ICCK Transactions on Advanced Computing and Systems
http://dx.doi.org/10.62762/TACS.2025.619794

RESEARCH ARTICLE

An Efficient Algorithm for Weather Forecasting Using Causal Graph Neural Network

Ahmad Ali[1,2,*], **H. M. Yasir Naeem**[1,2], **Riaz Ali**[3], **Mujtaba Asad**[4], **Md Belal Bin Heyat**[5] **and Tamam Alsarhan**[6]

[1] College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen 518060, China
[2] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
[3] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[4] Institute of Image Processing & Pattern Recognition, Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China
[5] Department of Computer Science, Westlake University, Hangzhou 310024, China
[6] King Abdullah II School of Information Technology, The University of Jordan, Amman, Jordan

Abstract

The rapid accumulation of large-scale, long-term meteorological data presents unprecedented opportunities for data-driven weather modeling and high-resolution numerical weather prediction. While various deep learning techniques—such as Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNNs), and Graph Neural Networks (GNNs)—have been explored for weather forecasting, the complex spatial dependencies within historical meteorological data, particularly dynamic spatial correlations, remain insufficiently addressed. To tackle this challenge, we propose a Dynamic Spatio-Temporal Fusion Graph Network (DSTFGN), a novel module that integrates multivariate time-series analysis with graph-based causal inference to capture intricate and time-varying interdependencies among weather variables. The DSTFGN module fuses real-time inputs (e.g., sensor data, live weather feeds, external events) with historical records to model the propagation of disruptions—such as accidents or road closures—through the meteorological network. By effectively capturing dynamic spatial-temporal interactions, our approach significantly enhances forecasting accuracy and supports adaptive weather management strategies. Experimental evaluations on two real-world datasets demonstrate that DSTFGN consistently outperforms existing baseline models across short, medium, and long-term forecasting horizons.

Keywords: intelligent transportation systems, weather forecasting, causal graph learning, spatio-temporal attention, GCN.

1 Introduction

The advancement of Intelligent Transportation Systems (ITS) is closely linked to the accuracy of

Submitted: 26 March 2025
Accepted: 09 July 2025
Published: 28 October 2025

Vol. 1, **No.** 4, 2025.
10.62762/TACS.2025.619794

***Corresponding author:**
Ahmad Ali
ahmadali@szu.edu.cn

Citation
Ali, A., Naeem, H. M. Y., Ali, R., Asad, M., Heyat, M. B. B., & Alsarhan, T. (2025). An Efficient Algorithm for Weather Forecasting Using Causal Graph Neural Network. *ICCK Transactions on Advanced Computing and Systems*, 1(4), 258–274.

urban weather forecasting, which has emerged as a critical area of research. Rapid urbanization and the growing number of vehicles have intensified urban congestion, adversely impacting the daily lives and productivity of city residents [1, 2]. Accurate weather forecasting—particularly in urban settings is essential, as it influences traffic flow, travel safety, and infrastructure planning. However, achieving reliable predictions remains a significant challenge due to the inherent spatio-temporal complexity of urban transportation networks and the dynamic interactions between weather conditions and traffic behavior.

Researchers have increasingly focused on time-series and spatial correlation analysis techniques to tackle the challenges of weather forecasting, particularly in capturing the spatio-temporal dependencies within urban road networks. Early approaches primarily relied on statistical time-series models to extract temporal patterns, such as the Historical Averaging (HA) model [3] and the Autoregressive Integrated Moving Average (ARIMA) model [4]. Subsequently, Convolutional Neural Networks (CNNs) were introduced to model spatial correlations in grid-based weather and traffic networks [5, 6]. However, these CNN-based methods are limited to Euclidean space representations and struggle to accurately model the irregular and non-Euclidean structure of transportation networks. To better represent the complex topology of road networks, Graph Neural Networks (GNNs) have gained traction, as they provide a more flexible and powerful framework for capturing intricate spatial relationships in graph-structured data [7–10]. GNNs enable more accurate modeling of interactions between road segments, making them highly suitable for weather forecasting in urban transportation systems.

Although Graph Neural Network (GNN)-based approaches have shown promising predictive performance in weather forecasting, several key challenges remain. One major challenge is dynamic spatial dependence: urban road networks are subject to sudden and inherent changes due to factors such as turning restrictions, points of interest, accidents, and road maintenance. These changes can alter the spatial topology of the network and subsequently affect weather-related conditions on adjacent roads. For example, Figure 1 illustrates the propagation of weather phenomena (e.g., heavy rain, fog) through an urban road network and their impact on traffic flow dynamics, depicting a directed graph where nodes represent intersections or sensor locations,

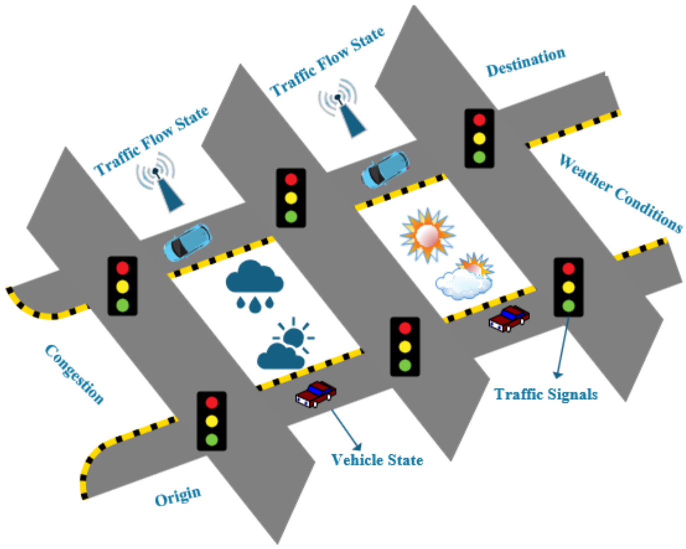

Figure 1. Weather propagation and impact on traffic flow.

and edges represent road segments, with arrows indicating traffic flow direction. Adverse weather conditions, such as heavy rain at Node A and fog at Node C, propagate across the network, leading to congestion and traffic slowdowns. Additionally, road constraints—such as one-way streets and no-turn zones—further influence these dynamics by redirecting traffic and amplifying congestion. Effectively modeling this complex causality requires an advanced framework capable of capturing dynamic spatio-temporal interactions, highlighting the need for a more adaptive and context-aware forecasting approach.

The need to more accurately capture the intricate and dynamic nature of weather within urban transportation networks has motivated the development of this approach. Traditional forecasting methods often fall short in modeling the complex interplay between spatial and temporal factors that influence urban weather patterns. Urban road networks are highly susceptible to abrupt and unpredictable events—such as accidents, road closures, and evolving weather regulations—that introduce non-trivial spatio-temporal dependencies. To address these challenges, we propose an integrated framework that combines gated dilated convolutions, spatio-temporal attention mechanisms, and a fusion graph learning module. The spatio-temporal attention mechanism selectively emphasizes the most critical temporal and spatial variations, while the fusion graph learning module captures fine-grained spatial interactions among road nodes. Together, these components enable the model to represent and learn complex, dynamic relationships across the network.

This approach provides a more robust and accurate framework for urban weather forecasting, which is essential for improving real-time decision-making in smart cities, enhancing weather-responsive traffic management, and promoting more efficient urban mobility. The major contributions of our work are as follows:

- We propose DSTFGN, a novel model that combines multivariate time-series analysis with graph-based causal inference to improve weather forecasting. Unlike traditional methods, DSTFGN captures both dynamic spatial correlations and temporal causality using adaptive graph learning, Granger causality, and spatio-temporal attention. This allows the model to adapt to real-time changes and accurately predict future weather conditions.
- Existing methods like RNNs and LSTMs focus on temporal correlations but lack causal understanding. We propose a causal graph learning framework using Granger causality to capture directionality and latency in weather propagation. This enhances DSTFGN accuracy and interpretability over traditional correlation-based models.
- We evaluate DSTFGN on two real-world weather datasets, where it outperforms ten state-of-the-art baselines across MAE, MAPE, and RMSE. The model excels at capturing long-term temporal dependencies and dynamic spatial correlations. It also maintains low computational cost, making it ideal for real-time forecasting.

The rest of the paper is organized as follows: In Section 2, we outline and model the problem statement. The GCN model and the proposed DSTFGN method are discussed in more depth in Section 3. Section 4 contains a discussion of extensive experiments and comparisons. A detailed literature is provided in Section 6. Finally, Section 7 concludes this work.

2 Problem Formulation

In this section, we formally present the weather forecasting problem and provide a mathematical explanation of the weather network concept.

