Air Quality, Atmosphere & Health
https://doi.org/10.1007/s11869-025-01713-8

Hybrid graph convolutional LSTM model for spatio-temporal air quality transfer learning

Sooraj Raj[1] · Jim Smith[1] · Enda Hayes[2]

Received: 21 July 2024 / Accepted: 3 February 2025

Abstract

The short-term air quality forecasting models serve as an early warning system for local agencies, aiding in preparing mitigation strategies against severe pollution episodes. This paper explores the application of Transfer Learning to enhance short-term air quality forecasting model accuracy when labelled data is limited or missing, as often occurs with newly installed monitoring stations or due to sensor malfunctions. These monitoring stations are typically installed in areas of high exposure, like roads or urban/industrial areas, due to recurrent peak episodes or to monitor background pollutant levels generally. Forecasts with greater reliability, even when there is limited historical data available due to the recent installation of the monitoring station for example, are expected to enable the swift implementation of proactive measures to prevent significant pollution episodes from happening. The proposed method leverages knowledge from spatially neighbouring air quality monitoring stations to achieve the multi-modal spatial-temporal transfer learning to the target station, exploring multivariate time series data available from neighbouring monitoring stations. This study employed historical air quality data from spatially adjacent monitoring stations identified in South Wales, UK. The study evaluates the predictive capabilities of four base models and their corresponding transfer learning variants for estimating NO_2 and PM_{10} pollutant levels, which are the most difficult pollutants to meet objectives and limit values in the UK's air quality strategy. The paper highlights the importance of capturing spatial patterns from different monitoring stations along with temporal trends when it comes to air quality prediction. Our experiments demonstrate that transfer learning models outperform models trained from scratch on air quality multivariate time series prediction problems in a low data environment. The proposed hybrid Graph Convolutional-LSTM model, making use of a novel Granger causality-based adjacency matrix for the new site, has significantly outperformed other baseline models in predicting pollutants, achieving notable improvements in prediction accuracy of approximately 8% for PM_{10} and 7% for NO_2 values, as reflected in the RMSE values. It has also demonstrated the potential for data-efficient approaches in spatial transfer learning by reducing the need for large datasets by incorporating prior causal information.

Keywords Graph convolutions · Granger causality · Transfer learning · Air quality prediction · Spatio-temporal forecast

Introduction

Combining more precise short-term forecasting with episode-specific air quality management is expected to enable the proactive deployment of mitigation strategies to prevent peak episodes from happening. These measures, for example, can include limiting traffic speed for reduced emissions, delaying high-emitting industrial activities, or scheduling them for lower background pollution levels. They can also help develop improved urban air quality information and forecasting systems, enhancing the capabilities of local authorities to successfully predict and describe air contamination episodes in advance on a day-to-day basis. A successful air

✉ Sooraj Raj
sooraj2.rajasekharan@live.uwe.ac.uk

Jim Smith
james.smith@uwe.ac.uk

Enda Hayes
Enda.Hayes@uwe.ac.uk

[1] Computer Science and Creative Technologies, University of the West of England, Bristol, UK

[2] Geography and Environmental Management, University of the West of England, Bristol, UK

Published online: 05 March 2025

quality forecasting model should be able to capture both temporal recurring patterns - short time variations, long-term periodicity and spatial correlations for accurate predictions. Deep Learning models enable accurate short-term air quality forecasting by learning complex trends influenced by various variables, including transportation, industrial and home emissions, industrial operations, and meteorological elements like wind direction and speed (Liao et al. 2020). However, in many practical cases, with the recent installation of the monitoring stations or information missing due to sensor failures, the scarcity or significant gaps in labelled training data is a typical issue with air quality forecasting. New Air Quality monitoring stations are typically placed in areas of concern where exposure exists due to frequent peak episodes or to monitor background pollutant levels in general. In both these scenarios, applying transfer learning techniques from nearby monitoring stations allows valuable insights about future air quality values to be gained rapidly with minimal new data from the target monitoring station.

Although Deep Learning architectures are proven effective in problems like multi-variate time series forecasting, their predictive capabilities are significantly reduced when the amount of data available is insufficient for effective training as the network fails to capture useful patterns and trends (Dhole et al. 2021). To alleviate this problem, the transfer learning methods leveraging knowledge from spatially neighbouring air quality monitoring stations to help accomplish the target prediction task are proposed in this paper. In general, transfer learning enables a machine learning model trained to address one problem to be modified or improved to address another – a portion of the model's knowledge from the previous task is applied to solve a new task. Transfer learning helps us solve the data insufficiency issue that arises when models are expected to learn patterns from sparse datasets by allowing us to apply the knowledge that a pre-trained model has learnt from a related but different dataset. Moreover, training a model from scratch takes a lot of computational resources and time. Compared to training from randomly initialised models, knowledge transfer minimises the number of training examples needed to complete a given task, cutting training time and improving accuracy (Otović et al. 2022). In this paper, we propose a hybrid GCN-LSTM model, which utilises a Granger causality-based adjacency matrix derived for the new site to incorporate prior causal information. This approach enables the model to efficiently handle spatial transfer learning scenarios in limited data environments, resulting in substantial reductions in time, energy, and environmental costs typically associated with training new models from scratch.

