MSAGAT-Net: Multi-Scale Adaptive Graph Attention Network for Efficient Spatiotemporal Epidemic Forecasting

Michael Ajao-olarinoye[a,*], Vasile Palade[a], Fei He[a], Petra A Wark[b], Seyed Mousavi[a] and Zindoga Mukandavire[c]

[a] Centre for Computational Science and Mathematical Modelling, Coventry University, Coventry, United Kingdom
[b] Research Methods and Evaluation Unit, Research Centre for Healthcare and Communities, Coventry University, Coventry, United Kingdom
[c] Institute of Applied Research and Technology, Emirates Aviation University, Dubai, United Arab Emirates

ARTICLE INFO

Keywords:
Graph attention networks
graph structure learning
epidemic forecasting
spatiotemporal forecasting
multi-horizon prediction
COVID-19

ABSTRACT

Accurate spatiotemporal epidemic forecasting is vital for preparedness and resource planning, yet many graph neural network approaches rely on predefined adjacency matrices, struggle to control how structural priors affect attention across varying graph densities, and become unstable at longer forecast horizons. We propose MSAGAT-Net, a computationally efficient multi-scale adaptive graph attention network that learns spatial dependencies directly from data while allowing optional adjacency priors. MSAGAT-Net introduces a self-regulating additive structural bias by augmenting attention logits with a learnable low-rank graph bias and an optional adjacency prior before softmax; this design ensures that the prior's influence adapts automatically with graph density, eliminating the need for hand-tuned thresholds. The architecture further features an adaptive multi-hop spatial diffusion module whose depth scales with graph size to reduce oversmoothing, depthwise separable temporal feature extraction, dilated multi-scale temporal features, and progressive prediction refinement for stable multi-horizon forecasts. Evaluated on diverse datasets spanning influenza, COVID-19, and ICU bed occupancy, MSAGAT-Net consistently outperforms strong baselines, achieving up to 23.5% RMSE reduction on LTLA COVID-19 forecasting and 22.2% on NHS time-series forecasting. Ablation results reveal horizon-dependent architectural importance: adaptive spatial attention is universally essential, while spatial refinement mechanisms grow increasingly critical as forecast difficulty increases.

1. Introduction

Timely and accurate epidemic forecasting underpins effective public health preparedness, enabling decision-makers to allocate clinical resources, plan interventions, and communicate risk before outbreaks escalate [8, 4]. The COVID-19 pandemic underscored how rapidly disease dynamics can evolve across hundreds or thousands of interconnected regions [35, 12, 39, 26], demanding computational methods that jointly model spatial propagation between regions and temporal patterns ranging from daily reporting fluctuations to seasonal waves [30, 36]. Traditional compartmental models (e.g., SIR/SEIR) provide valuable mechanistic insight but struggle with the heterogeneity and computational complexity required for real-time surveillance at this scale [15, 34, 8].

Deep learning approaches—originally demonstrated in traffic forecasting [22, 49] and environmental monitoring [43, 20]—have advanced epidemic prediction through recurrent, convolutional, and graph neural network (GNN) architectures [2, 50, 1, 17, 40]. GNNs are particularly well suited to this setting because they can learn complex spatiotemporal dependencies directly from data, outperforming traditional models in capturing regional heterogeneity and dynamic transmission mechanisms [2, 18].

Despite these advances, current GNN-based approaches face three fundamental limitations:

- **Uncontrolled structural bias in attention.** Most graph attention mechanisms are over-parameterised and offer no mechanism to regulate how much a structural prior (e.g., an adjacency matrix) influences the learned attention weights. This causes performance to vary unpredictably across graphs of different density and size.

- **Mandatory reliance on predefined adjacency matrices.** State-of-the-art models such as EpiGNN [44], Cola-GNN [9], and DCRNN [22] require adjacency matrices constructed from geographical proximity or mobility data, which may not reflect true transmission pathways and presuppose domain expertise or external data sources that are often unavailable.

- **Fixed spatial processing and unstable multi-horizon forecasts.** Existing methods employ fixed architectural choices for spatial aggregation, failing to adapt to diverse graph sizes and disease dynamics. Compounding this rigidity, error accumulation across extended forecast horizons limits the utility of long-range predictions essential for public health planning [4].

To address these limitations, we propose MSAGAT-Net (Multi-Scale Adaptive Graph Attention Network), whose central contribution is a *self-regulating additive structural bias* in graph attention that automatically adapts to graph density without manual tuning. The architecture comprises two novel modules—an Efficient Adaptive Graph Attention Module (EAGAM) with learnable low-rank graph bias and

[*] Corresponding author.
✉ olarinoyem@coventry.ac.uk (M. Ajao-olarinoye)
ORCID(s):

self-regulating adjacency prior, and a Multi-Scale Spatial Feature Module (MSSFM) with graph-size-adaptive multi-hop convolutions and locality-biased fusion—supported by depthwise separable temporal feature extraction (TFEM), progressive prediction refinement with a learnable decay rate (PPRM), and a highway autoregressive connection. Our key contributions are:

1. **Self-regulating additive structural bias attention (EAGAM).** We propose a graph attention mechanism that adds a learnable low-rank graph bias and an optional adjacency prior directly to attention scores before softmax normalisation. Softmax shift-invariance causes this additive prior to self-regulate across graph densities—negligible on dense graphs, meaningful on sparse ones—eliminating graph-size-dependent thresholds used in prior work and enabling the model to learn spatial relationships entirely from data. Unlike EpiGNN, Cola-GNN, and DCRNN, MSAGAT-Net does not require predefined adjacency matrices; ablation confirms comparable or better performance without any adjacency input.
2. **Adaptive multi-hop spatial refinement with anti-oversmoothing (MSSFM).** We introduce a spatial module that aggregates multi-hop graph convolutions with locality-biased learnable fusion weights and adapts the maximum diffusion depth to graph size, preventing oversmoothing on small graphs while retaining broader spatial context on larger ones. This graph-size-adaptive mechanism is absent from existing epidemic GNNs.
3. **Horizon-dependent architectural analysis on diverse epidemic data.** Through systematic evaluation on six datasets spanning influenza, COVID-19, and ICU bed occupancy, we demonstrate that MSAGAT-Net achieves the best RMSE in the majority of settings, with improvements of up to 23.5% on LTLA-Timeseries and 22.2% on NHS-Timeseries over the strongest baselines. Ablation studies reveal that component importance is fundamentally horizon-dependent: adaptive spatial attention is universally essential, while spatial refinement and prediction modules become increasingly critical as forecast difficulty increases.

The remainder of this paper is organised as follows: Section 2 reviews related work; Section 3 presents the problem formulation and proposed architecture; Section 4 describes the experimental setup; Section 5 reports results and ablation studies; and Section 6 concludes with limitations and future directions.

2. Related Work

Epidemic forecasting has evolved from classical compartmental models to neural network architectures that capture spatial coupling and temporal heterogeneity. This section reviews (i) spatiotemporal graph learning for epidemics, including attention-based, hybrid physics-informed, and transformer-based models that learn dynamic cross-regional influence, and (ii) multi-scale temporal modelling and multi-horizon forecasting strategies. We position MSAGAT-Net within this landscape by highlighting the efficiency limits of quadratic attention, the rigidity of fixed architectural choices, and the need for stable long-range forecasts.

2.1. Spatiotemporal Epidemic Modelling

The application of graph neural networks to epidemic forecasting has emerged as a dominant paradigm, fundamentally addressing limitations of traditional compartmental models that assume uniform mixing and fixed transmission parameters [23]. A comprehensive taxonomy by Liu et al. [24, 42] distinguishes between statistical epidemiology models, general machine learning models, deep neural network-based time series models and spatiotemporal approaches, with different preprocessing and modelling choices, revealing distinct trade-offs between interpretability and modelling flexibility.

Spatiotemporal approaches have demonstrated remarkable success in learning complex spatiotemporal patterns directly from data. Deng et al. [9] presented Cola-GNN, a cross-location attention-based graph neural network for long-term Influenza-Like Illness (ILI) prediction, where the dynamic cross-location attention mechanism replaced fixed geographic adjacency matrices with learnable attention weights that adapt to time-varying transmission patterns.

Building on these foundations, Xie et al. [44] developed EpiGNN, which combines transmission risk encoding with a Region-Aware Graph Learner that explicitly models both local clustering effects and global connectivity patterns. By incorporating human mobility data into the graph learning process, EpiGNN achieved substantial improvements on multiple epidemic forecasting tasks, reducing RMSE by approximately 9.5% compared to baseline methods.

Gao et al. [11] proposed STAN, a spatiotemporal attention network that uses graph attention mechanisms with patient electronic health records and geography-based features. Applied to COVID-19 forecasting in all US counties, STAN achieved up to 87% lower mean square error compared to classical SIR/SEIR models, demonstrating that attention-based spatial modelling can significantly outperform traditional compartmental approaches.

Recent developments have focused on unifying spatial and temporal modelling with more sophisticated dynamic mechanisms. Han et al. [14] developed DyGraphFormer, which integrates dynamic graph learning with Transformer architectures to capture evolving spatial-temporal dependencies through gated recurrent units that continuously update graph structure based on recent observations. Similarly, Pu et al. [31] proposed DASTGN with dual-scale attention mechanisms that adaptively fuse spatial and temporal effects at both fine and coarse-grained resolutions. Qiu et al. [33] proposed MSGNN, a multi-scale spatio-temporal graph neural network that decomposes epidemic signals across spatial and temporal resolutions via scale-specific graph convolutions.

However, MSGNN relies on predefined adjacency matrices, employs fixed-depth message passing without mechanisms to prevent oversmoothing on small graphs, and does not address multi-horizon forecast stability.

A related direction involves hybrid approaches that incorporate epidemiological knowledge into neural architectures to improve interpretability and long-range forecast stability. For example, Cao et al. [5] proposed MepoGNN, which combines region-level SEIR compartmental simulators with Graph Attention Networks, transforming static travel matrices into dynamic transmission adjacency matrices. Gao et al. [10] introduced HOIST, using Ising spin dynamics to regularise forecasting models based on the assumption that neighbouring regions' case counts evolve in correlated patterns. While such hybrid approaches provide theoretical grounding, they often require extensive domain expertise for model specification and may struggle to capture complex non-linear dynamics that deviate from assumed mechanistic forms.

Despite these advances, current spatiotemporal approaches face two critical limitations. First, they typically employ over-parameterised attention mechanisms that require predefined adjacency matrices and offer no mechanism to control how much structural prior influences the learned attention—the bias is either absent or applied at a fixed strength, which can degrade performance when graph density varies across datasets. Low-rank decomposition has been explored to address the computational cost of graph attention: Puny et al. [32] proposed Low-Rank Global Attention (LRGA) that replaces full dot-product attention with a factorised form, Kong et al. [19] introduced Global Representation Key attention using shared low-rank projections, and Yang et al. [45] embedded adaptive low-rank decomposition within ego-network propagation layers. More broadly, Wang et al. [41] demonstrated that self-attention can be approximated with linear complexity via low-rank projections. However, these approaches reduce parameter count without addressing the deeper challenge of adaptive structural bias in graph attention, which remains unaddressed in epidemic forecasting. Second, existing approaches often fail to maintain stability in multi-horizon forecasts, as error propagation compounds over extended forecast horizons [4].

2.2. Multi-Scale Temporal Modelling and Multi-Horizon Forecasting

Epidemic time series data exhibit complex, multi-scale temporal dynamics arising from a range of underlying processes. Short-term fluctuations are often driven by reporting practices, such as testing schedules and data collection delays [30], while longer-term patterns, including seasonal waves, are shaped by environmental factors and behavioural responses to disease spread [36]. Effective multi-horizon forecasting of such data typically falls into two principal methodological categories: (1) direct forecasting, in which models forecast multiple future time steps simultaneously, and (2) iterative (or autoregressive) forecasting, where forecasts are generated sequentially and recursively at each time step [4]. Capturing these temporal dependencies across multiple scales is therefore essential for designing forecasting models that remain robust under data irregularities and regime shifts.

Direct multi-horizon models, often implemented using sequence-to-sequence architectures with LSTM or CNN components, have demonstrated effectiveness in influenza forecasting. However, these models generally require substantial training data and exhibit sensitivity to the inclusion and quality of external covariates [43, 38]. Wang et al. [40] developed DEFSI, which integrates deep learning with compartmental models to improve long-range forecasts, but observed that performance deteriorates significantly beyond four-week horizons due to accumulating uncertainty.

Iterative strategies, while more data-efficient, are prone to error propagation across extended forecasting horizons. This inherent limitation has motivated the development of multi-module architectures designed to capture both high-frequency fluctuations and low-frequency trends simultaneously. Deng et al. [9] addressed these challenges using dilated convolutions for multi-scale temporal feature extraction, finding that incorporating seasonal trends improved forecast stability. However, their approach relies on fixed dilation patterns that may not adapt effectively to the changing dynamics of the epidemic in different diseases and regions.

Recognising the need to represent both short-term outbreaks and long-term epidemiological waves, recent research has explored the incorporation of external data sources, including climatic variables, demographic information, and digital surveillance indicators [25, 27]. Although such approaches can improve long-range predictive performance, they often require extensive feature engineering and may not generalise well in heterogeneous epidemic contexts [30].

These limitations across spatiotemporal modelling, physics-informed approaches, and multi-scale temporal processing underscore three critical gaps in contemporary epidemic forecasting research: (1) existing graph attention mechanisms are over-parameterised and lack adaptive control over how structural priors influence the learned attention, limiting robustness across diverse graph topologies; (2) the rigidity of fixed architectural designs and mandatory reliance on predefined adjacency matrices limits adaptability across diverse epidemic settings; and (3) the lack of stable multi-horizon forecasting capabilities required for effective public health planning. Our proposed framework, MSAGAT-Net, addresses these challenges by integrating structural-bias graph attention with low-rank projections, adaptive multi-hop spatial refinement, and progressive forecast refinement to enable stable and scalable multi-horizon epidemic forecasting.

3. Methodology

3.1. Problem Formulation

Consider a set of N geographical regions, such as cities, counties, states, countries, administrative health regions, or NHS regions in England, conceptualised as nodes within a graph framework. Historical epidemic data are structured in the matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T] \in \mathbb{R}^{N\times T}$, where T denotes the total length of the historical observation period, and each vector $\mathbf{x}_t \in \mathbb{R}^N$ (for $t = 1, 2, \dots, T$) corresponds to the observed data for all N regions at time step t. The individual component $x_{i,t}$ signifies the epidemic metric (such as case count, vaccination counts, hospital admissions, or ventilator occupancy) for region i at time step t.

For each specific region i, its temporal progression is represented by the vector $\mathbf{x}^i = [x_{i,1}, x_{i,2}, \dots, x_{i,T}] \in \mathbb{R}^T$. This dual representation facilitates the analysis of both spatial patterns (across different regions at a given time) and temporal patterns (within a single region over time).

The principal aim of this investigation is to forecast future epidemic values for all regions over a designated time horizon of h steps into the future. Mathematically, given the historical data available up to time t, the task is to predict:

$$\mathbf{x}_{t+h} = [x_{1,t+h}, x_{2,t+h}, \dots, x_{N,t+h}]^T \tag{1}$$

For forecasting, we employ a sliding window approach with a fixed-length look-back period w. At any current time step t, we use the most recent observations $[\mathbf{x}_{t-w+1}, \mathbf{x}_{t-w+2}, \dots, \mathbf{x}_t]$ to forecast $\mathbf{x}_{t+h}$.

