, and draws the ideal diagonal. The point cloud of the mixed model is closest to the reference line and has the lightest colour, which further verifies its superior performance. Taken together, the hybrid framework that combines the spatial sensitivity of a GNN and the feature interpretation ability of LightGBM is significantly better than a single model in terms of accuracy, stability and generalisation ability, and has good application value for regional spatio-temporal prediction.

Figure 7 Multi-model net flow prediction results and error analysis (see online version for colours)



(a)

(b)

(c)

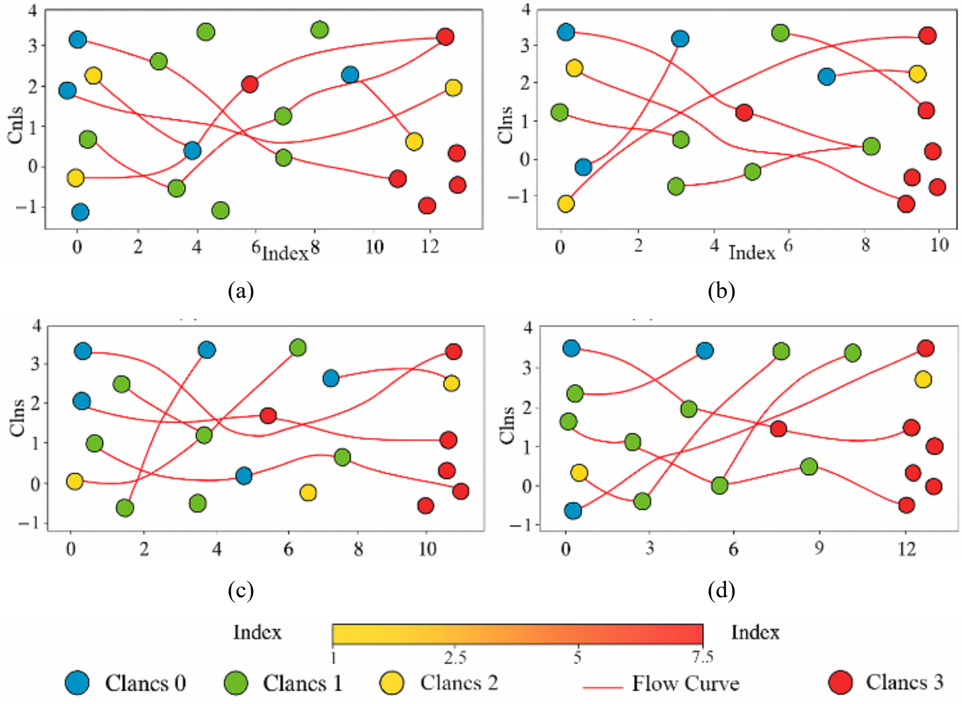
(d)

Figure 7 comprises four subgraphs that comprehensively compare the performance and error distribution of STGNN, LightGBM, and Hybrid models in net flow forecasts from 2010 to 2020. Figure 7(a) shows the measured net flow with a black solid line, and the blue dotted line, green dotted line and red solid line show the time series prediction trajectories of STGNN, LightGBM and Hybrid models, respectively, which intuitively reflect the ability of each model to capture the fluctuation trend and the deviation of inflection point timing. Figure 7(b) is the spatial distribution map of the average absolute error (MAE) of the LightGBM model in each region. The colour scale, ranging from light orange to dark red, is used to indicate the low error to the high error, revealing that the model's accuracy is higher in data-dense areas, such as provincial capitals and transportation hubs, while the error increases significantly in remote or data-scarce areas. Figure 7(c) draws the scatter plot of the predicted values and observed values of the three models. The convergence and dispersion of the three-colour point cloud around the dashed line (1 1 ideal fitting line) reflect the systematic deviation and fitting effect of each model. The LightGBM model has the most concentrated point cloud, followed by Hybrid, and STGNN has the widest dispersion. Figure 7(d) shows the probability density function (solid line) and cumulative distribution function (dashed line) of the three model errors, among which the error distribution curve of the Hybrid model is the steepest and left-skewed, indicating that most of its prediction errors are smaller and more stable; LightGBM performs in the middle, and STGNN error distribution is the most scattered.

The proposed hybrid model showed consistent performance across 17 cities in Henan Province, maintaining low error rates even in regions with sparse data. This indicates strong generalisation ability, as the model effectively captured migration patterns in both urban and rural areas. By leveraging spatial relationships through graph structures and temporal dynamics via rolling window prediction, STGNN learned complex flow behaviors. LightGBM complemented this by enhancing interpretability and adapting to structured features. Together, the model demonstrated robustness across heterogeneous environments, making it suitable for real-world forecasting where data availability varies.

