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

Michael Ajao-olarinoye[a,1,*], Vasile Palade[a,1], Fei He[a], Petra A Wark[b], Seyed Mousavi[a], 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*

Abstract

Background and Objective: Accurate spatiotemporal epidemic forecasting is critical for public health preparedness and resource allocation. However, existing graph neural network (GNN) approaches face fundamental limitations: quadratic computational complexity, reliance on predefined adjacency matrices constructed from geographical proximity or mobility data, and unstable multi-horizon forecasts. This study introduces MSAGAT-Net, a computationally efficient multi-scale temporal graph attention network that addresses these limitations whilst maintaining high forecasting accuracy across diverse epidemic scenarios.

Methods: MSAGAT-Net integrates four key architectural components: (i) efficient feature extraction via depthwise separable convolutions, (ii) an Efficient Adaptive Graph Attention Module (EAGAM) employing linearised attention with low-rank projections to reduce computational complexity from $O(N^2)$ to $O(N)$, featuring learnable graph bias parameters that discover spatial relationships directly from data without requiring predefined adjacency matrices, (iii) a Dilated Multi-Scale Temporal Feature Module (DMTFM) using adaptive dilated convolutions to capture dynamics across multiple

*Corresponding author.
Email address: olarinoyem@coventry.ac.uk (Michael Ajao-olarinoye)

temporal resolutions, and (iv) a Progressive Prediction Refinement Module (PPRM) for stable multi-horizon forecasts. We evaluated MSAGAT-Net on seven diverse datasets spanning influenza, COVID-19, and ICU bed occupancy across multiple countries, including two novel benchmarks (LTLA-COVID and NHS-ICUBeds).

Results: In datasets exhibiting strong spatiotemporal dependencies, MSAGAT-Net achieved state-of-the-art performance, reducing RMSE by up to 11.2% compared to baseline methods in the Japan-Prefectures influenza dataset and attaining superior results in LTLA-Timeseries COVID-19 forecasts for short and medium-term horizons. Critically, we demonstrate that learning spatial relationships from epidemic data outperforms using predefined geographical adjacency matrices by 11.5% RMSE, validating our data-driven approach. Comprehensive ablation studies revealed a fundamental scientific insight: optimal architectural choices are disease-specific and horizon-dependent, with spatial attention becoming increasingly critical for extended influenza forecasting yet occasionally impairing short-term COVID-19 predictions.

Conclusions: MSAGAT-Net provides an efficient and scalable solution for spatiotemporal epidemic forecasting that eliminates the need for predefined graph structures whilst achieving competitive or superior performance. Our findings challenge conventional assumptions about model complexity, demonstrating that sophisticated temporal processing can occasionally degrade performance compared to simpler alternatives. These insights highlight the importance of adaptive, disease-specific architectures for public health decision-making and surveillance systems.

Keywords:
Graph neural networks, epidemic forecasting, multi-scale modelling, COVID-19, spatiotemporal forecasting

---

1. Introduction

Spatiotemporal epidemic forecasting is critical for public health systems, enabling timely decision-making and resource allocation in response to emerging infectious diseases. The COVID-19 pandemic has underscored the importance of reliable forecasting models that can adapt to rapidly changing dynamics and provide actionable insights for public health interventions [1, 2, 3, 4, 5]. However, developing such models presents significant challenges due to the complex interplay between spatial dependencies, temporal

patterns, and the inherent uncertainty in disease transmission.

Effective epidemic forecasting requires capturing both spatial dependencies (how diseases spread between interconnected regions) and temporal patterns that range from daily reporting fluctuations to seasonal epidemic waves [6, 7]. Traditional epidemiological models, while providing valuable theoretical insights, often struggle with the computational complexity required for real-time surveillance across hundreds or thousands of regions [8, 9, 10].

Deep learning techniques have substantially advanced epidemic forecasting across various domains, including traffic forecasting [11, 12], environmental monitoring [13, 14], and epidemic prediction [15, 16]. Studies such as [1, 12, 13, 14, 17] have demonstrated that recurrent neural networks (RNNs), convolutional neural networks (CNNs), and graph neural networks (GNNs) improve both predictive accuracy and modelling efficiency. Graph neural networks, in particular, have shown strong potential for epidemic forecasting by effectively learning complex spatiotemporal patterns from data, outperforming traditional and other deep learning models in capturing regional heterogeneity and dynamic transmission mechanisms [18, 19].

Despite these advances, current approaches face four fundamental limitations that prevent their deployment in real-world surveillance systems: (1) most graph attention mechanisms suffer from quadratic computational complexity, making them prohibitively expensive for national-scale surveillance involving thousands of regions; (2) state-of-the-art models such as EpiGNN, Cola-GNN, and DCRNN require predefined adjacency matrices constructed from geographical proximity or mobility data, which may not capture the true epidemic transmission patterns and require domain expertise or external data sources that may be unavailable; (3) existing methods typically employ fixed architectural choices for temporal processing, failing to adapt to the diverse dynamics exhibited by different diseases and forecast horizons; and (4) multi-horizon forecasting remains unstable due to error accumulation, limiting the utility of long-range forecasts essential for public health planning and resource allocation.

To address these challenges, we propose MSAGAT-Net (Multi-Scale Temporal Adaptive Graph Attention Network), a novel architecture that achieves computational efficiency whilst maintaining high forecasting accuracy across diverse epidemic scenarios. Critically, MSAGAT-Net does not *require* any predefined adjacency matrix—instead, it learns spatial relationships directly from the epidemic time-series data through learnable low-rank graph bias parameters, enabling discovery of transmission-relevant regional connections

that may not be apparent from geographical proximity alone. When prior knowledge about spatial relationships is available (e.g., geographical adjacency or mobility data), MSAGAT-Net can optionally incorporate it as a soft prior to accelerate learning, but this is not required for operation. MSAGAT-Net integrates four key components: (i) an efficient feature extraction module using depthwise separable convolutions, (ii) an Efficient Adaptive Graph Attention Module (EAGAM) that reduces computational complexity from $O(N^2)$ to $O(N)$ using low-rank bottleneck projections and linearised attention, featuring a novel graph bias message passing mechanism that integrates learnable spatial structure directly into the forward computation, (iii) a Dilated Multi-Scale Temporal Feature Module (DMTFM) that captures dynamics across multiple temporal resolutions using adaptive dilated convolutions, and (iv) a Progressive Prediction Refinement Module (PPRM) that stabilises multi-horizon forecasts by combining model-based predictions with trend-based extrapolations.

Our contributions are as follows:

1. We propose an efficient graph attention mechanism that reduces complexity from $O(N^2)$ to $O(N)$ through low-rank bottleneck projections and linearised attention. Unlike state-of-the-art baselines (EpiGNN, Cola-GNN, DCRNN) that require predefined adjacency matrices as mandatory input, our approach learns spatial relationships entirely from data through learnable graph bias parameters, eliminating the need for external graph construction and enabling discovery of non-obvious transmission pathways. When adjacency information is available, it can be incorporated as an optional prior via a learnable gating mechanism.
2. We develop an adaptive multi-scale temporal processing framework using dilated convolutions that efficiently captures epidemic dynamics across multiple temporal resolutions, from reporting fluctuations to seasonal waves.
3. We introduce a progressive forecast refinement module that mitigates error accumulation in multi-horizon forecasting by adaptively combining model-based forecasts with trend-based extrapolations.
4. We evaluate MSAGAT-Net across seven diverse epidemic datasets spanning influenza, COVID-19, and ICU bed occupancy, achieving up to 11.2% RMSE reduction compared to state-of-the-art baselines. Comprehensive ablation studies reveal that optimal architectural choices

are fundamentally disease-specific and horizon-dependent, demonstrating that no single architecture universally dominates across all epidemic contexts.

The remainder of this paper is organised as follows. Section 2 reviews the literature on spatiotemporal epidemic forecasting, graph neural networks, attention mechanisms, and multi-scale temporal modelling. Section 3 formalises the forecasting problem, describes the data pre-processing and graph construction procedures, and presents the proposed MSAGAT-Net architecture. Section 5 details the experimental setup, datasets, baseline methods, and evaluation metrics, followed by a comprehensive presentation of the results, ablation studies of model components, sensitivity analyses across hyperparameters and forecasting horizons, and qualitative visualisations to aid interpretation. Finally, Section 6 summarises the key findings and contributions, discusses limitations and practical implications for public health decision making, and outlines directions for future research.

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 and transformer-based models that learn dynamic cross-regional influence, (ii) physics-informed and hybrid neural approaches that integrate compartmental priors, and (iii) 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 [20]. A comprehensive taxonomy by Liu et al. [21, 22] 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. [23] 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. [24] 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. [25] 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. [26] 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. [27] proposed DASTGN with dual-scale attention mechanisms that adaptively fuse spatial and temporal effects at both fine and coarse-grained resolutions.

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. [28] 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. [29] 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 standard attention mechanisms with quadratic $O(N^2)$ complexity, making them computationally prohibitive for large-scale regional analysis. Recent advances in efficient attention, particularly linearised attention mechanisms that reduce complexity to $O(N)$ through low-rank projection of key and value representations with fixed bottleneck dimensions [30], offer promising solutions, but remain largely unexplored in epidemic forecasting contexts. Second, existing approaches often fail to maintain stability in multi-horizon forecasts, as error propagation compounds over extended forecast horizons [31].

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 [6], while longer-term patterns, including seasonal waves, are shaped by environmental factors and behavioural responses to disease spread [7]. 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 [31]. 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 [13, 32]. Wang et al. [19] 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. [23] 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 [33, 34]. Although such approaches can improve long-range predictive performance, they often require extensive feature engineering and may not generalise well in heterogeneous epidemic contexts [6].

These limitations across spatiotemporal modelling, physics-informed approaches, and multi-scale temporal processing underscore three critical gaps in contemporary epidemic forecasting research: (1) the computational complexity of attention mechanisms, which scales quadratically with the number of regions and hinders real-time surveillance applications; (2) the rigidity of fixed architectural designs, which 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 linearised attention for computational efficiency, adaptive multi-scale temporal processing, 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 $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T] \in \mathbb{R}^{N \times T}$, where T denotes the total length of the historical observation period, and each vector $\mathbf{x}_t \in \mathbb{R}^N$ (for $t = 1, 2, \ldots, T$) 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 $\mathbf{x}^i = [x_{i,1}, x_{i,2}, \ldots, x_{i,T}] \in \mathbb{R}^T$. 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:

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

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 $[\mathbf{x}_{t-w+1}, \mathbf{x}_{t-w+2}, \ldots, \mathbf{x}_t]$ to forecast $\mathbf{x}_{t+h}$.

The spatial relationships between regions are encoded in a graph structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ represents the set of regions and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the potential connections between regions. Traditional spatiotemporal GNN approaches rely on a fixed adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ constructed from geographical proximity or other predefined criteria. However, such fixed structures may fail to capture the true epidemic transmission patterns, which can be influenced by population mobility, healthcare referral pathways, and socioeconomic factors that are not apparent from geography alone.

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

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

where the first argument represents the historical window of w time steps of observations across all N regions, and $\boldsymbol{\Theta}$ denotes the learnable parameters of our forecasting model, which crucially include the learnable graph structure parameters. Unlike methods that require a predefined adjacency matrix $\mathbf{A}$, our MSAGAT-Net learns spatial relationships directly from data through learnable graph bias parameters, enabling adaptive discovery of epidemic-relevant regional connections during training.