The spatial relationships between regions are encoded in a graph structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$ represents the set of regions and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the potential connections between regions. An optional adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ may encode known spatial relationships (e.g., geographical proximity), but is not required by our model.

The forecasting task can be formalised as learning a function f that maps recent historical data to future predictions:

$$\mathbf{x}_{t+h} = f([\mathbf{x}_{t-w+1}, \mathbf{x}_{t-w+2}, \dots, \mathbf{x}_t]; \boldsymbol{\Theta}) \tag{2}$$

where $\boldsymbol{\Theta}$ denotes the learnable parameters, including learnable graph structure parameters ($\mathbf{U}$, $\mathbf{V}$) that enable MSAGAT-Net to discover spatial relationships from data, optionally augmented by a predefined adjacency prior.

3.2. Temporal Feature Extraction Module (TFEM)

The first component of the MSAGAT-Net architecture is the Temporal Feature Extraction Module (TFEM), which transforms the raw time-series data into compact feature representations. Given input data $\mathbf{X} = [\mathbf{x}_{t-w+1}, \dots, \mathbf{x}_t] \in \mathbb{R}^{N\times w}$ for N regions over a look-back window of w time steps, TFEM extracts temporal features through depthwise separable convolutions [7] followed by low-rank projections, reducing parameter count and computational cost while maintaining expressive power [21, 46].

3.2.1. Depthwise Separable Convolutions

For each region's historical window $\mathbf{x}^i_{[t-w+1:t]} \in \mathbb{R}^w$, we apply a depthwise convolution followed by a pointwise convolution. The depthwise convolution applies a separate filter per channel:

$$\mathbf{z}^i_{\text{depth}} = \text{Conv1D}_{\text{depth}}(\mathbf{x}^i; \boldsymbol{\Theta}_{\text{depth}}) \tag{3}$$

where $\mathbf{z}^i_{\text{depth}} \in \mathbb{R}^{w\times 1}$ represents the output after depthwise convolution, maintaining the temporal dimension whilst processing each input channel independently.

Following the depthwise convolution, a pointwise convolution (implemented as a 1×1 convolution) is applied to expand the single channel to multiple feature channels:

$$\mathbf{z}^i_{\text{point}} = \text{Conv1D}_{\text{point}}(\mathbf{z}^i_{\text{depth}}; \boldsymbol{\Theta}_{\text{point}}) \tag{4}$$

where $\mathbf{z}^i_{\text{point}} \in \mathbb{R}^{w\times d_{\text{feat}}}$ and $d_{\text{feat}} = 16$ is the number of output feature channels. Batch normalisation and ReLU activation are applied after each convolution:

$$\mathbf{z}^i_{\text{norm}} = \text{ReLU}(\text{BatchNorm}(\mathbf{z}^i_{\text{point}})) \tag{5}$$

3.2.2. Low-Rank Feature Projection

After extracting features using depthwise separable convolutions, we apply a low-rank projection to further reduce dimensionality and capture the most salient features. This projection consists of two linear transformations with a bottleneck in between:

$$\mathbf{F}^i_{\text{low}} = \text{Linear}_{\text{low}}(\text{Flatten}(\mathbf{z}^i_{\text{norm}})) \tag{6}$$

$$\mathbf{F}^i = \text{Linear}_{\text{high}}(\mathbf{F}^i_{\text{low}}) \tag{7}$$

where $\mathbf{F}^i_{\text{low}} \in \mathbb{R}^{d_{\text{bottle}}}$ is the bottleneck representation and $\mathbf{F}^i \in \mathbb{R}^{d_{\text{hidden}}}$ is the final feature vector for region i. The flattening operation converts $\mathbf{z}^i_{\text{norm}} \in \mathbb{R}^{w\times d_{\text{feat}}}$ into a vector of dimension $w\times d_{\text{feat}}$, which is then compressed through the bottleneck ($d_{\text{bottle}} \ll w \times d_{\text{feat}}$) before expansion to d_{hidden}. This low-rank bottleneck reduces the parameter count from $\mathcal{O}(w \cdot d_{\text{feat}} \cdot d_{\text{hidden}})$ to $\mathcal{O}(w \cdot d_{\text{feat}} \cdot d_{\text{bottle}} + d_{\text{bottle}} \cdot d_{\text{hidden}})$ per region, and acts as an information bottleneck that regularises the representations by forcing the model to retain only the most predictive temporal patterns.

After applying this projection to all regions, we obtain the feature matrix $\mathbf{F} \in \mathbb{R}^{N\times d_{\text{hidden}}}$ with layer normalisation and ReLU activation:

$$\mathbf{F} = \text{ReLU}(\text{LayerNorm}(\mathbf{F})) \tag{8}$$

The TFEM pipeline is illustrated in Figure 2.

All TFEM hyperparameters (d_{feat}, d_{bottle}, d_{hidden}, kernel size) are reported in Table 2.

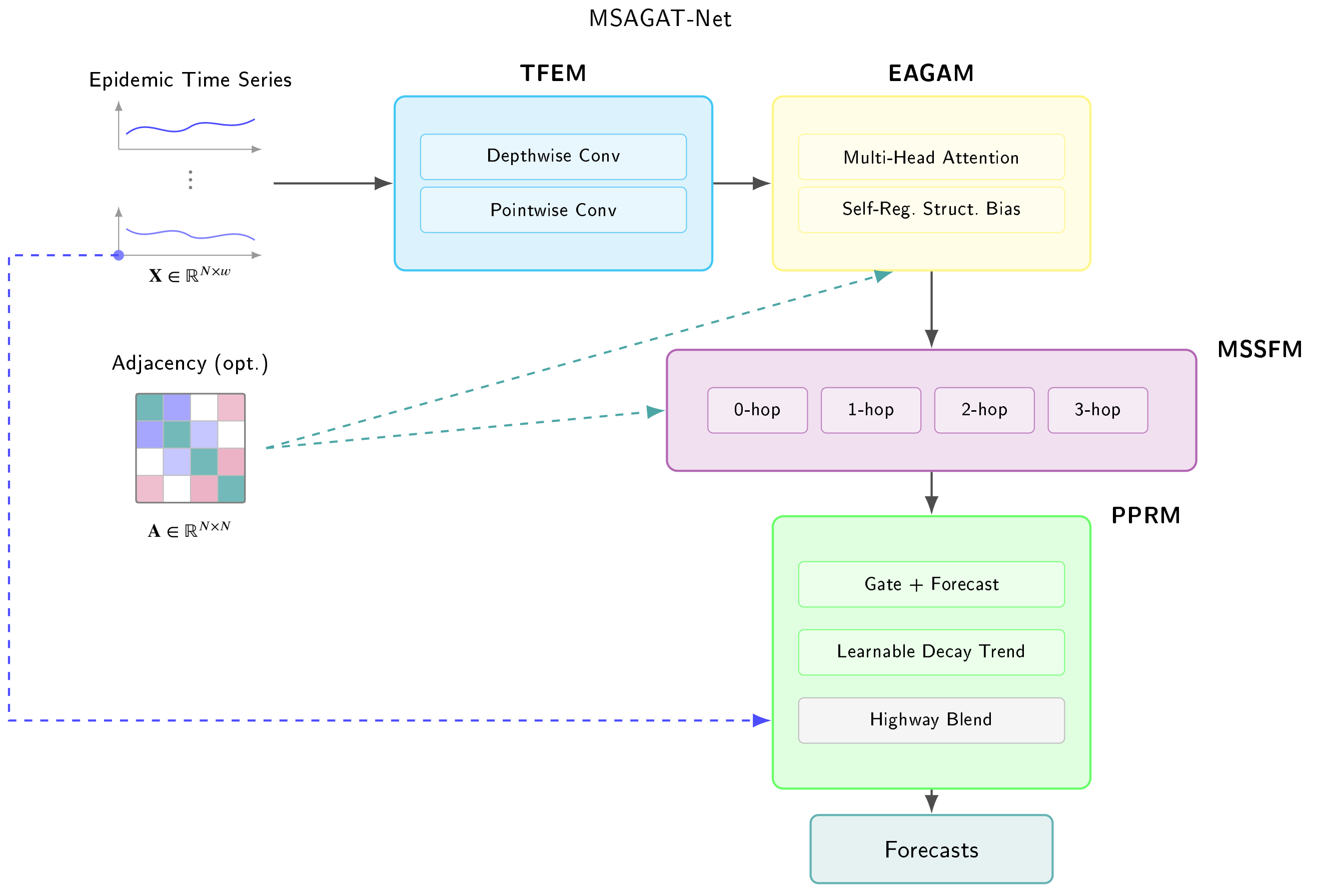


Figure 1: Overview of the MSAGAT-Net architecture. Epidemic time series for N regions over a look-back window w pass through TFEM, EAGAM, MSSFM, and PPRM before producing forecasts. PPRM incorporates the Highway Blend, which merges model predictions with an autoregressive baseline via a learnable gate λ. Teal dashed lines indicate the optional adjacency matrix feeding into EAGAM (as a self-regulating soft prior) and MSSFM (for multi-hop aggregation); the blue dashed line marks the autoregressive skip connection.

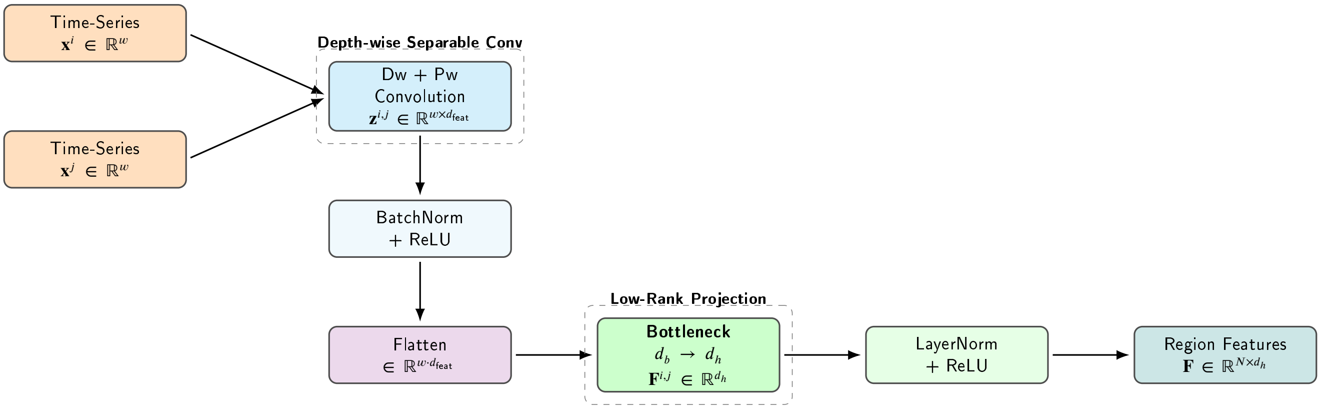

Time-Series
$\mathbf{x}^i \in \mathbb{R}^w$
Time-Series
$\mathbf{x}^j \in \mathbb{R}^w$
Depth-wise Separable Conv
Dw + Pw
Convolution
$\mathbf{z}^{i,j} \in \mathbb{R}^{w \times d_{\text{feat}}}$
BatchNorm
+ ReLU
Flatten
$\in \mathbb{R}^{w \cdot d_{\text{feat}}}$
Low-Rank Projection
Bottleneck
$d_b \rightarrow d_h$
$\mathbf{F}^{i,j} \in \mathbb{R}^{d_h}$
LayerNorm
+ ReLU
Region Features
$\mathbf{F} \in \mathbb{R}^{N \times d_h}$

Figure 2: Feature-extraction pipeline. Independent regional time-series $\mathbf{x}^i$ and $\mathbf{x}^j$ are processed in parallel by depth-wise and point-wise convolutions, normalised, flattened, passed through a bottleneck projection ($d_{\text{bottle}} \rightarrow d_{\text{hidden}}$), and normalised again to yield region-level feature vectors $\mathbf{F}$.

3.3. Efficient Adaptive Graph Attention with Low-Rank Decomposition

The second core component of our MSAGAT-Net architecture is EAGAM. Traditional approaches to spatial modelling often rely on fixed adjacency matrices based on geographical proximity or administrative boundaries, which do not capture the evolving nature of epidemic spread influenced by factors such as population mobility, healthcare referral patterns, and socioeconomic connections. Based on the principles of graph attention networks [37], our EAGAM adaptively learns the relationships between regions based on their feature representations, rather than being constrained by a predefined graph structure. This adaptive approach allows the model to discover and leverage spatial dependencies that may not be immediately apparent from geographical proximity alone, and to adjust these dependencies as the epidemic evolves.

Standard graph attention mechanisms incur $\mathcal{O}(N^2 \cdot d)$ complexity and offer no mechanism to control how much structural prior influences the learned attention. Low-rank decomposition has been shown to reduce this cost effectively [32, 19, 45], motivating the design of EAGAM.

EAGAM employs multi-head scaled dot-product attention with low-rank bottleneck projections for query, key, and value representations. Its key innovation is the integration of two complementary structural biases directly into the attention scores *before* softmax normalisation: (i) a learnable low-rank graph bias $\mathbf{B} = \mathbf{UV}$ that captures persistent spatial relationships from data, and (ii) an optional additive adjacency prior with learnable scale whose influence self-regulates based on graph density (detailed in Section 3.3.4). The module comprises five components described below.

3.3.1. Bottleneck Projection

Given the feature matrix $\mathbf{F} \in \mathbb{R}^{N \times d_{\text{hidden}}}$ from TFEM, where N is the number of regions and d_{hidden} is the hidden dimension, we first project these features into query, key, and value representations through an efficient bottleneck projection:

$$\mathbf{Q}_{\text{low}}, \mathbf{K}_{\text{low}}, \mathbf{V}_{\text{low}} = \text{Split}(\text{Linear}_{\text{low}}(\mathbf{F}), 3) \tag{9}$$

where $\text{Linear}_{\text{low}} : \mathbb{R}^{d_{\text{hidden}}} \to \mathbb{R}^{3 \times d_{\text{bottle}}}$ projects the features into a lower-dimensional space and Split divides the output into three separate tensors of dimension $\mathbb{R}^{N \times d_{\text{bottle}}}$.

These low-dimensional projections are then expanded back to the full hidden dimension:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \text{Split}(\text{Linear}_{\text{high}}([\mathbf{Q}_{\text{low}}; \mathbf{K}_{\text{low}}; \mathbf{V}_{\text{low}}]), 3) \tag{10}$$

where $\text{Linear}_{\text{high}} : \mathbb{R}^{3 \times d_{\text{bottle}}} \to \mathbb{R}^{3 \times d_{\text{hidden}}}$ and each of $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N \times d_{\text{hidden}}}$.

This bottleneck projection significantly reduces the parameter count from $\mathcal{O}(3 \times d_{\text{hidden}}^2)$ to $\mathcal{O}(3 \times d_{\text{hidden}} \times d_{\text{bottle}})$, where $d_{\text{bottle}} \ll d_{\text{hidden}}$.

