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

Michael Ajao-olarinoye, *Member, IEEE*, Vasile Palade, *Senior Member, IEEE*, Fei He, Petra Wark, and Zindoga Mukandavire

***Abstract*—Accurate and scalable spatiotemporal epidemic forecasting is a critical yet challenging task for public health systems, as traditional models often struggle with computational complexity and capturing multi-scale temporal dynamics. This paper introduces MSAGAT-Net, a novel Multi-Scale Temporal Graph Attention Network designed to overcome these limitations through three key innovations: linearised attention for computational efficiency, adaptive multi-scale temporal processing, and progressive prediction refinement for stable multi-horizon forecasts. Additionally, we contribute two novel epidemic forecasting datasets (LTLA-COVID and NHS-ICUBeds) following established graph construction methodologies. Extensive evaluations on seven diverse epidemic datasets, including influenza and COVID-19 across multiple countries, show that MSAGAT-Net outperforms strong baselines such as Cola-GNN, EpiGNN, achieving up to 11.2% lower root mean square error (RMSE) than the best prior model on the Japan-Prefectures dataset (3, 10 and 15-day horizons) and leading short-horizon results on LTLA. Comprehensive ablation studies reveal disease-specific and horizon-dependent importance of each component, highlighting the importance of adaptive architectural design for epidemic forecasting.**

***Index Terms*—Deep Learning, Graph Neural Networks, Multi-head Attention, Time Series Analysis, Epidemic Forecasting, Adaptive Graph Learning, Multi-scale Feature Fusion, Spatiotemporal Prediction**

I. Introduction

Spatio-temporal forecasting plays a critical role in addressing complex real-world problems, from urban traffic management and environmental monitoring to epidemic prediction and resource allocation. The COVID-19 pandemic has underscored the importance of reliable epidemic forecasting models that can adapt to rapidly changing dynamics and provide actionable insights for timely public health decision-making and resource allocation [1–5]. However, developing such models presents significant challenges due to the complex interplay between spatial dependencies, temporal patterns, and the inherent uncertainty in disease transmission.

The challenge of epidemic forecasting lies in capturing the complex interplay between spatial dependencies (how diseases spread between interconnected regions) and temporal patterns that range from daily reporting fluctuations to seasonal epidemic waves. Traditional epidemiological models, whilst providing valuable theoretical insights, often struggle with the computational complexity required for real-time surveillance across hundreds or thousands of regions.

A number of studies have explored forecasting in various domains, including traffic prediction [6, 7], environmental monitoring [8, 9], and epidemic forecasting [10, 11]. Using deep learning techniques, such as recurrent neural networks (RNN), convolutional neural networks (CNN), and graph neural networks (GNN), researchers have made significant strides in improving forecasting accuracy and efficiency, particularly in capturing complex patterns and relationships within the data. Recent advances in deep learning, particularly graph neural networks, have shown particular promise for epidemic forecasting by learning complex spatiotemporal patterns directly from data.

However, current approaches face three fundamental limitations that prevent their deployment in real-world surveillance systems. First, most graph attention mechanisms suffer from quadratic computational complexity, making them prohibitively expensive for national-scale surveillance involving thousands of regions. Second, existing methods typically employ fixed architectural choices for temporal processing, failing to adapt to the diverse dynamics exhibited by different diseases and forecasting horizons. Third, multi-step forecasting remains unstable due to error accumulation, limiting the utility of long-range predictions essential for public health planning and resource allocation.

To address these challenges, we propose MSAGAT-Net (Multi-Scale Temporal Graph Attention Network), a novel architecture that achieves computational efficiency whilst maintaining high forecasting accuracy across diverse epidemic scenarios. Our architecture integrates four complementary innovations: an efficient feature extraction module using depthwise separable convolutions; an Efficient Adaptive Graph Attention Module (EAGAM) that achieves linear O(N) complexity through low-rank bottleneck projections and linearised attention computation while learning interpretable inter-regional dependencies; a Dilated Multi-Scale Temporal Feature Module (DMTFM) that captures dynamics across multiple temporal resolutions using adaptive dilated convolutions; and a Progressive Prediction Refinement Module (PPRM) that stabilises multi-step predictions by combining model-based forecasts with trend-based extrapolations.

M. Ajao-olarinoye, V. Palade, and F. He are with the Centre for Computational Science and Mathematical Modelling, Coventry University, Coventry, U.K. (e-mail: olarinoyem@coventry.ac.uk).

P. Wark is with the Research Institute for Health and Wellbeing, Coventry University, Coventry, U.K.

Z. Mukandavire is with the Institute of Applied Research and Technology, Dubai International Academic City, Dubai, United Arab Emirates.

Our contributions are as follows:

1) We propose a computationally efficient graph attention mechanism that achieves linear O(N) complexity through low-rank bottleneck projections and linearised attention computation, enabling real-time surveillance across thousands of regions whilst maintaining modelling capacity for accurate epidemic forecasting.
2) We develop an adaptive multi-scale temporal processing framework using dilated convolutions that efficiently captures epidemic dynamics across multiple temporal resolutions, from immediate reporting fluctuations to seasonal epidemic waves.
3) We introduce a progressive prediction refinement module that stabilises multi-step forecasting by adaptively combining model-based predictions with trend-based extrapolations, addressing the critical challenge of error accumulation in long-range epidemic predictions.
4) We introduce two novel epidemic forecasting datasets (LTLA-COVID and NHS-ICUBeds) following established graph construction methodologies, providing new benchmarks for spatiotemporal epidemic forecasting research.
5) We demonstrate substantial performance improvements across seven diverse epidemic datasets, with comprehensive ablation studies revealing disease-specific and horizon-dependent architectural requirements that challenge conventional assumptions about spatiotemporal modelling.

II. Related Work

Epidemic forecasting has undergone a fundamental transformation from classical compartmental models to sophisticated neural architectures. This evolution reflects the growing recognition that real-world disease transmission involves complex spatiotemporal dependencies that resist analytical solutions. We organise our review around three key themes that directly inform our architectural choices: neural spatiotemporal modelling, multi-scale temporal processing, and computational scalability challenges.

A. 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 [12]. A comprehensive taxonomy by Liu et al. [13, 14] distinguishes between Statistical epidemiology models, General machine learning models, Deep-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. [15] pioneered dynamic cross-location attention through Cola-GNN, replacing fixed geographic adjacency matrices with learnable attention weights that adapt to time-varying transmission patterns. This innovation proved particularly effective for influenza forecasting, where seasonal migration patterns and behavioural changes create non-stationary spatial dependencies. However, Cola-GNN's reliance on standard attention mechanisms results in quadratic computational complexity, limiting scalability to fine-grained regional analysis.

Building upon these foundations, Xie et al. [16] developed EpiGNN with transmission risk encoding and 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 (approximately 9.5% RMSE reduction) across diverse epidemic scenarios. Crucially, EpiGNN's architecture revealed that optimal graph structures vary significantly between diseases and temporal scales, suggesting the need for adaptive rather than fixed spatial representations.

Gao et al. [17] proposed STAN, which leverages attention-based graph convolution enhanced with patient electronic health records and geography-based features. Applied to COVID-19 forecasting across all U.S. counties, STAN achieved up to 87% lower mean squared error compared to classical SIR/SEIR models. Notably, STAN incorporated physics-based regularisation derived from compartmental dynamics, demonstrating that domain knowledge can enhance rather than constrain neural learning.

Recent developments have focused on unifying spatial and temporal modelling more deeply. Han et al. [18] 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. [19] proposed DASTGN with dual-scale attention mechanisms that adaptively fuse spatial and temporal effects at both fine-grained and coarse-grained resolutions.

Despite these advances, current spatiotemporal approaches face two critical limitations. First, they typically employ standard attention mechanisms with quadratic complexity, making them computationally prohibitive for large-scale regional analysis. Recent advances in efficient attention, particularly linearised attention mechanisms that achieve $O(N)$ complexity through low-rank matrix decomposition [20], offer promising solutions but remain largely unexplored in epidemic forecasting contexts. Second, they often struggle to provide stable multi-horizon forecasts, as error accumulation compounds over extended prediction windows.

B. Physics-Informed and Hybrid Approaches

A particularly promising research direction involves integrating epidemiological domain knowledge into neural architectures to improve both interpretability and long-range forecast stability. Wang et al. [12] developed CausalGNN, which combines mechanistic ODE-based disease models with dynamic graph neural networks. The architecture employs an attention-based graph module for cross-regional influences whilst a causal module, grounded in SIR-type ordinary differential equations, injects epidemiological context into node embeddings. This mutually-informed design yields more robust predictions with reduced parameter requirements, particularly for scenarios with limited training data.

Cao et al. [21] proposed MepoGNN, employing multi-patch SEIR models within a metapopulation graph neural network framework. By merging region-level compartmental simulators with Graph Attention Networks that process real-world mobility and case data, MepoGNN transforms static travel matrices into dynamic transmission adjacency matrices. Applied to COVID-19 spread in South Korea, learned transmission rates closely aligned with actual policy interventions, demonstrating the value of mechanistic constraints for interpretable predictions.

Gao et al. [22] took a physics-inspired approach with HOIST, drawing analogies between disease spread and Ising spin systems. By treating counties as nodes on a lattice and using Ising dynamics to regularise forecasting models, they encoded the principle that neighbouring regions' case counts should evolve in correlated patterns. Applied to 2,299 U.S. counties for COVID-19 hospitalisation forecasting, HOIST provided policy-relevant insights whilst maintaining competitive accuracy.

Whilst hybrid approaches offer improved interpretability and 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. This suggests the need for architectures that can benefit from domain knowledge without being constrained by rigid mechanistic assumptions.

C. Multi-Scale Temporal Processing

Epidemic time series exhibit complex multi-scale temporal patterns, from daily fluctuations driven by testing schedules and reporting delays to seasonal waves influenced by behavioural changes and environmental factors. Current approaches to multi-step forecasting can be broadly categorised into direct methods, which predict multiple time steps simultaneously, and iterative methods, which generate forecasts sequentially.

Direct multi-horizon models, typically implemented using sequence-to-sequence architectures with LSTM or CNN components, have demonstrated effectiveness in influenza forecasting but require substantial training data and may depend heavily on external covariates [8, 23]. Wang et al. [24] developed DEFSI, combining deep learning with compartmental models to enhance long-range forecasts, but found that performance degrades significantly beyond 4-week horizons due to accumulating uncertainty.