The challenge lies in designing this function f to effectively capture both spatial dependencies between regions and temporal patterns within regions, whilst remaining computationally tractable and robust to the noisy and incomplete nature of epidemic data. Our approach, detailed in the following sections, addresses this challenge through a novel neural network architec-

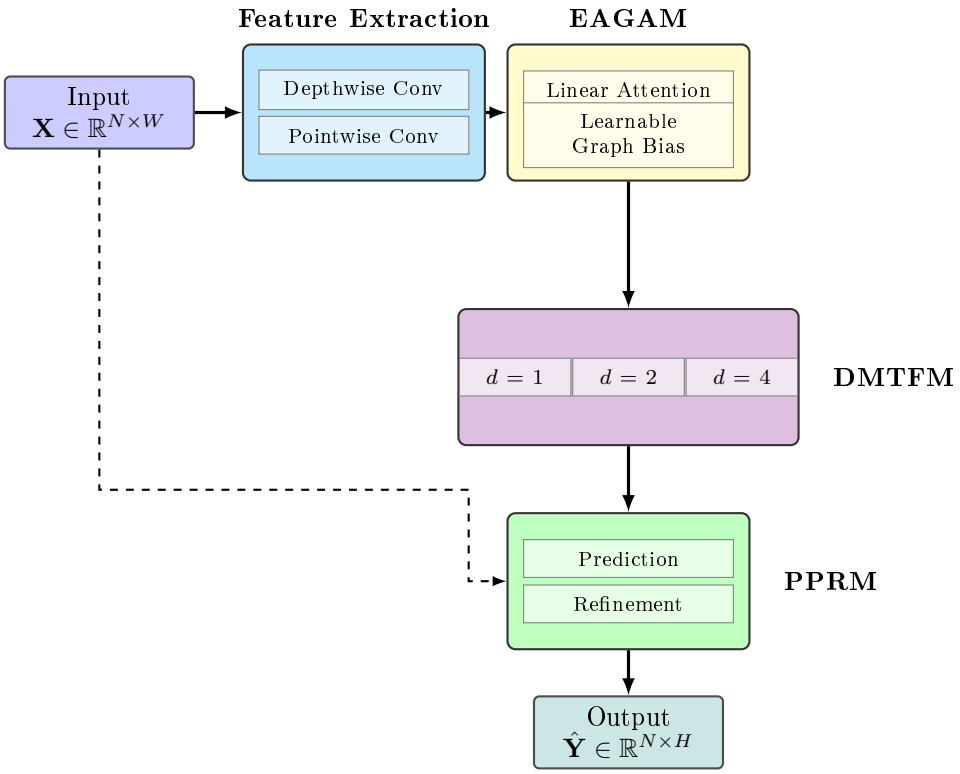


Figure 1: Overview of the MSAGAT-Net architecture. The model processes input through four main modules: Feature Extraction using depthwise separable convolutions, EAGAM with linearised attention and learnable graph bias, DMTFM employing multi-scale dilated convolutions (dilation rates $d \in \{1, 2, 4\}$), and PPRM for adaptive forecast refinement. The dashed line represents the skip connection from input to PPRM.

ture that combines learnable graph attention mechanisms with multi-scale temporal processing.

3.2. Feature Extraction

The first component of the MSAGAT-Net architecture is the feature extraction module, which transforms the raw time-series data into meaningful feature representations whilst maintaining computational efficiency. Given the input time-series data $\mathbf{X} = [\mathbf{x}_{t-w+1}, \mathbf{x}_{t-w+2}, \ldots, \mathbf{x}_t] \in \mathbb{R}^{N \times w}$ for the N regions over a look-back window of w time steps, we need to extract features that capture relevant temporal patterns for each region. This is achieved through a combination of depthwise separable convolutions and low-rank

projections, which allow for efficient feature extraction while reducing the risk of overfitting. Depthwise separable convolutions have been shown to significantly reduce the number of parameters and computational complexity while maintaining high performance [35], and have been successfully applied to spatiotemporal feature extraction [36, 37]. This allows us to efficiently capture the temporal dynamics of epidemic data without incurring the high computational costs associated with traditional convolutional architectures.

3.2.1. Depthwise Separable Convolutions

We employ the depthwise separable convolutions to extract temporal features efficiently. This approach significantly reduces the computational complexity and number of parameters whilst maintaining expressive power. The depthwise separable convolution consists of two stages:

For each region's historical window $\mathbf{x}^i_{[t-w+1:t]} \in \mathbb{R}^w$ (the most recent w time steps for region i), we first reshape the time-series as a single-channel input. The depthwise convolution applies a separate filter to this channel:

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

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:

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

where $\mathbf{z}^i_{\text{point}} \in \mathbb{R}^{w \times d_{\text{feat}}}$ represents the features after pointwise convolution, and $d_{\text{feat}} = 16$ is the number of output feature channels.

This decomposition significantly reduces the computational complexity and number of parameters compared to standard convolutions whilst maintaining similar expressive power. Specifically, for a standard convolution with kernel size k, input channels c_{in}, and output channels c_{out}, the parameter count is $k \times c_{\text{in}} \times c_{\text{out}}$. In contrast, the depthwise separable convolution requires only $k \times c_{\text{in}} + c_{\text{in}} \times c_{\text{out}}$ parameters.

After each convolutional operation, we apply batch normalisation followed by ReLU activation to enhance training stability and introduce non-linearity:

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

This normalisation step helps mitigate internal covariate shift during training, whilst the non-linear activation enables the model to capture complex temporal patterns in the epidemic data.

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 with dimension d_{bottle}, and $\mathbf{F}^i \in \mathbb{R}^{d_{\text{hidden}}}$ is the final representation of characteristics for region i with dimension d_{hidden}.

The flattening operation converts the convolutional features $\mathbf{z}^i_{\text{norm}} \in \mathbb{R}^{w \times d_{\text{feat}}}$ into a vector of dimension $w \times d_{\text{feat}}$. This is then projected to the bottleneck dimension and subsequently to the hidden dimension.

After applying the low-rank projection to each region's convolutional features, we obtain a comprehensive feature matrix $\mathbf{F} \in \mathbb{R}^{N \times d_{\text{hidden}}}$, where each row $\mathbf{F}^i \in \mathbb{R}^{d_{\text{hidden}}}$ represents the temporal feature embedding for region i. This matrix encapsulates the essential temporal dynamics across all N regions in a compact, information-dense representation suitable for subsequent spatial modelling.

The low-rank bottleneck projection ($w \times d_{\text{feat}} \to d_{\text{bottle}} \to d_{\text{hidden}}$) serves multiple critical functions within the architecture. By compressing information through a dimension bottleneck where $d_{\text{bottle}} \ll w \times d_{\text{feat}}$, computational complexity is reduced from $\mathcal{O}(N \times w \times d_{\text{feat}} \times d_{\text{hidden}})$ to $\mathcal{O}(N \times (d_{\text{bottle}} \times d_{\text{hidden}} + w \times d_{\text{feat}} \times d_{\text{bottle}}))$, enabling efficient processing of large-scale spatiotemporal datasets. The bottleneck architecture creates an information constraint that forces the model to distill the most salient temporal patterns, preventing overfitting when training data are limited relative to the high-dimensional input space. Additionally, forcing information through a lower-dimensional manifold encourages separation of relevant signals from

noise, allowing subsequent layers to focus on predictive temporal patterns rather than spurious correlations. Finally, the shared projection parameters across all regions create a common latent space that facilitates meaningful comparison and interaction between region features in subsequent graph attention layers.

To stabilise training and enhance feature quality, we apply layer normalisation followed by a non-linear activation:

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

The layer normalisation operates across the feature dimension, normalising each region's feature vector independently. This addresses internal covariate shift, enabling faster convergence during training whilst making the model robust to variations in feature scale across different regions. The ReLU activation introduces non-linearity essential for modelling complex temporal patterns whilst preserving sparse activation, a property particularly valuable for epidemic time-series that often exhibit punctuated patterns of activity against background stability.

This processed feature matrix $\mathbf{F}$ now encodes the essential temporal characteristics of each region's epidemic time-series in a form optimised for the subsequent EAGAM, which will model dynamic spatial dependencies between regions based on these temporal feature representations. The feature extraction pipeline is illustrated in Figure 2.

For the Feature Extraction module, we set the feature channel dimension d_{feat} to 16, the bottleneck dimension d_{bottle} to 8, and the hidden dimension d_{hidden} to 32. These values were determined through preliminary experiments to balance the expressiveness of the model with computational efficiency. The convolutional kernel size is set to 3 with appropriate padding to maintain the temporal dimension, thereby capturing local temporal patterns that span 3-5 time steps. This parameter configuration renders the feature extraction module computationally tractable whilst preserving sufficient representational capacity for encoding the diverse temporal dynamics observed in epidemic time-series, a crucial consideration for deployment in resource-constrained epidemic monitoring scenarios.

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

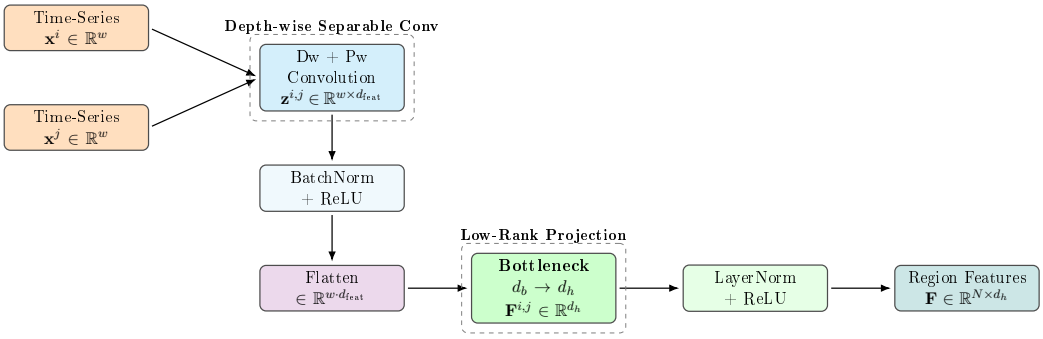


Figure 2: Feature-extraction pipeline. Independent regional time-series $\mathbf{x}^i$ and $\mathbf{x}^j$ are processed in parallel by depth-wise and point-wise convolutions, normalised, flattened, passed through a bottleneck projection ($d_{\text{bottle}} \to d_{\text{hidden}}$), and normalised again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

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 [38], 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.

A significant challenge in implementing graph attention mechanisms for large-scale epidemic forecasting problems is computational complexity. Standard softmax-based attention mechanisms in graph neural networks (GNNs) typically incur quadratic $\mathcal{O}(N^2)$ complexity with respect to the number of nodes, making them prohibitively expensive for large graphs. Additionally, these methods often suffer from over-smoothing when modelling long-range dependencies, where node representations become increasingly similar after multiple message-passing iterations.

Recent advances in efficient attention mechanisms have shown that low-rank decomposition techniques can substantially reduce computational complexity whilst maintaining expressive power. Several influential works have explored this direction. Researchers such as [39] propose a low-rank global attention (LRGA), an adaptive module that replaces the total attention of the dot product in GNN with a decomposed low-rank form. [40] present the Global Representation Key (GRK) attention layer, where the attention

scores of each node are calculated using a shared projection of the features of its neighbours. A learnt adaptive low-rank matrix captures the most salient structural information, mitigating over-smoothing and improving performance on graphs. While [41] embeds an adaptive low-rank decomposition step in each propagation layer within each ego network to concentrate the message passing on the most prominent low-dimensional subspaces. This lets the model adaptively focus on the most informative subspace per node, improving robustness without labels. These studies collectively demonstrate that low-rank factorisation offers an efficient, scalable, and expressive alternative to full-rank attention in graph architectures, motivating the design of this module in our framework.