3.3.2. Multi-Head Attention Mechanism

To enhance the model's capacity to capture different types of inter-regional relationships, we implement a multi-head attention mechanism where the hidden representations are split into h heads, each with dimension $d_{\text{head}} = d_{\text{hidden}}/h$:

$$\mathbf{Q}^{(i)}, \mathbf{K}^{(i)}, \mathbf{V}^{(i)} \in \mathbb{R}^{N \times d_{\text{head}}}, \quad i \in \{1, 2, \dots, h\} \tag{11}$$

For efficient computation, we reshape these tensors to explicitly represent the multiple heads:

$$\mathbf{Q}_h = \text{Reshape}(\mathbf{Q}, [N, h, d_{\text{head}}]) \tag{12}$$

$$\mathbf{K}_h = \text{Reshape}(\mathbf{K}, [N, h, d_{\text{head}}]) \tag{13}$$

$$\mathbf{V}_h = \text{Reshape}(\mathbf{V}, [N, h, d_{\text{head}}]) \tag{14}$$

We then transpose the first two dimensions to facilitate batch-wise processing across attention heads:

$$\mathbf{Q}_h = \text{Transpose}(\mathbf{Q}_h, 0, 1) \tag{15}$$

$$\mathbf{K}_h = \text{Transpose}(\mathbf{K}_h, 0, 1) \tag{16}$$

$$\mathbf{V}_h = \text{Transpose}(\mathbf{V}_h, 0, 1) \tag{17}$$

resulting in tensors of shape $[h, N, d_{\text{head}}]$.

Within each attention head, we compute standard scaled dot-product attention scores:

$$\mathbf{S}_h = \frac{\mathbf{Q}_h \mathbf{K}_h^T}{\sqrt{d_{\text{head}}}} \tag{18}$$

where $\mathbf{S}_h \in \mathbb{R}^{h \times N \times N}$. The two structural biases—a learnable graph bias $\mathbf{B}$ (Section 3.3.3) and an optional adjacency prior (Section 3.3.4)—are added before softmax normalisation:

$$\mathbf{A}_h = \text{softmax}(\mathbf{S}_h + \mathbf{B} + \alpha \cdot \tilde{\mathbf{A}}) \tag{19}$$

where $\mathbf{B} = \mathbf{UV}$ is the learnable graph bias, $\tilde{\mathbf{A}}$ is the row-normalised adjacency prior (when available), and $\alpha = \text{softplus}(\alpha_0)$ is a learnable positive scale parameter. The self-regulating properties of this additive formulation are analysed in Section 3.3.4.

The final attention output for each head is then computed as:

$$\mathbf{O}_h = \mathbf{A}_h \mathbf{V}_h \tag{20}$$

where $\mathbf{O}_h \in \mathbb{R}^{h \times N \times d_{\text{head}}}$ represents the attended features across all heads.

3.3.3. Learnable Graph Structure Bias

An important feature of our EAGAM is the incorporation of a learnable graph structure bias that is added directly to the attention scores before softmax normalisation. Unlike traditional graph attention networks that rely solely on node features for computing attention, we include a learnable bias term that captures persistent structural relationships between regions that may not be evident from the node features alone.

This bias is implemented as a low-rank decomposition for parameter efficiency:

$$\mathbf{B} = \mathbf{U}\mathbf{V} \quad (21)$$

where $\mathbf{U} \in \mathbb{R}^{h\times N\times d_{\text{bias}}}$ and $\mathbf{V} \in \mathbb{R}^{h\times d_{\text{bias}}\times N}$ are learnable parameters initialised with Xavier uniform [13], and $d_{\text{bias}} \ll N$ is the bottleneck dimension. The resulting bias $\mathbf{B} \in \mathbb{R}^{h\times N\times N}$ is added directly to the content-based attention scores before softmax, allowing it to reinforce or suppress specific region-to-region attention patterns as a first-class component of the spatial reasoning mechanism.

3.3.4. Self-Regulating Additive Adjacency Prior

Unlike state-of-the-art baselines that *require* predefined adjacency matrices, MSAGAT-Net learns spatial relationships entirely from data. However, when prior knowledge about regional connectivity is available (e.g., geographical proximity or mobility patterns), it can be optionally incorporated to accelerate learning and improve performance through an additive prior mechanism.

Given an adjacency matrix $\mathbf{A}$, we first compute a row-normalised prior:

$$\tilde{\mathbf{A}} = \mathbf{D}_{\text{row}}^{-1}\mathbf{A} \quad (22)$$

where $\mathbf{D}_{\text{row}} = \text{diag}(\mathbf{A}\mathbf{1})$ is the row-sum diagonal matrix. This prior is then added to the attention scores with a learnable positive scale:

$$\mathbf{S}'_h = \mathbf{S}_h + \mathbf{B} + \alpha \cdot \tilde{\mathbf{A}} \quad (23)$$

where $\alpha = \text{softplus}(\alpha_0)$ ensures positivity, and α_0 is a learnable scalar parameter initialised to 1.0 (yielding $\alpha \approx 1.31$ initially, providing a moderate structural nudge).

The key insight behind this additive formulation is its *self-regulating* behaviour across different graph topologies. Because the softmax function is shift-invariant—$\text{softmax}(\mathbf{x} + c\mathbf{1}) = \text{softmax}(\mathbf{x})$ for any constant vector $c\mathbf{1}$—the prior's effect naturally adapts to graph density:

- **Dense graphs:** When the adjacency matrix is dense, the row-normalised prior $\tilde{\mathbf{A}}$ approaches a uniform distribution across all columns. Adding a near-constant vector to each row of attention scores before softmax has negligible effect on the resulting attention distribution. The model thus relies primarily on content-based attention and the learned graph bias.

- **Sparse graphs:** When the adjacency matrix is sparse, $\tilde{\mathbf{A}}$ has peaked entries for connected nodes and zeros elsewhere. The additive prior meaningfully shifts attention towards geographically relevant neighbours, providing structural guidance where it is most needed.

This self-regulating property eliminates the need for graph-size-dependent thresholds or conditional gating mechanisms, which we found to be fragile in practice. When no adjacency is provided, the model operates using only content-based attention and the learned graph bias; when adjacency is available, the scale α is learned during training, and the prior is pre-computed and stored as a constant buffer with negligible overhead. This distinguishes MSAGAT-Net from baselines such as EpiGNN [44], Cola-GNN [?], and DCRNN [?], which mandate adjacency matrices as required input.

3.3.5. Attention Regularisation

To promote sparse and interpretable spatial relationships, we apply L1 regularisation to the attention weight matrix after softmax normalisation:

$$\mathcal{L}_{\text{attn}} = \lambda \|\mathbf{A}_h\|_1 \quad (24)$$

where $\mathbf{A}_h$ denotes the attention weights (the softmax output used for value aggregation) and λ is a learnable regularisation weight. By penalising the L1 norm of the attention distribution rather than the structural bias alone, this regularisation directly encourages each node to attend to a sparse subset of other nodes, producing interpretable spatial dependency patterns. The value of λ is initialised to 10^{-5} and adapted during training through gradient descent in log-domain (ensuring positivity), allowing the model to automatically balance forecast accuracy with attention sparsity.

After computing the attended values for each head, we combine them and project back to the original feature dimension:

$$\mathbf{O} = \text{Reshape}(\text{Transpose}(\mathbf{O}_h, 0, 1), [N, d_{\text{hidden}}]) \quad (25)$$

Similarly to the input projection, we employ a low-rank output projection for efficiency:

$$\mathbf{O}_{\text{low}} = \text{Linear}_{\text{out\_low}}(\mathbf{O}) \quad (26)$$

$$\mathbf{O}_{\text{final}} = \text{Linear}_{\text{out\_high}}(\mathbf{O}_{\text{low}}) \quad (27)$$

where $\mathbf{O}_{\text{low}} \in \mathbb{R}^{N\times d_{\text{bottle}}}$ and $\mathbf{O}_{\text{final}} \in \mathbb{R}^{N\times d_{\text{hidden}}}$.

The output of EAGAM, $\mathbf{O}_{\text{final}}$, represents the features of the region after incorporating spatial dependencies. This output, along with the attention regularisation loss $\mathcal{L}_{\text{attn}}$, is passed to the subsequent MSSFM for further processing.

All EAGAM hyperparameters (attention heads, bottleneck dimension, regularisation weight, adjacency prior scale) are reported in Table 2. Figure 3 presents the data flow through the EAGAM module.

3.4. Multi-Scale Spatial Feature Module

The third major component of the proposed MSAGAT-Net architecture is the Multi-Scale Spatial Feature Module (MSSFM), which refines spatial dependencies using multi-hop graph convolutions. Epidemic propagation involves interactions that extend beyond immediate neighbours, but

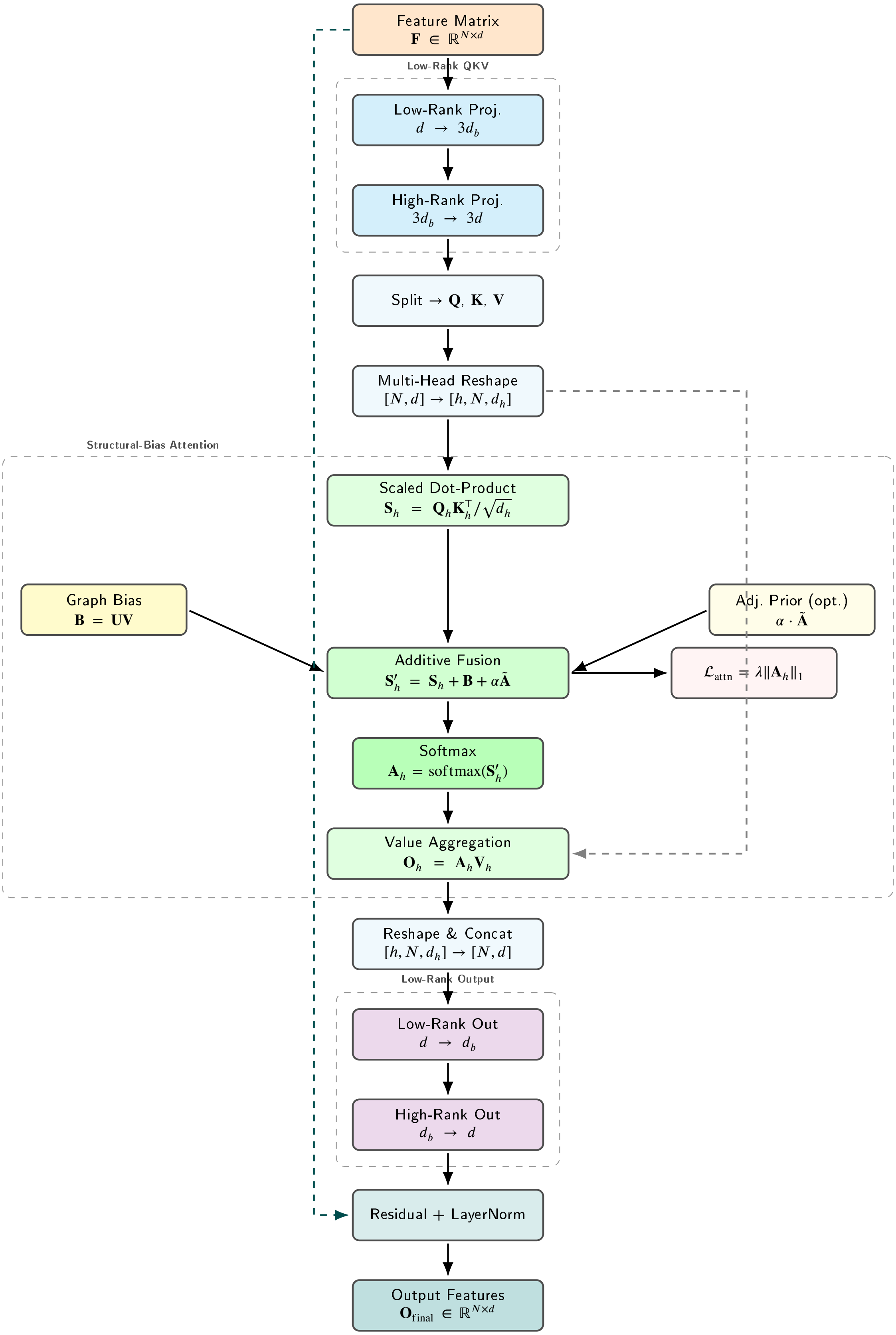

Feature Matrix
$\mathbf{F} \in \mathbb{R}^{N \times d}$
Low-Rank QKV
Low-Rank Proj.
$d \rightarrow 3d_b$
High-Rank Proj.
$3d_b \rightarrow 3d$
Split → $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$
Multi-Head Reshape
$[N, d] \rightarrow [h, N, d_h]$
Structural-Bias Attention
Scaled Dot-Product
$\mathbf{S}_h = \mathbf{Q}_h \mathbf{K}_h^{\top} / \sqrt{d_h}$
Graph Bias
$\mathbf{B} = \mathbf{UV}$
Adj. Prior (opt.)
$\alpha \cdot \tilde{\mathbf{A}}$
Additive Fusion
$\mathbf{S}'_h = \mathbf{S}_h + \mathbf{B} + \alpha \tilde{\mathbf{A}}$
$\mathcal{L}_{\text{attn}} = \lambda \|\mathbf{A}_h\|_1$
Softmax
$\mathbf{A}_h = \text{softmax}(\mathbf{S}'_h)$
Value Aggregation
$\mathbf{O}_h = \mathbf{A}_h \mathbf{V}_h$
Reshape & Concat
$[h, N, d_h] \rightarrow [N, d]$
Low-Rank Output
Low-Rank Out
$d \rightarrow d_b$
High-Rank Out
$d_b \rightarrow d$
Residual + LayerNorm
Output Features
$\mathbf{O}_{\text{final}} \in \mathbb{R}^{N \times d}$

Figure 3: Data flow in the EAGAM module. Input features $\mathbf{F}$ are projected through a low-rank bottleneck into Q, K, V representations. Scaled dot-product scores are augmented with a learnable graph bias $\mathbf{B} = \mathbf{UV}$ and an optional adjacency prior $\alpha \cdot \tilde{\mathbf{A}}$ before softmax. Attended values pass through a low-rank output stage with residual connection. $\mathcal{L}_{\text{attn}}$: L1 sparsity loss on attention weights.

excessive diffusion can oversmooth small graphs. MSSFM addresses this by aggregating information across multiple hop distances while adaptively limiting the maximum hop depth based on graph size and fusing scales with locality-biased weights.

3.4.1. Multi-Hop Graph Convolutions

Let $\mathbf{G} \in \mathbb{R}^{B \times N \times d_{\text{hidden}}}$ denote the spatial features output by EAGAM. Given an adjacency matrix $\mathbf{A}$, we construct a row-normalised matrix with self-loops:

$$\hat{\mathbf{A}} = \mathbf{D}^{-1}(\mathbf{A} + \mathbf{I}), \tag{28}$$

where $\mathbf{D}$ is the degree matrix of $(\mathbf{A}+\mathbf{I})$. For each hop $k \in \{0, 1, \dots, S-1\}$, MSSFM computes a k-hop aggregation:

$$\mathbf{H}^{(k)} = \hat{\mathbf{A}}^k \mathbf{G} \mathbf{W}^{(k)}, \tag{29}$$

with $\mathbf{W}^{(k)}$ implemented as a linear transform followed by LayerNorm and ReLU. The $k = 0$ scale uses the identity matrix to preserve self-features. When no adjacency is provided, MSSFM defaults to identity aggregation across all scales.