Related work

Various studies have proposed utilising graph structures to represent the spatial relationships between monitoring stations and incorporating them into the model. In a recent study, the authors (Qi et al. 2019) proposed a hybrid model integrating Graph Convolutional Networks (GCNs) and LSTMs to capture the spatiotemporal variations in PM2.5 levels. Within the proposed model's spatio-temporal block, the spatial weight matrix/adjacency matrix is derived based on the spatial distances between air quality monitoring stations. This adjacency matrix and graph signals consisting of the air quality historical observations are used to extract spatial features by a graph convolution layer. In the next step, the graph signals are concatenated to form the input of the LSTM layers. Finally, the output of LSTM is treated as the input of a fully connected layer, and the output of fully connected layers is the prediction of PM2.5 mass concentration at a desirable time. In this study, by just using physical distance to calculate the graph adjacency matrix, the model might ignore any topological/terrain features of the landscape. Another limitation is that, though the authors compared the performance of the models against several state-of-the-art methods across various time intervals, the proposed GCN-LSTM model is found to be not evaluated against other established hybrid models like CNN-LSTMs. In another study, the authors (Ge et al. 2021) proposed a multi-scale spatio-temporal graph convolution network consisting of a multi-scale block, several spatio-temporal blocks and a fusion block. The authors claim the multi-scale blocks and spatio-temporal blocks form a multi-scale spatio-temporal graph convolution network and capture the temporal dependencies and spatial correlations jointly. Within the proposed model, the graph convolution layer captures the spatial correlations of air quality by collecting neighbours' information. The temporal convolution layer captures the temporal dependencies of air quality by stacking multiple layers of dilated causal convolution. It uses the residual connection to expand the receptive field on the temporal dimension. The authors have evaluated the model against several state-of-the-art methods across various time intervals, and it demonstrated superior performance, with significant improvements in prediction accuracy. However, the adjacency matrix in this study is defined according to the spatial distance between pairs of stations.

The proposed method of utilising an adjacency matrix centred on causality in our study, as opposed to reliance on spatial distance between monitoring stations in contrast to earlier studies, is expected to provide several advantages when applying graph convolutional models to air quality data. Causality-based adjacency includes the directional relationship between stations, capturing how changes in one station's readings affect others. In air quality research, directional pollutant transport is especially significant, often influenced by wind patterns or geographical features. The distance-based adjacency, where the assumption of symmetric influence is made with closer stations having a stronger influence, may not accurately represent real-world dynamics like, for example, the pollutants being driven by wind patterns. A distance matrix overlooks intricate relationships, potentially the indirect links or placing excessive emphasis on nearby stations that are not causally significant. Temporal dependencies and actual interactions among variables can be reflected by causality matrices derived from data using methods such as Granger causality. Factors like weather, industrial emissions, and regional geography are taken into account here, which cannot be attributed solely to distance.

Convolutional Neural Network (CNN) models are trained using a dataset with enough observations, referred to as the source dataset. These models typically utilise earlier layers to learn simpler patterns and later layers to learn more complicated patterns from the dataset (Tajbakhsh et al. 2016). Therefore, the initial layers of CNNs are generally responsible for detecting and identifying generic characteristics, and it is assumed that these generic features useful for solving one problem can be leveraged to address another. This allows the freezing of parts of the initial layers (so that their weights don't change during further training) and fine-tuning of the remaining layers using a target dataset containing a small number of samples (Jmour et al. 2018; Ribani and Marengoni 2019; Soekhoe et al. 2016). However, with the time series problems, though several related datasets exist, the degree of relevance between them and the target dataset usually appears ambiguous. Blindly transferring knowledge from less relevant datasets to the target one is expected to drag the prediction performance. Using transfer learning techniques to create a forecast for a target time series has been explored in some previous studies. A deep LSTM model with fully connected layers for demonstrating transfer learning to predict future residential scale electricity loads at hourly granularity is used by authors (Laptev et al. 2018). Though LSTMs are good at exploring temporal trends, it is found that spatial dependencies are not effectively captured by this model (Yin et al. 2020). In another study, the authors (Fong et al. 2020) considered transfer learning using LSTMs to predict air pollutant concentrations at different air quality monitoring stations. While the approach involves transfer learning and reusing the pre-trained base LSTM network for related air quality datasets, it does not explicitly explore spatial dependencies between different monitoring stations. In air pollution forecasting, capturing spatial dependencies is important because specific geographical features of neighbouring locations can influence pollution levels at one location. A different approach with transfer learning to the air quality forecasting problem is applied to tackle the lack of enough labelled datasets for newly installed monitoring stations in another study (Dhole et al. 2021). The authors of this study offer a system that transfers the information gathered from several source stations to a specific station of interest, giving a cumulative forecast. This ensemble approach allows each model to generate a prediction on the test data of the target dataset, and then it combines these individual predictions using the Exponential Weights Algorithm (EWA) was proposed to investigate the impact of transferring knowledge from 10 models pre-trained on multiple source datasets to a given target. However, this method is found to be computationally expensive and inefficient as we need to re-train multiple deep learning models from different monitoring stations to achieve the ensemble approach and to achieve the goal of transferring spatial information from neighbouring stations to a specific station of interest.