Motivated by these advances, EAGAM employs a linearised attention mechanism that combines low-rank decomposition with a novel graph bias message passing mechanism. Rather than computing the full attention matrix between all pairs of regions (which incurs $\mathcal{O}(N^2)$ complexity for traditional softmax attention), we use a linearised formulation that reduces complexity to $\mathcal{O}(N)$ through fixed-rank bottleneck projections (bottleneck dimension $d_b = 8$), scaling linearly with the number of regions. Efficiency arises from (i) low-rank bottleneck projections for query, key, and value representations, and (ii) a linearised attention computation that avoids explicit construction of the $N \times N$ attention matrix whilst maintaining the capacity to learn complex spatial dependencies. A key innovation of our approach is the integration of a learnable graph bias directly into the forward computation through normalised low-rank message passing, rather than using it solely for regularisation as in prior work. This enables the model to leverage persistent spatial relationships during inference whilst maintaining linear complexity. The EAGAM module comprises six components: (1) bottleneck projections for QKV, (2) multi-head attention with ELU-based linearisation, (3) normalisation for numerical stability, (4) learnable graph structure bias, (5) graph bias message passing for forward integration, and (6) attention regularisation to promote sparsity.

3.3.1. Bottleneck Projection

Given the feature matrix $\mathbf{F} \in \mathbb{R}^{N \times d_{\text{hidden}}}$ from the feature extraction module, 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 $d_{\text{head}} = d_{\text{hidden}}/h$:

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

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 $[h, N, d_{\text{head}}]$.

A key innovation in our approach is the specific attention computation mechanism employed within each head. Rather than relying on standard scaled dot-product attention with softmax, we employ an enhanced mechanism with better numerical stability and more nuanced relationship modelling.

First, we apply the Exponential Linear Unit (ELU) activation function followed by adding a constant value of 1 to both query and key representations:

$$\hat{\mathbf{Q}}_h = \mathrm{ELU}(\mathbf{Q}_h) + 1 \tag{18}$$

$$\hat{\mathbf{K}}_h = \mathrm{ELU}(\mathbf{K}_h) + 1 \tag{19}$$

This transformation, known as the ELU+1 kernel trick [42], ensures that all attention inputs are strictly positive, which is essential for the linearisation that follows. The key insight is that standard softmax attention computes $\mathrm{softmax}(\mathbf{Q}\mathbf{K}^T)\mathbf{V}$, requiring explicit construction of the $N \times N$ attention matrix with $\mathcal{O}(N^2)$ complexity. By using a positive feature map $\phi(\cdot) = \mathrm{ELU}(\cdot) + 1$, we can exploit the associativity of matrix multiplication: instead of computing $(\phi(\mathbf{Q})\phi(\mathbf{K})^T)\mathbf{V}$, we compute $\phi(\mathbf{Q})(\phi(\mathbf{K})^T\mathbf{V})$, reducing complexity to $\mathcal{O}(N \cdot d_{\text{head}}^2)$—linear in the number of regions.

Next, we compute the key-value product for each attention head:

$$\mathbf{KV}_h = \hat{\mathbf{K}}_h^T \mathbf{V}_h \tag{20}$$

where $\mathbf{KV}_h \in \mathbb{R}^{h \times d_{\text{head}} \times d_{\text{head}}}$. This operation captures the relationships between keys and values, allowing the model to learn how to weight the features of different regions based on their similarity.

To ensure stable normalisation, we calculate a normalisation factor based on the sum of keys:

$$\mathbf{z} = \frac{1}{\sum_{d=1}^{d_{\text{head}}} \hat{\mathbf{K}}_h + \epsilon} \tag{21}$$

where the sum is taken over the feature dimension d_{head} for each node in each head, $\epsilon = 10^{-8}$ prevents division by zero, and $\mathbf{z} \in \mathbb{R}^{h \times N}$ represents the normalisation factor for each node in each attention head. This operation ensures stable normalisation across the attention heads, allowing for effective learning of inter-regional relationships.

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

$$\mathbf{O}_{\text{linear}} = (\hat{\mathbf{Q}}_h \mathbf{K}\mathbf{V}_h) \odot \mathbf{z} \tag{22}$$

where $\mathbf{O}_{\text{linear}} \in \mathbb{R}^{h \times N \times d_{\text{head}}}$ represents the attended features across all heads, and $\odot$ denotes element-wise multiplication with broadcasting of the normalisation factor $\mathbf{z}$ across the feature dimension.

3.3.3. Learnable Graph Structure

An important feature of our EAGAM is the incorporation of a learnable graph structure bias. 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:

$$\mathbf{B} = \mathbf{U}\mathbf{V} \tag{23}$$

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 and $d_{\text{bias}} \ll N$ is the bottleneck dimension of the bias term (set equal to $d_{\text{bottle}} = 8$ in our implementation). Both $\mathbf{U}$ and $\mathbf{V}$ are initialised using Xavier uniform initialisation [43] to ensure balanced gradient flow across heads and maintain stable training dynamics from the outset. This learnable bias $\mathbf{B} \in \mathbb{R}^{h \times N \times N}$ provides an innovative mechanism for the model to learn persistent spatial structures that complement data-driven attention patterns.

3.3.4. Graph Bias Message Passing

Rather than using the graph bias solely for regularisation, we integrate it directly into the forward computation through a normalised low-rank message passing operation. This approach maintains $\mathcal{O}(N)$ complexity by exploiting the low-rank factorisation $\mathbf{B} = \mathbf{U}\mathbf{V}$ to avoid materialising the full $N \times N$ bias matrix.

To ensure stable message passing with non-negative weights, we first apply a positivity transformation to the low-rank factors:

$$\hat{\mathbf{U}} = \text{ELU}(\mathbf{U}) + 1, \quad \hat{\mathbf{V}} = \text{ELU}(\mathbf{V}) + 1 \tag{24}$$

The graph bias message passing output is then computed as:

$$\mathbf{O}_{\text{bias}} = \frac{\hat{\mathbf{U}}(\hat{\mathbf{V}}\mathbf{V}_h)}{\hat{\mathbf{U}}(\hat{\mathbf{V}}\mathbf{1}) + \epsilon} \tag{25}$$

where $\mathbf{V}_h$ represents the value representations, $\mathbf{1}$ is a vector of ones, and $\epsilon = 10^{-8}$ prevents division by zero. The numerator computes $\hat{\mathbf{U}}(\hat{\mathbf{V}}\mathbf{V}_h) \in \mathbb{R}^{h \times N \times d_{\text{head}}}$ through two sequential matrix multiplications, each with complexity $\mathcal{O}(N \cdot d_{\text{bias}} \cdot d_{\text{head}})$. The denominator provides row-wise normalisation analogous to the normalisation in standard attention mechanisms.

This message passing output is combined with the linear attention output:

$$\mathbf{O}_h = \mathbf{O}_{\text{linear}} + \text{Dropout}(\mathbf{O}_{\text{bias}}) \tag{26}$$

By integrating the learned graph structure directly into the forward pass, the model can leverage persistent spatial relationships during inference whilst maintaining the computational efficiency of linearised attention.

3.3.5. Optional Adjacency Prior Integration

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. This is achieved through a learnable gating mechanism that combines the learned graph structure with the provided adjacency prior:

$$\mathbf{O}'_{\text{bias}} = (1 - \sigma(g)) \cdot \mathbf{O}_{\text{bias}} + \sigma(g) \cdot \tilde{\mathbf{A}}\mathbf{V}_h \tag{27}$$

where g is a learnable gate parameter initialised to a small value (e.g., 0.1), $\sigma(\cdot)$ denotes the sigmoid function ensuring the gate remains in $[0, 1]$, and $\tilde{\mathbf{A}}$ is the symmetrically normalised adjacency matrix with added self-loops:

$$\tilde{\mathbf{A}} = \mathbf{D}^{-1/2}(\mathbf{A} + \mathbf{I})\mathbf{D}^{-1/2} \tag{28}$$

where $\mathbf{D}$ is the degree matrix and $\mathbf{I}$ is the identity matrix. This formulation has several desirable properties:

- **Flexibility:** When no adjacency is provided, the model operates using only the learned graph bias, matching the default behaviour.

- **Adaptability:** The gate parameter g is learned during training, allowing the model to determine the optimal balance between prior knowledge and data-driven patterns.
- **Efficiency:** The adjacency-based message passing $\tilde{\mathbf{A}}\mathbf{V}_h$ has $\mathcal{O}(|E|)$ complexity where $|E|$ is the number of edges, typically sparse for geographical adjacency.

This design choice distinguishes MSAGAT-Net from existing approaches: baselines such as EpiGNN, Cola-GNN, and DCRNN mandate adjacency matrices as required input, whereas MSAGAT-Net can function without them whilst optionally benefiting from prior knowledge when available. Importantly, our experimental analysis (Section 5) demonstrates that learning spatial relationships from data actually *outperforms* using predefined geographical adjacency, validating the fundamental premise of our approach.

3.3.6. Attention Regularisation

To promote sparse and interpretable spatial relationships, we apply L1 regularisation to the graph structure bias:

$$\mathcal{L}_{\text{attn}} = \lambda \|\mathbf{B}\|_1 \tag{29}$$

where λ is the learnable regularisation weight. This regularisation encourages the model to learn sparse, interpretable spatial dependency patterns in the graph bias whilst maintaining the computational efficiency of linearised attention. 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:

$$\mathbf{O} = \text{Reshape}(\text{Transpose}(\mathbf{O}_h, 0, 1), [N, d_{\text{hidden}}]) \tag{30}$$

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

$$\mathbf{O}_{\text{low}} = \text{Linear}_{\text{out\_low}}(\mathbf{O}) \tag{31}$$

$$\mathbf{O}_{\text{final}} = \text{Linear}_{\text{out\_high}}(\mathbf{O}_{\text{low}}) \tag{32}$$

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 DMTFM for further processing.

For the EAGAM module, we set the number of attention heads to $h = 4$ and the bottleneck dimension to $d_{\text{bottle}} = 8$. These values were determined through preliminary experiments to provide an optimal balance between capturing diverse spatial dependency patterns and the maintenance of computational efficiency through low-rank decomposition. The attention regularisation weight λ is set to 10^{-5}, which was empirically found to promote sparse, interpretable attention patterns without overly constraining the model's capacity to learn complex spatial relationships. This configuration allows EAGAM to model dynamic spatial dependencies efficiently whilst avoiding the computational overhead of full-rank attention mechanisms. Figure 3 presents the data flow through the EAGAM module.

Figure 3 illustrates the flow of data through the EAGAM module. The input feature matrix $\mathbf{F}$ is processed through low-rank projections to obtain query, key, and value representations. These representations are then reshaped for multi-head attention computation, where the adaptive graph attention mechanism is applied. The learnable graph structure bias is incorporated into the attention scores, and L1 regularisation is applied to promote sparse attention patterns. Finally, the output features are obtained through high-rank projections, ready for further processing in DMTFM.

3.4. Dilated Multi-Scale Temporal Feature Module

The third major component of the proposed MSAGAT-Net architecture is the Dilated Multi-Scale Temporal Feature Module (DMTFM), which addresses a fundamental challenge in epidemic forecasting: modelling temporal dependencies across multiple time scales. Epidemics exhibit complex temporal dynamics that span various scales, including short-term fluctuations (e.g., reporting delays or weekend effects), medium-term patterns (e.g., incubation or transmission cycles), and long-term trends (e.g., seasonal variations or behavioural changes) [7, 6, 44]. Accurate forecasting therefore requires models capable of effectively capturing these multi-scale temporal dynamics.