- **Dynamic Spatial Dependencies**: Urban weather networks are influenced by a variety of dynamic factors, such as road closures, accidents, and sudden changes in weather conditions. These factors can alter the spatial structure of the network, leading to complex and time-varying interactions between nodes. Traditional GNNs often rely on static adjacency matrices, which fail to capture these dynamic spatial dependencies, resulting in suboptimal predictions.
- **Temporal Causality**: Weather patterns exhibit strong temporal causality, where changes in one node can propagate to downstream nodes with a time delay. Existing methods, such as RNNs and Long Short-Term Memory (LSTM) networks, primarily focus on capturing temporal correlations rather than causal relationships. This can lead to incorrect dependencies, as correlations do not necessarily imply causation. For example, a sudden drop in temperature at one location may cause a delayed response in humidity levels at a downstream location, which is not adequately captured by correlation-based models.
- **Heterogeneity in Spatio-Temporal Data**: Weather data is inherently heterogeneous, with different nodes exhibiting varying patterns of temporal and spatial influence. Some nodes may be more influenced by their historical data, while others may be more affected by interactions with neighboring nodes. Existing methods often treat all nodes uniformly, failing to account for this heterogeneity, which can lead to inaccurate forecasts.

We represent the urban weather network as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ denotes the set of sensor nodes or road intersection nodes distributed across the weather network, with N representing the total number of nodes. $\mathcal{E}$ represents the set of edges connecting these nodes. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ encodes the relationships or proximities between each pair of nodes.

Historical weather signal data can be represented as a weather map feature vector $\mathcal{Y} \in \mathbb{R}^{T\times N\times D}$, where D denotes the feature dimension (e.g., temperature, humidity, pressure). The historical graph signal at time t is expressed as $\mathcal{Y}_t \in \mathbb{R}^{N\times D}$.

The weather forecasting problem is framed as follows: based on the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ and the historical weather graph signals $\mathcal{Y}_{(t-T):t}$ over T time steps, the goal is to predict weather signals for the next P time steps by constructing a model $f(\cdot)$. Specifically, we have:

$$[\mathcal{Y}_{(t-T):t}, \mathcal{G}]\, f(\cdot) \rightarrow \hat{\mathcal{Z}}_{t,(t+P)}, \tag{1}$$

where $\mathcal{Y}_{(t-T):t} = (\mathcal{Y}_{t-T+1}, \mathcal{Y}_{t-T+2}, \ldots, \mathcal{Y}_t) \in \mathbb{R}^{T\times N\times D}$ and $\hat{\mathcal{Z}}_{t,(t+P)} = (\hat{\mathcal{Z}}_{t+1}, \hat{\mathcal{Z}}_{t+2}, \ldots, \hat{\mathcal{Z}}_{t+P}) \in \mathbb{R}^{P\times N\times D}$.

We define the weather forecasting task as follows: given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ and historical weather signals $\mathcal{Y}_{(t-\Delta T):t}$ over ΔT time steps, predict future weather signals for the next ΔP time steps by learning the function $g(\cdot)$. This is mathematically represented as:

$$[\mathcal{Y}_{(t-\Delta T):t}, \mathcal{G}]\, g(\cdot) \rightarrow \hat{\mathcal{Z}}_{t,(t+\Delta P)}, \tag{2}$$

where $\mathcal{Y}_{(t-\Delta T):t} = (\mathcal{Y}_{t-\Delta T+1}, \mathcal{Y}_{t-\Delta T+2}, \ldots, \mathcal{Y}_t) \in \mathbb{R}^{\Delta T\times N\times D}$ and $\hat{\mathcal{Z}}_{t,(t+\Delta P)} = (\hat{\mathcal{Z}}_{t+1}, \hat{\mathcal{Z}}_{t+2}, \ldots, \hat{\mathcal{Z}}_{t+\Delta P}) \in \mathbb{R}^{\Delta P\times N\times D}$.

3 Model Framework

In this section, we provide a detailed description of the spatio-temporal causal fusion graph neural network, as illustrated in Figure 2.

3.1 Modeling Temporal Correlation Module

RNN-based methods have generally been used to capture temporal correlations in earlier weather forecasting research. However, these techniques often face challenges such as gradient explosion when modeling long-term temporal dependencies and significant computational costs due to repetitive processing. In contrast, CNN-based approaches offer advantages such as gradient stability, parallel processing, and simpler model architectures. This research adopts an extended temporal convolution technique, to address these challenges. By increasing the depth of convolutional layers and gradually expanding the receptive field, this approach enhances the model's ability to capture long-range temporal correlations, which is crucial for accurate weather forecasting. This strategy reduces processing costs by enabling the temporal convolution to identify long-term dependencies in weather sequences with fewer layers. Moreover, non-recursive methods allow for parallel processing, further decreasing operational costs and temporal complexity. To improve the handling of sequential data, we also incorporate a gating mechanism to regulate the flow of information into the spatio-temporal convolution. The temporal gated convolutional network (TGCN) is defined as follows:

$$Z_l^T = \tanh(W_{a,l} * Z_{l-1}^{out} + b_1) \odot \sigma(W_{b,l} * Z_{l-1}^{out} + b_2) \tag{3}$$

$$Z_l^{out} = \mathrm{GRU}(Z_l^T, Z_{l-1}^{out}) \tag{4}$$

$$Y = \sum_{l=1}^{L} Z_l^T \tag{5}$$

where $Z_l^T \in \mathbb{R}^{T\times N\times D}$ denotes the output of the l-th layer of the Gated Temporal Convolutional Network (GatedTCN), where T is the number of time steps, N is the number of nodes, and D is the feature dimension. $W_{a,l}$ and $W_{b,l}$ are the learnable parameters at layer l, while b_1 and b_2 are the bias terms. The symbols a and b represent filters and gates, respectively. The symbol $\odot$ denotes the Hadamard product, and $*$ represents the convolution operator. Finally, $\sigma(\cdot)$ is the sigmoid activation function.

3.2 Spatio-Temporal Fusion Graph Learning

The integrated Spatio-Temporal Fusion Graph Learning (STFGL) is designed to capture the complex spatial correlations and causal dependencies between nodes. The STFGL consists of three sub-modules: the causal graph learning, the adaptive graph learning, and the spatial gated fusion module. The outputs from the adaptive graph learning module and causal graph learning at the l-th layer are represented as Z_l^{AG} and Z_l^{CG}, respectively. Following the spatial gating fusion process, the final output of the l-th layer from the STFGL is denoted as Z_l^{SG}.

In a road network, weather conditions at each node are influenced by weather information from neighboring nodes, leading to causal interactions between weather patterns across different locations. Specifically, changes in weather at upstream nodes can affect conditions downstream, potentially leading to congestion. To accurately capture and quantify these causal relationships in weather forecasting, we propose a novel causal graph learning framework based on **Granger Causality Analysis (GCA)**. In statistics, the Granger causality test [11] is used to determine whether one time series can predict another, revealing causal connections. In this study, we apply Granger causality analysis to identify and uncover the causal structure between weather conditions at different road nodes within transportation time series data, improving the accuracy of weather forecasts in urban environments.

Specifically, we construct two regression models: the partial model $\mathbf{Z}_p$ and the full model $\mathbf{Z}_f$, which predict the time series values $\mathbf{y}_i$ and $\mathbf{y}_j$, respectively. The

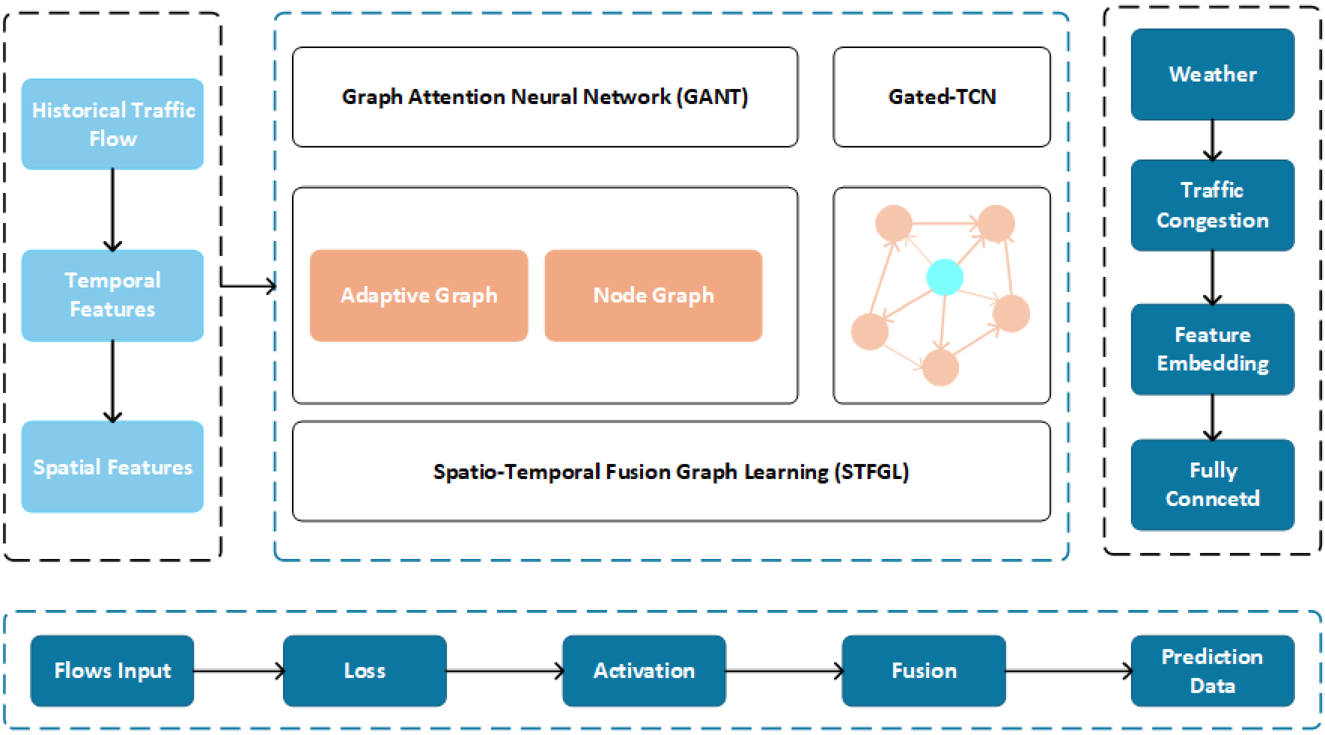

Figure 2. Framework of the proposed DSTFGN model.