Research objective and graph convolutional neural networks

In an ideal approach, we should be able to achieve the multi-modal spatial-temporal transfer learning to the target station, exploring multivariate time series data available from different neighbouring monitoring stations. To explicitly explore spatial dependencies, additional techniques can be incorporated into the approach. For example:

- Spatial Convolutional Layers: Integrate convolutional layers into the LSTM architecture to extract spatial features and capture spatial patterns in the data (Huang and Kuo 2018; Zhang et al. 2020).
- Graph Structures: Utilise graph structures to represent the spatial relationships between monitoring stations and incorporate them into the model (Qi et al. 2019; Get et al. 2021).

By incorporating these techniques, the approach can better capture and utilise the spatial dependencies in the data efficiently, leading to improved air pollutant concentration predictions at different monitoring stations. However, it's wrong to assume that every other variable influences a single variable's anticipated value, and many advanced deep learning models applied to multivariate time series forecasting problems tend to focus on the specific causal relationship among the variables (Duan et al. 2022). However, considering such prior causal information is crucial when deploying a spatial transfer learning paradigm to the new target dataset with limited/missing entries.

We can represent the causal relationships among multiple spatial pollutant variables by constructing a multivariate times series graph, with each variable as a node and each edge indicating a causal link between them. This study employs Graph Convolutional Neural Network (GCN) filters for time series spatial feature extraction (Fig. 1). They aggregate features from their own nodes and neighbours to generate the node's representation, learning feature embeddings and graph patterns using varying perception scales. Then, the GCN is used to resolve the forecasting issue from the generated graph based on the spatial multivariate times series.

Our work uses the weights trained from the base models as the initial weights for training the model on the new related dataset. Transfer learning is implemented as fine-tuning with a weight initialisation approach (Fong et al. 2020) in this paper, and the idea is that the weights from the base models are already trained with useful features and patterns from the initial task or dataset and can then be fine-tuned for the new task or dataset with less data and training time (Pinto et al. 2022; Bird et al. 2020). In order to assess how well the Transfer Learned models considered in this research performed, we chose to fine-tune the whole network with pre-trained weights of the base models, as it is found that re-training the whole network almost always leads to better results (Otović et al. 2022; Fawaz et al. 2018).

Hybrid GCN-LSTM model

The proposed hybrid model uses GCN and LSTM layers to perform forecasting over a graph consisting of multivariate time series.

The GCN layer consists of three main steps:

1. Computing node representations by multiplying the input features with a trainable weight matrix.

   Let X be the input feature matrix of shape (Num_N, Num_F), where Num_N is the total number of nodes and Num_F is the number of input features per node. Let W1 be the trainable weight matrix of shape (Num_F, D), where D is the desired dimensionality of the node representations. The node representations R1 are computed as follows:

   $$R1 = X * W1 \tag{1}$$

   This transformation applies learnable weights to the features of each node independently, without considering neighbours.

2. Computing aggregated messages for each node - by first gathering features from neighbouring nodes defined as the graph's edges by the adjacency matrix and then calculating the mean of these neighbour representations. Each element now corresponds to the aggregated message from the neighbouring nodes for each specific node in the graph. These aggregated messages are further multiplied with a trainable weight matrix.

   Let A be the adjacency matrix of shape (Num_N, Num_N), where Num_N is the total number of nodes and A (i, j) = 1 represents the link between nodes i to j. Let R1 be the node representations from the previous step. Let W2 be the trainable weight matrix of shape (D, H), where H is the dimensionality of the aggregated messages. The aggregated messages M2 are computed as follows with (A * R1) representing the element-wise multiplication between the adjacency matrix A and the node representations R1:

   $$M2 = (A * R1) * W2 \tag{2}$$

   This step lets the node listen to its neighbours and incorporate relational/contextual information.