3.4.2. Adaptive Hop Depth and Locality-Biased Fusion

To reduce oversmoothing on small graphs, the number of scales is set adaptively as

$$S = \min(S_{\text{max}}, \max(2, \lfloor N/5 \rfloor)), \tag{30}$$

ensuring that small graphs use fewer hops while larger graphs retain broader context. We fuse the multi-hop features with learnable weights initialised to favour locality:

$$\boldsymbol{\alpha} = \text{softmax}(\mathbf{w}), \quad \mathbf{H}_{\text{fused}} = \sum_{k=0}^{S-1} \alpha_k \mathbf{H}^{(k)}. \tag{31}$$

3.4.3. Bottleneck Projection and Residual Connection

The fused features are passed through a low-rank bottleneck projection and residual connection to stabilise training:

$$\mathbf{H}_{\text{proj}} = \text{Linear}_{\text{high}}(\text{Linear}_{\text{low}}(\mathbf{H}_{\text{fused}})), \tag{32}$$

$$\mathbf{H}_{\text{final}} = \text{LayerNorm}(\mathbf{H}_{\text{proj}} + \mathbf{G}). \tag{33}$$

This design preserves the hidden dimension while enabling adaptive spatial refinement across multiple hop distances. Figure 4 illustrates the data flow through the MSSFM module.

3.5. Progressive Multi-Horizon Forecast Refinement

The final component of the MSAGAT-Net architecture is the Progressive Prediction Refinement Module (PPRM). Since forecast errors accumulate over extended horizons [3, 6], PPRM incorporates an adaptive refinement mechanism—inspired by gating concepts in recurrent networks [16]—that balances model-based forecasts with trend-based extrapolations conditioned on recent observations.

3.5.1. Low-Rank Forecast Projection

Given the spatiotemporal feature tensor $\mathbf{H}_{\text{final}} \in \mathbb{R}^{B \times N \times d_{\text{hidden}}}$ from MSSFM, where B is the batch size, N is the number of regions, and d_{hidden} is the hidden dimension, we first apply a bottleneck projection to distil the most forecast-relevant information:

$$\mathbf{P}_{\text{low}} = \text{Linear}_{\text{pred\_low}}(\mathbf{H}_{\text{final}}) \tag{34}$$

where $\mathbf{P}_{\text{low}} \in \mathbb{R}^{B \times N \times d_{\text{bottle}}}$ is the bottleneck representation with dimension $d_{\text{bottle}} \ll d_{\text{hidden}}$. This projection reduces dimensionality before the final forecast layer, reducing the parameter count whilst encouraging compact feature representations.

We then apply layer normalisation, ReLU activation, and dropout to the bottleneck representation:

$$\mathbf{P}_{\text{mid}} = \text{Dropout}(\text{ReLU}(\text{LayerNorm}(\mathbf{P}_{\text{low}}))) \tag{35}$$

This intermediate processing enhances training stability and introduces non-linearity necessary for modelling complex forecast patterns.

3.5.2. Horizon-Specific Forecasting

From the processed bottleneck representation, we generate initial forecasts for all forecast horizons using a linear projection:

$$\mathbf{P}_{\text{initial}} = \text{Linear}_{\text{pred\_high}}(\mathbf{P}_{\text{mid}}) \tag{36}$$

where $\mathbf{P}_{\text{initial}} \in \mathbb{R}^{B \times N \times h}$ represents the raw model forecasts for each region across all forecast horizons h.

To improve multi-horizon forecasting stability, we incorporate an adaptive refinement mechanism that combines these model-based forecasts with trend-based extrapolations from recent observations.

3.5.3. Adaptive Refinement Mechanism

The PPRM incorporates an adaptive refinement gate that balances model-based forecasts with trend-based extrapolations conditioned on the most recent observations.

We first compute an adaptive gate based on the spatiotemporal features:

$$\mathbf{R} = \sigma(\text{Linear}_{\text{gate\_high}}(\text{ReLU}(\text{Linear}_{\text{gate\_low}}(\mathbf{H}_{\text{final}})))) \tag{37}$$

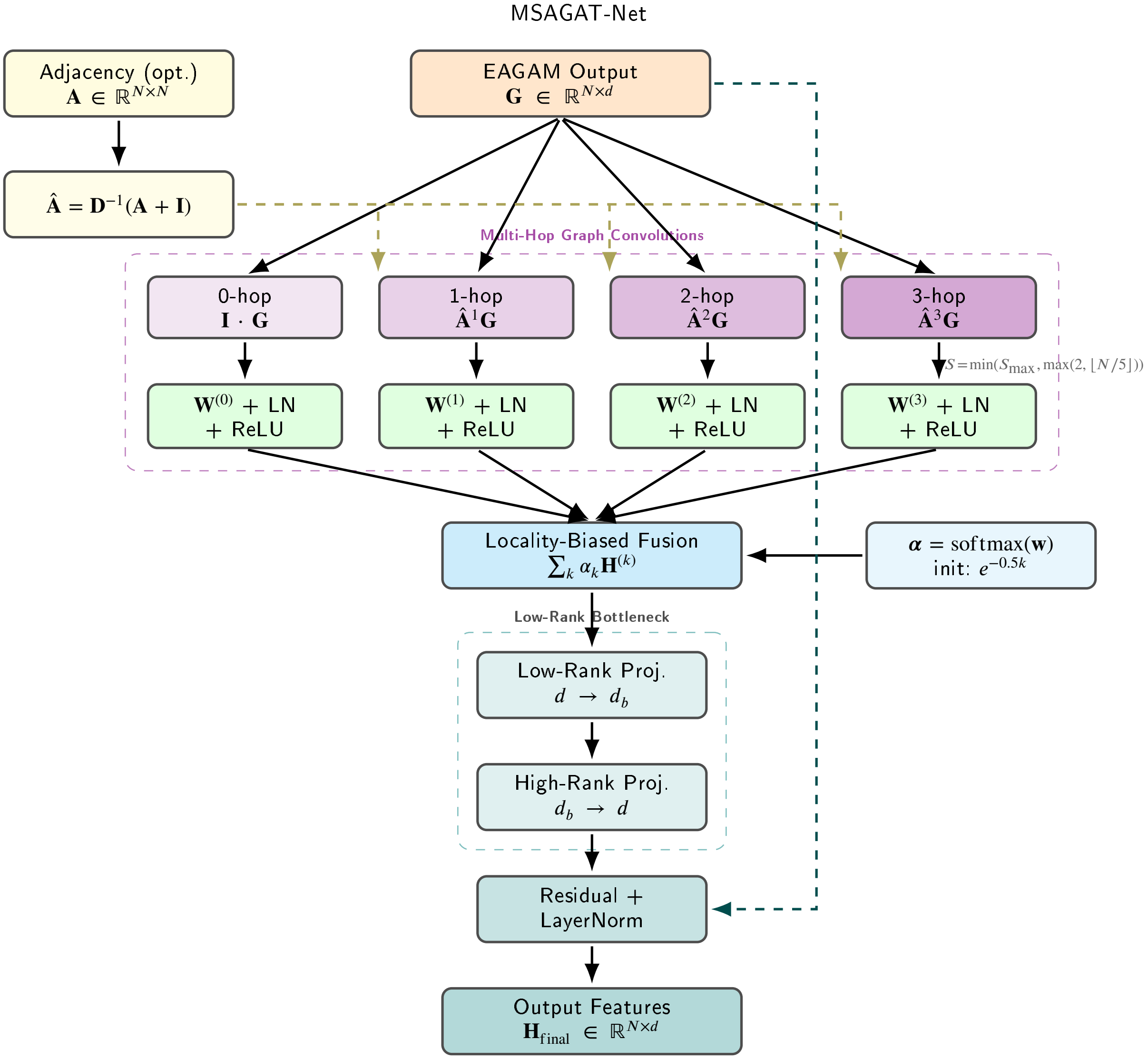

MSAGAT-Net
Adjacency (opt.) $\mathbf{A} \in \mathbb{R}^{N\times N}$
$\hat{\mathbf{A}} = \mathbf{D}^{-1}(\mathbf{A}+\mathbf{I})$
EAGAM Output $\mathbf{G} \in \mathbb{R}^{N\times d}$
Multi-Hop Graph Convolutions
0-hop $\mathbf{I}\cdot\mathbf{G}$
1-hop $\hat{\mathbf{A}}^1\mathbf{G}$
2-hop $\hat{\mathbf{A}}^2\mathbf{G}$
3-hop $\hat{\mathbf{A}}^3\mathbf{G}$
$S = \min(S_{\max}, \max(2, \lfloor N/5 \rfloor))$
$\mathbf{W}^{(0)}$ + LN + ReLU
$\mathbf{W}^{(1)}$ + LN + ReLU
$\mathbf{W}^{(2)}$ + LN + ReLU
$\mathbf{W}^{(3)}$ + LN + ReLU
Locality-Biased Fusion $\sum_k \alpha_k \mathbf{H}^{(k)}$
$\boldsymbol{\alpha} = \mathrm{softmax}(\mathbf{w})$ init: $e^{-0.5k}$
Low-Rank Bottleneck
Low-Rank Proj. $d \rightarrow d_b$
High-Rank Proj. $d_b \rightarrow d$
Residual + LayerNorm
Output Features $\mathbf{H}_{\text{final}} \in \mathbb{R}^{N\times d}$

Figure 4: Data flow in the MSSFM module. EAGAM output features $\mathbf{G}$ are processed through parallel multi-hop graph convolution branches using pre-computed powers of the normalised adjacency matrix ($\hat{\mathbf{A}}^0$ to $\hat{\mathbf{A}}^{S-1}$). Each branch applies a learnable linear transform with LayerNorm and ReLU. The multi-hop features are fused with locality-biased learnable weights (initialised to favour lower hops), then projected through a low-rank bottleneck with a residual connection. The hop depth S adapts to graph size to prevent oversmoothing.

where $\mathbf{R} \in \mathbb{R}^{B\times N\times h}$ represents gate values between 0 and 1 for each region and forecast horizon, and σ denotes the sigmoid activation function.

We use the most recent observation $\mathbf{x}_{\text{last}} \in \mathbb{R}^{B\times N}$ to generate a trend-based forecast using an exponential decay projection with a *learnable* decay rate:

$$\mathbf{T} = \mathbf{x}_{\text{last}} \odot \exp(-\gamma \cdot \mathbf{d}), \quad \gamma = \exp(\gamma_0) \tag{38}$$

where $\mathbf{x}_{\text{last}}$ is expanded to the shape $[B, N, h]$, $\mathbf{d} \in \mathbb{R}^h$ is a vector of increasing horizon indices $[1, 2, \dots, h]$, γ_0 is a learnable scalar parameter stored in log-domain to ensure positivity (initialised to -2.3, yielding $\gamma \approx 0.1$ at the start of training), and $\odot$ represents element-wise multiplication.

This exponential decay formulation is inspired by epidemiological models that exhibit exponential growth or decay patterns. Crucially, the decay rate γ is learned during training rather than fixed, allowing the model to adapt the trend extrapolation speed to each dataset's dynamics—faster decay for rapidly changing epidemics and slower decay for more persistent trends.

The final forecasts are then computed as a weighted combination of the model-based forecasts and the trend-based projections:

$$\mathbf{P}_{\text{final}} = \mathbf{R} \odot \mathbf{P}_{\text{initial}} + (1 - \mathbf{R}) \odot \mathbf{T} \tag{39}$$

where $\mathbf{P}_{\text{final}} \in \mathbb{R}^{B\times N\times h}$ represents the refined forecasts for each region across all forecast horizons.

All PPRM hyperparameters (d_{bottle}, γ_0, dropout) are reported in Table 2. The forecast horizon h is configurable; we evaluate $h \in \{3, 5, 7, 10, 14, 15\}$ across our experiments.

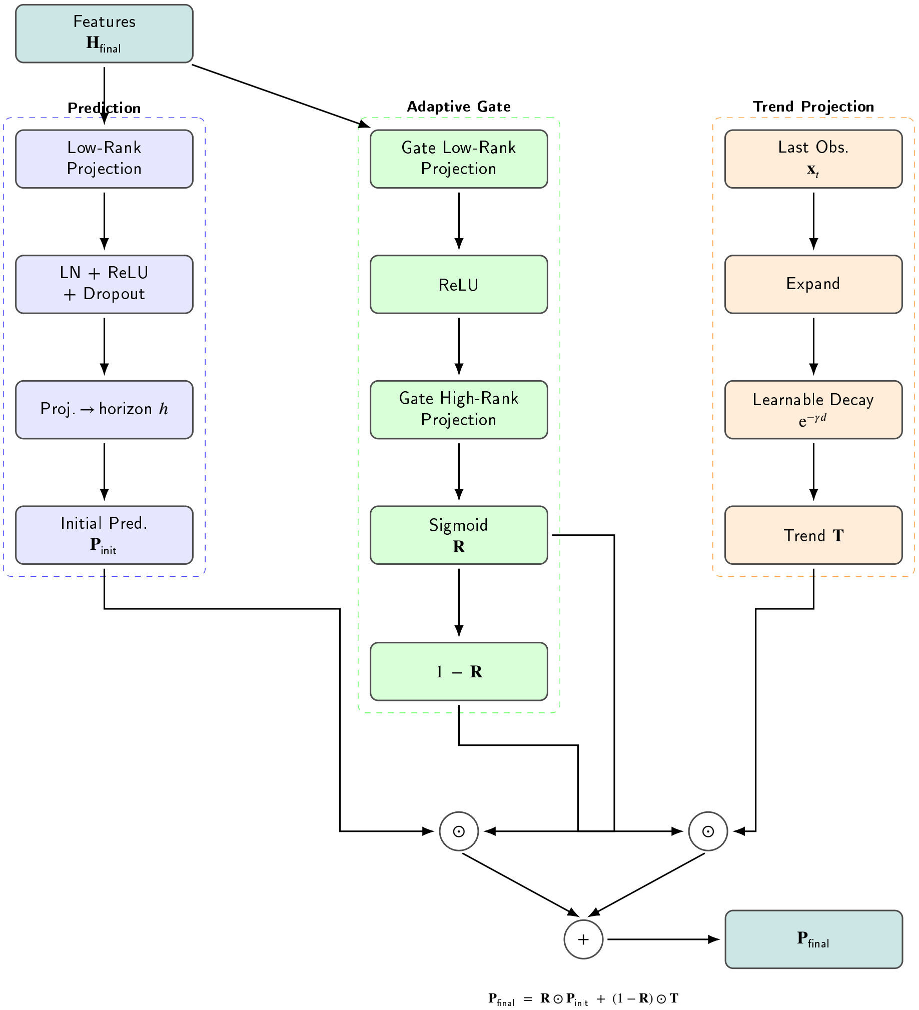


Figure 5: Data flow in the PPRM module. Features are projected through a low-rank bottleneck to generate initial predictions. An adaptive gate learns to balance these model-based forecasts with exponential trend extrapolations based on recent observations.