primary difference between the two models is that the full model includes the historical data from both time series $\mathbf{y}_i$ and $\mathbf{y}_j$, whereas the partial model only incorporates the historical data of $\mathbf{y}_j$ to predict $\mathbf{y}_j(t)$ $(t = 1, 2, \ldots, T)$. The regression models are defined as follows:

$$\mathbf{Z}_f = \alpha_0 + \sum_{i=1}^{T} \alpha_i \cdot \mathbf{y}_i(t) + \epsilon_f(t) \tag{6}$$

$$\mathbf{Z}_p = \gamma_0 + \sum_{j=1}^{T} \gamma_j \cdot \mathbf{y}_j(t) + \epsilon_p(t) \tag{7}$$

where Z_f represent the output of the full regression model, Z_p denotes the partial regression model, y_i, y_j represent the weather data at nodes i and j, while T denote total number of time steps.

Next, by comparing the prediction errors of the two models ($\mathcal{E}_p$, $\mathcal{E}_f$), we evaluate the impact of the historical data of variable $\mathbf{a}_i$ on the prediction of variable $\mathbf{a}_j$. According to the Granger causality hypothesis, if the lagged values of variable $\mathbf{a}_i$ help predict the future values of variable $\mathbf{a}_j$, then $\mathbf{a}_i$ is considered the Granger cause of $\mathbf{a}_j$. We employ the F-distribution to test the statistical significance of causal relationships between pairs of nodes as follows:

$$\mathcal{F} = \frac{(\mathcal{Q}_p - \mathcal{Q}_f)}{(\nu_f - \nu_p)} \cdot \frac{\mathcal{Q}_f}{(\mathcal{T} - \nu_f - 1)} \tag{8}$$

where the variables Q_p and Q_f represent the degrees of freedom for the regression parameters in the two models, which define the number of parameters in each model.

To reduce the time consumption of the Granger causality test while preserving the temporal characteristics of the raw weather data, this study employs K-means clustering for preprocessing. We use the raw weather data, after clustering, to perform the Granger causality test and construct the causal adjacency matrix A_G, which is represented as follows:

$$A_G = \begin{cases} 1, & \text{if } p < \text{sig} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

where p denotes the p-value obtained from the Granger causality test, which is used to determine whether the null hypothesis in the hypothesis test can be rejected. Additionally, sig represents the significance threshold.

In this study, we employ the causal adjacency matrix A_G as a prior graph structure to guide the Graph Attention Neural Network (GANT) in effectively utilizing causal insights and identifying critical relationships between nodes. Additionally, we incorporate a multi-head attention mechanism to capture dependencies between nodes from different subspaces. This approach enhances the model expressive power by enabling parallel computation, which significantly reduces temporal complexity. The process is outlined as follows:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T[\mathbf{h}_i, \mathbf{h}_j]\right)\right)}{\sum_{k\in N_i}\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T[\mathbf{h}_i, \mathbf{h}_k]\right)\right)}, \tag{10}$$

$$\tilde{\mathbf{h}}_i = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N_i}\alpha_{ij}^k \mathbf{h}_j^k\right), \tag{11}$$

where $\mathbf{h}_i$ and $\mathbf{h}_j$ represent the feature vectors of nodes i and j, $\mathbf{a}$ is the learnable attention vector, N_i denotes the neighbors of node i, LeakyReLU$(\cdot)$ is the Leaky ReLU activation function, K is the number of attention heads, α_{ij}^k is the attention coefficient from the k-th attention head, $\mathbf{h}_j^k$ is the feature vector of node j from the k-th attention head, $\sigma(\cdot)$ is a non-linear activation function (such as ReLU or sigmoid).

The weather states of different road nodes exhibit significant dynamic interconnections in the spatial dimension, which are crucial for understanding the evolution of weather and making accurate predictions. However, traditional graph neural network methods typically rely on a predefined adjacency matrix based on distance computation. While this matrix can represent the spatial relationships between nodes, it often overlooks the complex and dynamically evolving spatial dependencies present in the node attributes. We propose a node-adaptive learning mechanism designed to thoroughly investigate and understand potential associations among road nodes at a fine-grained level. Specifically, instead of relying on a fixed graph structure, this mechanism dynamically adjusts and optimizes the connection weights between nodes based on their individual weather data (e.g., flow, speed, weather, density, etc.) and their interactions. Through this process, we generate an adaptive adjacency matrix $\tilde{A}_{\text{adm}}$, which more accurately reflects the real-time and dynamic spatial dependencies between nodes.

$$A_{\text{adm}} = \text{softmax}\left(\text{ReLU}\left(\mathbf{Z}_1 \cdot \mathbf{Z}_2^T\right)\right) \tag{12}$$

where $\mathbf{Z}_1 \in \mathbb{R}^{N\times d}$ represents the embedding matrix for the source nodes, while $\mathbf{Z}_2$ corresponds to the embedding matrix for the target nodes. The spatial dependency weights between these nodes are obtained by computing the product of $\mathbf{Z}_1$ and $\mathbf{Z}_2$. The activation function used is ReLU, and softmax is applied for normalization.

We employ the GCN method to perform feature aggregation on adaptive graphs. According to the literature [31], the graph convolution process can be accurately approximated in the spectral domain using the expansion of the first-degree Chebyshev polynomial. This approximation can also be applied to GCNs in high-dimensional settings. Therefore, our graph convolutional network can be represented as follows:

$$\mathbf{H}^{l+1} = \left(\mathbf{I}_N + \tilde{\mathbf{A}}_{\text{adm}}\right)\mathbf{H}^l\Theta \tag{13}$$

where H^{l+1} represent feature matrix, $I - N$ denotes the identity matrix of size N, while A_{adm} represent the adjacency matrix.

To select important spatial features more effectively and enhance the model expressiveness, we employ a spatially gated fusion mechanism to integrate two input values, $\mathbf{H}_{\text{AG}}^l$ and $\mathbf{H}_{\text{CG}}^l$, which are the outputs of the AGL and CGL modules in the l-th layer, respectively. The spatially gated fusion mechanism is represented as follows:

$$\mathbf{H}_{\text{S}}^l = \sigma\left(\mathbf{G}^l\right)\odot\mathbf{H}_{\text{AG}}^l + \left(1-\sigma\left(\mathbf{G}^l\right)\right)\odot\mathbf{H}_{\text{CG}}^l \tag{14}$$

$$\begin{aligned}\mathbf{H}_{\text{S}}^{l+1} = \text{LayerNorm}\Bigg(\sum_{k=1}^{K}\mathbf{W}_k\Big(&\mathbf{H}_{\text{AG}}^l\odot\alpha_k \\ &+(1-\alpha_k)\odot\mathbf{H}_{\text{CG}}^l\Big)\Bigg) + \mathbf{B}_{\text{S}}\end{aligned} \tag{15}$$

where H_{S}^{l+1} represent the updated feature matrix at next layer, H_{AG}^l denotes output feature matrix from the adaptive graph learning, K represent the number of attention heads, while H_{CG}^l represent the output feature matrix from the convolutional graph learning.

3.3 Attention-Based Spatio-Temporal Module

In weather forecasting, we observe significant variations in how different road nodes are affected by weather flows across both the temporal and spatial domains. Specifically, some nodes may be more influenced by their own historical weather patterns than by direct interactions with neighboring nodes. To more accurately capture the unique temporal and spatial change patterns at each node, we designed an Attention-Based Spatio-Temporal (ABST) module. As shown in Figure 3, the overall structure of the ABST fully considers the individualized characteristics of each node, enabling the model to capture spatio-temporal adaptive trends with high precision at the node level. This allows for a more accurate forecasting of weather conditions in urban transportation networks. Additionally, we employ the same graph node embedding method in both HGLM and ABST to ensure consistency in modeling the weather-related interactions across nodes.