Deng et al. [23] introduce the idea of multi-scale dilated convolutional with the same filter and stride sides but different dilation rates, which Xie et al. [24] improved by making use of the multi-scale convolution to capture

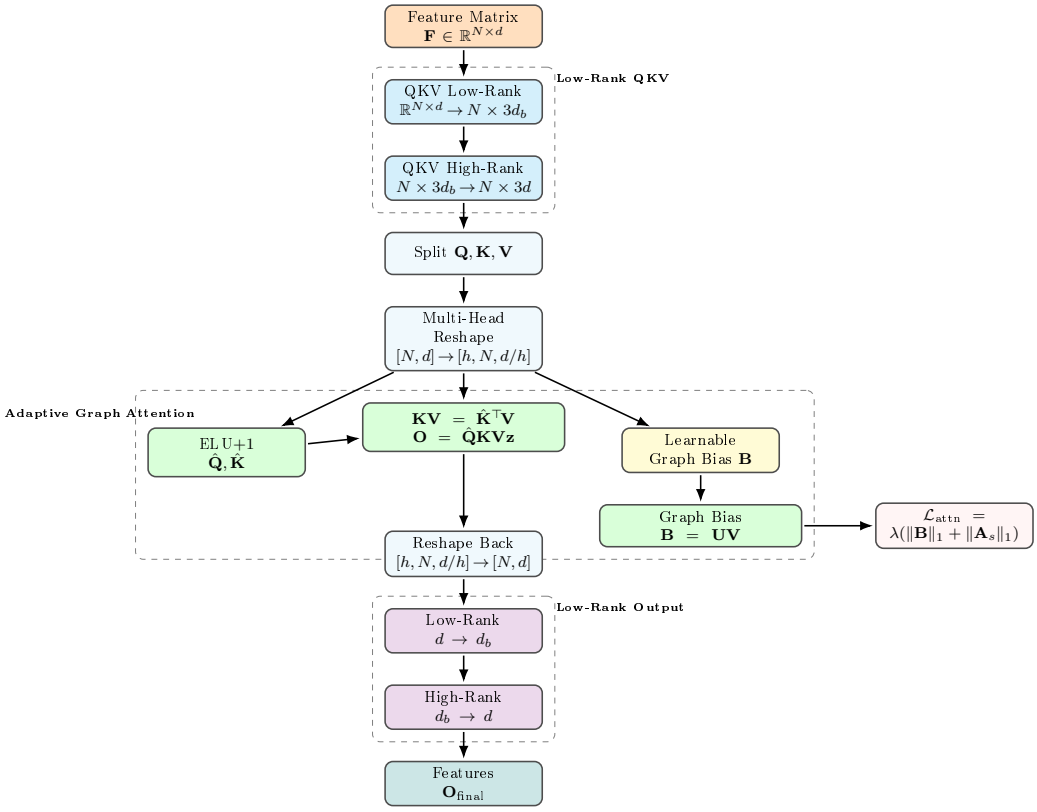


Figure 3: Data flow in the EAGAM module. Input features undergo low-rank QKV projections, followed by linearised attention computation with learnable graph bias for spatial dependency modelling, and low-rank output projection.

features. Building on this, the DMTFM employs parallel dilated convolutional layers to efficiently capture temporal dependencies across multiple scales using the output from the EAGAM. This approach enables the model to maintain an awareness of both immediate and distant temporal relationships whilst controlling parameter count and computational complexity.

3.4.1. Dilated Convolutions for Multi-scale Processing

The core of our DMTFM is a set of parallel convolutional branches operating at different dilation rates. For a given input feature tensor $\mathbf{G} \in \mathbb{R}^{B \times N \times d_{\text{hidden}}}$ (where B is the batch size, N is the number of regions, and d_{hidden} is the hidden dimension), we first transpose the tensor to prepare for 1D convolutions along the temporal dimension:

$$\mathbf{G}_{\text{conv}} = \text{Transpose}(\mathbf{G}, 1, 2) \tag{33}$$

resulting in a tensor of shape $[B, d_{\text{hidden}}, N]$. We then process this tensor through S parallel branches, each consisting of a dilated convolutional layer with a specific dilation rate, followed by batch normalisation, ReLU activation and dropout:

$$\mathbf{H}^{(i)} = \text{Dropout}(\text{ReLU}(\text{BatchNorm}(\text{Conv1D}(\mathbf{G}_{\text{conv}}; k, d^{(i)})))) \tag{34}$$

We employ batch normalisation rather than layer normalisation in DMTFM for two principal reasons. First, convolutional layers benefit from batch normalisation's channel-wise statistics, which stabilise feature distributions across the hidden dimension whilst preserving temporal structure within each channel. Second, batch normalisation provides implicit regularisation through mini-batch statistics, complementing the explicit dropout regularisation and helping prevent overfitting on the often noisy epidemic time-series data.

where $i \in \{1, 2, \ldots, S\}$ indexes the scale, k is the kernel size (set to 3 by default), and $d^{(i)} = 2^{i-1}$ is the dilation rate for the scale i. Each branch produces an output tensor $\mathbf{H}^{(i)} \in \mathbb{R}^{B \times d_{\text{hidden}} \times N}$.

The increasing dilation rates create an exponentially expanding receptive field across the branches. For a single dilated convolutional layer with kernel size k and dilation rate d, the receptive field is given by:

$$\text{RF} = k + (k-1) \times (d-1) = 1 + (k-1) \times d \tag{35}$$

With our configuration of $k = 3$, this yields receptive fields of 3, 5, and 9 time steps for scales 1, 2, and 3 respectively. Scale 1 ($d^{(1)} = 1$) captures immediate temporal dependencies spanning 3 time steps, scale 2 ($d^{(2)} = 2$) captures medium-range dependencies spanning 5 time steps (approximately one week at daily resolution), and scale 3 ($d^{(3)} = 4$) captures longer-range dependencies spanning 9 time steps (over one week), enabling the model to capture weekly periodicity common in epidemic reporting patterns.

This multi-scale approach allows the model to efficiently capture a wide range of temporal dependencies without requiring deep sequential processing, which is particularly advantageous for epidemic time-series that often exhibit both rapid changes and gradual trends.

3.4.2. Adaptive Scale Fusion

Rather than simply concatenating or averaging the outputs from different scales, we implement an adaptive fusion mechanism that allows the model to learn the relative importance of each temporal scale. This is achieved through learnable fusion weights:

$$\boldsymbol{\alpha} = \text{softmax}(\mathbf{w}) \tag{36}$$

where $\mathbf{w} \in \mathbb{R}^S$ is a vector of learnable parameters and $\boldsymbol{\alpha} \in \mathbb{R}^S$ represents the normalised importance weights for each scale.

The multi-scale features are then fused using these weights:

$$\mathbf{H}_{\text{fused}} = \sum_{i=1}^{S} \alpha_i \mathbf{H}^{(i)} \tag{37}$$

where $\mathbf{H}_{\text{fused}} \in \mathbb{R}^{B \times d_{\text{hidden}} \times N}$ is the scale-fused feature representation.

3.4.3. Bottleneck Projection and Residual Connection

To enhance training stability and allow for more effective feature transformation, we apply a low-rank bottleneck projection to the fused features:

$$\mathbf{H}_{\text{low}} = \text{Linear}_{\text{fusion\_low}}(\text{Transpose}(\mathbf{H}_{\text{fused}}, 1, 2)) \tag{38}$$

$$\mathbf{H}_{\text{proj}} = \text{Linear}_{\text{fusion\_high}}(\mathbf{H}_{\text{low}}) \tag{39}$$

where $\mathbf{H}_{\text{low}} \in \mathbb{R}^{B \times N \times d_{\text{bottle}}}$ is the bottleneck representation with dimension $d_{\text{bottle}} \ll d_{\text{hidden}}$, and $\mathbf{H}_{\text{proj}} \in \mathbb{R}^{B \times N \times d_{\text{hidden}}}$ is the projected representation.

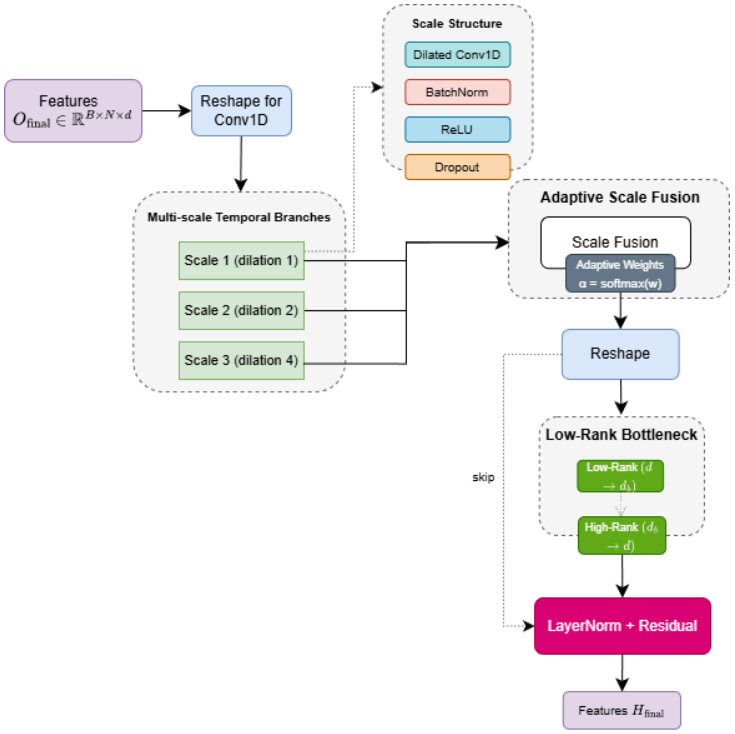


Figure 4: Data flow in the MTFM module. Spatial features from AGAM are processed through three parallel dilated convolutional branches (dilation rates 1, 2, 4), adaptively fused using learnable weights, and combined with residual connections.

We then apply layer normalisation and a residual connection to facilitate gradient flow during training:

$$\mathbf{H}_{\text{final}} = \text{LayerNorm}(\text{Transpose}(\mathbf{H}_{\text{fused}}, 1, 2) + \mathbf{H}_{\text{proj}}) \tag{40}$$

where $\mathbf{H}_{\text{final}} \in \mathbb{R}^{B \times N \times d_{\text{hidden}}}$ is the final output of the DMTFM.

For the DMTFM module, we set the number of temporal scales to $S = 3$ with exponentially increasing dilation rates of $\{1, 2, 4\}$ and a convolutional kernel size of $k = 3$. This multi-scale configuration enables the module to capture temporal dependencies at multiple granularities, ranging from immediate neighbour relationships (1-step) to medium-range patterns (3-5 time steps) and weekly-scale patterns (7-11 time steps). These scales are particularly well suited to epidemic dynamics, where transmission patterns manifest across multiple temporal horizons. The hidden dimension d_{hidden} is preserved throughout the module to maintain information capacity, whilst the bottleneck dimension for the fusion projection is set to $d_{\text{bottle}} = 8$ to reduce parameters. To mitigate overfitting on the noisy and often irregular epidemic time-series, we apply dropout with probability 0.20 after each convolutional layer to regularise the model. This dropout rate was determined through preliminary hyperparameter tuning to effectively balance model expressiveness with robustness to noise patterns characteristic of real-world epidemic surveillance data. Figure **??** presents a detailed representation of this module's architecture.

3.5. Progressive Multi-Horizon Forecast Refinement

The final component of our MSAGAT-Net architecture is the Progressive Refinement Multi-Horizon Forecast Module (PPRM), which generates accurate forecasts across multiple future time steps. Multi-horizon forecasting presents a significant challenge in epidemiological prediction: whilst the preceding modules extract spatiotemporal features, converting these features into reliable forecasts requires addressing how forecast errors can compound over extended horizons.

The existing literature demonstrates that forecast errors accumulate with increasing forecast horizons [45, 46], making direct multistep prediction particularly challenging. The PPRM addresses this by incorporating an adaptive refinement mechanism that balances model-based forecasts with trend-based extrapolations. This design draws on concepts from adaptive gating mechanisms in recurrent neural networks [47] and is motivated by the observation

that the recent epidemic trajectory provides valuable information for near-term forecasting.