3.6. Highway Autoregressive Connection

The final output of MSAGAT-Net blends the model's spatiotemporal predictions with a simple autoregressive baseline via a learnable highway connection. Given the last w_h observations (with $w_h = \min(4, w)$), we compute a linear autoregressive forecast:

$$\mathbf{Z}_{\text{ar}} = \text{Linear}(\mathbf{X}_{[t-w_h+1:t]}) \tag{40}$$

where Linear : $\mathbb{R}^{w_h} \rightarrow \mathbb{R}^{h}$ projects recent observations to the forecast horizon for each region. The final prediction is a sigmoid-gated blend:

$$\hat{\mathbf{Y}} = \sigma(\lambda) \cdot \mathbf{P}_{\text{final}} + (1 - \sigma(\lambda)) \cdot \mathbf{Z}_{\text{ar}} \tag{41}$$

where $\mathbf{P}_{\text{final}}$ is the PPRM output, λ is a learnable scalar initialised to 0.5, and σ denotes the sigmoid function. This highway mechanism provides a direct gradient path from input to output, stabilises early training by anchoring predictions to recent observations, and allows the model to gracefully degrade to a simple autoregressive baseline when the learned spatiotemporal features are uninformative.

4. Experimental Setup

We evaluate MSAGAT-Net against strong baselines on six epidemic datasets spanning influenza, COVID-19 case counts, and ICU bed occupancy, using root mean square error (RMSE), Pearson correlation coefficient (PCC), mean absolute error (MAE), and coefficient of determination (R^2).

4.1. Computing Environment

All experiments were conducted on a desktop workstation equipped with an AMD Ryzen 7 7700 8-core processor (3.80 GHz), 32 GB DDR5 RAM, and an NVIDIA GeForce RTX 5060 Ti GPU (16 GB GDDR7), ensuring consistent hardware conditions across all model evaluations.

4.2. Datasets

To comprehensively evaluate the performance and generalisability of our proposed MSAGAT-Net framework, we performed experiments on several real-world epidemic

Table 1
Overview of the epidemic datasets used in our experimental evaluation. "Granularity" indicates the temporal resolution of the epidemic data, whilst "Size" represents the product of the number of locations and the number of time steps.

Dataset	Size	Min	Max	Mean	Granularity
Japan-Prefecture	348×47	0	26,635	655	Weekly
US-Region	785×10	0	16,526	1,009	Weekly
US-State	360×49	0	9,716	223	Weekly
Australia-COVID	556×8	0	9,987	539	Daily
LTLA-COVID	839×372	0	4,170	85	Daily
NHS-ICUBeds	895×7	0	1,215	102	Daily

datasets spanning various geographical regions, time periods, and disease types. This approach enables a thorough assessment of the model's versatility and robustness across varying spatiotemporal characteristics and epidemic scenarios.

Our experimental evaluation encompasses six distinct datasets, each offering unique challenges and characteristics for epidemic forecasting. These datasets represent different geographical scales (from local authorities to national regions), temporal resolutions (daily and weekly measurements), and disease contexts (seasonal influenza and COVID-19). Table 1 provides a statistical overview of these datasets, summarising their key characteristics and numerical properties.

4.2.1. Influenza Datasets

We used three established influenza datasets from different regions to evaluate our model's performance on seasonal patterns:

- **Japan-Prefecture Dataset:** This dataset is derived from the Infectious Disease Weekly Report (IDWR) published by the Japanese government[1]. It comprises weekly statistics of ILI cases from August 2012 to March 2019 in all 47 prefectures in Japan.
- **US-Region Dataset:** Extracted from the ILINet surveillance system maintained by the US Health and Human Services (US-HHS)[2], this dataset includes weekly influenza activity levels in ten HHS regions across the continental United States from 2002 to 2017.
- **US-State Dataset:** Obtained from the Centres for Disease Control and Prevention (CDC), this dataset consists of weekly numbers of visits to healthcare providers with influenza-like illnesses from 2010 to 2017 for 49 states in the US (one state was excluded due to incomplete data).

4.2.2. COVID-19 Datasets

To assess the adaptability of our model to new epidemic scenarios, we incorporated three COVID-19 datasets that span different countries and healthcare metrics:

- **Australia-COVID Dataset:** Compiled from the Johns Hopkins University Centre for Systems Science and Engineering (JHU-CSSE) repository, this dataset contains daily new confirmed cases of COVID-19 from 27 January 2020 to 4 August 2021 across all eight Australian jurisdictions (six states and two territories).
- **LTLA-COVID Dataset:** Derived from the UK Health Security Agency[3], this dataset contains daily data from COVID-19 cases from March 2020 to February 2022 for 372 Lower-Tier Local Authority districts in England. We constructed spatial graph structures for this dataset using geographic proximity, providing a spatiotemporal benchmark for COVID-19 forecasting at the local authority level.
- **NHS-ICUBeds Dataset:** Obtained from the National Health Service (NHS) England[28], this dataset provides daily counts of mechanical ventilator beds occupied in seven regions of the NHS from March 2020 to February 2022. Unlike the other datasets that focus on case counts, this dataset offers an opportunity to evaluate the model's capability to predict healthcare resource utilisation, which is critical for effective epidemic response and management. We constructed spatial connectivity structures for this dataset, addressing the gap in spatially-structured healthcare resource forecasting benchmarks.

4.3. Spatial Graph Construction

Following the established methodology of STAN [11], we construct spatial graph structures to capture epidemic transmission patterns between geographic regions using geographic proximity as the primary criterion for establishing spatial relationships. For the LTLA-COVID and NHS-ICUBeds datasets, which previously lacked predefined spatial connectivity structures, we developed these spatial graphs to enable spatiotemporal modeling of these publicly available epidemic data sources.

For our implementation, we constructed the adjacency matrix based on geographic proximity, using the Haversine formula to calculate the great circle distance between regions, consistent with established practices in spatiotemporal epidemic modelling. Two regions are considered connected if the distance between them falls below a threshold $d_{\text{threshold}}$ (set to 150 km in our experiments):

[1] https://tinyurl.com/y5dt7stm

[2] https://tinyurl.com/y39tog3h

[3] https://ukhsa-dashboard.data.gov.uk/respiratory-viruses/covid-19

$$a_{ij} = \begin{cases} 1, & \text{if Haversine}(\text{region}_i, \text{region}_j) \leq d_{\text{threshold}} \\ 0, & \text{otherwise} \end{cases} \quad (42)$$

This threshold-based connectivity captures the intuition that epidemic spread is influenced by the movement of people between nearby regions. Although more sophisticated connectivity measures could be employed, this approach provides a straightforward and interpretable baseline for spatial relationship modelling. The noise in the dataset was smoothed using the rolling mean of 7 days established in previous studies [2, 29, 47, 17], and normalisation was performed to ensure that the data are on a similar scale in different regions.

The diverse nature of these datasets, spanning different geographic regions, temporal resolutions, and epidemic contexts, allows us to comprehensively evaluate the performance and generalisability of our proposed MSAGAT-Net model across a range of epidemic forecasting scenarios.

4.4. Training and Optimisation Strategy

The MSAGAT-Net model is trained using the Adam optimiser with a learning rate of 1×10^{-3} and a batch size of 32, which were determined through preliminary hyperparameter tuning to provide optimal convergence speed and stability. The model is trained for a maximum of 1500 epochs, with early stopping criteria based on validation loss to prevent overfitting. The training process is monitored using a patience parameter of 100 epochs, which means that if the validation loss does not improve for 100 consecutive epochs, the training will be stopped. The loss function for the MSAGAT-Net model is a combination of forecast error and regularisation terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{forecast}} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}} + \lambda_{12}\|\Theta\|_2 \quad (43)$$

where $\mathcal{L}_{\text{forecast}}$ is the mean squared error measuring discrepancies between the model forecasts and the observed data, and $\mathcal{L}_{\text{attn}}$ represents the attention regularisation term that enforces sparsity and interpretability in spatial relationships. The hyperparameters λ_{attn} and λ_{12} control the strength of attention and L2 regularisation, respectively. Following prior work on graph attention and spatiotemporal forecasting networks [37, 48, 44, 9, 41], we initialise λ_{attn} at 10^{-5} and optimise it as a learnable parameter during training, allowing the model to adaptively balance forecast accuracy with attention sparsity. In contrast, $\lambda_{12} = 5 \times 10^{-4}$ remains fixed throughout training, consistent with established practices that balance generalisation and numerical stability.

For all datasets, we employ a sliding window of 20 time steps and split the data into training, validation, and test sets with a ratio of 60%:20%:20%. The complete training procedure is formalised in Algorithm 1. Table 2 summarises all architectural hyperparameters.

Table 2
MSAGAT-Net architectural hyperparameters. All values are shared across datasets. Parameters marked † are learnable; the listed value is the initialisation.

Module	Hyperparameter	Value
TFEM	Feature channels d_{feat}	16
	Kernel size	3
	Bottleneck dim d_{bottle}	8
	Hidden dim d_{hidden}	32
EAGAM	Attention heads	4
	Bottleneck dim d_{bottle}	8
	Attn. reg. weight $\lambda^{\dagger}_{\text{attn}}$	10^{-5}
	Adj. prior scale $\alpha^{\dagger}_{0}$	1.0
MSSFM	Max hops $S_{\max}$	4
PPRM	Bottleneck dim d_{bottle}	8
	Decay $\gamma^{\dagger}_{0}$	−2.3
	Dropout	0.2
Highway	Window w_h	$\min(4, w)$
	Gate $\lambda^{\dagger}$	0.5

4.5. Baseline Models

We compare MSAGAT-Net against several state-of-the-art baseline models widely used in epidemic forecasting:

- **DCRNN** [22]: A diffusion convolution recurrent neural network that integrates graph convolutions with recurrent neural networks in an encoder-decoder architecture to capture both spatial dependencies and temporal dynamics. It models spatial dependencies using a diffusion process on graphs and temporal dependencies through recurrent units. *Requires a predefined adjacency matrix to perform diffusion convolutions.*

- **LSTNet** [20]: A model that combines convolutional neural networks and recurrent neural networks to extract short-term local dependency patterns and discover long-term patterns for time-series trends. It employs a convolutional component to extract local dependency patterns and a recurrent component to capture long-term temporal dependencies. *Does not model explicit spatial structure.*

- **CNNRNN-Res** [43]: A deep learning framework that combines convolutional neural networks, recurrent neural networks, and residual connections to solve epidemiological prediction problems. It uses CNNs to extract spatial features, RNNs to capture temporal dependencies, and residual connections to enhance gradient flow during training. *Does not model explicit spatial structure.*

- **Cola-GNN** [9]: A graph neural network model that leverages cross-location attention mechanisms to capture dynamic spatial relationships between regions. It employs location-aware attention to model the impact

Algorithm 1: MSAGAT-Net Training Algorithm

```
Input: Training data $\mathcal{D}_{\text{train}}$, validation data $\mathcal{D}_{\text{val}}$
Output: Optimized model parameters $\mathbf{\Theta}^*$
Initialize model parameters $\mathbf{\Theta}$ (including graph bias
  $\mathbf{U}, \mathbf{V}$) and Adam optimizer with weight decay $\lambda_{l2}$;
$L_{\text{best}} \leftarrow \infty, p \leftarrow 0$ ;   // Best validation loss and
  patience counter
for epoch $e = 1$ to $E_{\max}$ do
    foreach mini-batch $(\mathbf{X}, \mathbf{y})$ in $\mathcal{D}_{train}$ do
        $\mathbf{F} \leftarrow \text{TFEM}(\mathbf{X})$ ;  // Depthwise sep. conv +
          bottleneck
        $\mathbf{G}, \mathcal{L}_{\text{attn}} \leftarrow \text{EAGAM}(\mathbf{F})$ ;       // Scaled
          dot-product + structural bias
        $\mathbf{H} \leftarrow \text{MSSFM}(\mathbf{G})$ ; // Multi-hop graph conv
          + fusion
        $\mathbf{P} \leftarrow \text{PPRM}(\mathbf{H}, \mathbf{x}_{\text{last}})$ ;       // Progressive
          refinement
        $\mathbf{Z}_{\text{ar}} \leftarrow \text{Linear}(\mathbf{X}_{[t-w_h+1:t]})$ ;
          // Autoregressive baseline
        $\hat{\mathbf{Y}} \leftarrow \sigma(\lambda)\mathbf{P} + (1 - \sigma(\lambda))\mathbf{Z}_{\text{ar}}$ ;     // Highway
          blend
        $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{MSE}}(\hat{\mathbf{Y}}, \mathbf{y}) + \mathcal{L}_{\text{attn}}$ ; // L2 handled
          by optimizer
        Update $\mathbf{\Theta}$ using gradient descent on $\mathcal{L}_{\text{total}}$;
    $L_{\text{val}} \leftarrow \text{Evaluate}(\mathcal{D}_{\text{val}}, \mathbf{\Theta})$ ;         // Compute
      validation loss
    if $L_{val} < L_{best}$ then
        $\mathbf{\Theta}^* \leftarrow \mathbf{\Theta}$, $L_{\text{best}} \leftarrow L_{\text{val}}$, $p \leftarrow 0$;
    else
        $p \leftarrow p + 1$;
        if $p \geq P_{max}$ then
            break;
return $\mathbf{\Theta}^*$
```

of each region on others, allowing for adaptive and context-dependent spatial dependency learning. *Requires a predefined adjacency matrix to initialise and constrain the attention mechanism.*

- **EpiGNN** [44]: A model based on graph neural networks specifically designed for epidemic forecasting. It incorporates a transmission risk encoding module to characterise local and global spatial effects, and features a Region-Aware Graph Learner (RAGL) that considers transmission risk, geographical dependencies, and temporal information to explore spatiotemporal dependencies. *Requires a predefined adjacency matrix (specified via the* `-sim_mat` *parameter) to encode geographical dependencies.*

These baselines represent a diverse range of approaches to spatiotemporal forecasting. A key distinction is that all graph-based baselines (DCRNN, Cola-GNN, EpiGNN) require predefined adjacency matrices as mandatory input, whereas MSAGAT-Net learns spatial relationships through learnable graph bias parameters **U** and **V** and can optionally incorporate adjacency as a self-regulating soft prior (Section 3.3.4). In our experiments, MSAGAT-Net uses the adjacency prior when available, but ablation results (Section 5) confirm competitive performance without it.

5. Results and Discussion

Table 3 presents a comprehensive comparison of our proposed MSAGAT-Net model against state-of-the-art baseline approaches across three influenza datasets (Japan-Prefectures, US-Regions, and US-States) and four forecast horizons (3, 5, 10, and 15 days ahead). Furthermore, Table 4 shows the performance comparison on three COVID-19 datasets (Australia-COVID, LTLA-TimeSeries, and NHS-TimeSeries) for horizons of 3, 7, and 14 days ahead.

5.0.1. Performance on Influenza Datasets

The Japan-Prefectures dataset provides strong validation of MSAGAT-Net's architectural design. Whilst EpiGNN achieves the best short-term performance (3-day RMSE: 1272), MSAGAT-Net delivers the best results for medium- and long-term horizons (5-day RMSE: 1437, 10-day: 1584, 15-day: 1550), consistently outperforming all baselines from horizon 5 onwards. MSAGAT-Net also achieves the highest correlation coefficients at these extended horizons (PCC: 0.852, 0.831, 0.816), demonstrating that the model's learnable graph structure, additive structural-bias attention, and multi-hop spatial refinement effectively capture the complex spatiotemporal dynamics of influenza transmission between Japanese prefectures, particularly for longer-range forecasting where spatial dependencies become increasingly important.