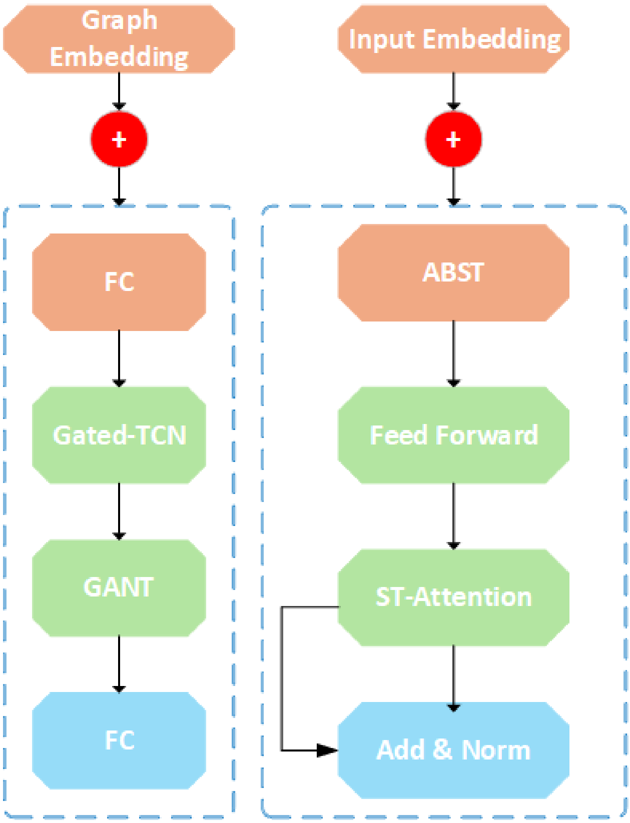

Figure 3. Spatio-temporal attention module.

To obtain the graph node query vector $\mathbf{Y} = \mathbf{W}_Y \mathbf{Z}_G$, we first apply a linear transformation to the graph node embeddings through full connectivity. The weight matrix $\mathbf{W}_Y \in \mathbb{R}^{d \times D}$, and the query vector $\mathbf{Q} \in \mathbb{R}^{N \times D}$. Next, the output of the gated temporal convolution network (TCN) $\mathbf{H}_T^l$ is transformed using the formula $\mathbf{K}_T^l = \mathbf{H}_T^l \mathbf{W}_{k_t}$, where $\mathbf{W}_{k_t} \in \mathbb{R}^{D \times D}$ is the weight matrix that projects the output $\mathbf{H}_T^l$ from the gated TCN into a D-dimensional space, obtaining the node's key in the time dimension. Similarly, the spatial dimension key values $\mathbf{W}_{k_s} \in \mathbb{R}^{D \times D}$ are obtained by performing a linear transformation on $\mathbf{K}_S^l = \mathbf{W}_{k_s} \mathbf{H}_S^l$, where $\mathbf{H}_S^l$ is the output of the HGLM. We then compute their attention scores as follows:

$$\mathbf{X}_{\mathrm{T}}^l = \frac{\exp(\mathbf{X}_{\mathrm{T}}^l)}{\sum_{r \in \{\mathrm{T},\mathrm{S}\}} \exp(\mathbf{X}_r^l)} \tag{16}$$

$$\mathbf{X}_{\mathrm{S}}^l = \frac{\exp(\mathbf{X}_{\mathrm{S}}^l)}{\sum_{r \in \{\mathrm{T},\mathrm{S}\}} \exp(\mathbf{X}_r^l)} \tag{17}$$

$$\mathbf{H}^l = \mathbf{A}_{\mathrm{S}}^l \cdot \mathbf{H}_{\mathrm{S}}^l + \mathbf{A}_{\mathrm{T}}^l \cdot \mathbf{H}_{\mathrm{T}}^l \tag{18}$$

where $\mathbf{X}_{\mathrm{T}}^l \in \mathbb{R}^{T \times N \times 1}$ and $\mathbf{X}_{\mathrm{S}}^l \in \mathbb{R}^{T \times N \times 1}$ represent the attention scores of nodes in the spatio-temporal dimensions, respectively. Both the output of the l-th layer model and the output of ABST are denoted as $\mathbf{H}^l \in \mathbb{R}^{T \times N \times d}$, where d is the dimensionality of the feature space.

3.4 Output Layer

To directly connect the output of each layer module to the output layer, we employ skip connections. The stacking of model layers expands the temporal receptive field of the DSTFGN. Higher-layer blocks capture long-term temporal weather information, while lower-layer blocks focus on temporally adjacent aspects. Skip connections are utilized to manage spatial dependencies at various time scales. The outputs of each layer in our model, denoted as $\mathbf{H}_{\mathrm{out}} \in \mathbb{R}^{T \times N \times D}$, are combined through summation after applying the skip connections. Additionally, we apply dimension-specific linear transformations to the output sequence using two fully connected layers as the output layers. The output layer is structured as follows:

$$\mathbf{Z} = \mathrm{ReLU}(\mathbf{X}_{\mathrm{out}} W_a + b_a) \cdot W_b + b_b \tag{19}$$

where the variables $W_1 \in \mathbb{R}^{(T \times D) \times C}$ and $W_2 \in \mathbb{R}^{C \times (M \times F)}$ represent the weight matrices. b_1 and b_2 are the bias terms. The output of the entire model is $\mathbf{Z}$.

We use the mean absolute error (MAE) as the training objective to minimize the prediction loss. This metric evaluates the difference between predicted and actual values, and is optimized via back-propagation using the following equation:

$$\mathrm{loss} = \frac{1}{N} \sum_{i=1}^{N} \left| Y_i - \hat{Y}_i \right| \tag{20}$$

where the term Y_i represents the ground truth, while $\hat{Y}_i$ denotes the value predicted by our model.

3.5 Training Process of the DSTFGN model

The training procedure for DSTFGN is described in Algorithm 1. Back-propagation is used to randomly initialize and optimize the trainable parameters of the DSTFGN model. We apply a stochastic gradient descent method to minimize the model loss function through back-propagation. Furthermore, the dropout technique is incorporated to enhance the efficiency and overall performance of the proposed approach.

Algorithm 1: Training the DSTFGN Model

Input: Training data $\mathbf{X}$, labels $\mathbf{Y}$, learning rate η, epochs E, batches B, dropout rate p
Output: Optimized parameters θ
Initialize θ randomly;
for *epoch* $= 1$ ***to*** E **do**
  for *batch* $= 1$ ***to*** B **do**
    Forward pass:
     $\hat{\mathbf{Y}} = \text{ReLU}(\mathbf{H}_{\text{out}} W_1 + b_1) W_2 + b_2$;
    Apply dropout with rate p;
    Loss: loss $= \frac{1}{N} \sum_{i=1}^{N} |\mathbf{Y}_i - \hat{\mathbf{Y}}_i|$;
    Back-propagation: Compute gradients ∇_θ of the loss;
    Update: $\theta \leftarrow \theta - \eta \nabla_\theta \text{loss}$;
  end
end
return *Optimized* θ

4 Experiment

We evaluated the performance of our DSTFGN model using real-world weather network data obtained from Kaggle. This dataset includes weather data from 30 regions across the United States and Canada, covering various parameters such as temperature, humidity, and atmospheric pressure, with a sampling frequency of one hour. As a result, there are 24 data samples per day for each region. To standardize the input data, we applied Z-score normalization. In our experiments, we used 48 hours of historical data to predict the next 24 hours, meaning we leveraged the past 48 time steps to forecast the next 24 time steps. Specifically, we used humidity and temperature as the primary features to assess the model performance. The training set consisted of data from July to August 2017, while the test set was derived from data from September 2017. The details of the datasets are presented in Table 1. Table 2 provides the system specifications used in our experiments.

We use three widely recognized evaluation metrics root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) to assess the model performance. The formulas for these metrics are provided below:

Table 1. Dataset details for DSTFGN evaluation.

Information	Details
Source	Kaggle
Regions	30 (US and Canada)
Parameters	Temperature, Humidity, Pressure
Sampling	1 hour (24 samples/day)
Normalization	Z-score
Training	July-Aug 2017
Testing	Sept 2017
Features	Humidity, Temperature
Prediction	48 hours history, 24 hours forecast

$$\text{MAPE} = \frac{1}{T} \sum_{t=1}^{T} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \tag{21}$$

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^{T} |y_t - \hat{y}_t| \tag{22}$$

$$\text{RMSE} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{y}_t)^2} \tag{23}$$

where T represents the observed sample index, and we have set $T = 12$ for our experiment.

Table 2. System specifications.

Component	Specification
CPU	8x Intel Xeon E5-2680 v4 @ 3.80GHz
CUDA	12.0
cuDNN	8.0
RAM	512 GB
GPU	4x NVIDIA Tesla P100

4.1 Baselines Comparison

To conduct a comparative analysis of the proposed model, we incorporated the following state-of-the-art baseline methods into the evaluation.

- **FC-LSTM [12]:** Fully-Connected LSTM (FC-LSTM) is a variant of the LSTM that incorporates fully-connected layers, making it highly effective for time series prediction tasks. By utilizing a structure of fully-connected hidden units, FC-LSTM improves the model's memory and expressive capabilities.
- **DHSTNet [13]:** This model proposed a unique dynamic deep spatio-temporal neural network model, called DHSTNet, which combines

convolutional neural networks and long short-term memory to simultaneously predict crowd flows across different regions of a city.