Iterative strategies, whilst more data-efficient, suffer from error accumulation over extended forecasting horizons. This fundamental limitation has motivated the development of multi-module architectures that can simultaneously capture high-frequency fluctuations and low-frequency trends. Deng et al. [15] addressed these challenges through dilated convolutions for multi-scale temporal feature extraction, finding that incorporating seasonal trends significantly improved forecast stability. However, their approach relies on fixed dilation patterns that may not adapt to varying epidemic dynamics across different diseases and regions.

The recognition of both short-term outbreaks and long-term epidemiological waves has led researchers to incorporate external data sources, including climate variables, demographic information, and digital surveillance indicators. Whilst these approaches can improve long-range predictions, they often require extensive feature engineering and may not generalise across different epidemic contexts. Moreover, current multi-scale temporal processing approaches rely on fixed architectural choices (e.g., predetermined dilation patterns) that may not adapt to varying epidemic dynamics across diseases and regions, whilst the computational scalability of complex architectures with multiple dilated convolution branches remains prohibitively expensive for large-scale regional analysis.

These limitations across spatiotemporal modelling, physics-informed approaches, and multi-scale temporal processing highlight three critical gaps in current epidemic forecasting: the quadratic complexity bottleneck of attention mechanisms that prevents real-time surveillance applications; the rigidity of fixed architectural choices that limit effectiveness across diverse epidemic scenarios; and the lack of stable multi-horizon forecasting capabilities essential for public health planning. Our work addresses these challenges through MSAGAT-Net, which integrates linearised attention for computational efficiency, adaptive multi-scale temporal processing, and progressive prediction refinement for stable multi-horizon forecasts.

III. Methodology

A. Problem Formulation

Let us consider N geographical regions (e.g., cities, counties, states, or NHS regions in England) as nodes in a graph. Historical epidemic data are represented as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T] \in \mathbb{R}^{N \times T}$, where $\mathbf{x}_t \in \mathbb{R}^N$ denotes the observed values in all N regions at time step t. Each individual element $x_{i,t}$ represents the epidemic measure (e.g. ventilator bed occupancy or positive case count) for the region i at time t.

For each specific region i, its temporal sequence is represented as $\mathbf{x}^i = [x_{i,1}, x_{i,2}, \ldots, x_{i,T}] \in \mathbb{R}^T$. This dual representation allows us to analyse both the spatial patterns (across regions at a specific time) and temporal patterns (within a region across time).

Our primary objective is to predict future epidemic values for all regions on a specific time horizon h steps ahead. Formally, given the historical data up to time t, we want to predict:

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

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

The spatial relationships between regions are encoded in a graph structure $\mathscr{G} = (\mathscr{V}, \mathscr{E}, \mathbf{A})$, where $\mathscr{V} = \{v_1, v_2, \ldots, v_N\}$ represents the set of regions, $\mathscr{E} \subseteq \mathscr{V} \times \mathscr{V}$ denotes the connections between regions and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix. Each element a_{ij} of $\mathbf{A}$ quantifies the strength of the relationship between regions v_i and v_j. The adjacency matrix can be constructed based on various criteria, such as geographical proximity, transportation networks, or healthcare referral patterns. For example, in the context of epidemic forecasting, the adjacency matrix can reflect the mobility patterns of individuals between regions or the referral pathways of patients between healthcare facilities.

The forecasting task can be formalised as learning a function f that maps recent historical data and graph structure to future predictions.

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

where Θ represents the learnable parameters of our forecasting model.

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 architecture that combines graph attention mechanisms with multi-scale processing.

B. 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 lookback 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. The idea of using depthwise separable convolutions is inspired by the success of this approach in computer vision tasks, where it has been shown to significantly reduce the number of parameters and computational complexity while maintaining high performance [25]. Our adaptation of this technique draws from the work of [26] and [27], who have successfully applied depthwise separable convolutions to feature extraction rather than the standard convolutional approach. This allows us to efficiently capture the temporal dynamics of epidemic data without incurring the high computational costs associated with traditional convolutional architectures.

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 time-series $\mathbf{x}^i \in \mathbb{R}^w$, the depthwise convolution applies a separate filter to each input channel (in this case, we treat the time-series as a single-channel input):

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

where $\mathbf{z}^i_{\text{depth}} \in \mathbb{R}^{w\times d_{\text{mid}}}$ represents the intermediate features after depthwise convolution, and d_{mid} is the number of intermediate feature channels.

Following the depthwise convolution, a pointwise convolution (implemented as a 1×1 convolution) is applied to combine the features across channels:

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

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

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 separable convolution in depth 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.

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:

1) By compressing information through a dimension bottleneck $d_{\text{bottle}} \ll w \times d_{\text{feat}}$, we reduce computational complexity from $\mathscr{O}(N \times w \times d_{\text{feat}} \times d_{\text{hidden}})$ to $\mathscr{O}(N \times (d_{\text{bottle}} \times d_{\text{hidden}} + w \times d_{\text{feat}} \times d_{\text{bottle}}))$, enabling efficient processing of large-scale spatio-temporal datasets.

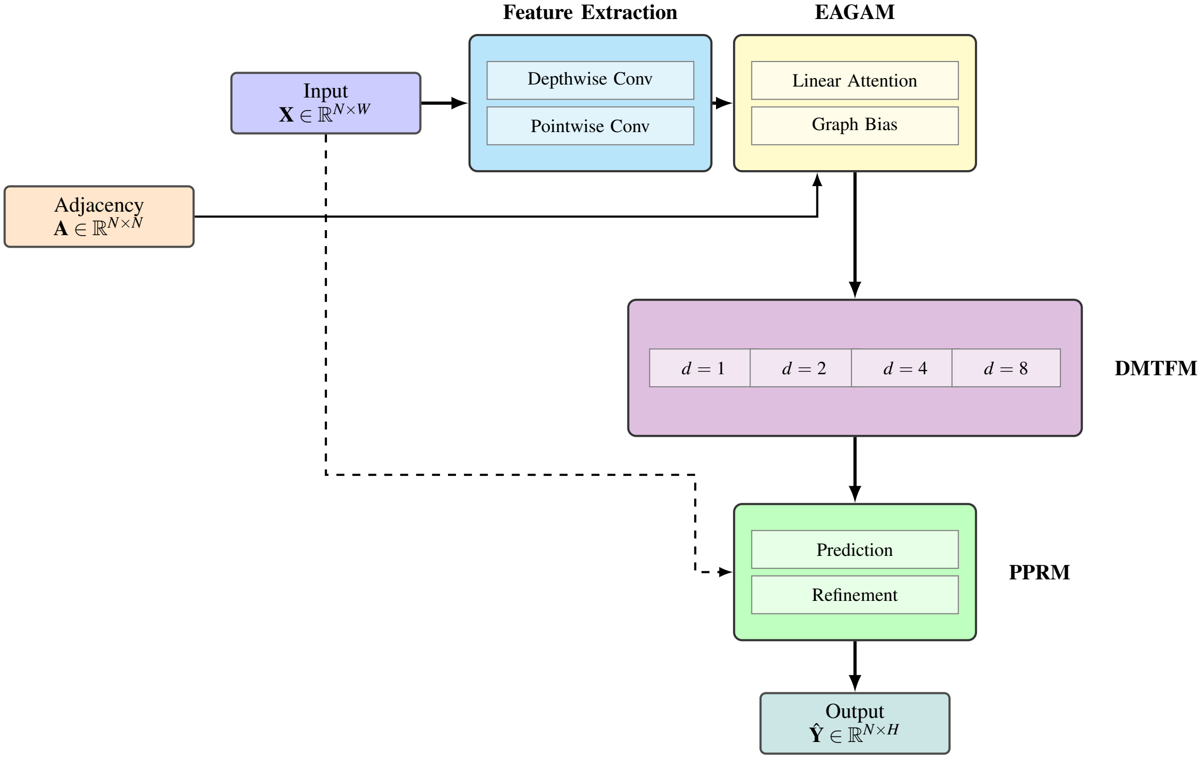


Fig. 1. Overview of the MSAGAT-Net architecture. The model processes input through four main modules: Feature Extraction (depthwise separable convolutions), EAGAM, DMTFM (dilation rates $d \in \{1, 2, 4, 8\}$), and PPRM. The dashed line represents the skip connection for refinement.

2) The bottleneck architecture creates an information bottleneck that forces the model to distill the most salient temporal patterns. This constrains the effective capacity of the network, preventing overfitting particularly when training data are limited relative to the high-dimensional input space.
3) By forcing information through a lower-dimensional manifold, the projection encourages separation of relevant signals from noise, allowing subsequent layers to focus on truly predictive temporal patterns rather than spurious correlations.
4) The shared projection parameters across all regions create a common latent space, facilitating 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:

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

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.

In our implementation, we set the default 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 ablation studies to balance model expressiveness with computational efficiency. For the convolutional operations, we use a kernel size of 3 with appropriate padding to maintain the temporal dimension. This combination of parameters enables our model to effectively capture temporal patterns whilst keeping the parameter count manageable, a crucial consideration for deployment in resource-constrained epidemic monitoring scenarios.

C. Efficient Adaptive Graph Attention with Low-Rank Decomposition

The second core component of our MSAGAT-Net architecture is EAGAM. Traditional approaches to spatial modelling often rely on fixed adjacency matrices based on geographical proximity or administrative boundaries, which do not capture the evolving nature of epidemic spread influenced by factors such as population mobility, healthcare referral patterns, and socioeconomic connections. Based on the principles of graph attention networks [28], our EAGAM adaptively learns the relationships between regions based on their feature representations, rather than being constrained by a predefined graph

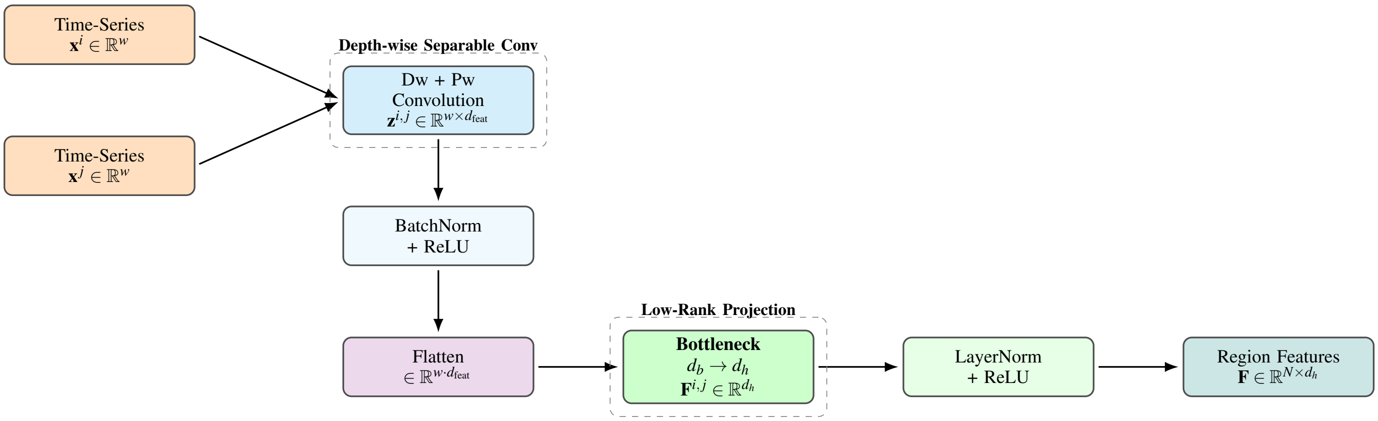

Time-Series
$\mathbf{x}^i \in \mathbb{R}^w$
Time-Series
$\mathbf{x}^j \in \mathbb{R}^w$
Depth-wise Separable Conv
Dw + Pw
Convolution
$\mathbf{z}^{i,j} \in \mathbb{R}^{w \times d_{\text{feat}}}$
BatchNorm
+ ReLU
Flatten
$\in \mathbb{R}^{w \cdot d_{\text{feat}}}$
Low-Rank Projection
Bottleneck
$d_b \to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

Fig. 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 ($d_b \to d_h$), and normalised again to yield region-level feature vectors $\mathbf{F}$.