The attention visualisation in Figure 7 reveals why MSAGAT-Net succeeds on this dataset: the learned attention patterns diverge substantially from both geographical adjacency and simple correlation structures, identifying epidemic-relevant connections that likely reflect commuter flows, air travel routes, and socioeconomic linkages rather than mere proximity. This data-driven discovery of transmission pathways, enabled by the learnable graph bias parameters (**U**, **V**), represents a key advantage over baselines that rely on predefined spatial structures.

Performance on US-Regions and US-States reveals nuanced patterns. On US-Regions, MSAGAT-Net achieves the best short-term RMSE (3-day: 659) and competitive 5-day performance (902), whilst Cola-GNN dominates at longer horizons (10-day: 888, 15-day: 1063). On US-States, MSAGAT-Net achieves the best RMSE across all four horizons (3-day: 163, 5-day: 197, 10-day: 226, 15-day: 245), consistently outperforming EpiGNN and Cola-GNN. However, PCC analysis reveals that Cola-GNN achieves the highest correlation for 10-day and 15-day US-States forecasts (0.896, 0.907), suggesting that when explicit graph structure is available, Cola-GNN's attention mechanism can produce better-correlated predictions even when absolute errors are higher. The US-Regions 15-day horizon remains a weakness (RMSE: 1334, PCC: 0.562), likely due to the small node

Table 3
RMSE and PCC performance of different methods on three datasets (horizon = 3, 5, 10, 15). Bold = best, underline = second best.

Method	Metric	Japan–Prefectures				US–Regions				US–States			
		3	5	10	15	3	5	10	15	3	5	10	15
DCRNN	RMSE	2345	2610	2579	2284	1116	1381	1652	1666	235	292	340	366
	PCC	0.424	0.238	0.555	0.515	0.791	0.684	0.573	0.565	0.899	0.843	0.823	0.804
LSTNet	RMSE	2360	2595	2470	1999	965	1139	1224	1143	247	297	354	380
	PCC	0.392	0.190	0.416	0.567	0.802	0.730	0.636	0.683	0.855	0.775	0.777	0.750
CNNRNN-Res	RMSE	2289	2653	2710	1943	874	1154	1373	1277	288	328	313	360
	PCC	0.445	0.102	0.255	0.611	0.834	0.696	0.579	0.632	0.801	0.744	0.817	0.765
Cola-GNN	RMSE	1548	1772	1802	1836	774	1030	**888**	**1063**	203	228	271	280
	PCC	0.845	0.798	0.792	0.705	0.882	0.804	**0.840**	**0.800**	0.914	0.902	**0.896**	**0.907**
EpiGNN	RMSE	**1272**	1647	1806	1692	689	852	1169	1224	187	207	240	248
	PCC	**0.880**	0.813	0.724	0.732	0.898	**0.860**	0.723	0.716	0.923	**0.907**	0.884	0.888
MSAGAT-Net	RMSE	1385	**1437**	**1584**	**1550**	**659**	**902**	962	1334	**163**	**197**	**226**	**245**
	PCC	0.836	**0.852**	**0.831**	**0.816**	**0.907**	0.816	0.800	0.562	**0.936**	0.907	0.885	0.876

Table 4
RMSE and PCC performance of different methods on three COVID-19 datasets (horizon = 3, 7, 14). Bold = best, underline = second best.

Method	Metric	Australia-COVID			LTLA-Timeseries			NHS-Timeseries		
		3	7	14	3	7	14	3	7	14
DCRNN	RMSE	230	360	634	64.5	109	145	5.10	9.07	15.7
	PCC	0.996	0.985	0.972	0.908	0.707	0.440	0.977	0.932	0.830
LSTNet	RMSE	367	**248**	**298**	57.5	115	149	3.66	10.2	17.9
	PCC	0.984	**0.994**	**0.992**	0.927	0.663	0.430	0.988	0.907	0.669
CNNRNN-Res	RMSE	**153**	300	352	121	131	140	4.20	9.77	**13.0**
	PCC	0.998	0.991	0.986	0.612	0.528	0.449	0.984	0.942	0.844
Cola-GNN	RMSE	265	362	575	70.6	247	143	**3.15**	9.56	14.2
	PCC	0.997	0.993	0.985	0.886	0.164	0.453	0.991	0.945	0.840
EpiGNN	RMSE	318	387	463	155	144	157	6.74	8.11	13.2
	PCC	0.995	0.992	0.987	0.530	0.501	0.448	0.962	0.944	**0.869**
MSAGAT-Net	RMSE	226	372	650	**49.8**	**83.0**	**116**	3.45	**6.31**	13.9
	PCC	**0.998**	0.994	0.986	**0.945**	**0.842**	**0.653**	**0.993**	**0.965**	0.869

count (10 regions) limiting the benefit of MSAGAT-Net's adaptive multi-hop spatial learning.

5.0.2. Performance on COVID-19 Datasets

COVID-19 datasets reveal context-dependent performance patterns that illuminate the conditions under which MSAGAT-Net's architectural choices confer advantages. On LTLA-Timeseries (372 nodes, highest spatial resolution), MSAGAT-Net achieves the best performance across all horizons (3-day RMSE: 49.8, 7-day: 83.0, 14-day: 116), substantially outperforming all baselines. The improvements are particularly striking: 13.5% lower RMSE than LSTNet (second-best) at 3-day, 23.5% lower than DCRNN (second-best) at 7-day, and 17.3% lower than CNNRNN-Res (second-best) at 14-day. MSAGAT-Net also achieves the highest PCC across all horizons (0.945, 0.842, 0.653), confirming that structural-bias attention and multi-hop spatial refinement are highly effective at larger spatial scales where the graph structure encodes meaningful transmission patterns between local authorities.

The Australia-COVID dataset (8 nodes, 556 timesteps) reveals a different dynamic. CNNRNN-Res achieves the best 3-day RMSE (153), whilst LSTNet dominates at 7-day (248) and 14-day (298) horizons. MSAGAT-Net performs competitively (3-day RMSE: 226, second-best) but falls behind at longer horizons (7-day: 372, 14-day: 650). This pattern is informative: Australia's COVID-19 response involved strict state border closures and aggressive containment, effectively decoupling spatial transmission. With only 8 nodes and minimal inter-regional coupling, the dataset offers insufficient spatial complexity to benefit from graph-based modelling. The success of simpler models (LSTNet, CNNRNN-Res) confirms that purely temporal approaches are optimal when spatial dependencies are suppressed by policy interventions.

On NHS-Timeseries (7 nodes, ICU bed occupancy), MSAGAT-Net achieves the best RMSE at 7-day (6.31) and strong performance at the other horizons: 3-day RMSE 3.45 (second to Cola-GNN's 3.15) and 14-day RMSE 13.9 (competitive with CNNRNN-Res's 13.0 and EpiGNN's 13.2). MSAGAT-Net achieves the highest PCC at 3-day (0.993) and 7-day (0.965), whilst EpiGNN leads at 14-day (0.869 vs 0.869, effectively tied). This strong performance on a very small graph (7 nodes) suggests that MSAGAT-Net's adaptive architecture effectively handles both large and small spatial scales, leveraging temporal dynamics when spatial complexity is limited. The shift from case counts to resource utilisation introduces different temporal dynamics (administrative delays, capacity constraints), but MSAGAT-Net's progressive prediction refinement module appears well-suited to capture these patterns.

5.0.3. Synthesis: When MSAGAT-Net Succeeds

Cross-dataset analysis reveals that MSAGAT-Net achieves the best or second-best RMSE in the majority of experimental settings across all six datasets. The model dominates on LTLA-Timeseries (best at all horizons), US-States (best RMSE at all four horizons), and Japan-Prefectures (best from 5-day onward), with strong performance on NHS-Timeseries (best at 7-day, competitive at 3- and 14-day). This demonstrates strong generalisation across both influenza and COVID-19 contexts. Performance is weakest on Australia-COVID, where the combination of few nodes (8) and suppressed spatial coupling due to border closures limits the benefit of graph-based spatial learning. The consistent strength on datasets with moderate-to-large spatial scale (47–372 nodes) confirms that MSAGAT-Net's adaptive spatial learning and multi-hop refinement are most effective when sufficient spatial complexity exists to discover meaningful transmission pathways.

5.0.4. Qualitative Forecast Analysis

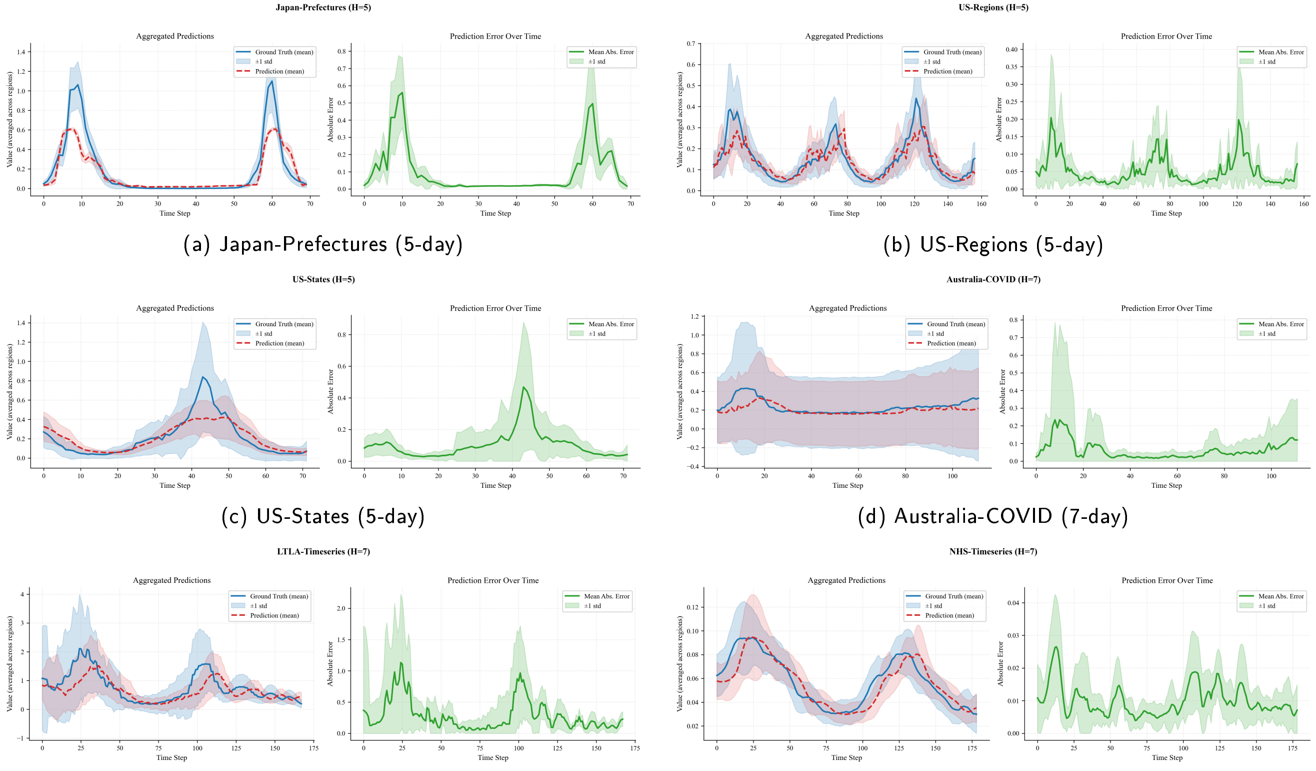
Figure 6 presents an aggregated view of MSAGAT-Net's forecasting performance across all six datasets, showing the mean prediction across all regions with ±1 standard deviation confidence bands (left panels) and the temporal distribution of absolute prediction errors (right panels). Across all datasets, the predicted mean closely follows the ground truth, with narrow error bands during quiescent periods and wider bands during epidemic peaks—consistent with the inherent difficulty of forecasting rapid case surges. The model accurately tracks sharp seasonal peaks in Japan-Prefectures and US-States influenza, irregular COVID-19 waves in Australia and LTLA, and ICU bed occupancy patterns in NHS-Timeseries. The error plots reveal that prediction errors are temporally concentrated around peak periods rather than uniformly distributed, suggesting that MSAGAT-Net captures the overall epidemic trajectory well but faces challenges during the most dynamic phases of outbreaks. The close correspondence between predicted and actual values confirms that the quantitative improvements reported in Tables 3 and 4 translate into visually meaningful forecast quality.

5.0.5. Learned Spatial Representations

Figures 7, 8, and 9 provide visual evidence of EAGAM's role in discovering epidemic-relevant spatial structure. The complete model (Figure 7) learns attention patterns that exhibit three key characteristics: (1) strong diagonal dominance indicating local temporal persistence, (2) structured off-diagonal patterns capturing inter-prefectural dependencies, and (3) substantial divergence from both geographical adjacency (left panel) and simple input correlation (center panel). This divergence is critical—it demonstrates that EAGAM does not merely replicate predefined structures but discovers latent transmission pathways informed by actual epidemic dynamics rather than geographic or statistical proxies.

Qualitatively, the learned attention matrix shows low similarity to the adjacency matrix, suggesting that epidemic transmission does not align with simple geographical proximity in modern Japan where bullet trains, air travel, and economic corridors create complex mobility patterns. The attention patterns reveal long-range connections between major metropolitan areas (Tokyo, Osaka, Nagoya) that would be missed by adjacency-based methods, whilst also capturing local clustering in rural prefectures—a nuanced spatial structure that emerges from the data rather than being imposed a priori.

The ablation variants confirm EAGAM's architectural necessity. Removing PPRM (Figure 8) produces nearly identical attention patterns, validating that PPRM operates on prediction refinement rather than spatial learning—the attention mechanism remains intact because EAGAM is preserved. In stark contrast, removing EAGAM (Figure 9) forces the model to use an identity spatial module (diagonal pattern in right panel), where each prefecture attends only to itself with zero inter-regional information flow. This ablation variant replaces EAGAM with a minimal identity pass-through (see Section 5.1), eliminating all spatial dependency modelling. The ablation results (Table 5) show that EAGAM removal produces consistent degradation across all horizons, with the strongest impact at 7-day (+17.11% RMSE, −22.47% R^2), confirming that adaptive spatial attention is the most critical component of MSAGAT-Net.



(a) Japan-Prefectures (5-day)

(b) US-Regions (5-day)

(c) US-States (5-day)

(d) Australia-COVID (7-day)

(e) LTLA-Timeseries (7-day)

(f) NHS-Timeseries (7-day)

Figure 6: Aggregated prediction performance across all regions for each dataset. Left panels: mean prediction (red dashed) vs. ground truth (blue solid) with ±1 standard deviation bands. Right panels: temporal distribution of absolute prediction errors, showing that errors concentrate around epidemic peaks.