- **Transformer [14]:** Due to its impressive sequence modeling ability, the Transformer can also be applied to time series forecasting tasks. By utilizing self-attention mechanisms and positional encoding, the Transformer effectively captures long-range dependencies in time series, resulting in outstanding performance in time series prediction.
- **STGCN [15]:** Spatial-Temporal Graph Convolutional Networks (STGCN) are specifically designed for weather prediction.
- **DMFGNet:** The objective of the proposed DMFGNet model is to capture dynamic spatio-temporal relationships between different regions.
- **Att-DHSTNet [16]:** It addresses the dynamic spatio-temporal dependencies of weather flows. Temporal attention is employed to capture the dynamic temporal aspects across various time steps, while spatial attention is used to emphasize the spatial relationships between different locations.
- **STSGCN [17]:** Spatial-Temporal Synchronous Graph Convolutional Networks (STSGCN) is a model that combines data from multiple time steps simultaneously and uses graph convolution networks to effectively capture spatial and temporal correlations in graph-structured data.
- **DSTAGNN [18]:** Dynamic Spatial-Temporal Aware Graph Neural Network (DSTAGNN) is a model that uses a data-driven dynamic spatial-temporal perception graph, replacing the traditional static graph typically used in conventional graph convolutions. Furthermore, the model introduces an improved multi-head attention mechanism, integrated with multi-scale gated convolutional layers, to capture both temporal and spatial dependencies.
- **STGSA [19]:** The Spatial-Temporal Graph Synchronous Aggregation (STGSA) model is an innovative deep learning approach that effectively captures both localized and long-term dependencies through a specialized graph aggregation method, improving the extraction of spatial-temporal features. Additionally, STGSA utilizes a multi-stream module to process information from various representation projections, aggregating the most relevant features for precise forecasting
- **AAtt-DHSTNet [20]:** This model proposes a method for aggregating data to anticipate weather flows over the whole city in real-time.
- **DSTFGN:** This study proposes a model that combines multivariate time-series analysis and graph-based fusion causal inference to reveal complex interdependencies between weatherpatterns.

4.2 Hyper-parameters Settings

The *PyCharm* packages (version 3.1.0) are implemented using the *Keras* library. Moreover, all convolutions, fully connected layers are initialized using the Xavier initialization. We also used batch normalization and selected 64 as the mini-batch size. There is a fixed learning rate (LR) of 0.001. To lessen the issue of overfitting, the dropout rate is set at 0.25. To enhance the optimization of the proposed method, we employ the renowned Adam optimization technique to minimize the Euclidean loss. In addition, 70% of these datasets are divided into training sets, 10% into validation sets, and the remaining 20% are used for test sets. We also select the optimal model parameters based on their performance on the validation set and subsequently apply them to the test set to obtain the final prediction results. The training process aimed for maximum accuracy, but without validation data, there is a risk of overfitting.

4.3 Performance Comparison

Table 3 presents a comparison of the performance of our DSTFGN model against several baseline models in forecasting temperature and humidity data for the next 24 hours using 48 hours of historical data. The experimental results demonstrate that the DSTFGN model outperforms the baseline models across the temperature and humidity datasets, consistently achieving lower MAE, MAPE, and RMSE values, as indicated by the Improvement column (relative to the best baseline AAtt-DHSTNet). For temperature forecasting, these represent improvements of 8.18% (MAE), 10.58% (MAPE), and 5.31% (RMSE). For the humidity dataset, the DSTFGN model similarly shows superior performance, with improvements of 0.19% (MAE), 5.22% (MAPE), and 8.88% (RMSE). These results highlight the DSTFGN model's significant advantage in capturing the spatial-temporal dependencies within weather time series data.

Additionally, the experimental results show that baseline models like Transformer and FC-LSTM, which only consider temporal connections without taking advantage of a spatio-temporal network spatial dependencies, have worse predictive accuracy on the temperature and humidity datasets. Table 3 illustrates that baseline models that take spatial dependencies into account, such STGCN and STSGCN, perform better than the FC-LSTM and Transformer models, which have higher values for MAE, MAPE, and RMSE. In terms of experimental performance, models with spatial awareness outperform models that merely capture temporal dependencies because they can capture the spatio-temporal linkages within the data. Furthermore, FC-LSTM has drawbacks, including the incapacity to identify spatial links in the data, which limits its precision in situations where spatial dependencies are essential, like as weather forecasting. When working with long-term series data, its performance may be impacted by its sensitivity to input sequence length. Additionally, when processing large datasets, FC-LSTM models often have a high computational complexity. Conversely, the Transformer model has its own set of difficulties. The self-attention mechanism quadratic scaling with sequence length causes it to struggle with excessive memory consumption, despite its superior ability to capture long-range temporal relationships. When working with huge datasets or long-term series, this reduces its effectiveness. Furthermore, Transformer models application in tasks requiring both spatial and temporal knowledge is limited since, like FC-LSTM, they are not naturally able to record spatial relationships. Finally, transformers are less appropriate for issues with sparse or limited data since they frequently need big datasets for efficient training.

Figure 4 visualizes the performance of eight models across three standard metrics (MAE, MAPE, and RMSE) for temperature forecasting. Each group of bars represents a model, and each bar shows the corresponding error or accuracy metric. It is evident that DSTFGN consistently achieves the lowest errors across all metrics, demonstrating its superiority in capturing both spatio-temporal patterns. This clear visualization helps in comparing model robustness and selecting the most effective one for deployment in real-time weather forecasting systems.

The results demonstrate that although the Transformer and FC-LSTM models are excellent at capturing temporal dependencies, their predictive capabilities are constrained when spatial connections are not taken into account. However, models that combine temporal and spatial information, such as STGCN and STSGCN, are more equipped to identify intricate patterns in the data, which improves the accuracy of practical forecasting. This emphasizes how important it is to take into account the geographical dependencies of the monitoring station when forecasting time series. The observed performance discrepancy underscores the potential of spatio-temporal networks to increase forecasting accuracy and the need to include spatial dependencies when creating weather forecasting models. Furthermore, FC-LSTM has drawbacks, such as its inability to account for spatial links in data, which might impair performance in tasks that need both spatio-temporal knowledge. Furthermore, FC-LSTM is sensitive to input sequence length, making it less effective for handling long-term relationships in big datasets. The Transformer model is effective at capturing long-range temporal dependencies, but it consumes a lot of memory due to the self-attention mechanism, which scales quadratically with sequence length. This can be especially challenging for large-scale datasets. Furthermore, like FC-LSTM, the Transformer does not capture spatial relationships by default, limiting its efficacy in tasks requiring both spatial and temporal awareness. Lastly, Transformer models typically need large datasets to perform well, making them less suitable for applications with limited or sparse data.

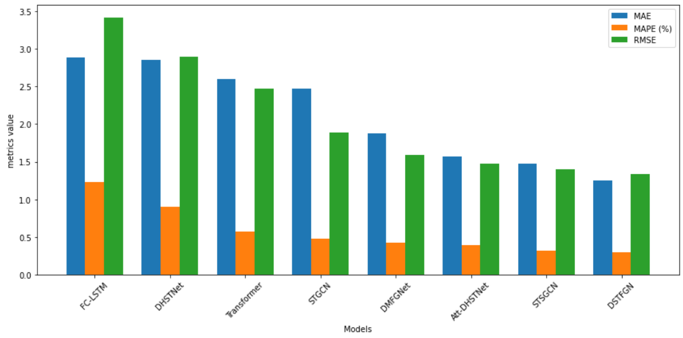

Figure 4. Comparison of different models.

While STSGCN generates spatio-temporal synchronized networks to capture spatial-temporal connections concurrently, it only considers local spatial-temporal dependencies and ignores long-term temporal dependencies. In contrast, our DSTFGN model takes into account both short-term and long-term temporal dependencies. Furthermore, our

Table 3. Performance comparison of different models on the dataset (Improvement indicates % improvement of DSTFGN over the best baseline AAtt-DHSTNet).

Dataset	Metric	FC-LSTM	DHSTNet	Transformer	STGCN	DMFGNet	Att-DHSTNet	STSGCN	AAtt-DHSTNet	Improvement (%)
	MAE	2.882	2.653	2.201	2.090	1.879	1.564	1.473	1.235	**8.18**
Temperature	MAPE (%)	1.231	0.897	0.568	0.473	0.424	0.394	0.332	0.293	**10.58**
	RMSE	3.604	2.894	2.456	1.892	1.587	1.475	1.398	1.338	**5.31**
	MAE	11.298	10.219	10.165	9.904	9.542	9.186	8.673	7.921	**0.19**
Humidity	MAPE (%)	23.899	26.981	15.867	11.508	20.987	13.764	12.660	15.234	**5.22**
	RMSE	14.247	11.349	11.955	10.321	9.998	9.887	9.657	9.432	**8.88**

model shows an enhanced construction of both the static spatial adjacency matrix and the fusion graph adjacency matrix. As a result, the experimental results suggest that our DSTFGN model is strong and effective at representing spatio-temporal interdependence.

4.4 Ablation Study

To better demonstrate the usefulness of each component in our DSTFGN model, we conducted a comparative analysis with different variation models using a series of experimental assessments. The settings for each variant are described in Table 4.