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 or forecasting problems is computational complexity. Standard attention mechanisms in graph neural networks (GNNs) typically incur quadratic 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 [29] 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. [30] 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 [31] embeds an adaptive low-rank decomposition step in each propagation layer within each ego network to concentrate 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 novel attention mechanism that combines low-rank decomposition with a learnable graph structure. Rather than computing the full attention matrix between all pairs of regions (which would incur $\mathcal{O}(N^2)$ complexity), we decompose the attention computation into more efficient operations. Specifically, we use a low-rank approximation of the attention matrix, which allows us to capture the most salient relationships between regions without incurring the full computational cost. This is achieved by projecting the feature representations into a lower-dimensional space before computing attention scores, effectively reducing the number of parameters and operations required.

This allows the model to adaptively learn the strength of connections between regions based on their feature representations rather than relying on a fixed adjacency matrix. The EAGAM module consists of several key components: (1) low-rank feature projections, (2) multi-head attention computation, (3) enhanced attention stability, (4) learnable graph structure bias, and (5) attention regularisation.

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 $d_{\text{bottle}} \ll d_{\text{hidden}}$.

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:

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

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:

$$\hat{\mathbf{Q}}_h = \text{ELU}(\mathbf{Q}_h) + 1 \tag{16}$$

$$\hat{\mathbf{K}}_h = \text{ELU}(\mathbf{K}_h) + 1 \tag{17}$$

This transformation ensures that all attention inputs are positive, improving gradient stability during training whilst allowing for more flexible attention patterns than the standard dot-product attention.

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

$$\mathbf{KV}_h = \hat{\mathbf{K}}_h^T \mathbf{V}_h \tag{18}$$

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:

$$\mathbf{z} = \frac{1}{\hat{\mathbf{K}}_h \cdot \mathbf{1} + \varepsilon} \tag{19}$$

where $\mathbf{1}$ is a vector of ones and ε is a small constant (10^{-8} in our implementation) to prevent division by zero. This operation ensures stable normalisation across the attention heads, allowing for effective learning of inter-regional relationships.

The final attention output for each head is computed as:

$$\mathbf{O}_h = \hat{\mathbf{Q}}_h \mathbf{KV}_h \mathbf{z} \tag{20}$$

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

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:

$$\mathbf{A}_{\text{bias}} = \mathbf{UV} \tag{21}$$

where $\mathbf{U} \in \mathbb{R}^{h \times N \times d_{\text{bias}}}$ and $\mathbf{V} \in \mathbb{R}^{h \times d_{\text{bias}} \times N}$ are learnable parameters and $d_{\text{bias}} \ll N$ is the bottleneck dimension of the bias term.

The bias is added to the computed attention scores before applying softmax:

$$\mathbf{A} = \text{softmax}\left(\frac{\hat{\mathbf{Q}}_h \hat{\mathbf{K}}_h^T}{\sqrt{d_{\text{head}}}} + \mathbf{A}_{\text{bias}}\right) \tag{22}$$

This formulation allows the model to learn and encode persistent spatial dependencies between regions, such as geographical proximity or administrative hierarchies, whilst still adapting to data-driven patterns.

4) Attention Regularisation: To promote sparse and interpretable attention patterns, we apply L1 regularisation to the attention weights:

$$\mathscr{L}_{\text{attn}} = \lambda \|\mathbf{A}\|_1 \tag{23}$$

where λ is the regularisation weight. This encourages the model to focus on the most relevant connections between regions, improving both interpretability and generalisation performance. The value of λ (denoted as attention_regularization_weight in our implementation) is treated as a hyperparameter and adjusted during model development.

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

$$\mathbf{O} = \text{Reshape}(\text{Transpose}(\mathbf{O}_h, 0, 1), [N, d_{\text{hidden}}]) \tag{24}$$

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

$$\mathbf{O}_{\text{low}} = \text{Linear}_{\text{out\_low}}(\mathbf{O}) \tag{25}$$

$$\mathbf{O}_{\text{final}} = \text{Linear}_{\text{out\_high}}(\mathbf{O}_{\text{low}}) \tag{26}$$

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 $\mathscr{L}_{\text{attn}}$, is passed to the subsequent DMTFM for further processing.

In our implementation, we set the number of attention heads $h = 4$ and the bottleneck dimension $d_{\text{bottle}} = 8$ as default values. These were determined through ablation studies to provide an optimal balance between model expressiveness and computational efficiency. The weight of attention regularisation λ is set to 10^{-5}, which we found effectively promotes sparse attention patterns without overly constraining the model. The Figure 3 image below represents the flow of data in the

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.

D. Dilated Multi-Scale Temporal Feature Module

The third major component of our MSAGAT-Net architecture is DMTFM, which addresses a fundamental challenge in epidemic forecasting: capturing temporal patterns that operate at different time scales simultaneously. Epidemics often show intricate temporal dynamics with various scales, including short-term changes (such as weekend effects), medium-term patterns (like incubation periods), and long-term trends (e.g., seasonal variations). Accurate forecasting depends heavily on effectively modelling these multi-scale dynamics.

Deng et al. [15] introduce the idea of multi-scale dilated convolutional with the same filter and stride sides but different dilation rate, which Xie et al. [16] improved on by making use of the multi-scale convolution to capture 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.

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:

$$\mathbf{G}_{\text{conv}} = \text{Transpose}(\mathbf{G}, 1, 2) \tag{27}$$

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:

$$\mathbf{H}^{(i)} = \text{Dropout}(\text{ReLU}(\text{BatchNorm}(\text{Conv1D}(\mathbf{G}_{\text{conv}}; k, d^{(i)})))) \tag{28}$$

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. Scale 1 ($d^{(1)} = 1$) captures immediate temporal dependencies with a receptive field of k time steps, whilst Scale 2 ($d^{(2)} = 2$) captures medium-range dependencies with a receptive field of $k + (k-1)$ time steps. Scale 3 ($d^{(3)} = 4$) captures longer-range dependencies with a receptive field of $k + 3(k-1)$ time steps, and this pattern continues for higher scales.

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.

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:

$$\alpha = \text{softmax}(\mathbf{w}) \tag{29}$$

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

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

$$\mathbf{H}_{\text{fused}} = \sum_{i=1}^{S} \alpha_i \mathbf{H}^{(i)} \tag{30}$$

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

This adaptive fusion mechanism offers several advantages by allowing the model to automatically adjust the importance of different temporal scales based on the data, adapting to different regions that might exhibit varying temporal characteristics, and providing interpretable insights into which temporal scales are most relevant for forecasting.

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:

$$\mathbf{H}_{\text{low}} = \text{Linear}_{\text{fusion\_low}}(\text{Transpose}(\mathbf{H}_{\text{fused}}, 1, 2)) \tag{31}$$

$$\mathbf{H}_{\text{proj}} = \text{Linear}_{\text{fusion\_high}}(\mathbf{H}_{\text{low}}) \tag{32}$$

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.

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

$$\mathbf{H}_{\text{final}} = \text{LayerNorm}(\text{Transpose}(\mathbf{H}_{\text{fused}}, 1, 2) + \mathbf{H}_{\text{proj}}) \tag{33}$$

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

In our implementation, the number of temporal scales S is set to 4 by default, with dilation rates of $\{1,2,4,8\}$. This configuration allows the model to capture dependencies