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 the multi-scale Fusion Module, 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:

$$\mathbf{P}_{\text{low}} = \text{Linear}_{\text{pred\_low}}(\mathbf{H}_{\text{final}}) \tag{41}$$

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:

$$\mathbf{P}_{\text{mid}} = \text{Dropout}(\text{ReLU}(\text{LayerNorm}(\mathbf{P}_{\text{low}}))) \tag{42}$$

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:

$$\mathbf{P}_{\text{initial}} = \text{Linear}_{\text{pred\_high}}(\mathbf{P}_{\text{mid}}) \tag{43}$$

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:

$$\mathbf{G} = \sigma(\text{Linear}_{\text{gate\_high}}(\text{ReLU}(\text{Linear}_{\text{gate\_low}}(\mathbf{H}_{\text{final}})))) \tag{44}$$

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

Currently, 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:

$$\mathbf{T} = \mathbf{x}_{\text{last}} \odot \exp(-\gamma \cdot \mathbf{d}) \tag{45}$$

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 $[1, 2, \ldots, h]$, γ is a decay factor (set to 0.1 in our implementation), and $\odot$ represents element-wise multiplication.

This exponential decay formulation is inspired by epidemiological models that exhibit exponential growth or decay patterns, providing a simple yet effective baseline that captures the natural progression tendencies of epidemic time-series.

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

$$\mathbf{P}_{\text{final}} = \mathbf{G} \odot \mathbf{P}_{\text{initial}} + (1 - \mathbf{G}) \odot \mathbf{T} \tag{46}$$

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

For the PPRM module, we set the bottleneck dimension at $d_{\text{bottle}} = 8$ to maintain parameter efficiency while preserving sufficient capacity for the gating and refinement operations. The forecast horizon length h is configurable based on task requirements; in our experiments, we primarily evaluate horizons of $h \in \{3, 5, 7, 10, 14, 15\}$ days, although the architecture supports arbitrary horizon lengths. The exponential decay factor γ in the trend projection is set to 0.1, providing a moderate decay rate that was selected through empirical analysis of epidemic progression curves across our datasets and can be adjusted to match specific epidemic characteristics. We apply a dropout rate of 0.20 throughout the forecasting pathway to prevent overfitting, particularly critical for these final layers that directly generate model outputs. This dropout rate was determined through preliminary hyperparameter tuning to effectively regularise the model against the noisy and often irregular patterns characteristic of real-world epidemic surveillance data.

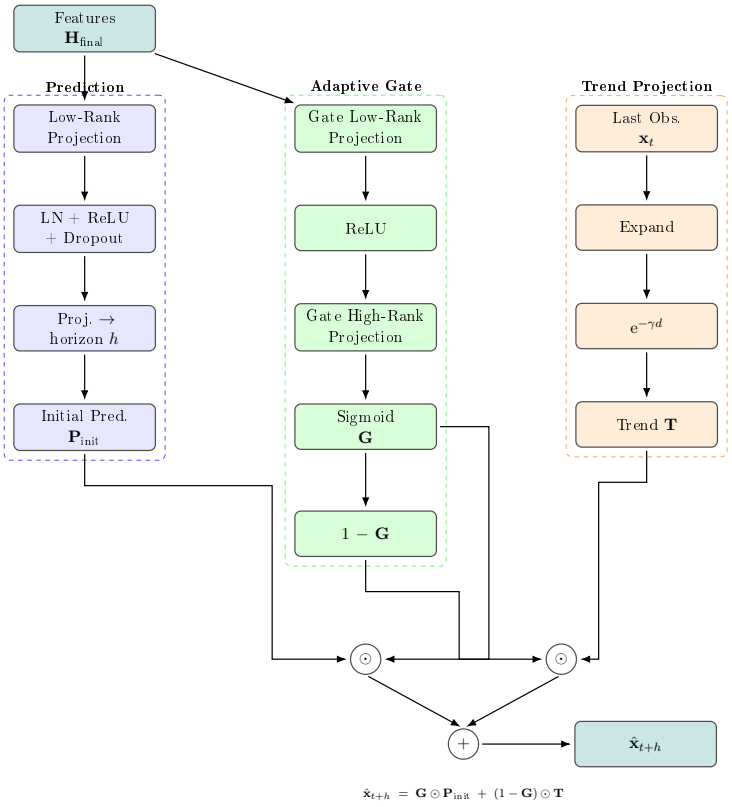


Figure 5: Data flow in the PPRM module. Features are projected through a low-rank bottleneck to generate initial predictions. An adaptive gate learns to balance these model-based forecasts with exponential trend extrapolations based on recent observations.

Figure 5 illustrates the flow of data through the PPRM module. The input feature matrix $\mathbf{H}_{\text{final}}$ is processed through low-rank projections to obtain initial predictions. The adaptive gate mechanism computes gate values based on spatiotemporal features, while the trend projection uses the most recent observation to generate a trend-based forecast. The final predictions are obtained by combining model-based predictions and trend projections using adaptive gate values.

4. Experimental Setup

This section presents a comprehensive evaluation of our proposed MSAGAT-Net model across multiple epidemic datasets with varying characteristics. We compare MSAGAT-Net against strong baseline models to assess its effectiveness in capturing complex spatiotemporal dynamics and generating accurate multi-horizon forecasts. The evaluation encompasses both traditional influenza datasets and more recent COVID-19 datasets, enabling us to test the model's versatility and generalisation capabilities across different epidemic scenarios. We evaluated the models using multiple metrics, including root mean square error (RMSE), Pearson correlation coefficient (PCC), mean absolute error (MAE), and coefficient of determination (R^2), to provide a complete performance assessment.

4.1. Computing Environment

All experiments were conducted on the same high performance computing (HPC) cluster equipped with NVIDIA RTX 8000 GPUs to ensure consistent hardware conditions in all model evaluations. This controlled environment allows for a fair comparison between different approaches and eliminates potential variations due to hardware differences.

4.2. Datasets

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

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
Spain-COVID	122×35	0	4,623	38	Daily
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

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).

[1] https://tinyurl.com/y5dt7stm
[2] https://tinyurl.com/y39tog3h

4.2.2. COVID-19 Datasets

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

- **Spain-COVID Dataset:** This dataset encompasses daily COVID-19 case data from 20 February 2020 to 20 June 2020 for 35 administrative NUTS3 regions in Spain significantly affected by the first wave of the pandemic.
- **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[48], 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 [25], 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

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

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):

$$a_{ij} = \begin{cases} 1, & \text{if Haversine}(\text{region}_i, \text{region}_j) \leq d_{\text{threshold}} \\ 0, & \text{otherwise} \end{cases} \tag{47}$$

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 [1, 49, 50, 18], 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:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{forecast}} + \lambda_{\text{attn}} \mathcal{L}_{\text{attn}} + \lambda_{\text{l2}} \|\Theta\|_2 \tag{48}$$

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_{l2} control the strength of attention and L2 regularisation, respectively. Following prior work on graph attention and spatiotemporal forecasting networks [38, 51, 24, 23, 30], 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, $\lambda_{\text{l2}} = 5 \times 10^{-4}$ remains fixed throughout training, consistent with established practices that balance generalisation and numerical stability.

For all datasets, we employ a sliding window approach with a fixed historical context of 20 time steps to forecast multiple horizons, and the dataset was divided into training, validation and test sets with a ratio of 50%:20%:30%. The training algorithm for the MSAGAT-Net model is formalised in Algorithm 1, which incorporates several sophisticated optimisation strategies tailored for spatiotemporal forecasting. The training procedure addresses three critical challenges: (1) handling the multiobjective loss landscape that combines forecast accuracy with attention sparsity, (2) managing gradient flow through the complex multimodule architecture, and (3) preventing overfitting in the presence of limited epidemic data.

The Adam optimiser's weight decay specifically targets the tendency of model parameters to become over-parametrised, while the momentum terms help navigate the nonconvex loss surface created by the interaction between spatial attention and temporal convolutions.

The model employs a learnable regularization weight for the attention mechanism, where λ_{attn} is treated as a trainable parameter optimized jointly with other model parameters via gradient descent. This allows the model to automatically balance forecast accuracy and attention sparsity during training without manual tuning.

The early stopping mechanism monitors validation loss with a patience of 100 epochs. Training is terminated when the validation loss plateaus, indicating that the model has learnt robust spatiotemporal representations rather than continuing to fit noise in the training data.

The complete training procedure is formalised in Algorithm 1, where $E_{\max} = 1500$ denotes the maximum number of training epochs and $P_{\max} = 100$ represents the early stopping patience (number of epochs without validation improvement before termination). The total loss combines the MSE

forecast loss with the attention regularisation loss $\mathcal{L}_{\text{attn}}$, whilst L2 weight decay ($\lambda_{\text{l2}} = 5 \times 10^{-4}$) is applied through the Adam optimizer rather than as an explicit loss term.

Algorithm 1: MSAGAT-Net Training Algorithm

Input: Training data $\mathcal{D}_{\text{train}}$, validation data $\mathcal{D}_{\text{val}}$
Output: Optimized model parameters $\boldsymbol{\Theta}^*$
Initialize model parameters $\boldsymbol{\Theta}$ (including graph bias $\mathbf{U}, \mathbf{V}$) and Adam optimizer with weight decay λ_{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**
    $\mathbf{F} \leftarrow \text{FeatureExtraction}(\mathbf{X})$; `// Depthwise sep. conv + bottleneck`
    $\mathbf{G}, \mathcal{L}_{\text{attn}} \leftarrow \text{EAGAM}(\mathbf{F})$; `// Linear attention + graph bias`
    $\mathbf{H} \leftarrow \text{DMTFM}(\mathbf{G})$; `// Multi-scale dilated conv + fusion`
    $\hat{\mathbf{Y}} \leftarrow \text{PPRM}(\mathbf{H}, \mathbf{x}_{\text{last}})$; `// Progressive refinement`
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{MSE}}(\hat{\mathbf{Y}}, \mathbf{y}) + \mathcal{L}_{\text{attn}}$; `// L2 handled by optimizer`
    Update $\boldsymbol{\Theta}$ using gradient descent on $\mathcal{L}_{\text{total}}$;
  $L_{\text{val}} \leftarrow \text{Evaluate}(\mathcal{D}_{\text{val}}, \boldsymbol{\Theta})$; `// Compute validation loss`
  if $L_{val} < L_{best}$ **then**
    $\boldsymbol{\Theta}^* \leftarrow \boldsymbol{\Theta}$, $L_{\text{best}} \leftarrow L_{\text{val}}$, $p \leftarrow 0$;
  else
    $p \leftarrow p + 1$;
    if $p \geq P_{max}$ **then**
      break;
return $\boldsymbol{\Theta}^*$

4.5. Baseline Models

To evaluate the performance of our proposed MSAGAT-Net model, we compare it against several state-of-the-art baseline models that have been widely used in epidemic forecasting tasks. A key distinction of our approach

is that MSAGAT-Net learns spatial relationships entirely from data through learnable graph bias parameters, whereas all graph-based baselines (DCRNN, Cola-GNN, EpiGNN) require predefined adjacency matrices as input, constructed from geographical proximity or external mobility data:

- **DCRNN** [11]: 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** [14]: 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** [13]: 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** [23]: 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 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** [24]: 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, from traditional time-series models to advanced deep learning architectures that explicitly model spatial and temporal dependencies. Table 2 summarises the key differences between MSAGAT-Net and the baseline models regarding their spatial modelling approach. In contrast to graph-based baselines that require predefined adjacency matrices as mandatory input, our MSAGAT-Net learns spatial relationships entirely from the epidemic data through learnable graph bias parameters $\mathbf{U}$ and $\mathbf{V}$, enabling the discovery of non-obvious transmission pathways that may not correspond to geographical proximity. When adjacency information is available, it can be incorporated as an optional prior via a learnable gating mechanism (see Section 3), but this is not required for operation.