Based on the simulation coordinates of 30 random nodes and four cluster labels, Figure 8 visually compares the predicted results of OD network traffic in the 4-hour and 8-hour periods with the actual data from the proposed model. The image is divided into four subgraphs: Figure 8(a) 4h predicted traffic; Figure 8(b) 4h true flow; Figure 8(c) 8h predicted flow; Figure 8(d) 8h true flow. All nodes are randomly distributed on the plane, and Gaussian jitter is added to each subgraph to break the regular pattern. The node size is dynamically mapped according to the total departure traffic, and the nodes in the cluster use uniform colour blocks, adding black borders to distinguish the spatial clustering relationship. The connection line retains only the trunk flow above the 90th percentile, and the line width is linearly scaled according to the flow intensity. The colour from light orange to dark red corresponds to the flow from low to high. The backbone trend of subgraphs Figures 8(a) and 8(c) reflects the capture ability of the model in the core area of clusters and the exudation flow between clusters; Figures 8(b) and 8(d) show the key channels of real traffic. Comparing the two sets of images can demonstrate the accuracy of the model's proposed prediction data.

Figure 8 Visualisation of OD network traffic simulation in Henan Province: 4h/8h forecast and real comparison, (a) 4h predicted flow (b) 4h ground truth (c) 8h predicted flow (b) 8h ground truth (see online version for colours)



(a)

(b)

(c)

(d)

To evaluate the feasibility of model deployment, we recorded the training time and resource consumption of the STGNN-LightGBM framework. On our workstation, the STGNN component required approximately 2.5 hours for training, while LightGBM completed training within 15 minutes. Peak memory usage remained below 20 GB, indicating that the model is suitable for efficient deployment in real-world regional planning scenarios.

5 Conclusions

The experimental results of the STGNN-LightGBM hybrid model on the 2019–2020 test set of Henan Province demonstrate clear performance advantages. Compared with the standalone STGNN model, the hybrid model reduced the mean absolute error (MAE) from 0.28 to 0.24, the root mean square error (RMSE) from 0.35 to 0.31, and the mean absolute percentage error (MAPE) from 11.9% to 10.2%. Relative to the single LightGBM model, improvements of approximately 12.5%, 11.4%, and 10.8% were observed in MAE, RMSE, and MAPE, respectively. At the city level, the average prediction error in Zhengzhou and other transportation hubs was below 8.5%, while errors in remote areas remained within 12.3%, indicating strong regional generalisation.

This hybrid framework effectively combines the spatial sensitivity of STGNN with the interpretability and efficiency of LightGBM, achieving enhanced accuracy and robustness in spatio-temporal forecasting. Beyond Henan Province, the model's architecture is adaptable to other provinces and sectors where migration, mobility, or resource allocation patterns exhibit spatial-temporal complexity. Its modular design allows for integration with diverse datasets, making it suitable for applications in urban planning, healthcare workforce distribution, and industrial labour forecasting. This study is the first to integrate STGNN and LightGBM for agricultural talent migration prediction, offering a transferable and scalable solution for broader policy-oriented forecasting tasks.

Declarations

All data generated or analysed during the study are available from the corresponding author by request.

The authors declare no conflict of interest.

References

Al-Selwi, S.M. et al. (2024) 'RNN-LSTM: from applications to modeling techniques and beyond-systematic review', *Journal of King Saud University-Computer and Information Sciences*, Vol. 36, No. 5, p.102068.

Assis, M.V.O. et al. (2021) 'A GRU deep learning system against attacks in software defined networks', *Journal of Network and Computer Applications*, Vol. 177, No. 2021, p.102942.

Bashir, F. and Wei, H.-L. (2018) 'Handling missing data in multivariate time series using a vector autoregressive model-implementation (VAR-IM) algorithm', *Neurocomputing*, Vol. 276, pp.23–30.

Beg, R., Pateriya, R.K. and Tomar, D.S. (2024) 'Design of an iterative method for malware detection using autoencoders and hybrid machine learning models', *IEEE Access*, Vol. 12, pp.175032–175055.

Chen, P. et al. (2025a) 'New-type professional farmers: how to make use of different types of social capital to engage in agriculture specialization', *Journal of Rural Studies*, Vol. 114, No. 2025, p.103545.

Chen, X. et al. (2025b) 'Intelligent network-level energy saving strategy with STGNN-driven traffic prediction and path optimization in transport networks and field trial', *IEEE Access*, Vol. 13, pp.116118–116129.