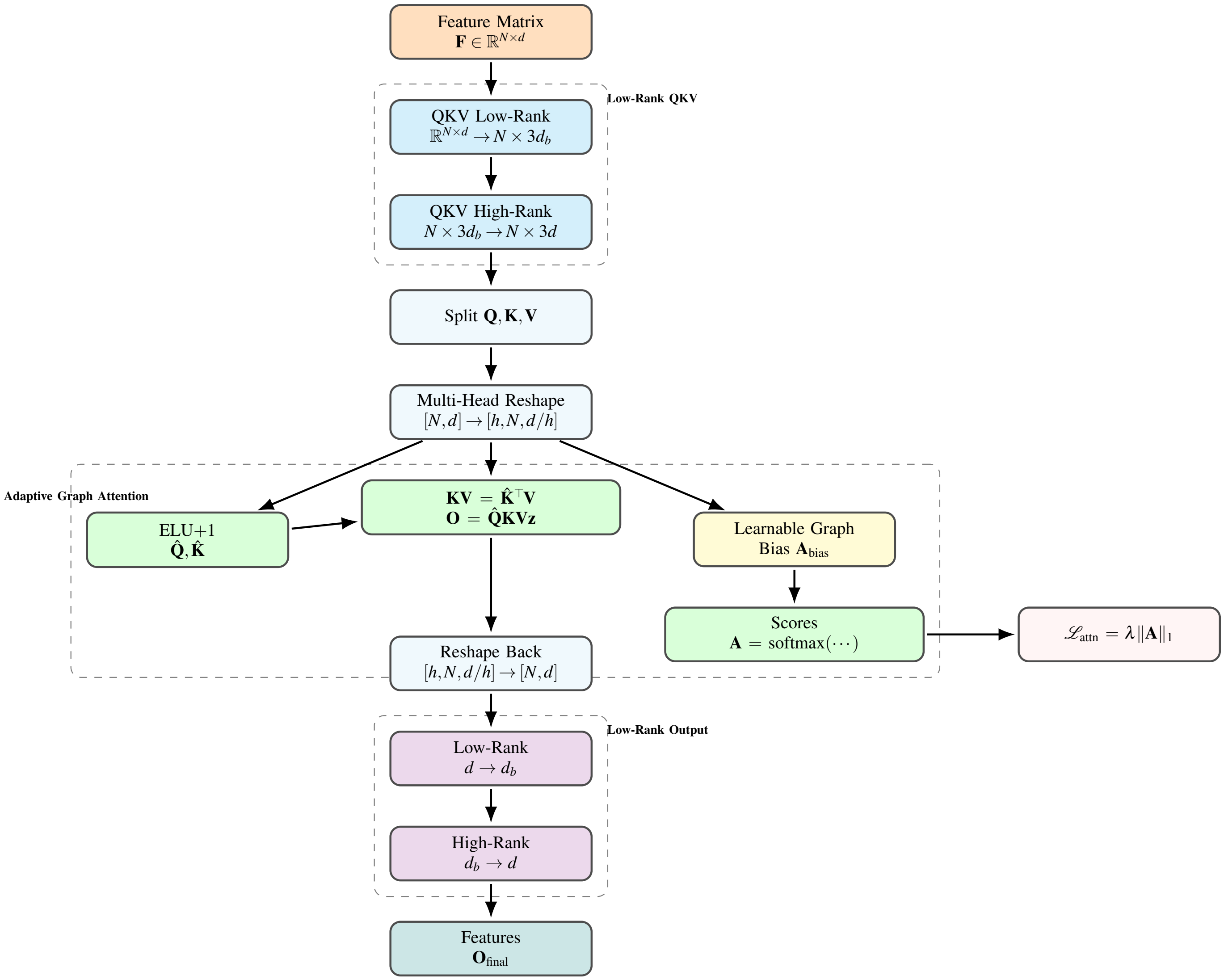

Feature Matrix
$\mathbf{F} \in \mathbb{R}^{N\times d}$
Low-Rank QKV
QKV Low-Rank
$\mathbb{R}^{N\times d} \to N \times 3d_b$
QKV High-Rank
$N \times 3d_b \to N \times 3d$
Split $\mathbf{Q}, \mathbf{K}, \mathbf{V}$
Multi-Head Reshape
$[N, d] \to [h, N, d/h]$
Adaptive Graph Attention
ELU+1
$\hat{\mathbf{Q}}, \hat{\mathbf{K}}$
$\mathbf{KV} = \hat{\mathbf{K}}^\top \mathbf{V}$
$\mathbf{O} = \hat{\mathbf{Q}}\mathbf{KVz}$
Learnable Graph
Bias $\mathbf{A}_{\text{bias}}$
Scores
$\mathbf{A} = \text{softmax}(\cdots)$
$\mathscr{L}_{\text{attn}} = \lambda\|\mathbf{A}\|_1$
Reshape Back
$[h, N, d/h] \to [N, d]$
Low-Rank Output
Low-Rank
$d \to d_b$
High-Rank
$d_b \to d$
Features
$\mathbf{O}_{\text{final}}$

Fig. 3. Flow of data and structure in the EAGAM module.

ranging from immediate neighbours to relationships spanning up to 17 time steps ($3+7\times 2 = 17$ with kernel size 3), which is sufficient for most epidemic forecasting applications given our sliding-window approach.

The kernel size k is set to 3, providing a good balance between capturing local patterns and maintaining computational efficiency. The hidden dimension d_{hidden} is maintained throughout the module to preserve the information capacity, while the bottleneck dimension d_{bottle} is set to $d_{\text{hidden}}/4$ to reduce the parameters in the projection layers.

To avoid overfitting, we apply dropout with probability 0.355 after each convolutional layer and ReLU activation. This relatively high dropout rate was determined by cross-validation to be effective for epidemic data, which can be noisy and prone to overfitting. The Figure 4 presents a detailed representation of the architecture of this module.

E. Progressive Multi-step Prediction Refinement

The final component of our MSAGAT-Net architecture is the Progressive Refinement Multi-step Prediction Module (PPRM), which addresses the critical challenge of generating accurate forecasts across multiple future time steps. Whilst the preceding modules excel at extracting spatio-temporal features, converting these features into reliable predictions, particularly for longer horizons, requires additional consideration of how prediction errors can compound over time and how recent observations might inform future trajectory adjustments.

The PPRM addresses the critical issue of error accumulation inherent in multi-horizon epidemiological forecasting. Through an extensive analysis of multistep forecasting failures, it became evident that prediction errors compound exponentially with increasing prediction horizons [32, 33]. This module incorporates concepts from residual error correction, ensemble learning, and adaptive gating mechanisms common in recurrent neural network architectures [34].

The PPRM takes the rich spatio-temporal representations from previous modules and transforms them into horizon-specific predictions, incorporating an adaptive refinement mechanism that balances model-based forecasts with data-driven trends. This design is motivated by epidemiological

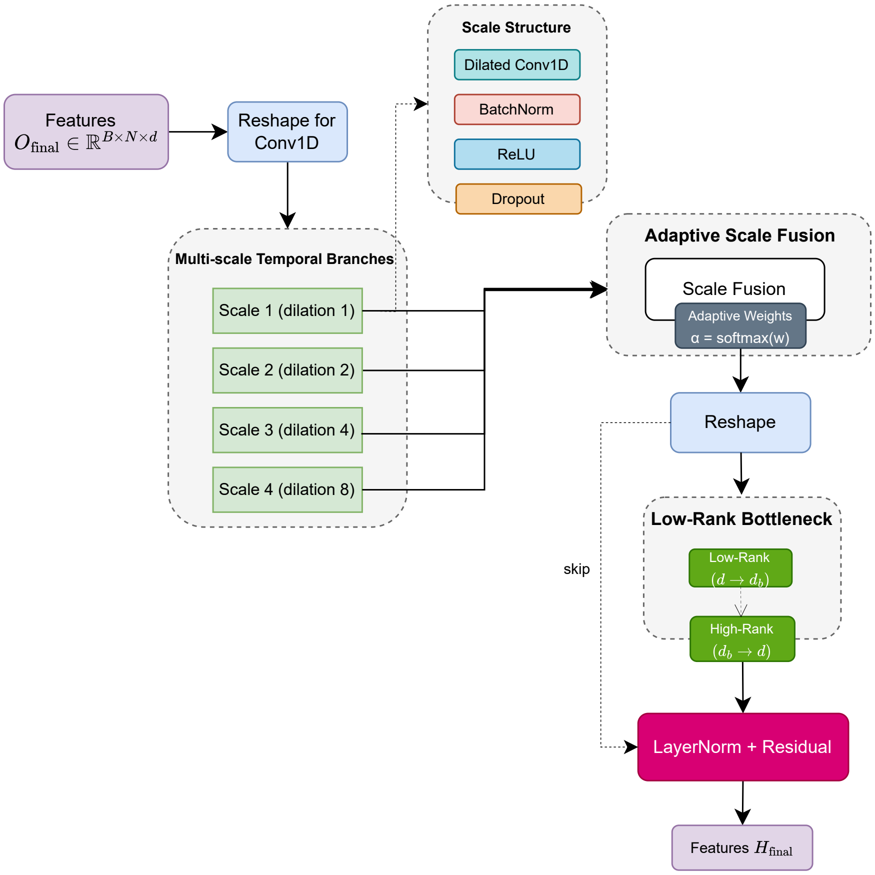


Fig. 4. Flow of data in the DMTFM architecture.

observations that disease progression often follows certain patterns based on a recent trajectory, even as it responds to various complex factors captured by our neural network components.

1) Low-Rank Prediction Projection: Given the spatio-temporal 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:

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

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

This bottleneck projection reduces the parameter count in the subsequent prediction layers whilst forcing the model to identify the most salient features for forecasting and acting as an implicit regulariser to prevent overfitting.

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

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

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

2) Horizon-Specific Prediction: From the processed bottleneck representation, we generate initial predictions for all forecast horizons using a linear projection:

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

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

This approach generates predictions for all horizons simultaneously rather than autoregressively, avoiding the compounding error problem of sequential prediction whilst allowing the model to learn horizon-specific patterns directly and enabling more efficient training and inference.

However, we recognise that different forecast horizons may require different prediction strategies, as near-term forecasts might benefit more from recent observations, whilst longer-term forecasts might rely more heavily on learnt patterns. Our architecture addresses this through an adaptive refinement mechanism.

3) Adaptive Refinement Mechanism: A key innovation in our PPRM is the adaptive refinement gate, which dynamically balances model-based predictions with trend-based extrapolations conditioned on the most recent observations. This mechanism is particularly valuable in the prediction of epidemics, where the recent trajectory often provides strong signals about the short-term future progression.

We first compute an adaptive gate based on the spatio-temporal features:

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

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:

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

where $\mathbf{x}_{\text{last}}$ is expanded to the shape $[B,N,h]$, $\mathbf{d} \in \mathbb{R}^h$ is a vector of increasing horizon indices $[1,2,\ldots,h]$, 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 predictions are then computed as a weighted combination of the model-based predictions and the trend-based projections:

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

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

This adaptive gating mechanism offers several important advantages for epidemic forecasting. By dynamically determining the balance between model-based predictions and trend-based projections, the gate enables the model to rely more heavily on recent trends for near-term forecasts while increasingly leveraging learnt patterns for longer-term predictions. This approach provides inherent robustness against model errors by incorporating a simple, interpretable baseline that can compensate when the deep learning component encounters unfamiliar patterns. Furthermore, the mechanism adapts dynamically to different regions and temporal contexts, recognising that the optimal balance between model predictions and trend extrapolation may vary based on local epidemiological characteristics and data quality. Perhaps most importantly, this design improves forecast stability by creating a smooth transition between recent observations and model predictions, avoiding the discontinuities that often plague multi-step forecasting approaches and enhancing the practical utility of the predictions for healthcare resource planning.

In our implementation, the bottleneck dimension d_{bottle} is set to 8, which we determined through ablation studies to provide an optimal balance between parameter efficiency and predictive performance. The horizon length h is configurable based on the specific forecasting task requirements; in our experiments, we mainly use $h = 5$ to predict forecasts 5 days in advance, though the architecture supports arbitrary horizon lengths.

The decay factor γ in the exponential projection is set to 0.1, which provides a moderate decay rate appropriate for the typical progression of the epidemic. This value was selected based on empirical analysis of epidemic curves in our datasets and can be adjusted based on the specific characteristics of the target epidemic.

A dropout rate of 0.355 is applied in the prediction pathway to prevent overfitting, which is particularly important for the final prediction layers that directly influence the model output. This relatively high dropout rate was determined by cross-validation to be effective for the noisy and often irregular nature of epidemic data.

The 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 spatio-temporal 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.