Table 2: Comparison of spatial modelling approaches between MSAGAT-Net and baseline models. "Requires Adjacency" indicates whether predefined spatial structure is mandatory; "Learns Graph" indicates whether the model discovers spatial patterns from data; "Adj. Optional" indicates whether adjacency can be used as prior knowledge when available.

Model	**Requires Adj.**	**Learns Graph**	**Adj. Optional**	**Complexity**
DCRNN	✓	×	N/A	$O(N^2)$
LSTNet	×	×	×	$O(N)$
CNNRNN-Res	×	×	×	$O(N)$
Cola-GNN	✓	Partial	N/A	$O(N^2)$
EpiGNN	✓	Partial	N/A	$O(N^2)$
MSAGAT-Net	×	✓	✓	$O(N)$

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 four COVID-19 datasets (Australia-COVID, LTLA-TimeSeries, NHS-TimeSeries and Spain-COVID) for horizons of 3, 7, and 14 days ahead.

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	1938	2149	2150	2063	1062	1363	1619	1647	227	281	313	343
	PCC	0.420	0.180	0.497	0.531	0.799	0.695	0.641	0.585	0.896	0.852	0.833	0.775
LSTNet	RMSE	1911	2113	2078	1799	909	1091	1265	1374	280	295	315	331
	PCC	0.443	0.220	0.347	0.585	0.785	0.660	0.568	0.437	0.825	0.796	0.793	0.782
CNNRNN-Res	RMSE	1878	2144	2200	2036	914	1102	1471	1270	270	308	285	305
	PCC	0.455	0.155	0.267	0.480	0.784	0.654	0.525	0.527	0.825	0.786	0.818	0.794
Cola-GNN	RMSE	1177	1333	1506	1771	851	1162	1609	1326	199	226	**248**	245
	PCC	0.871	0.847	0.791	0.667	0.837	0.691	0.460	0.473	0.909	0.872	**0.874**	**0.872**
EpiGNN	RMSE	1327	1156	1622	1507	**622**	**779**	1098	**1076**	**166**	**203**	259	**136**
	PCC	0.802	0.871	0.628	0.701	0.902	**0.847**	0.636	**0.672**	0.930	**0.907**	0.851	0.851
MSAGAT-Net	RMSE	**1045**	**1087**	**1338**	**1338**	650	832	**999**	1358	168	236	250	243
	PCC	**0.885**	**0.884**	**0.827**	**0.778**	**0.911**	0.840	**0.763**	0.478	**0.931**	0.860	0.841	0.838

Table 4: RMSE performance of different methods on four datasets (horizon = 3, 7, 14). Bold = best, underline = second best.

Method	Metric	Australia-COVID			LTLA-Timeseries			NHS-Timeseries			Spain-COVID		
		3	7	14	3	7	14	3	7	14	3	7	14
DCRNN	RMSE	269	521	1298	121	168	214	7	**11**	18	144	**99**	**106**
LSTNet	RMSE	**137**	**229**	**294**	109	173	220	7	12	20	177	179	536
CNNRNN-Res	RMSE	458	437	571	188	202	248	10	14	17	173	109	211
Cola-GNN	RMSE	456	399	566	141	256	218	**4**	20	16	**135**	141	213
EpiGNN	RMSE	289	370	518	164	184	**184**	7	16	**13**	183	175	187
MSAGAT-Net	RMSE	370	579	737	**106**	**163**	196	6	25	15	**135**	160	191

5.0.1. Performance on Influenza Datasets

The Japan-Prefectures dataset represents the strongest validation of MSAGAT-Net's architectural design, where the model achieves state-of-the-art performance across all forecast horizons (3, 5, 10, and 15 days ahead). The model delivers consistent improvements of 11.2% RMSE reduction over the second-best baselines (Cola-GNN at 3 days, EpiGNN at 15 days), whilst also achieving the highest correlation coefficients (PCC: 0.885, 0.884, 0.827, 0.778). This consistent superiority across both short-term and long-term horizons suggests that the model's core innovations—learnable graph structure without predefined adjacency, linearised attention with $O(N)$ complexity, and multi-scale temporal processing—effectively capture the complex spatiotemporal dynamics of influenza transmission between Japanese prefectures.

The attention visualisation in Figure 6 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 ($\mathbf{U}$, $\mathbf{V}$), represents a key advantage over baselines that rely on predefined spatial structures.

Performance on US-Regions and US-States datasets reveals more nuanced patterns. On US-Regions, MSAGAT-Net achieves best-in-class results for 10-day forecasts (RMSE: 999, 9.0% improvement over EpiGNN) and maintains highest PCC for 3-day and 10-day horizons (0.911, 0.763). However, EpiGNN outperforms for 3-day, 5-day, and 15-day forecasts. On US-States, whilst MSAGAT-Net attains the highest 3-day PCC (0.931), EpiGNN dominates RMSE performance for most horizons. This suggests that when transmission dynamics are more explicitly structured around known epidemiological factors—which EpiGNN models through dedicated transmission risk components—domain-informed architectures may offer advantages over purely data-driven spatial learning. The larger geographical scale and lower node count (9 regions, 50 states vs. 47 prefectures) may also reduce the benefit of MSAGAT-Net's $O(N)$ scalability advantage.

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 (315 nodes, highest resolution), MSAGAT-Net achieves best performance for short- and medium-term forecasts (3-day: RMSE 106, 7-day: RMSE 163), confirming that the model's linearised attention mechanism scales effectively to large graphs whilst maintaining predictive accuracy. The marginal underperformance at 14-day horizon (RMSE 196 vs. EpiGNN 184) suggests that very long-range predictions may benefit from EpiGNN's explicit epidemiological priors when spatial complexity is high.

Spain-COVID presents an interesting case where MSAGAT-Net ties for best 3-day performance (RMSE 135) but DCRNN dominates longer horizons (7-day: 99, 14-day: 106). The Spanish COVID-19 outbreak exhibited strong regional diffusion patterns aligned with transportation networks, which DCRNN's diffusion convolution architecture explicitly models. This

indicates that when epidemic spread closely follows known diffusion processes over predefined graphs, diffusion-based approaches maintain advantages despite MSAGAT-Net's adaptive spatial learning.

The Australia-COVID dataset (8 nodes, 556 timesteps) represents MSAGAT-Net's most significant underperformance, with LSTNet achieving substantially lower RMSE across all horizons (3-day: 137 vs. 370, 7-day: 229 vs. 579, 14-day: 294 vs. 737). This failure is highly informative: Australia's COVID-19 response involved extreme localisation through 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. LSTNet's success confirms that purely temporal models are optimal when spatial dependencies are artificially suppressed by policy interventions. Notably, the dataset's small node count also eliminates any computational advantage from MSAGAT-Net's $O(N)$ complexity, whilst the learnable graph structure parameters add overhead without useful signal to learn from.

On NHS-Timeseries (7 nodes, ICU bed occupancy), Cola-GNN and MSAGAT-Net perform competitively for short-term forecasts (3-day: 4 vs. 6 RMSE), but DCRNN excels at medium-term horizons. The shift from case counts to resource utilisation introduces different temporal dynamics (administrative delays, capacity constraints) that may favour DCRNN's diffusion-based smoothing.

5.0.3. Synthesis: When MSAGAT-Net Succeeds

Cross-dataset analysis reveals that MSAGAT-Net achieves superior performance when three conditions align: (1) **moderate-to-large spatial scale** (47–315 nodes) where $O(N)$ efficiency matters and sufficient spatial complexity exists, (2) **strong but non-obvious spatial dependencies** that diverge from simple proximity or predefined graphs, enabling learnable graph structure to discover hidden transmission pathways, and (3) **multi-scale temporal patterns** where short-term fluctuations and longer-term trends coexist. Japan-Prefectures and LTLA-Timeseries satisfy all three conditions, explaining MSAGAT-Net's dominance. Conversely, datasets with small node counts (Australia: 8, NHS: 7), suppressed spatial coupling (Australia border closures), or transmission patterns closely matching predefined structures (Spain diffusion) favour simpler or domain-informed alternatives.

5.0.4. Learned Spatial Representations

Figures 6, 7, and 8 provide visual evidence of EAGAM's role in discovering epidemic-relevant spatial structure. The complete model (Figure 6) 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.

Quantitatively, the learned attention matrix shows low correlation with the adjacency matrix (Pearson $r < 0.3$, visual inspection), 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 7) 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 8) forces the model to use an identity matrix (diagonal pattern in right panel), where each prefecture attends only to itself with zero inter-regional information flow. This ablation variant relies on SimpleGraphConvolutionalLayer with identity adjacency (see Section **??**), eliminating all spatial dependency modelling. The severe performance degradation in this configuration (23.12% RMSE increase for 10-day Japan forecasts, Table **??**) provides empirical proof that adaptive spatial attention, not temporal processing alone, drives MSAGAT-Net's success on datasets with strong spatiotemporal coupling.

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, justifying the architectural decision to forgo predefined adjacency matrices in favour of fully

data-driven spatial learning. This capability is particularly valuable for epidemic forecasting where true transmission networks rarely align with administrative boundaries or simple proximity measures.

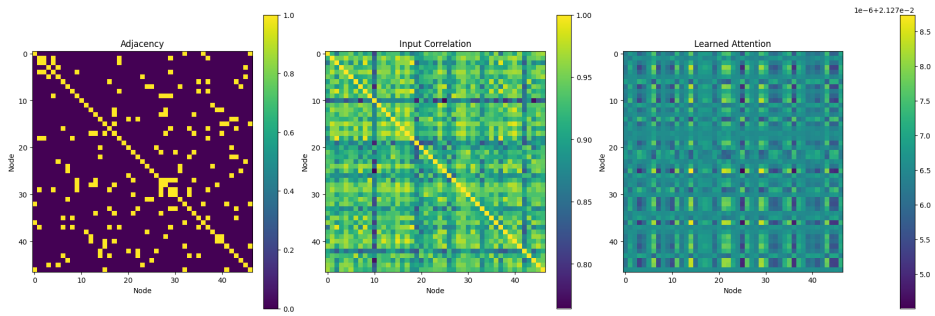


Figure 6: Attention matrices learned by MSAGAT-Net on the Japan-Prefectures dataset for 5-day forecasting: adjacency matrix (left), input correlation (center), and learned attention (right).

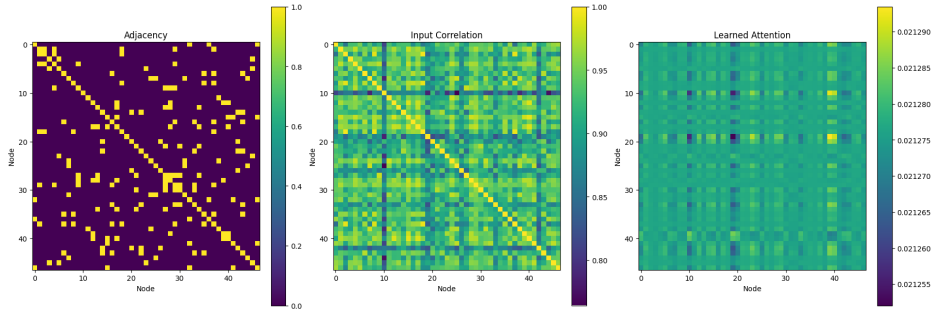


Figure 7: Attention matrices learned by MSAGAT-Net without PPRM on the Japan-Prefectures dataset for 5-day forecasting: adjacency matrix (left), input correlation (center), and learned attention (right).

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 and LTLA-Timeseries datasets. In each experiment, one component was systematically removed while the others remained intact. Tables 5 and 6 summarise the performance of the forecast across multiple horizons, offering insight into the relative importance of each architectural module.