- Case 1: This variant does not use the maximum information coefficient (MIC) to construct the static spatial adjacency matrix. Instead, it solely relies on Euclidean distance to build the adjacency matrix.
- Case 2: This variant does not use the Transformer-based self-attention mechanism to capture short-term temporal dependencies (STD).
- Case 3: This form creates a spatial-temporal function graph without using the dynamic time warping (DTW) technique. Specifically, the spatial-temporal fusion graph is unique. The structure is based on both the temporal self-connection matrix and the static spatial adjacency matrix.
- Case 4: This variant does not use spatial-temporal graph fusion (STGF) to capture spatial-temporal dependencies simultaneously.
- Case 5: This approach avoids using graph attention networks to capture spatial dependencies (SD) between individual nodes.
- Case 6: Our DSTFGN model includes all the aforementioned modules.

According to the experimental results provided in Table 4, each module helps to improve the effectiveness of our approach. This is further confirmed by the visual representations in Figures 5 and 6, which show a graphical study of the model performance across various configurations. Figure 5 depicts the ablation study findings for the temperature dataset, highlighting the effect of each component on the model predicted accuracy. Figure 6 is a thorough graphical representation of the ablation study results for the humidity dataset.

When comparing different cases, particularly between Case 2 and Case 5, it is observed that in the temperature dataset, Case 2 outperforms Case 5 despite not accounting for long-term temporal dependencies, while Case 5 neglects spatial dependencies. This suggests that, for temperature forecasting, the importance of spatial dependencies for accurate predictions outweighs that of long-term temporal dependencies. When comparing multiple situations, particularly Case 2 and 5, it is clear that, while Case 2 does not account for long-term temporal dependencies, it outperforms Case 5, which does not include geographical connections, on the temperature dataset. This shows that local dependencies are more important for accurate temperature forecasting than long-term temporal dependencies.

Case 3 significantly outperforms Case 4 in all parameters, with lower MAE, MAPE, and RMSE values. Using spatio-temporal fusion graphs can help the model capture spatio-temporal relationships and improve prediction performance. Based on the results above, the following conclusions can be drawn: When considering the construction of the spatial adjacency matrix using MIC, the Transformer-based self-attention mechanism, the creation of the spatial-temporal fusion graph with the DTW algorithm, and the use of graph attention networks, our DSTFGN model delivers the best performance in terms of prediction accuracy, achieving the lowest MAE, MAPE, and RMSE values. This demonstrates the effectiveness of incorporating these modules to capture spatio-temporal dependencies and enhance prediction performance. In contrast, models that rely solely on Euclidean distance, ignore long-term temporal dependencies, do not use the DTW algorithm, or overlook spatial dependencies perform

Table 4. Ablation experiment on different configurations of modules.

Name	MIC	STD	DTW	STGF	SD	Temperature			Humidity		
						MAE	MAPE (%)	RMSE	MAE	MAPE (%)	RMSE
case1	✗	✓	✓	✓	✓	1.492	0.487	1.794	10.198	14.998	10.986
case2	✓	✗	✓	✓	✓	1.693	0.487	1.986	9.982	17.189	11.197
case3	✓	✓	✗	✓	✓	1.582	0.435	0.998	7.957	14.975	9.983
case4	✓	✓	✓	✗	✓	1.701	0.542	1.992	8.967	17.992	10.893
case5	✓	✓	✓	✓	✗	1.896	0.567	1.984	11.091	15.987	11.991
case6	✓	✓	✓	✓	✓	0.923	1.061	1.476	7.982	13.994	8.897

poorly. Therefore, our DSTFGN model proves to be a promising and effective choice for weather forecasting tasks.

5 Discussion

5.1 Application Scenarios

The DSTFGN model, with its innovative approach to weather forecasting, holds significant potential for application across various industries. Its ability to capture complex spatio-temporal relationships makes it especially valuable in the agricultural sector, where accurate weather forecasts are essential for crop planting, growth, and harvesting. Farmers can leverage this information to make informed decisions about irrigation management and harvest scheduling. Additionally, precise temperature and humidity forecasts are crucial for energy demand forecasting, as they enable energy companies to optimize energy distribution and manage supply in the heating, cooling, and power sectors.

5.2 Limitations

Despite its great potential, the DSTFGN model has certain limitations. The performance of the model heavily depends on the quality of the meteorological data it receives. If the data contains numerous outliers or missing values, its predictive capabilities can be significantly compromised. In regions with sparse distribution of monitoring stations, the model may struggle to accurately capture spatial correlations of weather features and may fail to reflect local weather patterns with precision. Additionally, due to the complexity of its self-attention and graph attention network components, the model may require significant computational resources, which could be a constraint in resource-limited environments.

6 Related Work

We have divided this section into two sections: (i) spatio-temporal prediction approaches, (ii) Causal discovery.

6.1 Spatio-temporal models for weather forecasting

Weather forecasting involves analyzing atmospheric data, such as temperature, wind speed, and humidity [34]. Due to the significant temporal and regional fluctuations in atmospheric conditions, weather forecasting has always been a critical task [35]. Traditional forecasting methods rely on numerical weather prediction (NWP), which uses complex equations to model the atmospheric environment. NWP requires large amounts of data and substantial computational resources, and its ability to predict rare events is limited [36]. Furthermore, NWP has shown suboptimal performance in predicting short-term weather conditions [37]. Today, weather forecasting is primarily driven by computational methods, reducing the labor-intensive nature of earlier approaches [38]. In recent years, artificial intelligence-based data-driven models have been widely used in weather forecasting [39].

Recently, the task of urban weather forecasting has garnered significant attention and has been widely researched [21, 22, 40]. Early studies used convolutional neural networks (CNN) and recurrent neural networks (RNN) to independently analyze the spatial and temporal dependencies of road networks. Later works have sought to integrate these methods to capture the overall dynamics of weather flow. For instance, [5] employed a spatio-temporal feature selection method to extract weather flow features, with CNNs used for learning and prediction tasks. Similarly, [23] combined convolutional LSTMs and bidirectional LSTMs to adaptively capture the dynamic evolution of

Figure 5. Results of ablation experiment on the temperature dataset.

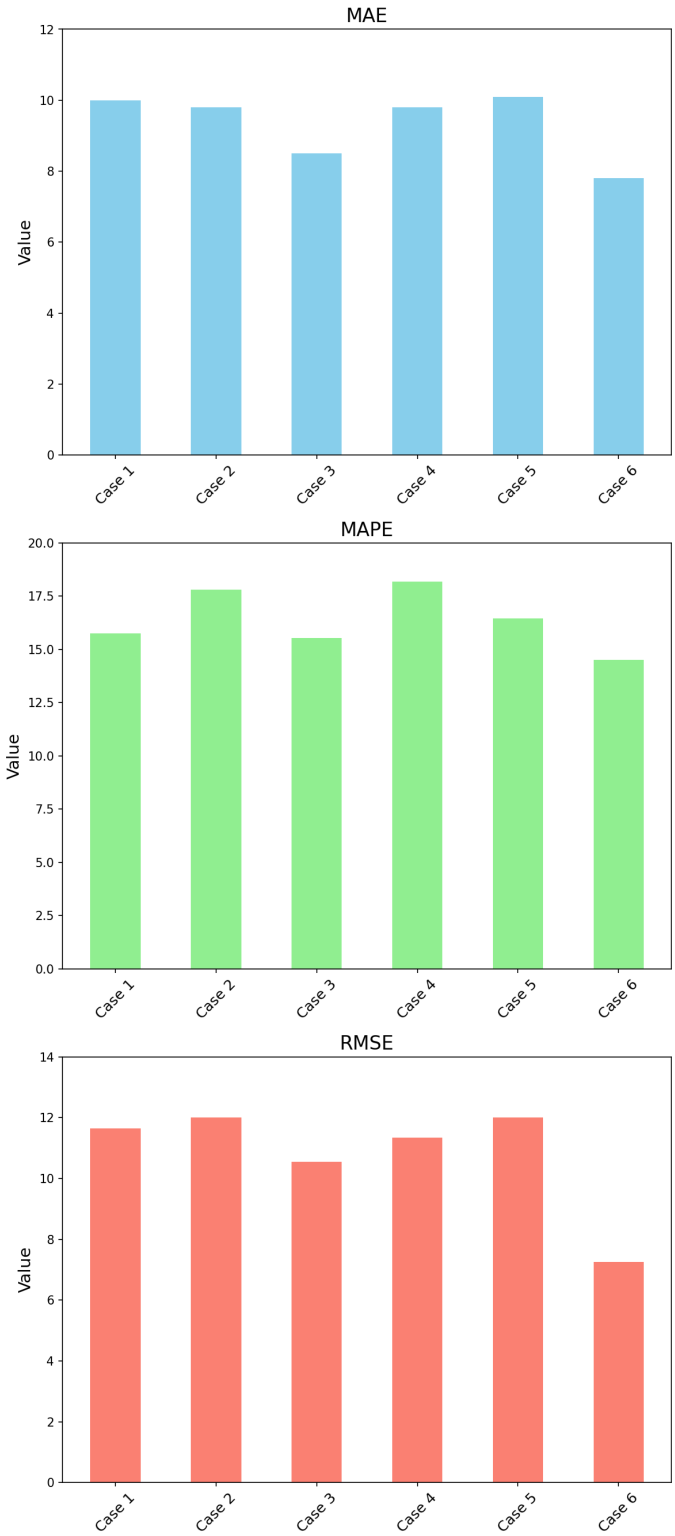

Figure 6. Results of ablation experiment on the humidity dataset.

weather flow. However, these approaches mainly focus on local spatio-temporal features and may not fully capture the complex spatio-temporal dependencies inherent in transportation network structures. In [1], this work proposes a model with four components: (i) closeness, (ii) period influence, (iii) weekly influence, and (iv) external branch, each with varying weights, which are fused to predict weather flow.