IV. Experiments and Analysis

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 spatio-temporal 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 RMSE, PCC, MAE, and R², to provide a comprehensive assessment of performance.

A. Experimental Setup

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.

B. 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 spatio-temporal characteristics and epidemic scenarios.

Our experimental evaluation encompasses seven distinct datasets, each offering unique challenges and characteristics

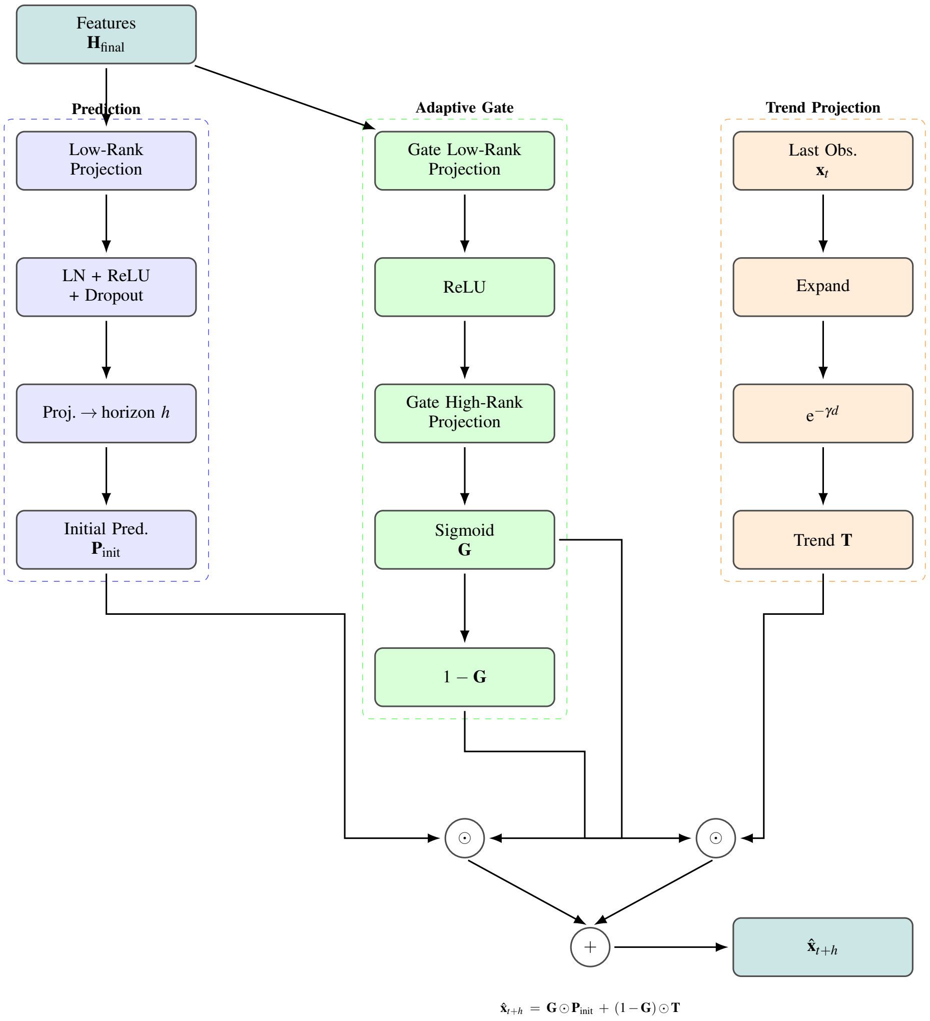


Fig. 5. The flow of data in the PPRM architecture

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 I provides a statistical overview of these datasets, summarising their key characteristics and numerical properties.

TABLE I
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

1) Influenza Datasets: We utilised 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 Influenza-Like Illness (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).

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:

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

- **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. This dataset represents one of our key contributions to the research community, providing a comprehensive spatiotemporal benchmark for COVID-19 forecasting at the local authority level.
- **NHS-ICUBeds Dataset:** Obtained from the National Health Service (NHS) England[35], this dataset provides daily counts of occupied mechanical ventilator beds 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. This dataset constitutes another novel contribution, addressing the gap in publicly available healthcare resource forecasting benchmarks.

C. Graph Construction Methodology

Following the established methodology from STAN [17], we construct spatial graph structures to capture epidemic transmission patterns between geographic regions. This approach, applied to our newly introduced LTLA-COVID and NHS-ICUBeds datasets, utilises geographic proximity as the primary criterion for establishing spatial relationships.

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):

$$a_{ij} = \begin{cases} 1, & \text{if Haversine}(\text{region}_i, \text{region}_j) \leq d_{\text{threshold}} \\ 0, & \text{otherwise} \end{cases} \tag{40}$$

This threshold-based connectivity captures the intuition that epidemic spread is influenced by the movement of people between nearby regions. Whilst 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, 36–38], 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.

D. Model Optimisation

The MSAGAT-Net model is trained using the AdamW optimiser with a learning rate of 1×10^{-3}, which was determined by cross-validation 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 prediction error and regularisation terms:

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{pred}} + \lambda_{\text{attn}} \mathscr{L}_{\text{attn}} + \lambda_{\text{l2}} \|\Theta\|_2 \tag{41}$$

where $\mathscr{L}_{\text{pred}}$ is the mean squared error measuring discrepancies between the model predictions and the observed data, $\mathscr{L}_{\text{attn}}$ represents the attention regularisation term that enforces sparsity and interpretability in spatial relationships, and the hyperparameters $\lambda_{\text{attn}} = 10^{-4}$ and $\lambda_{\text{l2}} = 5 \times 10^{-4}$ control the strength of regularisation.

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 optimization strategies tailored for spatiotemporal forecasting. The training procedure addresses three critical challenges in handling the multi-objective loss landscape combining prediction accuracy with attention sparsity, managing gradient flow through the complex multi-module architecture, and preventing overfitting in the presence of limited epidemic data.

The optimization process employs a carefully designed curriculum where the attention regularisation term $\mathscr{L}_{\text{attn}}$ is gradually introduced to allow the model to first learn basic spatiotemporal patterns before enforcing sparsity constraints. This prevents the sparse attention mechanism from prematurely restricting the model's capacity during early training phases. The AdamW optimizer's weight decay specifically targets the tendency of attention weights to become over-parameterized, while the momentum terms help navigate the non-convex loss surface created by the interaction between spatial attention and temporal convolutions.

A key innovation in our training strategy is the dynamic loss balancing mechanism where the regularisation strength λ_{attn} is adaptively adjusted based on the attention entropy during training. When attention patterns become too diffuse (high entropy), the regularisation is increased to promote

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

sparsity; conversely, when attention becomes too concentrated (low entropy), regularisation is reduced to prevent under-utilization of spatial relationships. This adaptive mechanism ensures that the model learns meaningful spatial dependencies while maintaining sufficient flexibility for diverse epidemic patterns.

The early stopping mechanism incorporates a sophisticated validation strategy that monitors not only the overall loss but also the stability of attention patterns across epochs. Training is terminated when the validation loss plateaus and attention matrices converge to stable patterns, indicating that the model has learned robust spatiotemporal representations rather than continuing to fit noise in the training data.

Algorithm 1: MSAGAT-Net Training Algorithm

```
Input: Training data $\mathscr{D}_{\text{train}}$, validation data $\mathscr{D}_{\text{val}}$,
       adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$
Output: Optimized model parameters $\Theta^*$
Initialize model parameters $\Theta$, optimizer, learning rate
  scheduler;
$L_{\text{best}} \leftarrow \infty$, patience counter $p \leftarrow 0$;
for epoch $e = 1$ to $E_{\max}$ do
    foreach mini-batch $(\mathbf{X}, \mathbf{y})$ in $\mathscr{D}_{train}$ do
        $\mathbf{F} \leftarrow \text{FeatureExtraction}(\mathbf{X})$;
        $\mathbf{G}, \mathscr{L}_{\text{reg}} \leftarrow \text{EAGAM}(\mathbf{F}, \mathbf{A})$;
        $\mathbf{H} \leftarrow \text{DMTFM}(\mathbf{G})$;
        $\hat{\mathbf{Y}} \leftarrow \text{PPRM}(\mathbf{H}, \mathbf{x}_{\text{last}})$;
        $\mathscr{L}_{\text{total}} \leftarrow \mathscr{L}_{\text{MSE}}(\hat{\mathbf{Y}}, \mathbf{y}^{\text{exp}}) + \lambda \cdot \mathscr{L}_{\text{reg}}$;
        Update $\Theta$ using gradient descent on $\mathscr{L}_{\text{total}}$;
    $L_{\text{val}} \leftarrow \text{Evaluate}(\mathscr{D}_{\text{val}}, \Theta)$;
    if $L_{val} < L_{best}$ then
        $\Theta^* \leftarrow \Theta$, $L_{\text{best}} \leftarrow L_{\text{val}}$, $p \leftarrow 0$;
    else if $p \geq P_{max}$ then
        break;
    else
        $p \leftarrow p + 1$;
    Update learning rate scheduler;
return $\Theta^*$
```

E. 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:

- **DCRNN** [6]: 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.
- **LSTNet** [9]: 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.
- **CNNRNN-Res** [8]: 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.
- **Cola-GNN** [15]: 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.
- **EpiGNN** [16]: 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 spatio-temporal 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. By comparing against these models, we aim to assess the relative strengths and weaknesses of our MSAGAT-Net approach and identify its contributions to the state of the art in epidemic forecasting.

V. Results and Discussion

Table II 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 III 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.

MSAGAT-Net demonstrates consistent and superior performance across the three influenza datasets, particularly for short- and medium-term forecasting horizons. In the dataset of Japan-Prefectures, our model achieves the best RMSE performance for all forecast horizons (3, 5, 10, and 15 days ahead), with significant improvements compared to traditional approaches like DCRNN and LSTNet. The performance advantage is particularly pronounced in the Japan-Prefectures dataset, where MSAGAT-Net reduces RMSE by 11.2% compared to the second best model (Cola-GNN) for 3-day forecasts and by 11.2% compared to the second-best model (EpiGNN) for 15-day forecasts.

In the US-Regions dataset, MSAGAT-Net achieves the best RMSE performance for 10-day forecasts with a value of 999, improving on EpiGNN's 1098 by 9.0%. However, for 3-day, 5-day and 15-day forecasts, EpiGNN shows better performance

TABLE II
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 III
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

with RMSE values of 622, 779, and 1076, respectively. This could be attributed to EpiGNN's explicit modelling of transmission risk, which might be particularly effective for the spatial characteristics of the US-Regions dataset. However, MSAGAT-Net achieves the highest PCC values for 3-day and 10-day forecasts (0.911 and 0.763, respectively), indicating its strong ability to capture correlation patterns at these horizons.