These visualisations collectively demonstrate that EAGAM's learnable graph bias mechanism ($\mathbf{B} = \mathbf{UV}$ with low-rank factorisation) successfully discovers and exploits non-obvious spatial relationships that predefined adjacency matrices alone cannot capture. Notably, the magnitude of the learned graph bias varies across datasets: on Japan-Prefectures, $\mathbf{B}$ develops pronounced structure reflecting inter-prefectural transmission pathways, whereas on datasets where content-based attention ($\mathbf{QK}^{\top}$) and the adjacency prior already capture the dominant spatial dependencies, the graph bias remains near its initialisation. This behaviour reflects the model's ability to adaptively allocate representational capacity—relying more on structural bias when data-driven spatial relationships are informative, and defaulting to content-based attention when they are not. This capability is particularly valuable for epidemic forecasting where true transmission networks rarely align with administrative boundaries or simple proximity measures.

5.1. Ablation Study

To evaluate the contribution of each key component in MSAGAT-Net, we conducted a comprehensive ablation study on the Japan-Prefectures dataset. In each experiment, one component was systematically replaced with a minimal identity pass-through module while the others remained intact. Table 5 summarises the performance across multiple horizons, offering insight into the relative importance of each architectural module.

The ablation results reveal clear horizon-dependent patterns in component importance.

EAGAM removal produces the most consistent and substantial degradation across all horizons, confirming spatial attention as the most critical architectural component. RMSE increases by +6.43% at 3-day, +17.11% at 7-day, and +3.85% at 14-day, with corresponding PCC drops of −3.07%, −8.86%, and −3.90%. The 7-day horizon shows the most dramatic impact (−22.47% R^2), indicating that EAGAM's adaptive spatial attention is most valuable at medium-range forecasting horizons where inter-regional transmission dynamics are strongest. Replacing EAGAM with an identity spatial module forces each prefecture to attend only to itself, eliminating all inter-regional information flow and confirming that the learned spatial structure is essential for capturing epidemic diffusion patterns.

MSSFM removal has a modest but consistent impact at medium and long horizons. At 3-day, removing MSSFM marginally improves RMSE (−1.34%), suggesting that multi-scale spatial aggregation is unnecessary for very short-term forecasts. However, at 7-day (+1.71% RMSE) and 14-day (+1.19% RMSE, −1.73% PCC), MSSFM contributes measurable improvements, confirming that multi-hop spatial feature extraction becomes more valuable as

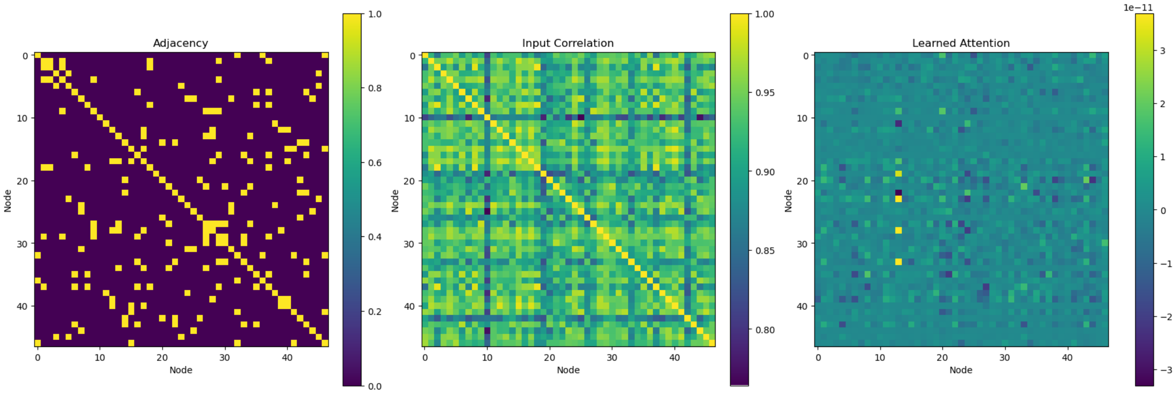


Figure 7: Attention matrices learned by MSAGAT-Net on the Japan-Prefectures dataset for 7-day forecasting: adjacency matrix (left), input correlation (center), and learned attention (right). The learned attention diverges substantially from both geographical adjacency and simple input correlation, discovering epidemic-relevant transmission pathways.

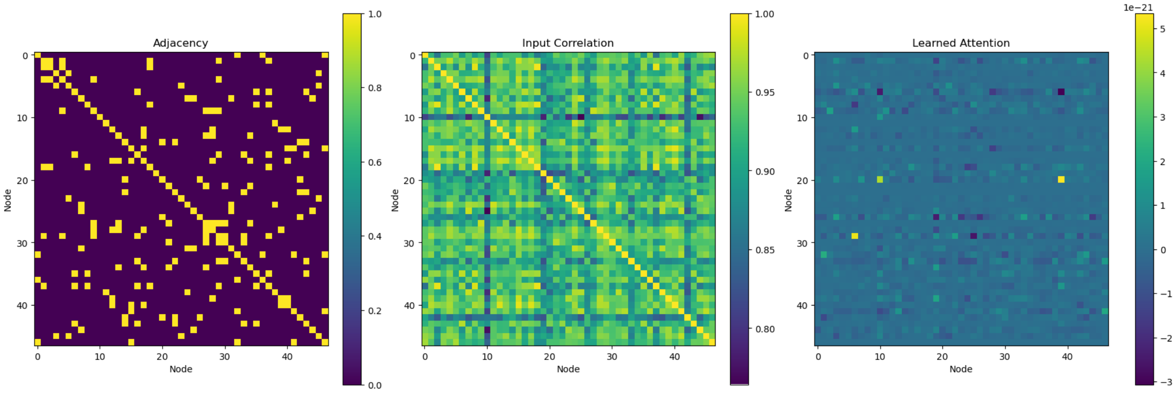


Figure 8: Attention matrices learned by MSAGAT-Net without PPRM on the Japan-Prefectures dataset for 7-day forecasting. The attention patterns remain nearly identical to the full model, confirming that PPRM operates on prediction refinement rather than spatial learning.

the forecast horizon extends and spatial patterns at different graph distances must be captured.

PPRM removal reveals a clear horizon-dependent pattern. At short horizons, removing PPRM and replacing it with a single linear layer actually *improves* performance (−5.74% RMSE at 3-day, −2.23% at 7-day), suggesting that progressive refinement adds unnecessary complexity when the forecast horizon is short enough for direct prediction

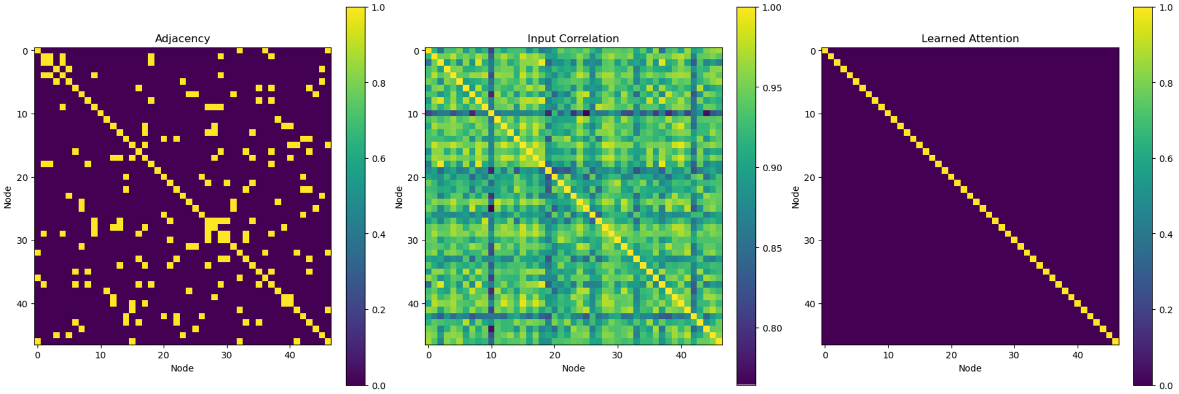


Figure 9: Attention matrices when EAGAM is replaced with an identity spatial module on the Japan-Prefectures dataset for 7-day forecasting. The learned attention collapses to a diagonal identity pattern (right panel), confirming zero inter-regional information flow and explaining the +17.1% RMSE degradation.

Table 5
Ablation study results on the Japan-Prefectures dataset, showing the impact of removing key components of MSAGAT-Net on forecasting performance across different horizons.

Model Variant	Metric	3-day Horizon		7-day Horizon		14-day Horizon	
		Value	% Change	Value	% Change	Value	% Change
Full Model	MAE	504.95	–	635.80	–	627.47	–
	RMSE	1385.14	–	1502.17	–	1560.24	–
	PCC	0.836	–	0.855	–	0.824	–
	R^2	0.680	–	0.623	–	0.593	–
Without EAGAM	MAE	571.08	(+13.10%)	690.49	(+8.60%)	690.19	(+10.00%)
	RMSE	1474.16	(+6.43%)	1759.26	(+17.11%)	1620.27	(+3.85%)
	PCC	0.810	(−3.07%)	0.779	(−8.86%)	0.792	(−3.90%)
	R^2	0.637	(−6.26%)	0.483	(−22.47%)	0.562	(−5.37%)
Without MSSFM	MAE	475.61	(−5.81%)	597.69	(−6.00%)	652.86	(+4.05%)
	RMSE	1366.58	(−1.34%)	1527.80	(+1.71%)	1578.81	(+1.19%)
	PCC	0.841	(+0.59%)	0.869	(+1.58%)	0.810	(−1.73%)
	R^2	0.688	(+1.26%)	0.610	(−2.08%)	0.584	(−1.64%)
Without PPRM	MAE	465.86	(−7.74%)	561.89	(−11.63%)	636.21	(+1.39%)
	RMSE	1305.60	(−5.74%)	1468.71	(−2.23%)	1612.34	(+3.34%)
	PCC	0.856	(+2.42%)	0.867	(+1.38%)	0.792	(−3.91%)
	R^2	0.715	(+5.26%)	0.640	(+2.66%)	0.566	(−4.65%)

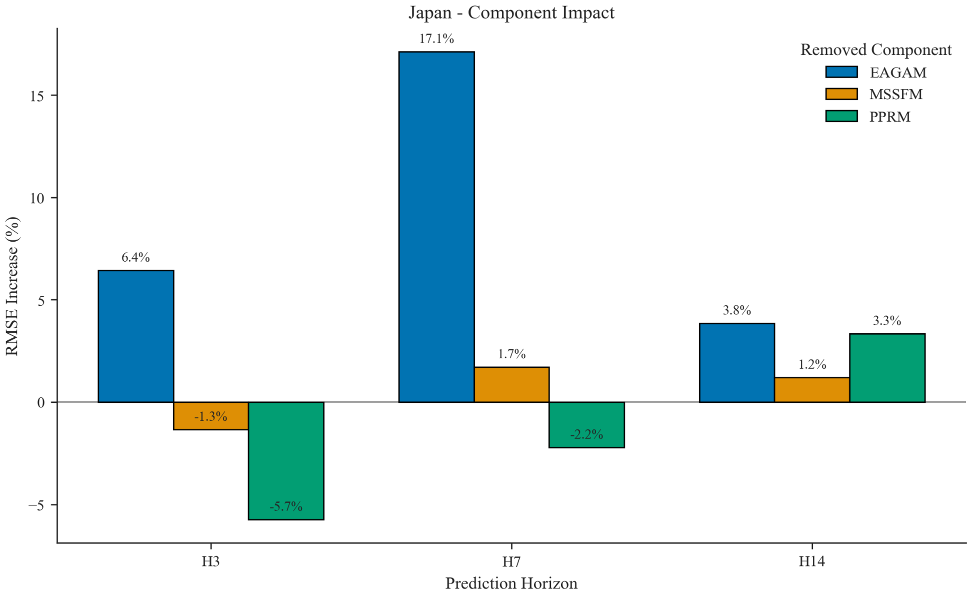


Figure 10: Component impact on RMSE across three forecast horizons for the Japan-Prefectures dataset. Positive bars indicate degradation upon removal; negative bars indicate improvement.

to suffice. However, at 14-day horizon, PPRM removal degrades RMSE by +3.34% and PCC by −3.91%, confirming that progressive prediction refinement becomes important for longer-range forecasts where iterative correction of initial trend estimates produces more accurate predictions.

These patterns demonstrate that component importance is fundamentally horizon-dependent. EAGAM is consistently essential across all horizons, PPRM's value increases with forecast length, and MSSFM provides modest but consistent benefits at medium-to-long horizons. The results suggest that all three components contribute to MSAGAT-Net's overall effectiveness, with their relative importance shifting predictably with the forecasting task difficulty. Figure 10 provides a visual summary that clearly illustrates these contrasting horizon-dependent patterns across all three components.

A key design choice distinguishing MSAGAT-Net from baselines is its ability to learn spatial relationships directly from epidemic data. To validate this, we compared MSAGAT-Net with and without the geographical adjacency prior on the Japan-Prefectures dataset (10-day horizon). The model *without* the adjacency prior achieved RMSE of 1562 and PCC of 0.850, compared with 1584 and 0.831 for the full model that includes it. That is, removing the adjacency prior yields marginally *better* performance on this dataset, confirming that the learnable graph bias parameters (**U**, **V**) are sufficient to capture spatial dependencies without geographical priors. This is a practical advantage over baselines such as EpiGNN, Cola-GNN, and DCRNN, which *require* predefined adjacency matrices.

6. Conclusion

This paper introduces MSAGAT-Net, a multi-scale adaptive graph attention network for epidemic forecasting with two core novel mechanisms: (1) *self-regulating additive structural bias attention* (EAGAM), where a learnable low-rank graph bias and an optional adjacency prior are added directly to attention scores before softmax normalisation—softmax shift-invariance causes this prior to automatically modulate its influence based on graph density, eliminating fragile graph-size thresholds without manual tuning; and (2) *adaptive multi-hop graph convolutions* (MSSFM) with graph-size-dependent hop depth and locality-biased fusion to prevent oversmoothing. These are complemented by progressive prediction refinement with a learnable exponential decay rate and a highway autoregressive connection that together stabilise multi-horizon forecasts.

A key contribution of our work is demonstrating that effective spatiotemporal epidemic forecasting is achievable without predefined graph structures. Unlike state-of-the-art baselines (EpiGNN, Cola-GNN, DCRNN) that require adjacency matrices constructed from geographical proximity or external mobility data, MSAGAT-Net learns spatial relationships through learnable graph bias parameters (**U**, **V**), while optionally incorporating adjacency priors when available. This eliminates the need for domain expertise or external data sources to construct spatial graphs, and enables discovery of non-obvious transmission pathways that may not correspond to geographical proximity.

Comprehensive evaluation across six diverse epidemiological datasets demonstrates that MSAGAT-Net achieves the best RMSE in the majority of experimental settings. On the LTLA-Timeseries dataset (372 local authorities), MSAGAT-Net reduces RMSE by 13.5–23.5% over the second-best baselines across all horizons. On NHS-Timeseries, the model achieves the best 7-day RMSE (6.31) and competitive performance at other horizons. On US-States, MSAGAT-Net achieves the lowest RMSE at all four horizons (3, 5, 10, 15 days), and on Japan-Prefectures it delivers the best results from 5-day horizons onward. However, our results reveal a critical insight: optimal forecasting architectures are inherently context-dependent. On datasets with weak spatial coupling (e.g., Australia-COVID with 8 nodes and border closures), simpler models such as LSTNet and CNNRNN-Res can outperform more complex spatiotemporal architectures.