Recently, Spatio-Temporal Graph Neural Networks (STGNNs) have shown exceptional effectiveness in capturing intricate spatio-temporal correlations, significantly advancing research in weather forecasting. Li et al. [24] used diffusion convolution to accurately capture spatio-temporal correlations and introduced the STGCN, which applies spatial graph convolution and temporal convolution to capture neighborhood relationships between nodes and time-varying trends. However, these methods are designed for predefined static adjacency matrices and fail to capture complex, dynamic spatio-temporal dependencies. To address this limitation [25] developed the Dynamic Graph Convolutional Recurrent Network (DGCRN), which uses a super network to generate dynamic graphs and combines them with static graphs to dynamically represent the road network structure. Similarly, [26] designed a spatial sentinel module that dynamically adjusts information extraction by reducing the influence of irrelevant road data based on temporal and spatial attention mechanisms. In [27] proposed the Randomized Graph Diffusion Attention Network (RGDAN), which incorporates a graph diffusion attention module to dynamically adjust spatial relationship weights and a temporal attention mechanism to extract temporal relationship weights. While existing studies have yielded promising results, they do not simultaneously address the influences of temporal causal relationships and spatial dynamics in the propagation of weather between road nodes.

6.2 Causal discovery

The subject of causal discovery has received great interest, with pioneering contributions from multiple disciplines. This includes statistical approaches based on apriori knowledge and data-driven correlation-based modeling. Granger causality testing [11] is a well-established method for examining causal relationships between variables through time series forecasting. In [28] provided a summary of various algorithmic variants used in Granger causality studies in recent years, and outlined potential future research directions. In weather forecasting, [29] introduced a transmission mechanism based on spatio-temporal Granger causality and a spatio-temporal arrangement algorithm to model global transmission causal relationships (TCR). Another noteworthy algorithm is the PC algorithm [30], which uses independently and identically distributed data for causal discovery. PCMCI [31] extends this approach to large, nonlinear time-series datasets by incorporating nonlinear conditional independence tests. More recently, TCDF [32] combined attention mechanisms with temporal convolutional neural networks for data-driven causal analysis. Causal analysis has been widely studied across various fields [33]. However, few studies have integrated causal analysis results into models to enhance their performance. There is still a need to develop effective methodologies for utilizing and analyzing the results of causal discovery.

7 Conclusions and Future Work

In this paper, we introduce a Spatio-Temporal Dynamic Fusion Graph Network (DSTFGN) model, which leverages causal analysis to address urban weather forecasting challenges. To tackle the dynamic spatial dependencies and the temporal causal effects present in weather scenarios, DSTFGN integrates Granger causality theory with deep learning techniques. This allows the model to effectively capture spatial correlations and causal relationships between road nodes using node attribute features. Specifically, DSTFGN learns the causal graph structure through Granger causality testing with historical node lags, while concurrently learning the adaptive graph structure based on node embeddings. Additionally, the model addresses the heterogeneity of road nodes through a spatial-temporal attention module. Experimental results show that our model performs exceptionally well on real-world datasets, demonstrating its effectiveness in predicting complex weather. In comparison to existing spatio-temporal graph models utilizing graph neural networks, our approach not only reduces the reliance on predefined adjacency matrices but also explores dynamic characteristics embedded in node data via dynamic causal analysis, resulting in more accurate predictions.

Our future research could focus on improving the model adaptability to real-time data, allowing it to dynamically respond to emerging weather patterns. This would be particularly beneficial for generating more accurate short-term forecasts. Additionally, exploring the application of the model in other spatial-temporal forecasting domains, such as public health, could be valuable. In this context, spatial-temporal predictions could help forecast the spread and velocity of infectious diseases, thus aiding in the development of disease prevention and control strategies. By addressing current limitations and expanding into new application areas, the DSTFGN model could evolve further to meet the changing demands of various industries and contribute to a broader range of fields.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 62373255; in part by the Natural Science Foundation of

Guangdong Province under Grant 2024A1515011204; in part by the Shenzhen Natural Science Fund through the Stable Support Plan Program under Grant 20220809175803001; in part by the Open Fund of National Engineering Laboratory for Big Data System Computing Technology under Grant SZU-BDSC-OF2024-15.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Ali, A., Zhu, Y., & Zakarya, M. (2022). Exploiting dynamic spatio-temporal graph convolutional neural networks for citywide traffic flows prediction. *Neural networks*, *145*, 233-247. [CrossRef]

[2] Dai, F., Huang, P., Mo, Q., Xu, X., Bilal, M., & Song, H. (2022). ST-InNet: Deep spatio-temporal inception networks for traffic flow prediction in smart cities. *IEEE Transactions on Intelligent Transportation Systems*, *23*(10), 19782-19794. [CrossRef]

[3] Khan, M. M. H., Mustafa, M. R. U., Hossain, M. S., Shams, S., & Julius, A. D. (2023). Short-term and long-term rainfall forecasting using arima model. *International Journal of Environmental Science and Development*, *14*(5), 292-298. [CrossRef]

[4] Jin, M., Koh, H. Y., Wen, Q., Zambon, D., Alippi, C., Webb, G. I., ... & Pan, S. (2024). A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. [CrossRef]

[5] Zhang, W., Yu, Y., Qi, Y., Shu, F., & Wang, Y. (2019). Short-term traffic flow prediction based on spatio-temporal analysis and CNN deep learning. *Transportmetrica A: Transport Science*, *15*(2), 1688-1711. [CrossRef]

[6] Tang, C., Sun, J., Sun, Y., Peng, M., & Gan, N. (2020). A general traffic flow prediction approach based on spatial-temporal graph attention. *IEEE Access*, *8*, 153731-153741. [CrossRef]

[7] Qi, T., Chen, L., Li, G., Li, Y., & Wang, C. (2023). FedAGCN: A traffic flow prediction framework based on federated learning and Asynchronous Graph Convolutional Network. *Applied Soft Computing*, *138*, 110175. [CrossRef]

[8] Choi, J., & Park, N. (2023). Graph neural rough differential equations for traffic forecasting. *ACM Transactions on Intelligent Systems and Technology*, *14*(4), 1-27. [CrossRef]

[9] Huo, G., Zhang, Y., Wang, B., Gao, J., Hu, Y., & Yin, B. (2023). Hierarchical spatio–temporal graph convolutional networks and transformer network for traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems*, *24*(4), 3855-3867. [CrossRef]

[10] Chen, J., Zheng, L., Hu, Y., Wang, W., Zhang, H., & Hu, X. (2024). Traffic flow matrix-based graph neural network with attention mechanism for traffic flow prediction. *Information Fusion*, *104*, 102146. [CrossRef]

[11] Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424-438. [CrossRef]

[12] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.

[13] Ali, A., Zhu, Y., Chen, Q., Yu, J., & Cai, H. (2019, December). Leveraging spatio-temporal patterns for predicting citywide traffic crowd flows using deep hybrid neural networks. In *2019 IEEE 25th international conference on parallel and distributed systems (ICPADS)* (pp. 125-132). IEEE. [CrossRef]

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[15] Yu, B., Yin, H., & Zhu, Z. (2017). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*.

[16] Ali, A., Zhu, Y., & Zakarya, M. (2021). Exploiting dynamic spatio-temporal correlations for citywide traffic flow prediction using attention based neural networks. *Information Sciences*, *577*, 852-870. [CrossRef]

[17] Song, C., Lin, Y., Guo, S., & Wan, H. (2020, April). Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 01, pp. 914-921). [CrossRef]

[18] Lan, S., Ma, Y., Huang, W., Wang, W., Yang, H., & Li, P. (2022, June). Dstagnn: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting. In *International conference on machine learning* (pp. 11906-11917). PMLR.