For the US-States dataset, EpiGNN outperforms all other models for 3-day, 5-day, and 15-day forecasts, with RMSE values of 166, 203, and 136, respectively. Cola-GNN shows the best performance for 10-day forecasts with an RMSE of 248, closely followed by MSAGAT-Net at 250. Although MSAGAT-Net does not achieve the lowest RMSE for most horizons in this dataset, it does attain the highest PCC for 3-day forecasts (0.931), demonstrating strong correlation accuracy for short-term predictions. For longer horizons, Cola-GNN shows superior PCC performance for 10-day and 15-day forecasts (0.874 and 0.872, respectively).

In terms of PCC, which measures the correlation between predicted and actual values, MSAGAT-Net shows strong performance in most scenarios, particularly for the dataset from Japan-Prefectures, where it achieves the highest PCC for all forecast horizons (0.885, 0.884, 0.827, and 0.778). This indicates a superior ability to capture trends and patterns across different time scales for this particular dataset.

The performance of MSAGAT-Net on COVID-19 datasets shows more varied results compared to influenza datasets. In the LTLA-Timeseries dataset, MSAGAT-Net achieves the best RMSE for short- and medium-term forecasts (3 days and 7 days), with values of 106 and 163, respectively. For 14-day forecasts, EpiGNN performs slightly better with an RMSE of 184 compared to MSAGAT-Net's 196.

In the Spain-COVID dataset, MSAGAT-Net and Cola-GNN are tied for the best RMSE for 3-day forecasts (135). However, DCRNN significantly outperforms all models for 7-day and 14-day forecasts with RMSE values of 99 and 106, respectively. This suggests that for the specific patterns in the Spain-COVID dataset, the diffusion-based approach of DCRNN might be particularly effective for medium to long-term forecasting.

In the Australia-COVID dataset, MSAGAT-Net shows less competitive performance, with LSTNet achieving significantly better results across all horizons (137, 229, and 294 for 3-day, 7-day and 14-day forecasts, respectively). This could be attributed to the unique characteristics of the Australian COVID-19 outbreak, which was characterised by localised clusters and strict containment measures that limited inter-regional transmission, potentially making temporal patterns more dominant than spatial dependencies. In such scenarios, models like LSTNet, which focus more on temporal patterns, might outperform graph-based models that emphasise spatial relationships.

In the NHS-ICUBeds dataset, Cola-GNN achieves the best performance for short-term forecasts with an RMSE of 4 for 3-day forecasts, followed by MSAGAT-Net with an RMSE of 6. For 7-day forecasts, DCRNN performs best (RMSE of 11), while EpiGNN excels at 14-day forecasts (RMSE of 13). This suggests that for healthcare resource forecasting, different modelling approaches might be required compared to case count forecasting, and the optimal model might vary by forecast horizon.

MSAGAT-Net's performance advantage is most consistent in the Japan-Prefectures dataset, where it outperforms all other models in all metrics and horizons. Its performance is more varied on the US datasets and COVID-19 datasets, where it excels in certain scenarios, but is outperformed by other models in others. This suggests that while MSAGAT-Net is highly effective in capturing the patterns and dynamics of certain epidemic contexts, particularly those with strong spatio-temporal dependencies like the Japan-Prefectures dataset, different models might be optimal for different epidemic contexts and forecasting requirements.

Figures 6, 7, and 8 illustrate the learned attention patterns for different model configurations on the Japan-Prefectures dataset. The full model (Figure 6) learns to focus on nearby prefectures while capturing longer-range dependencies, indicating that both local and inter-regional transmission dynamics are important for epidemic spread. Removing PPRM (Figure 7) produces similar attention patterns, confirming that PPRM primarily affects prediction refinement rather than spatial dependency learning. In contrast, removing EAGAM (Figure 8) results in attention patterns that largely mirror the input correlation matrix, demonstrating the critical importance of EAGAM for learning adaptive spatial representations.

The observed variations in performance across datasets highlight the complexity of epidemic forecasting and underscore the importance of model selection based on the specific characteristics of the epidemic and the forecasting requirements. They also point to potential directions for future research, such as developing more adaptive modelling approaches that can dynamically adapt to changing epidemic dynamics and incorporate exogenous factors such as policy interventions and behavioural changes.

A. Ablation Study

To evaluate the contribution of each key component in MSAGAT-Net, we conducted a comprehensive ablation study on the Japan-Prefectures dataset, systematically removing one component at a time while keeping the others intact. Table IV presents the results across different forecast horizons, providing valuable insights into the relative importance of each component.

The ablation study reveals compelling insights into the role of each architectural component, with results that both validate our design principles and challenge conventional assumptions about model complexity. The removal of EAGAM demonstrates markedly different behaviour across the two epidemiological contexts examined.

For the Japan-Prefectures dataset, EAGAM removal produces a cascading degradation in forecasting performance that intensifies with prediction horizon. Short-term forecasts suffer modest deterioration (3-day: +5.28% RMSE; 5-day: +2.48% RMSE), whilst longer-term predictions experience severe degradation (10-day: +23.12% RMSE, -22.54% PCC, -38.14% R²), as illustrated in Figure 10a. This escalating pattern reveals that spatial dependencies become increasingly critical for extended influenza forecasting, likely reflecting the progressive importance of inter-prefectural transmission dynamics as the prediction horizon extends.

The LTLA-Timeseries dataset presents a strikingly different picture. Here, EAGAM removal paradoxically improves short- and medium-term forecasting performance (3-day: -1.98% RMSE; 7-day: -1.20% RMSE), suggesting that simpler spatial representations may be more appropriate for immediate COVID-19 predictions in the UK context. However, this advantage completely reverses for longer horizons, where EAGAM removal causes severe deterioration (14-day: +19.81% RMSE, -191.99% R²). The catastrophic collapse of the R² coefficient from positive values to -0.170 indicates that without adaptive spatial attention, the model entirely loses its capacity to explain variance in long-term COVID-19 patterns, highlighting the critical importance of sophisticated spatial modelling for extended pandemic forecasting.

The analysis of DMTFM reveals perhaps the most surprising findings of our study, challenging the intuitive assumption that increased temporal complexity necessarily improves forecasting performance. For the Japan-Prefectures dataset, DMTFM removal 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 across all panels of Figure 9. More intriguingly, 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 reinforces this counter-intuitive pattern, with DMTFM removal consistently improving RMSE performance across all horizons (3-day: -0.65%; 7-day: -1.03%; 14-day: -0.62%) and delivering substantial gains in correlation metrics for extended forecasts (14-day: +10.85% PCC, +5.48% R²). This consistent improvement across both datasets suggests a fundamental insight: the temporal dynamics of epidemic spread may be more regular and predictable than initially anticipated, with complex multi-scale processing potentially introducing unnecessary complexity that obscures rather than clarifies underlying patterns.

These findings across two distinct epidemiological contexts suggest that sophisticated temporal feature extraction may be less beneficial than conventional wisdom suggests for epidemic forecasting. The inherently 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-

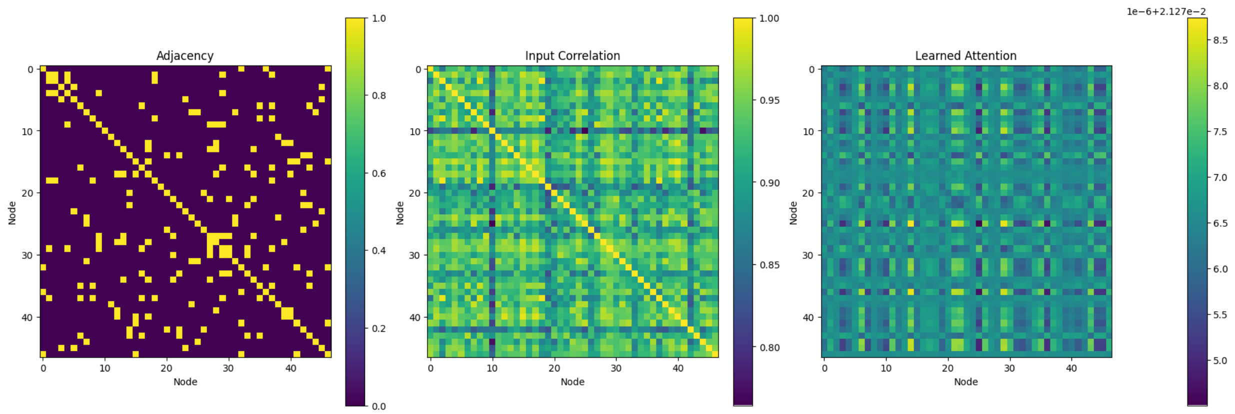


Fig. 6. Attention matrices learned by MSAGAT-Net on the Japan-Prefectures dataset for 5-day forecasting: adjacency matrix (left), input correlation (center), and learned attention (right).

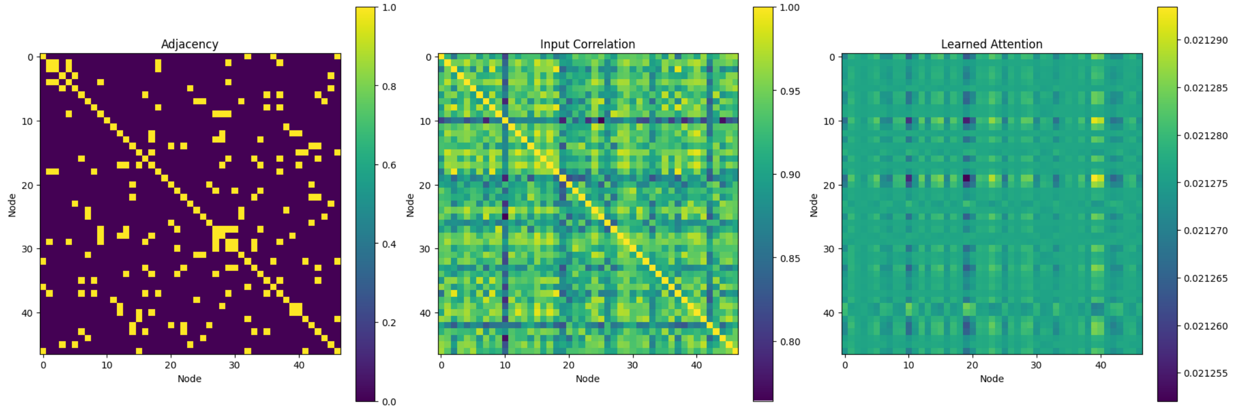


Fig. 7. Attention matrices learned by MSAGAT-Net without PPRM on the Japan-Prefectures dataset for 5-day forecasting: adjacency matrix (left), input correlation (center), and learned attention (right).