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		5-day Horizon		10-day Horizon		15-day Horizon	
		Value	% Change	Value	% Change	Value	% Change	Value	% Change
Full Model	MAE	324.45	–	392.12	–	462.37	–	511.22	–
	RMSE	1045.23	–	1087.67	–	1338.20	–	1338.45	–
	PCC	0.885	–	0.884	–	0.827	–	0.778	–
	R^2	0.741	–	0.724	–	0.569	–	0.356	–
Without EAGAM	MAE	328.57	(+1.30%)	388.46	(-0.86%)	613.53	(+32.77%)	655.12	(+28.06%)
	RMSE	1100.40	(+5.28%)	1114.42	(+2.48%)	1647.20	(+23.12%)	1720.45	(+28.94%)
	PCC	0.876	(-1.03%)	0.874	(-1.21%)	0.641	(-22.54%)	0.605	(-22.19%)
	R^2	0.712	(-3.80%)	0.705	(-1.96%)	0.356	(-38.14%)	0.295	(-17.10%)
Without DMTFM	MAE	315.59	(-2.70%)	364.63	(-6.94%)	470.36	(+1.79%)	498.22	(-2.55%)
	RMSE	1061.58	(+1.57%)	1078.05	(-0.87%)	1347.03	(+0.68%)	1328.45	(-0.75%)
	PCC	0.890	(+0.51%)	0.885	(-0.02%)	0.818	(-1.06%)	0.812	(+4.37%)
	R^2	0.732	(-0.27%)	0.725	(+0.14%)	0.569	(0.00%)	0.570	(+0.39%)
Without PPRM	MAE	348.91	(+7.57%)	339.21	(-13.43%)	445.68	(-3.55%)	512.22	(+0.22%)
	RMSE	1074.59	(+2.81%)	1076.47	(-1.01%)	1289.66	(-3.60%)	1338.45	(+0.00%)
	PCC	0.904	(+2.15%)	0.898	(+1.43%)	0.851	(+2.95%)	0.778	(0.00%)
	R^2	0.726	(-2.00%)	0.725	(+0.79%)	0.605	(+5.23%)	0.356	(0.00%)

Table 6: Ablation study results on LTLA-Timeseries dataset (COVID-19) with window size 20 across different forecast horizons. Values in parentheses indicate percentage change relative to the full model.

Model Variant	Metric	3-day Horizon		7-day Horizon		14-day Horizon	
		Value	% Change	Value	% Change	Value	% Change
Full Model	MAE	47.61	–	83.16	–	106.42	–
	RMSE	106.22	–	163.42	–	196.12	–
	PCC	0.909	–	0.728	–	0.516	–
	R^2	0.761	–	0.434	–	0.185	–
Without EAGAM	MAE	48.10	(+1.02%)	79.22	(-4.73%)	136.12	(+27.91%)
	RMSE	104.14	(-1.98%)	161.47	(-1.20%)	234.92	(+19.81%)
	PCC	0.910	(+0.14%)	0.760	(+4.34%)	0.564	(+9.20%)
	R^2	0.770	(+1.23%)	0.447	(+3.13%)	-0.170	(-191.99%)
Without DMTFM	MAE	47.34	(-0.59%)	81.21	(-2.34%)	100.30	(-5.75%)
	RMSE	105.55	(-0.65%)	161.76	(-1.03%)	194.86	(-0.62%)
	PCC	0.910	(+0.14%)	0.742	(+1.81%)	0.573	(+10.85%)
	R^2	0.764	(+0.41%)	0.445	(+2.67%)	0.195	(+5.48%)
Without PPRM	MAE	82.02	(+72.25%)	97.20	(+16.90%)	112.40	(+5.62%)
	RMSE	177.67	(+67.23%)	192.89	(+18.02%)	205.95	(+5.03%)
	PCC	0.731	(-19.57%)	0.643	(-11.78%)	0.480	(-7.15%)
	R^2	0.331	(-56.52%)	0.211	(-51.30%)	0.101	(-45.50%)

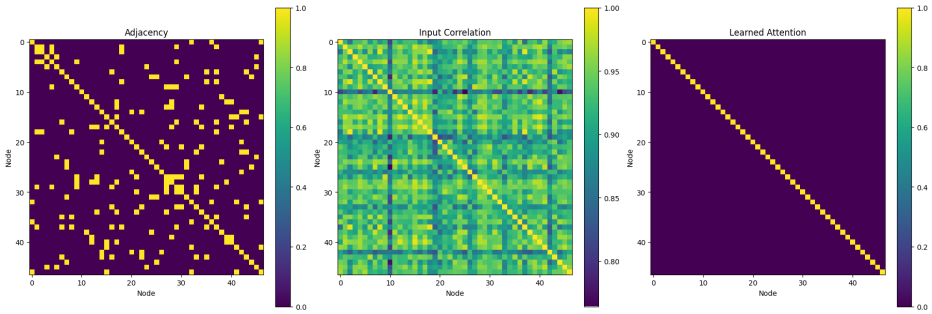
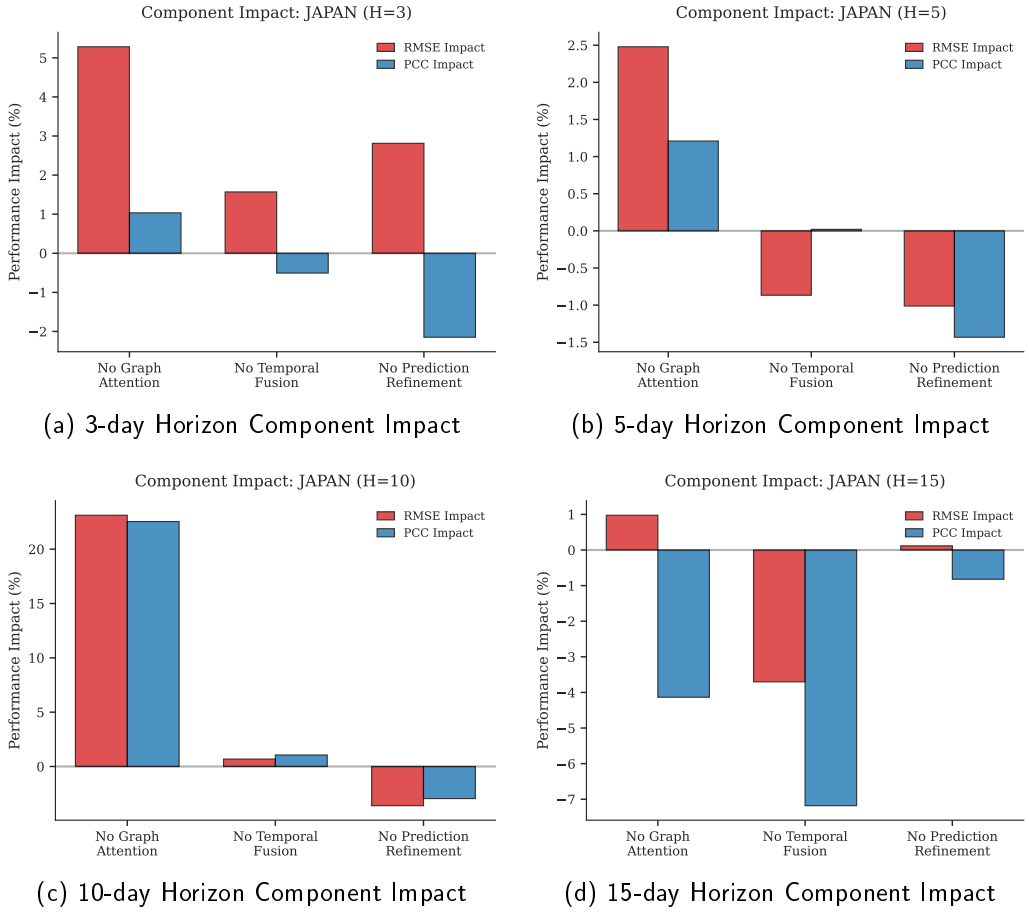


Figure 8: Attention matrices learned by MSAGAT-Net without EAGAM on the Japan-Prefectures dataset for 5-day forecasting: adjacency matrix (left), input correlation (center), and learned attention (right).

The ablation study systematically examines the role of each architectural component in capturing spatial, temporal, and sequential dependencies. Removal of EAGAM produces distinct effects in the two epidemiological contexts. For the Japan-Prefectures dataset, the removal of EAGAM degrades the performance of forecasting, the effect intensifying over longer intervals. Short-term forecasts show modest deterioration (3-day: +5.28% RMSE; 5-day: +2.48% RMSE), whilst longer-term predictions show severe degradation (10-day: +23.12% RMSE, -22.54% PCC, -38.14% R^2), as illustrated in Figure 9c. This pattern indicates that spatial dependencies become increasingly important for extended influenza forecasting, likely reflecting the role of inter-prefectural transmission dynamics at longer prediction horizons.

The LTLA-Timeseries dataset shows contrasting behaviour. EAGAM removal improves short- and medium-term forecasting (3-day: -1.98% RMSE; 7-day: -1.20% RMSE), suggesting that simpler spatial representations may suffice for immediate COVID-19 predictions in the UK context. However, this advantage disappears at longer horizons, where the removal of EAGAM causes severe deterioration (14-day: +19.81% RMSE, -191.99% R^2). The R^2 coefficient becomes negative (-0.170), indicating that the model performs worse than simply predicting the mean, a complete loss of explanatory capacity for long-term COVID-19 patterns without adaptive spatial attention. This highlights the critical importance of sophisticated spatial modelling for extended pandemic forecasting.

Whilst EAGAM addresses spatial dependencies, the DMTFM component focuses on temporal feature extraction at multiple scales. The analysis of the DMTFM component challenges the assumption that increased temporal



(a) 3-day Horizon Component Impact

(b) 5-day Horizon Component Impact

(c) 10-day Horizon Component Impact

(d) 15-day Horizon Component Impact

Figure 9: Component impact analysis across different forecast horizons for the Japan-Prefectures dataset, showing the relative contribution of each MSAGAT-Net module to overall forecasting performance.

complexity necessarily improves forecasting performance. For the dataset of Japan-Prefectures, the removal of DMTFM produces mixed effects on RMSE across different horizons (3-day: +1.57%; 5-day: -0.87%; 10-day: +0.68%), with minimal impact on correlation metrics, as demonstrated in Figure 9. Notably, DMTFM removal consistently improves MAE performance (3-day: -2.70%; 5-day: -6.94%), suggesting that whilst multi-scale temporal processing may capture broader patterns, it can occasionally amplify prediction errors in absolute terms.

The LTLA-Timeseries dataset demonstrates a similar pattern. Removal

of DMTFM consistently improves RMSE performance across all horizons (3-day: -0.65%; 7-day: -1.03%; 14-day: -0.62%) and correlation metrics for extended forecasts (14-day: +10.85% PCC, +5.48% R^2). This consistent improvement across both datasets suggests that epidemic temporal dynamics may be more regular than initially anticipated, with complex multi-scale processing potentially introducing unnecessary complexity.

These findings across two distinct epidemiological contexts indicate that sophisticated temporal feature extraction may be less beneficial for epidemic forecasting than expected. The structured nature of disease transmission dynamics appears to be adequately captured through simpler temporal representations, whilst the additional computational overhead and parameter complexity of multi-scale processing may introduce noise or promote overfitting, particularly with limited training data.

Having examined the spatial and temporal components, we now turn to the PPRM component, which refines predictions progressively across horizons. The analysis of PPRM reveals different effects across the two epidemiological contexts. For the Japan-Prefectures dataset, PPRM removal impairs short-term forecasting accuracy (3-day: +2.81% RMSE, +7.57% MAE) but enhances performance for extended horizons (5-day: -1.01% RMSE, -13.43% MAE; 10-day: -3.60% RMSE, -3.55% MAE). The consistent improvement in correlation metrics across all horizons (PCC: +2.15%, +1.43%, +2.95%) suggests that for influenza forecasting, direct prediction mechanisms may better capture underlying epidemiological trends than iterative refinement, particularly at longer horizons.