[19] Wei, Z., Zhao, H., Li, Z., Bu, X., Chen, Y., Zhang, X., ... & Wang, F. Y. (2023). STGSA: A novel spatial-temporal graph synchronous aggregation model for traffic prediction. *IEEE/CAA Journal of Automatica Sinica*, *10*(1), 226-238. [CrossRef]

[20] Jin, W., Lin, Y., Wu, Z., & Wan, H. (2018, March). Spatio-temporal recurrent convolutional networks for citywide short-term crowd flows prediction. In *Proceedings of the 2nd International Conference on Compute and Data Analysis* (pp. 28-35). [CrossRef]

[21] Liu, M., Zhu, T., Ye, J., Meng, Q., Sun, L., & Du, B. (2023). Spatio-temporal autoencoder for traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems, 24*(5), 5516-5526. [CrossRef]

[22] Ma, Y., Lou, H., Yan, M., Sun, F., & Li, G. (2024). Spatio-temporal fusion graph convolutional network for traffic flow forecasting. *Information Fusion, 104*, 102196. [CrossRef]

[23] Zheng, H., Lin, F., Feng, X., & Chen, Y. (2020). A hybrid deep learning model with attention-based conv-LSTM networks for short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems, 22*(11), 6910-6920. [CrossRef]

[24] Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2017). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*.

[25] Li, F., Feng, J., Yan, H., Jin, G., Yang, F., Sun, F., ... & Li, Y. (2023). Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution. *ACM Transactions on Knowledge Discovery from Data, 17*(1), 1-21. [CrossRef]

[26] Park, C., Lee, C., Bahng, H., Tae, Y., Jin, S., Kim, K., ... & Choo, J. (2020, October). ST-GRAT: A novel spatio-temporal graph attention networks for accurately forecasting dynamically changing road speed. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 1215-1224). [CrossRef]

[27] Fan, J., Weng, W., Tian, H., Wu, H., Zhu, F., & Wu, J. (2024). RGDAN: A random graph diffusion attention network for traffic prediction. *Neural networks, 172*, 106093. [CrossRef]

[28] Shojaie, A., & Fox, E. B. (2022). Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application, 9*(1), 289-319. [CrossRef]

[29] He, S., Luo, Q., Du, R., Zhao, L., He, G., Fu, H., & Li, H. (2023). STGC-GNNs: A GNN-based traffic prediction framework with a spatial–temporal Granger causality graph. *Physica A: Statistical Mechanics and its Applications, 623*, 128913. [CrossRef]

[30] Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., & Aragam, B. (2020, June). Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics* (pp. 1595-1605). Pmlr.

[31] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances, 5*(11), eaau4996. [CrossRef]

[32] Nauta, M., Bucur, D., & Seifert, C. (2019). Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction, 1*(1), 19. [CrossRef]

[33] Guo, R., Cheng, L., Li, J., Hahn, P. R., & Liu, H. (2020). A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR), 53*(4), 1-37. [CrossRef]

[34] Fathi, M., Haghi Kashani, M., Jameii, S. M., & Mahdipour, E. (2022). Big data analytics in weather forecasting: A systematic review. *Archives of Computational Methods in Engineering, 29*(2), 1247-1275. [CrossRef]

[35] Al Sukhni, H., Sakr, F., Alsmadi, M. K., Abd-Elghany, M., Gomaa, I. A., & Abdallah, S. (2025). Data-Driven Weather Prediction: Integrating Deep Learning and Ensemble Models for Robust Weather Forecasting. *Journal of Cybersecurity & Information Management, 15*(2). [CrossRef]

[36] Brotzge, J. A., Berchoff, D., Carlis, D. L., Carr, F. H., Carr, R. H., Gerth, J. J., ... & Wang, X. (2023). Challenges and opportunities in numerical weather prediction. *Bulletin of the American Meteorological Society, 104*(3), E698-E705. [CrossRef]

[37] Brotzge, J. A., Berchoff, D., Carlis, D. L., Carr, F. H., Carr, R. H., Gerth, J. J., ... & Wang, X. (2023). Challenges and opportunities in numerical weather prediction. *Bulletin of the American Meteorological Society, 104*(3), E698-E705. [CrossRef]

[38] Haupt, S. E., Cowie, J., Linden, S., McCandless, T., Kosovic, B., & Alessandrini, S. (2018, October). Machine learning for applied weather prediction. In *2018 IEEE 14th international conference on e-science (e-Science)* (pp. 276-277). IEEE. [CrossRef]

[39] Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., ... & Stadtler, S. (2021). Can deep learning beat numerical weather prediction?. *Philosophical Transactions of the Royal Society A, 379*(2194), 20200097. [CrossRef]

[40] Hu, N., Zhang, D., Xie, K., Liang, W., Li, K. C., & Zomaya, A. Y. (2024). Dynamic multi-scale spatial–temporal graph convolutional network for traffic flow prediction. *Future Generation Computer Systems, 158*, 323-332. [CrossRef]

Ahmad Ali received the master's and Ph.D. degrees from Shanghai Jiao Tong University, China, where he developed a solid foundation in computer science and engineering. Currently, he is a Postdoctoral Researcher with the College of Mechatronics and Control Engineering, Shenzhen University, China. In addition to his reviewing activities, he has been involved in several cutting-edge research projects. His work has led to numerous publications in top-tier journals and conferences, showcasing his expertise and contributions to the field. He is passionate about leveraging technology to solve real-world problems, particularly in urban environments where efficient data processing and intelligent decision-making are crucial. His research interests include deep learning, big data analytics, data mining, urban computing, cloud computing, and fog computing. He has reviewed over a thousand research articles for prestigious academic journals, such as Information Sciences, Information

Fusion, Neural Networks, IEEE Internet of Things Journal, Neural Computing and Applications, Multimedia Tools and Application, Wireless Networks, and IEEE Access. (Email: ahmadali@szu.edu.cn)

Hafiz Muhammad Yasir Naeem received the B.S. degree in Electrical Engineering from Bahria University Islamabad, Pakistan in 2013, M.S. degree in control systems from University of Engineering and Technology (UET), Taxila, Pakistan, in 2016 and Ph.D. degree from the Capital University of Science and Technology, Islamabad, Pakistan in 2023. His research interests include energy-efficient optimization strategies for EVs, Model Predictive Control and optimal control. Currently, He is a postdoctoral fellow at the College of Mechatronics and Control Engineering, and the College of Computer Science and Software Engineering at Shenzhen University, Shenzhen, China. (Email: yasirnaeem945@szu.edu.cn)

Riaz Ali is a distinguished CAS-TWAS PhD President's Fellow in Computer Applied Technology (Data Mining) with a focus on Meteorology. He has authored over 12 publications in renowned journals and conferences, including AGU Earth and Space Science, Atmospheric Research, Sustainability, and others. His research primarily centers on the quantitative forecasting of rainfall and storm surges associated with both landfalling and offshore tropical cyclones along the Guangdong and Hong Kong coasts. His broader academic interests include the spatio-temporal analysis of typhoon hazards, the application of artificial intelligence in weather forecasting, and natural hazard and risk assessment. (Email: riazbabuzai@sjtu.edu.cn)

Mujtaba Asad completed his PhD from Shanghai Jiao Tong university, China. Currently he is Postdoctoral Researcher at Shanghai Jiao Tong university, China. His research interest deal with programming languages, artificial neural networks, image processing, computer engineering etc. Dr. Asad has published a couple of research articles in well reputable journals including Advanced Engineering Informatics, Applied Intelligence etc. (Email: asadmujtaba@sjtu.edu.cn)

Md. Belal Bin Heyat received his B.Tech in Electronics and Instrumentation (EI) and M.Tech in Electronics Circuit and System (ECS) from Integral University, Lucknow, India, in 2014 and 2016. He completed his Ph.D. in Electronic Science and Technology from the University of Electronic Science and Technology of China (UESTC), Chengdu, China. During his Ph.D., he won nine awards and served as a Research Associate, Vice-Chairman of the UESTC Country League, and Country Representative for India. He worked as a Postdoctoral Fellow at the IoT Research Center, College of Computer Science and Software Engineering, Shenzhen University, China, from 2021 to 2023. Currently, he is a Postdoctoral Fellow at the CenBRAIN Neurotech Center of Excellence, Westlake University, China. He is also a visiting postdoctoral researcher at the Center for Visual Information Technology (CVEST), IIIT Hyderabad, India, and a faculty member in the Department of Science and Engineering at Novel Global Community Educational Foundation, NSW, Australia. Dr. Heyat's research focuses on detecting diseases such as sleep, neurological, and psycho-neurological disorders, as well as cardiovascular conditions. He specializes in signal processing and medical machine learning. He is a Guest Editor for journals like Life, Journal of Integrative Neuroscience, and Applied Sciences. He also reviews research for publishers such as Oxford University Press, IEEE, Elsevier, Hindawi, Wiley, Springer, MDPI, and Bentham Science. Dr. Heyat has published more than 70 articles (h-index: 25) in respected international journals and conferences. His work aims to improve healthcare using AI, advanced signal processing techniques, and traditional system of medicine. (Email: belalheyat@gmail.com)

Tamam Alsarhan received the Ph.D. degree from Shanghai Jiao Tong University, specializing in computer vision, and deep learning. During her academic career, she has contributed significantly to the computer vision area. She was a Visiting Researcher at Khalifa University. She is currently an Assistant Professor with The University of Jordan. Her research interests include computer vision, human action recognition, machine learning, and pattern recognition. In addition to her research, she is dedicated to teaching and mentoring students, aiming to inspire the next generation of scientists and researchers. (Email: t_alsarhan@ju.edu.jo)