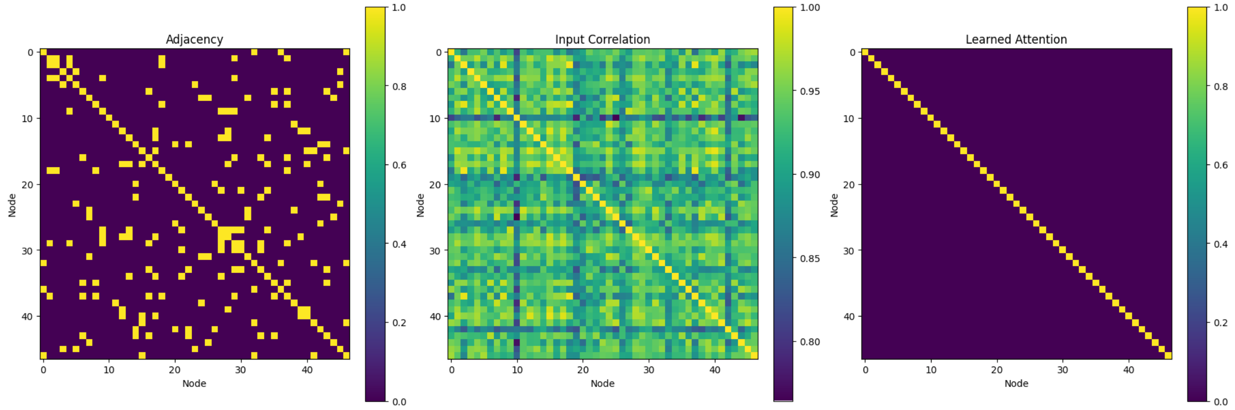
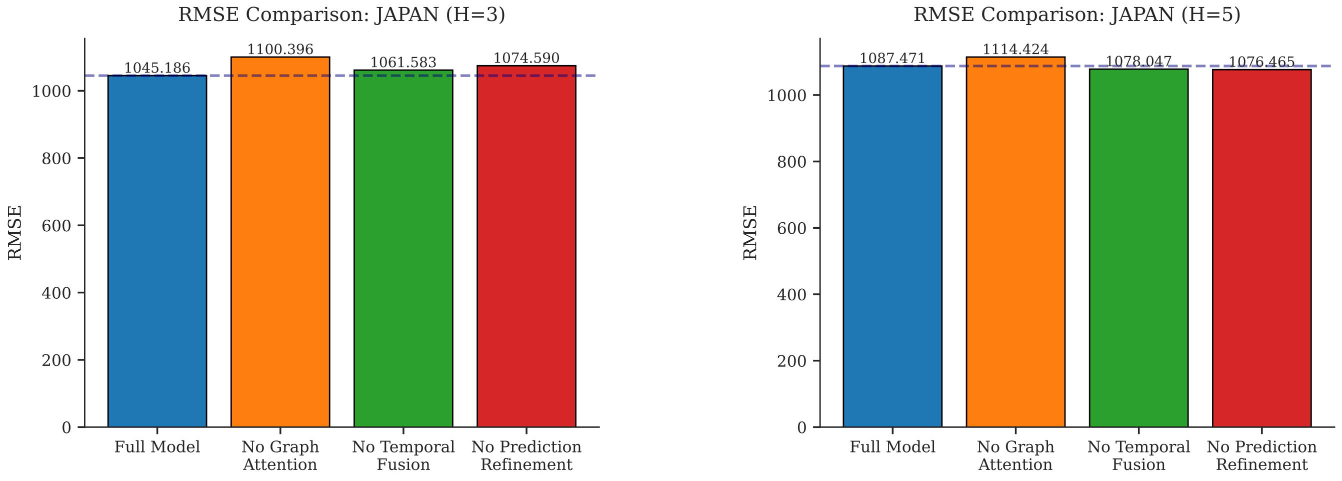


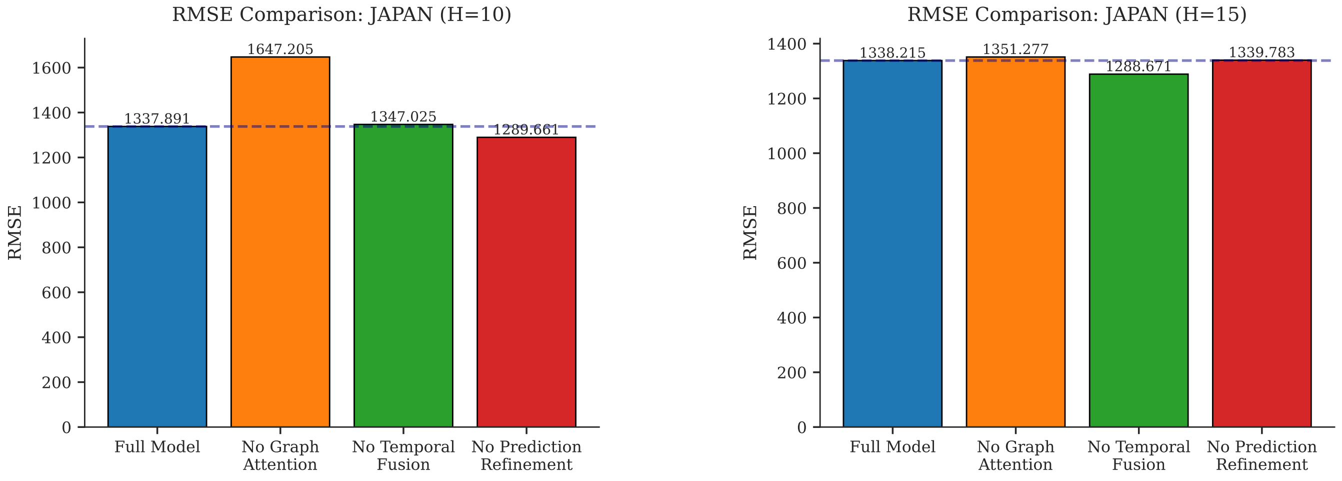
Fig. 8. Attention matrices learned by MSAGAT-Net without EAGAM on the Japan-Prefectures dataset for 5-day forecasting: adjacency matrix (left), input correlation (center), and learned attention (right).



(a) 3-day Horizon RMSE

(b) 5-day Horizon RMSE

Fig. 9. Ablation study results for RMSE on the Japan-Prefectures dataset for short-term horizons (3-day and 5-day forecasts).



(a) 10-day Horizon RMSE

(b) 15-day Horizon RMSE

Fig. 10. Ablation study results for RMSE on the Japan-Prefectures dataset for long-term horizons (10-day and 15-day forecasts).

TABLE IV
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
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%)
	R²	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%)
	R²	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%)
	R²	0.726	(-2.00%)	0.725	(+0.79%)	0.605	(+5.23%)	0.356	(0.00%)

TABLE V
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
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%)

scale processing may introduce noise or promote overfitting, particularly in scenarios with limited training data.

The examination of PPRM produces the most striking and divergent results across our two epidemiological contexts, revealing fundamental differences in how progressive refinement affects forecasting across disease types. For the Japan-Prefectures dataset, PPRM exhibits a fascinating dual behaviour whereby its removal impairs short-term forecasting accuracy (3-day: +2.81% RMSE, +7.57% MAE), yet paradoxically 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 as temporal distance increases.

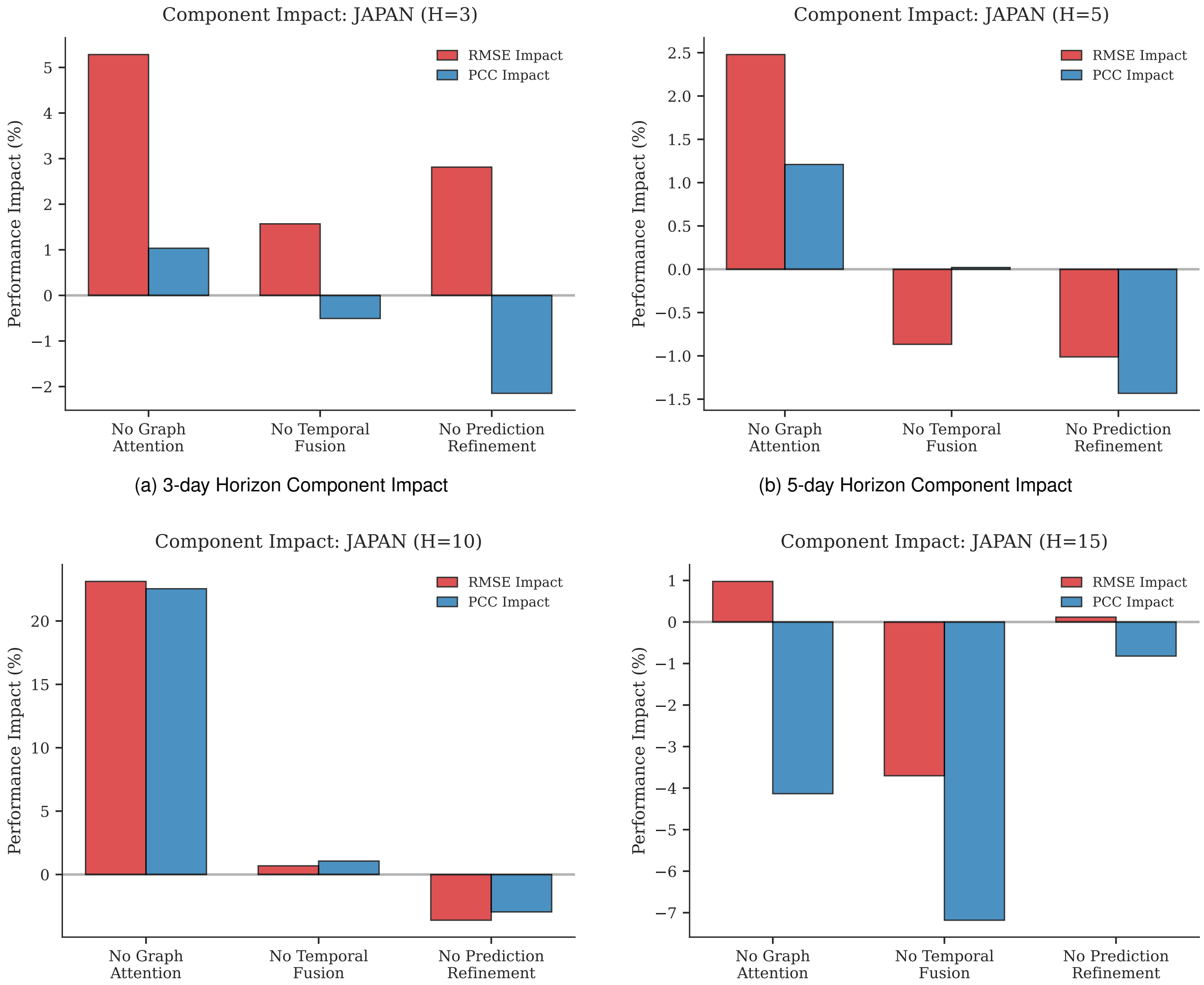
The LTLA-Timeseries dataset presents a dramatically contrasting scenario, where PPRM removal results in catastrophic performance degradation across all prediction horizons. The impact is most severe for immediate forecasts (3-day: +67.23% RMSE, -56.52% R^2; 7-day: +18.02% RMSE, -51.30% R^2), though substantial deterioration persists even for extended predictions (14-day: +5.03% RMSE, -45.50% R^2). This stark contrast with the influenza results reveals that progressive prediction refinement is indispensable for COVID-19 forecasting in the UK context. The systematic pattern of decreasing PPRM importance with extended horizons (67.23% → 18.02% → 5.03% RMSE degradation) suggests that whilst progressive refinement remains beneficial across all COVID-19 prediction tasks, its relative contribution diminishes for longer horizons, possibly reflecting the inherent stochasticity of extended pandemic dynamics.