Ablation studies on the Japan-Prefectures dataset reveal clear hierarchy and horizon-dependent patterns in component importance. EAGAM is the most critical component, with removal causing consistent degradation across all horizons (up to +17.11% RMSE at 7-day), confirming the essential role of adaptive spatial attention. PPRM shows a horizon-dependent contribution: unnecessary for short-term forecasts but important at longer horizons (+3.34% RMSE at 14-day upon removal). MSSFM provides modest but consistent improvements at medium-to-long horizons. These findings demonstrate that each architectural component contributes meaningfully, with EAGAM being universally essential and the spatial refinement (MSSFM) and prediction (PPRM) modules becoming increasingly important as forecast difficulty increases.

The practical implications extend beyond technical contributions. Effective forecasting systems should employ disease-specific and horizon-adaptive architectures rather than one-size-fits-all approaches. Future research directions include automatic neural architecture search for epidemic-specific optimisation, integration of dynamic external factors such as mobility patterns and policy interventions, incorporation of physics-informed constraints and disease-specific mechanistic knowledge to improve interpretability and long-range forecast stability, and extension to multivariate forecasting encompassing hospitalisation rates and healthcare resource utilisation.

The parameter efficiency of MSAGAT-Net through low-rank projections, combined with its flexible spatial modelling, makes it well-suited for diverse deployment scenarios. While all main experiments incorporate the adjacency prior as a soft structural guide, ablation experiments confirm that the model achieves comparable or marginally better performance without any predefined adjacency input, demonstrating the sufficiency of data-driven spatial learning. This flexibility is particularly valuable in settings where geographical proximity may not reflect true transmission patterns, or where external data for graph construction are unavailable. As the global community continues to face emerging infectious disease threats, forecasting systems that can automatically discover relevant spatial dependencies—while optionally incorporating domain knowledge—will become increasingly important for pandemic preparedness and response.

Ethics Statement

Ethical approval was not required for this study in accordance with local legislation and institutional requirements, as only publicly available de-identified epidemiological surveillance datasets were used. No individual patient data or human subjects were involved.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Open access funding provided by Coventry University within the Coventry-Elsevier Agreement.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the author(s) used Claude (Anthropic) and ChatGPT (OpenAI) in order to improve language and readability, and for grammar checking of the manuscript text. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Data Availability

The datasets used in this study are publicly available and are included in the `data` folder of the MSAGAT-Net repository: https://github.com/michaelajao/MSAGAT-Net. The PyTorch code used for the experiments is also publicly available in the same repository.

CRediT authorship contribution statement

Michael Ajao-olarinoye: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft. **Vasile Palade:** Supervision, Writing – review & editing. **Fei He:** Supervision, Writing – review & editing. **Petra A Wark:** Supervision, Writing – review & editing. **Seyed Mousavi:** Writing – review & editing. **Zindoga Mukandavire:** Writing – review & editing.

References

[1] Ahmadini, A.A.H., Raghav, Y.S., Mahnashi, A.M., Islam Rather, K.U., Ali, I., 2025. Neural networks to model covid-19 dynamics and allocate healthcare resources. Scientific Reports 15, 15326. URL: https://doi.org/10.1038/s41598-025-00153-9, doi:10.1038/s41598-025-00153-9.

[2] Ajao-Olarinoye, M., Palade, V., Mousavi, S., He, F., Wark, P.A., 2023. Deep learning based forecasting of covid-19 hospitalisation in england: A comparative analysis, in: 2023 International Conference on Machine Learning and Applications (ICMLA), IEEE. pp. 1344–1349.

[3] Ben Taieb, S., Bontempi, G., Atiya, A.F., Sorjamaa, A., 2012. A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition. Expert Systems with Applications 39, 7067–7083. doi:https://doi.org/10.1016/j.eswa.2012.01.039.

[4] Brooks, L., Farrow, D.C., Hyun, S., Tibshirani, R., Rosenfeld, R., 2018. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. PLoS Computational Biology 14. doi:10.1371/journal.pcbi.1006134.

[5] Cao, Q., Jiang, R., Yang, C., Fan, Z., Song, X., Shibasaki, R., 2022. Mepognn: Metapopulation epidemic forecasting with graph neural networks, in: Joint European conference on machine learning and knowledge discovery in databases, Springer. pp. 453–468.

[6] Chandra, R., Goyal, S., Gupta, R., 2021. Evaluation of deep learning models for multi-step ahead time series prediction. Ieee Access 9, 83105–83123.

[7] Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251–1258. doi:10.1109/CVPR.2017.625.

[8] De Angelis, D., Presanis, A.M., Birrell, P.J., Tomba, G.S., House, T., 2015. Four key challenges in infectious disease modelling using data from multiple sources. Epidemics 10, 83–87. URL: https://www.sciencedirect.com/science/article/pii/S175543651400053X, doi:https://doi.org/10.1016/j.epidem.2014.09.004. challenges in Modelling Infectious DIsease Dynamics.

[9] Deng, S., Wang, S., Rangwala, H., Wang, L., Ning, Y., 2020. Cola-GNN: Cross-location Attention based Graph Neural Networks for Long-term ILI Prediction, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Association for Computing Machinery, New York, NY, USA. pp. 245–254. doi:10.1145/3340531.3411975.

[10] Gao, J., Heintz, J., Mack, C., Glass, L., Cross, A., Sun, J., 2023. Evidence-driven spatiotemporal covid-19 hospitalization prediction with ising dynamics. Nature communications 14, 3093.

[11] Gao, J., Sharma, R., Qian, C., Glass, L.M., Spaeder, J., Romberg, J., Sun, J., Xiao, C., 2021. Stan: spatio-temporal attention network for pandemic prediction using real-world evidence. Journal of the American Medical Informatics Association 28, 733–743. doi:10.1093/jamia/ocaa322.

[12] Giuliani, D., Dickson, M.M., Espa, G., Santi, F., 2020. Modelling and predicting the spatio-temporal spread of covid-19 in italy. BMC infectious diseases 20, 1–10. doi:10.1186/s12879-020-05415-7.

[13] Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, PMLR. pp. 249–256. URL: https://proceedings.mlr.press/v9/glorot10a.html.

[14] Han, S., Xun, Y., Cai, J., Yang, H., Li, Y., 2025. Dygraphformer: Transformer combining dynamic spatio-temporal graph network for multivariate time series forecasting. Neural Networks 181, 106776. doi:https://doi.org/10.1016/j.neunet.2024.106776.

[15] Heltberg, M.L., Michelsen, C., Martiny, E.S., Christensen, L.E., Jensen, M.H., Halasa, T., Petersen, T.C., 2022. Spatial heterogeneity affects predictions from early-curve fitting of pandemic outbreaks: a case study using population data from denmark. Royal Society Open Science 9, 220018. doi:10.1098/rsos.220018.

[16] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[17] Kamalov, F., Rajab, K., Cherukuri, A., Elnagar, A., Safaraliev, M., 2022. Deep learning for covid-19 forecasting: State-of-the-art review. Neurocomputing 511, 142–154. doi:10.1016/j.neucom.2022.09.005.

[18] Kim, M., Kim, J.H., Jang, B., 2025. Forecasting Epidemic Spread With Recurrent Graph Gate Fusion Transformers. IEEE Journal of Biomedical and Health Informatics 29, 1546–1559. doi:10.1109/JBHI.2024.3488274.

[19] Kong, L., Ojha, V., Gao, R., Suganthan, P.N., Snášel, V., 2023. Low-rank and global-representation-key-based attention for graph transformer. Information Sciences 642, 119108. doi:10.1016/j.ins.2023.119108.

[20] Lai, G., Chang, W.C., Yang, Y., Liu, H., 2018. Modeling long-and short-term temporal patterns with deep neural networks, in: The 41st international ACM SIGIR conference on research & development in

information retrieval, pp. 95–104.

[21] Li, T., Liu, L., Li, M., 2023. Multi-scale residual depthwise separable convolution for metro passenger flow prediction. Applied Sciences 13, 11272. doi:10.3390/app132011272.

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

[23] Lijing Wang, Lijing Wang, Aniruddha Adiga, Aniruddha Adiga, Jiangzhuo Chen, Jiangzhuo Chen, Adam Sadilek, Adam Sadilek, Srinivasan Venkatramanan, Srinivasan Venkatramanan, Madhav V. Marathe, Madhav Marathe, 2022. CausalGNN: Causal-Based Graph Neural Networks for Spatio-Temporal Epidemic Forecasting. Proceedings of the ... AAAI Conference on Artificial Intelligence 36, 12191–12199. doi:10.1609/aaai.v36i11.21479.

[24] Liu, Z., Wan, G., Prakash, B.A., Lau, M.S.Y., Jin, W., 2024. A Review of Graph Neural Networks in Epidemic Modeling. doi:10.48550/arXiv.2403.19852, arXiv:2403.19852.

[25] Luo, L., Li, B., Wang, X., Cui, L., Liu, G., 2023. Interpretable spatial identity neural network-based epidemic prediction. Scientific Reports 13. doi:10.1038/s41598-023-45177-1.

[26] Ma, L., Qiu, Z., Van Mieghem, P., Kitsak, M., 2024. Reporting delays: A widely neglected impact factor in covid-19 forecasts. PNAS nexus 3, pgae204.

[27] Moss, R., Zarebski, A.E., Dawson, P., Franklin, L.J., Birrell, F.A., McCaw, J.M., 2020. Anatomy of a seasonal influenza epidemic forecast. Communicable Diseases Intelligence 43. URL: https://ojs.cdi.cdc.gov.au/index.php/cdi/article/view/553, doi:10.33321/cdi.2019.43.7.

[28] NHS England, 2024. Covid-19 hospital activity. https://www.england.nhs.uk/statistics/statistical-work-areas/covid-19-hospital-activity/. Accessed 2025-05-10.

[29] Oluwasakin, E.O., Khaliq, A.Q., 2023. Data-driven deep learning neural networks for predicting the number of individuals infected by COVID-19 omicron variant. Epidemiologia 4, 420–453. doi:10.3390/epidemiologia4040037.

[30] Panja, M., Chakraborty, T., Kumar, U., Liu, N., 2022. Epicasting: An ensemble wavelet neural network for forecasting epidemics. Neural Networks 165, 185–212. doi:https://doi.org/10.1016/j.neunet.2023.05.049.

[31] Pu, X., Zhu, J., Wu, Y., Leng, C., Bo, Z., Wang, H., 2024. Dynamic adaptive spatio–temporal graph network for covid-19 forecasting. CAAI Transactions on Intelligence Technology 9, 769–786. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cit2.12238, doi:https://doi.org/10.1049/cit2.12238, arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/cit2.12238.

[32] Puny, O., Ben-Hamu, H., Lipman, Y., 2020. Global attention improves graph networks generalization. arXiv preprint arXiv:2006.07846 doi:10.48550/arXiv.2006.07846.

[33] Qiu, M., Tan, Z., Bao, B., 2024. MSGNN: Multi-scale Spatio-temporal Graph Neural Network for Epidemic Forecasting. Data Mining and Knowledge Discovery 38, 2348–2376. URL: https://doi.org/10.1007/s10618-024-01035-w, doi:10.1007/s10618-024-01035-w.

[34] Shi, Y., Zhu, X., Zhu, X., Cheng, B., Zhong, Y., 2025. Kalman filter-based epidemiological model for post-covid-19 era surveillance and prediction. Sensors 25. URL: https://www.mdpi.com/1424-8220/25/8/2507, doi:10.3390/s25082507.

[35] da Silva, C.C., de Lima, C.L., da Silva, A.C.G., Silva, E.L., Marques, G.S., de Araújo, L.J.B., Albuquerque Júnior, L.A., de Souza, S.B.J., de Santana, M.A., Gomes, J.C., et al., 2021. Covid-19 dynamic monitoring and real-time spatio-temporal forecasting. Frontiers in public health 9, 641253. doi:10.3389/fpubh.2021.641253.

[36] Stone, L., Olinky, R., Huppert, A., 2007. Seasonal dynamics of recurrent epidemics. Nature 446, 533–536. doi:10.1038/nature05638.

[37] Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., et al., 2017. Graph attention networks. stat 1050, 10–48550. doi:https://doi.org/10.48550/arXiv.1710.10903.

[38] Venna, S.R., Tavanaei, A., Gottumukkala, R.N., Raghavan, V.V., Maida, A.S., Nichols, S., 2019. A novel data-driven model for real-time influenza forecasting. IEEE Access 7, 7691–7701. doi:10.1109/ACCESS.2018.2888585.

[39] Verma, H., Mandal, S., Gupta, A., 2022. Temporal deep learning architecture for prediction of covid-19 cases in india. Expert Systems with Applications 195, 116611.

[40] Wang, L., Chen, J., Marathe, M., 2019. Defsi: Deep learning based epidemic forecasting with synthetic information, in: Proceedings of the AAAI conference on artificial intelligence, pp. 9607–9612. doi:10.1609/aaai.v33i01.33019607.

[41] Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H., 2020. Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768 .

[42] Wang, Y., Zhu, Y., Liang, L., Wang, Y., Harrison, E.M., Ma, L., Gao, J., 2024. DeepEST: A Python Library for Spatio-Temporal Epidemiology Prediction, in: KDD'24 Workshop: Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare. URL: https://openreview.net/forum?id=YzWC6huGQq.

[43] Wu, Y., Yang, Y., Nishiura, H., Saitoh, M., 2018. Deep learning for epidemiological predictions, in: The 41st international ACM SIGIR conference on research & development in information retrieval, pp. 1085–1088.

[44] Xie, F., Zhang, Z., Li, L., Zhou, B., Tan, Y., 2023. Epignn: Exploring spatial transmission with graph neural network for regional epidemic forecasting, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer. pp. 469–485. doi:10.1007/978-3-031-26422-1_29.

[45] Yang, L., Shi, R., Zhang, Q., Wang, Z., Cao, X., Wang, C., et al., 2023. Self-supervised graph neural networks via low-rank decomposition. Advances in Neural Information Processing Systems 36, 34295–34307.

[46] Yu, Y., Sun, W., Liu, J., Zhang, C., 2022. Traffic flow prediction based on depthwise separable convolution fusion network. Journal of Big Data 9, 83. doi:10.1186/s40537-022-00637-9.

[47] Zeroual, A., Harrou, F., Dairi, A., Sun, Y., 2020. Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. Chaos, solitons & fractals 140, 110121. doi:10.1016/j.chaos.2020.110121.

[48] Zhang, C., James, J., Liu, Y., 2019. Spatial-temporal graph attention networks: A deep learning approach for traffic forecasting. Ieee Access 7, 166246–166256. doi:10.1109/ACCESS.2019.2953888.

[49] Zhang, S., Guo, Y., Zhao, P., Zheng, C., Chen, X., 2021. A graph-based temporal attention framework for multi-sensor traffic flow forecasting. IEEE Transactions on Intelligent Transportation Systems 23, 7743–7758. doi:10.1109/TITS.2021.3072118.

[50] Zhiwei Ding, Feng Sha, Yi Zhang, Zhouwang Yang, 2023. Biology-Informed Recurrent Neural Network for Pandemic Prediction Using Multimodal Data. Biomimetics 8, 158–158. doi:10.3390/biomimetics8020158.