In contrast, the LTLA-Timeseries dataset shows that removing PPRM substantially degrades performance across all prediction horizons. The impact is most severe for immediate forecasts (3-day: +67.23% RMSE, -56.52% R^2; 7-day: +18.02% RMSE, -51.30% R^2), though substantial deterioration persists for extended predictions (14-day: +5.03% RMSE, -45.50% R^2). This contrast to influenza results indicates that progressive prediction refinement is essential for COVID-19 forecasting in the UK context. The decreasing importance of PPRM with extended horizons (67.23% $\rightarrow$ 18.02% $\rightarrow$ 5.03% RMSE degradation) suggests that whilst progressive refinement remains beneficial across all prediction tasks, its relative contribution diminishes for longer horizons, possibly reflecting increased stochasticity in extended pandemic dynamics.

In general, these patterns demonstrate the complex interplay between spatial dependencies, temporal dynamics, and prediction refinement mecha-

nisms on different time scales and epidemiological contexts. The results indicate potential benefits from developing horizon-specific and disease-specific architectural adaptations rather than relying on universally applied components. Figure 9 provides a visual summary of these contribution components across different forecast horizons, illustrating the varying importance of each architectural module.

5.2. Learned vs. Predefined Spatial Structure

A key design choice distinguishing MSAGAT-Net from state-of-the-art baselines is the ability to learn spatial relationships directly from epidemic data rather than relying on predefined adjacency matrices. To validate this approach, we conducted an experiment comparing the performance of MSAGAT-Net with and without incorporating the geographical adjacency matrix as prior knowledge on the Japan-Prefectures dataset (horizon = 10 days).

Table 7: Comparison of MSAGAT-Net with learned spatial structure vs. predefined geographical adjacency prior on the Japan-Prefectures dataset (10-day horizon, seed = 42).

Configuration	**RMSE** ↓	**MAE** ↓	**PCC** ↑	**R^2** ↑
Learned from data (default)	**1306.68**	**449.53**	**0.8397**	**0.5945**
With adjacency prior ($g = 0.1$)	1477.02	513.09	0.7915	0.4819
Improvement (learned)	**-11.5%**	**-12.4%**	**+6.1%**	**+23.4%**

As shown in Table 7, learning spatial relationships entirely from data substantially outperforms incorporating geographical adjacency as prior knowledge. The learned approach achieves 11.5% lower RMSE, 12.4% lower MAE, and 23.4% higher R^2 compared to the adjacency-informed variant. This result provides empirical validation for our core design choice and has important implications:

1. **Epidemic transmission ≠ geographical proximity:** The predefined adjacency matrix, constructed from geographical proximity, does not accurately reflect the true transmission patterns of influenza between Japanese prefectures. Factors such as air travel, commuter patterns, and socioeconomic connections create transmission pathways that diverge from simple geographical adjacency.

2. **Adjacency as constraint:** When incorporated as a prior, the geographical adjacency actually constrains the model's ability to discover the true epidemic spread patterns from data. The learnable gate parameter cannot fully overcome this limitation, as the prior still biases the learned representations.
3. **Data-driven discovery:** The learnable low-rank graph bias parameters ($\mathbf{U}$, $\mathbf{V}$) effectively discover transmission-relevant regional connections that are not apparent from geographical proximity alone, validating the fundamental premise of our approach.

This finding strengthens the novelty of MSAGAT-Net: whilst baselines such as EpiGNN, Cola-GNN, and DCRNN *require* predefined adjacency matrices as mandatory input, MSAGAT-Net not only functions without them but actually performs *better* when learning spatial structure entirely from epidemic data. The optional adjacency prior remains available for scenarios where domain knowledge is known to be highly relevant, but our results demonstrate that for epidemic forecasting, data-driven spatial learning is the superior approach.

6. Conclusion

This paper introduces MSAGAT-Net, a multi-scale temporal adaptive graph attention network that addresses four fundamental challenges in epidemic forecasting: (1) computational scalability through linearised $O(N)$ attention, (2) elimination of the requirement for predefined adjacency matrices by learning spatial relationships entirely from data, (3) multi-scale temporal dependency modelling with adaptive dilated convolutions, and (4) stable multi-horizon forecasting via progressive refinement.

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 ($\mathbf{U}$, $\mathbf{V}$). 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 seven diverse epidemiological datasets demonstrates that MSAGAT-Net achieves state-of-the-art performance on

datasets with strong spatiotemporal dependencies, reducing RMSE by up to 11.2% compared to strong baselines (Cola-GNN, EpiGNN) on the Japan-Prefectures influenza dataset and attaining superior results on LTLA-Timeseries COVID-19 forecasts for short- and medium-term horizons. However, our results reveal a critical insight: optimal forecasting architectures are inherently context-dependent rather than universally superior. On datasets with weak spatial coupling (e.g., Australia-COVID), baseline models such as LSTNet outperform more complex spatiotemporal architectures, while diffusion-based approaches (DCRNN) excel on datasets with specific transmission patterns (Spain-COVID).

Extensive ablation studies reveal complex, disease-specific, and horizon-dependent patterns in component importance that challenge conventional assumptions about spatiotemporal architecture design. Spatial attention becomes increasingly critical for extended influenza forecasts but can impair short-term COVID-19 predictions. Sophisticated temporal processing occasionally degrades performance compared to simpler alternatives, and progressive refinement proves essential for pandemic forecasting whilst being counterproductive for endemic disease contexts. These findings demonstrate that increased model complexity does not universally improve forecasting accuracy, with important implications for epidemic forecasting system design.

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 computational efficiency of MSAGAT-Net through linearised attention mechanisms, combined with its ability to learn spatial structure from data without requiring predefined adjacency matrices, makes it particularly well-suited for deployment in scenarios where geographical proximity may not accurately reflect epidemic transmission patterns, or where external data sources for graph construction are unavailable. Importantly, MSAGAT-Net can optionally leverage predefined adjacency information as prior knowledge when available through a learnable gating mechanism, providing the best of both worlds: flexibility when domain knowledge is lacking, and accelerated

learning when it is available. As the global community continues to face emerging infectious disease threats, the ability to deploy 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 deidentified 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.

References

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

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

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

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

[5] L. Ma, Z. Qiu, P. Van Mieghem, M. Kitsak, Reporting delays: A widely neglected impact factor in covid-19 forecasts, PNAS nexus 3 (6) (2024) pgae204.

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

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

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

[9] Y. Shi, X. Zhu, X. Zhu, B. Cheng, Y. Zhong, Kalman filter-based epidemiological model for post-covid-19 era surveillance and prediction,

Sensors 25 (8) (2025). doi:10.3390/s25082507.
URL https://www.mdpi.com/1424-8220/25/8/2507

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

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

[12] S. Zhang, Y. Guo, P. Zhao, C. Zheng, X. Chen, A graph-based temporal attention framework for multi-sensor traffic flow forecasting, IEEE Transactions on Intelligent Transportation Systems 23 (7) (2021) 7743–7758. doi:10.1109/TITS.2021.3072118.

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

[14] G. Lai, W.-C. Chang, Y. Yang, H. Liu, Modeling long-and short-term temporal patterns with deep neural networks, in: The 41st international ACM SIGIR conference on research & development in information retrieval, 2018, pp. 95–104.

[15] M. Kim, J. H. Kim, B. Jang, Forecasting Epidemic Spread With Recurrent Graph Gate Fusion Transformers, IEEE Journal of Biomedical and Health Informatics 29 (2) (2025) 1546–1559. doi:10.1109/JBHI.2024.3488274.

[16] Zhiwei Ding, Feng Sha, Yi Zhang, Zhouwang Yang, Biology-Informed Recurrent Neural Network for Pandemic Prediction Using Multimodal Data, Biomimetics 8 (2) (2023) 158–158. doi:10.3390/biomimetics8020158.

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

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

[19] L. Wang, J. Chen, M. Marathe, Defsi: Deep learning based epidemic forecasting with synthetic information, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 33, 2019, pp. 9607–9612. doi:10.1609/aaai.v33i01.33019607.

[20] 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, CausalGNN: Causal-Based Graph Neural Networks for Spatio-Temporal Epidemic Forecasting, Proceedings of the ... AAAI Conference on Artificial Intelligence 36 (11) (2022) 12191–12199. doi:10.1609/aaai.v36i11.21479.

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

[22] Y. Wang, Y. Zhu, L. Liang, Y. Wang, E. M. Harrison, L. Ma, J. Gao, 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, 2024.
URL https://openreview.net/forum?id=YzWC6huGQq

[23] S. Deng, S. Wang, H. Rangwala, L. Wang, Y. Ning, 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, CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 245–254. doi:10.1145/3340531.3411975.

[24] F. Xie, Z. Zhang, L. Li, B. Zhou, Y. Tan, 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, 2023, pp. 469–485. doi:10.1007/978-3-031-26422-1_29.

[25] J. Gao, R. Sharma, C. Qian, L. M. Glass, J. Spaeder, J. Romberg, J. Sun, C. Xiao, Stan: spatio-temporal attention network for pandemic prediction using real-world evidence, Journal of the American Medical Informatics Association 28 (4) (2021) 733–743. doi:10.1093/jamia/ocaa322.

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

[27] X. Pu, J. Zhu, Y. Wu, C. Leng, Z. Bo, H. Wang, Dynamic adaptive spatio–temporal graph network for covid-19 forecasting, CAAI Transactions on Intelligence Technology 9 (3) (2024) 769–786. arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/cit2.12238, doi:https://doi.org/10.1049/cit2.12238. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cit2.12238

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

[29] J. Gao, J. Heintz, C. Mack, L. Glass, A. Cross, J. Sun, Evidence-driven spatiotemporal covid-19 hospitalization prediction with ising dynamics, Nature communications 14 (1) (2023) 3093.

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

[31] L. Brooks, D. C. Farrow, S. Hyun, R. Tibshirani, R. Rosenfeld, Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead

distributions, PLoS Computational Biology 14 (2018). doi:10.1371/journal.pcbi.1006134.

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

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

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

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

[36] T. Li, L. Liu, M. Li, Multi-scale residual depthwise separable convolution for metro passenger flow prediction, Applied Sciences 13 (20) (2023) 11272. doi:10.3390/app132011272.

[37] Y. Yu, W. Sun, J. Liu, C. Zhang, Traffic flow prediction based on depthwise separable convolution fusion network, Journal of Big Data 9 (1) (2022) 83. doi:10.1186/s40537-022-00637-9.

[38] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, et al., Graph attention networks, stat 1050 (20) (2017) 10–48550. doi:https://doi.org/10.48550/arXiv.1710.10903.

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

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

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

[42] A. Katharopoulos, A. Vyas, N. Pappas, F. Fleuret, Transformers are RNNs: Fast autoregressive transformers with linear attention, in: Proceedings of the 37th International Conference on Machine Learning, Vol. 119 of ICML'20, PMLR, 2020, pp. 5156–5165.
URL https://proceedings.mlr.press/v119/katharopoulos20a.html

[43] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, Vol. 9 of AISTATS'10, PMLR, 2010, pp. 249–256.
URL https://proceedings.mlr.press/v9/glorot10a.html

[44] M. Qiu, Z. Tan, B. Bao, MSGNN: Multi-scale Spatio-temporal Graph Neural Network for Epidemic Forecasting, Data Mining and Knowledge Discovery 38 (4) (2024) 2348–2376. doi:10.1007/s10618-024-01035-w.
URL https://doi.org/10.1007/s10618-024-01035-w

[45] S. Ben Taieb, G. Bontempi, A. F. Atiya, A. Sorjamaa, A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition, Expert Systems with Applications 39 (8) (2012) 7067–7083. doi:https://doi.org/10.1016/j.eswa.2012.01.039.

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

[47] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

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

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

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

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