Geng, W. and Yang, G. (2017) 'Partial correlation between spatial and temporal regularities of human mobility', *Scientific Reports*, Vol. 7, No. 1, p.6249.

Greff, K. et al. (2016) 'LSTM: a search space odyssey', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 28, No. 10, pp.2222–2232.

Han, Z. et al. (2019) 'A review of deep learning models for time series prediction', *IEEE Sensors Journal*, Vol. 21, No. 6, pp.7833–7848.

Hensher, D.A. and W.H. (2003) 'The mixed logit model: the state of practice', *Transportation*, Vol. 30, No. 2, pp.133–176.

Ju, Y. et al. (2019) 'A model combining convolutionary neural network and LightGBM algorithm for ultra-short-term wind power forecasting', *IEEE Access*, Vol. 7, pp.28309–28318.

Li, H. et al. (2022) 'OOD-GNN: out-of-distribution generalized graph neural network', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 35, No. 7, pp.7328–7340.

Li, L. et al. (2018) 'Towards effective network intrusion detection: a hybrid model integrating gini index and GBDT with PSO', *Journal of Sensors*, Vol. 2018, No. 1, p.1578314.

Limouni, T. et al. (2023) 'Accurate one step and multistep forecasting of very short-term PV power using LSTM-TCN model', *Renewable Energy*, Vol. 205, pp.1010–1024.

Mao, W. et al. (2024) 'STGNN-LMR: a spatial-temporal graph neural network approach based on sEMG lower limb motion recognition', *Journal of Bionic Engineering*, Vol. 21, No. 1, pp.256–269.

Mardani, A. et al. (2017) 'Application of structural equation modeling (SEM) to solve environmental sustainability problems: a comprehensive review and meta-analysis', *Sustainability*, Vol. 9, No. 10, p.1814.

Millo, G. and Piras, G. (2012) 'SPLM: spatial panel data models in R', *Journal of Statistical Software*, Vol. 47, No. 1, pp.1–38.

Mim, T.R. et al. (2023) 'GRU-INC: an inception-attention-based approach using GRU for human activity recognition', *Expert Systems with Applications*, Vol. 216, p.119419.

Parzen, E. (2003) 'Some recent advances in time series modeling', *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, pp.723–730.

Qu, H. et al. (2022) 'Spatio-temporal evolution of the agricultural eco-efficiency network and its multidimensional proximity analysis in China', *Chinese Geographical Science*, Vol. 32, No. 4, pp.724–744.

Rabbani, M.B.A. et al. (2021) 'A comparison between seasonal autoregressive integrated moving average (SARIMA) and exponential smoothing (ES) based on time series model for forecasting road accidents', *Arabian Journal for Science and Engineering*, Vol. 46, No. 11, pp.11113–11138.

Simaiya, S. et al. (2024) 'A transfer learning-based hybrid model with LightGBM for smart grid short-term energy load prediction', *Energy Exploration & Exploitation*, Vol. 42, No. 5, pp.1853–1876.

Valentini, M., dos Santos, G.B. and Vieira, B.M. (2021) 'Multiple linear regression analysis (MLR) applied for modeling a new WQI equation for monitoring the water quality of Mirim Lagoon, in the state of Rio Grande do Sul-Brazil', *SN Applied Sciences*, Vol. 3, No. 1, p.70.

Wang, D-n., Li, L. and Zhao, D. (2022) 'Corporate finance risk prediction based on LightGBM', *Information Sciences*, Vol. 602, pp.259–268.

Wójcik, M., Jeziorska-Biel, P. and Czapiewski, K. (2019) 'Between words: a generational discussion about farming knowledge sources', *Journal of Rural Studies*, Vol. 67, pp.130–141.

Yang, J. et al. (2024) 'Geographical big data and data mining: a new opportunity for water-energy-food nexus analysis', *Journal of Geographical Sciences*, Vol. 34, No. 2, pp.203–228.

Yin, J. et al. (2025) 'STGNN: a novel spatial-temporal graph neural network for predicting complicated business process performance under multi-event parallelism', *Expert Systems with Applications*, Vol. 291, No. 2025, pp.128391–128391.

Zhang, S. et al. (2025) 'Hybrid deep learning for gas price prediction using multi-factors and temporal features', *IEEE Access*, Vol. 13, pp.11989–12001.

Zhao, L. et al. (2019) 'T-GCN: a temporal graph convolutionary network for traffic prediction', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 21, No. 9, pp.3848–3858.