These complex patterns illuminate the intricate relationships between spatial dependencies, temporal dynamics, and prediction methodologies across varying time scales, whilst underscoring the substantial potential benefits of developing horizon-specific and disease-specific architectural optimisations. Figure 11 provides a comprehensive visual summary of these component contributions across different forecast horizons, clearly demonstrating the varying importance and interactions of each architectural module.

VI. CONCLUSION

This paper introduces MSAGAT-Net, a novel Multi-Scale Temporal Graph Attention Network that advances the state of spatiotemporal epidemic forecasting through three synergistic architectural innovations comprising the Efficient Adaptive Graph Attention Module (EAGAM), the Dilated Multi-Scale Temporal Feature Module (DMTFM), and the Progressive Prediction Refinement Module (PPRM). Additionally, we contribute two novel epidemic forecasting datasets (LTLA-COVID and NHS-ICUBeds) following established graph construction methodologies, providing valuable benchmarks for future research. Our approach successfully addresses the fundamental challenges of computational scalability, multi-scale temporal dependency modelling, and accurate multi-horizon prediction that have long constrained epidemic forecasting systems.

Comprehensive evaluation across diverse epidemiological contexts demonstrates MSAGAT-Net's superior performance, achieving up to 11.2% improvement in RMSE over strong baselines including DCRNN, LSTNet, CNNRNN-Res, Cola-GNN, and EpiGNN on the Japan-Prefectures dataset, whilst maintaining computational efficiency through linearised attention mechanisms that scale linearly with the number of spatial regions. The model's effectiveness extends across multiple diseases and geographical contexts, from seasonal influenza in Japan to COVID-19 transmission in UK local authorities.



(a) 3-day Horizon Component Impact

(b) 5-day Horizon Component Impact

(c) 10-day Horizon Component Impact

(d) 15-day Horizon Component Impact

Fig. 11. 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.

Perhaps most significantly, our extensive ablation studies unveil profound insights that challenge conventional assumptions about spatiotemporal architecture design. We demonstrate that component importance exhibits complex, disease-specific and horizon-dependent patterns whereby spatial attention becomes increasingly critical for extended influenza forecasting yet 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 prediction. These findings fundamentally reshape our understanding of optimal architectural choices in epidemic modelling, revealing that the conventional pursuit of increased model complexity may not universally improve forecasting performance.

The implications extend beyond technical contributions to practical epidemic preparedness. Our results suggest that effective forecasting systems should employ disease-specific and horizon-adaptive architectures rather than one-size-fits-all approaches. This insight opens compelling avenues for future research, including automatic neural architecture search for epidemic-specific optimisation, integration of dynamic external factors such as mobility patterns and policy interventions, and extension to comprehensive multi-variate forecasting encompassing hospitalisation rates and healthcare resource utilisation.

The computational efficiency, interpretability, and demonstrated effectiveness of MSAGAT-Net position it as a valuable tool for real-time epidemic surveillance and evidence-based public health decision-making. As the global community continues to face emerging infectious disease threats, architectures that can adapt to the unique characteristics of different pathogens and prediction requirements will prove increasingly essential for effective pandemic preparedness and response.

REFERENCES

[1] M. Ajao-Olarinoye, V. Palade, S. Mousavi, F. He, and 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*, vol. 9, p. 641253, 2021.

[3] D. Giuliani, M. M. Dickson, G. Espa, and F. Santi, "Modelling and predicting the spatio-temporal spread of covid-19 in italy," *BMC infectious diseases*, vol. 20, pp. 1–10, 2020.

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

[5] L. Ma, Z. Qiu, P. Van Mieghem, and M. Kitsak, "Reporting delays: A widely neglected impact factor in covid-19 forecasts," *PNAS nexus*, vol. 3, no. 6, p. pgae204, 2024.

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

[7] S. Zhang, Y. Guo, P. Zhao, C. Zheng, and X. Chen, "A graph-based temporal attention framework for multi-sensor traffic flow forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 7743–7758, 2021.

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

[9] G. Lai, W.-C. Chang, Y. Yang, and 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.

[10] M. Kim, J. H. Kim, and B. Jang, "Forecasting Epidemic Spread With Recurrent Graph Gate Fusion Transformers," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 2, pp. 1546–1559, Feb. 2025.

[11] Zhiwei Ding, Feng Sha, Yi Zhang, and Zhouwang Yang, "Biology-Informed Recurrent Neural Network for Pandemic Prediction Using Multimodal Data," *Biomimetics*, vol. 8, no. 2, pp. 158–158, 2023.

[12] Lijing Wang, Lijing Wang, Aniruddha Adiga, Aniruddha Adiga, Jiangzhuo Chen, Jiangzhuo Chen, Adam Sadilek, Adam Sadilek, Srinivasan Venkatramanan, Srinivasan Venkatramanan, Madhav V. Marathe, and Madhav Marathe, "CausalGNN: Causal-Based Graph Neural Networks for Spatio-Temporal Epidemic Forecasting," *Proceedings of the ... AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, pp. 12 191–12 199, Jun. 2022.

[13] Z. Liu, G. Wan, B. A. Prakash, M. S. Y. Lau, and W. Jin, "A Review of Graph Neural Networks in Epidemic Modeling," Sep. 2024.

[14] Y. Wang, Y. Zhu, L. Liang, Y. Wang, E. M. Harrison, L. Ma, and 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*, Jun. 2024. [Online]. Available: https://openreview.net/forum?id=YzWC6huGQq

[15] S. Deng, S. Wang, H. Rangwala, L. Wang, and 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*, ser. CIKM '20. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 245–254.

[16] F. Xie, Z. Zhang, L. Li, B. Zhou, and 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.

[17] J. Gao, R. Sharma, C. Qian, L. M. Glass, J. Spaeder, J. Romberg, J. Sun, and C. Xiao, "Stan: spatio-temporal attention network for pandemic prediction using real-world evidence," *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 733–743, 2021.

[18] S. Han, Y. Xun, J. Cai, H. Yang, and Y. Li, "Dygraphformer: Transformer combining dynamic spatio-temporal graph network for multivariate time series forecasting," *Neural Networks*, vol. 181, p. 106776, 2025.

[19] X. Pu, J. Zhu, Y. Wu, C. Leng, Z. Bo, and H. Wang, "Dynamic adaptive spatio–temporal graph network for covid-19 forecasting," *CAAI Transactions on Intelligence Technology*, vol. 9, no. 3, pp. 769–786, 2024. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cit2.12238

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

[21] Q. Cao, R. Jiang, C. Yang, Z. Fan, X. Song, and 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.

[22] J. Gao, J. Heintz, C. Mack, L. Glass, A. Cross, and J. Sun, "Evidence-driven spatiotemporal covid-19 hospitalization prediction with ising dynamics," *Nature communications*, vol. 14, no. 1, p. 3093, 2023.

[23] S. R. Venna, A. Tavanaei, R. N. Gottumukkala, V. V. Raghavan, A. S. Maida, and S. Nichols, "A novel data-driven model for real-time influenza forecasting," *IEEE Access*, vol. 7, pp. 7691–7701, 2019.

[24] L. Wang, J. Chen, and M. Marathe, "Defsi: Deep learning based epidemic forecasting with synthetic information," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9607–9612.

[25] 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.

[26] T. Li, L. Liu, and M. Li, “Multi-scale residual depthwise separable convolution for metro passenger flow prediction,” *Applied Sciences*, vol. 13, no. 20, p. 11272, 2023.

[27] Y. Yu, W. Sun, J. Liu, and C. Zhang, “Traffic flow prediction based on depthwise separable convolution fusion network,” *Journal of Big Data*, vol. 9, no. 1, p. 83, 2022.

[28] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio *et al.*, “Graph attention networks,” *stat*, vol. 1050, no. 20, pp. 10–48 550, 2017.

[29] O. Puny, H. Ben-Hamu, and Y. Lipman, “Global attention improves graph networks generalization,” *arXiv preprint arXiv:2006.07846*, 2020.

[30] L. Kong, V. Ojha, R. Gao, P. N. Suganthan, and V. Snášel, “Low-rank and global-representation-key-based attention for graph transformer,” *Information Sciences*, vol. 642, p. 119108, 2023.

[31] 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*, vol. 36, pp. 34 295–34 307, 2023.

[32] S. Ben Taieb, G. Bontempi, A. F. Atiya, and 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*, vol. 39, no. 8, pp. 7067–7083, 2012.

[33] R. Chandra, S. Goyal, and R. Gupta, “Evaluation of deep learning models for multi-step ahead time series prediction,” *Ieee Access*, vol. 9, pp. 83 105–83 123, 2021.

[34] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

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

[36] E. O. Oluwasakin and A. Q. Khaliq, “Data-driven deep learning neural networks for predicting the number of individuals infected by COVID-19 omicron variant,” *Epidemiologia*, vol. 4, no. 4, pp. 420–453, 2023.

[37] A. Zeroual, F. Harrou, A. Dairi, and Y. Sun, “Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study,” *Chaos, solitons & fractals*, vol. 140, p. 110121, 2020.

[38] F. Kamalov, K. Rajab, A. Cherukuri, A. Elnagar, and M. Safaraliev, “Deep learning for covid-19 forecasting: State-of-the-art review,” *Neurocomputing*, vol. 511, pp. 142–154, 2022.