3. Generating node embeddings by concatenating the weighted node representations derived from step 1 with the weighted aggregated messages from neighbouring nodes derived in step 2.

   Let R1 be the node representations from step 1, and M2 be the aggregated messages from step 2. The node embeddings E are generated by concatenating the weighted node representations with the weighted aggregated messages:

   $$E = [R1, M2] \tag{3}$$

   This concatenation step enriches the final node embedding by combining local features from the node and its neighbours' contextual features.

In order to improve the prediction accuracy, it's equally important to consider the temporal patterns in the data

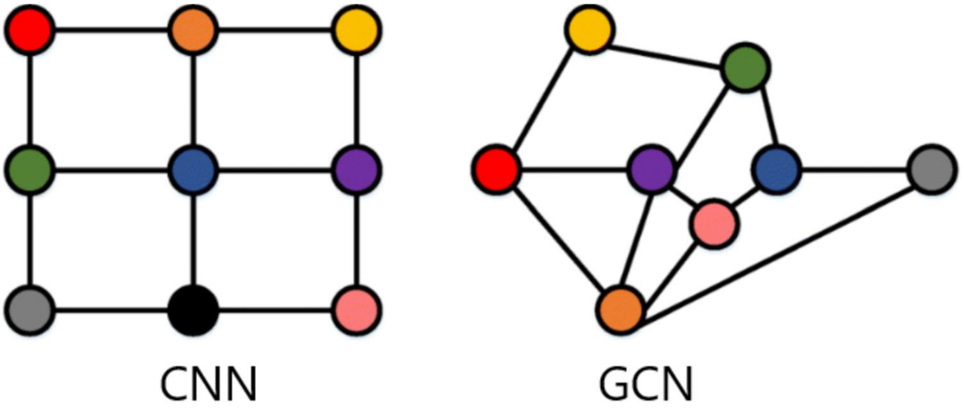


Fig. 1 CNN kernel operates on regular structures and graph convolutional kernels operate on graph structures (Lin et al. 2021)

along with the spatial patterns from neighbouring nodes. To achieve this, the hybrid model - consisting of graph convolution layers applied to the input features and then the output from the graph convolutional layers, i.e., the node embeddings (from step 3) fed to the LSTM layer is considered in this paper. The LSTM layer is expected to capture any sequential/temporal patterns, along with graph convolutional layers, which explore spatial patterns in the data. The model topology is shown below in Fig. 2.

Let E be the node embeddings from the GCN layers. The LSTM layer is used to capture temporal patterns. Assuming a single-layer LSTM, the hidden activation states $A_{<t>}$ and cell states $C_{<t>}$ at each time step t are computed as follows:

$$A_{<t>,}C_{<t>} = LSTM\left(E_{<t>,}A_{<t-1>,}C_{<t-1>,}\right) \tag{4}$$

Here, E_t represents the node embeddings at time step t, $A_{(t-1)}$ the previous hidden activation state, and $C_{(t-1)}$ the previous cell states.

Figure 2 shows the proposed hybrid GCN-LSTM model annotated with tensor shapes used in our experiment.

Deriving adjacency matrix using granger causality

An adjacency matrix is used to define the neighbourhood of each node in a Graph Neural Network. By using the adjacency matrix, a GCN can learn to aggregate information from the neighbouring nodes of each node in the graph. The group of nodes that are immediately connected to a node is referred to as its neighbourhood. One common strategy is to treat a monitoring station that greatly aids in forecasting the air quality at the target area if it is situated in close proximity to the destination. However, the geographic distance loses information when monitoring stations are spread off from one another, and all distances could have comparable values then. It is significant to remember that a variety of factors, each with a unique pattern of influence on air quality values, affect the spatial correlations of air quality. These factors include industrial, urban, rural areas where these monitoring stations are located and road traffic variations, and more. Accurately predicting air quality can be difficult if geographical correlations are represented by a single component, like geographic distance. (Ge et al. 2021). In order to tackle this, we explored the novel idea of applying Granger Causality tests to derive the adjacency matrix for the air quality values from spatially neighbouring monitoring stations in this paper.

Granger Causality is a statistical technique that is widely used to determine the causal relationships between time series (Shojaie and Fox 2022). It is used to test whether one time series, denoted as Xi, Granger-causes another time series, denoted as Xj. In mathematical terms, the Granger causality test can be defined as follows:

Given two time series, Xi and Xj, with observations at time steps t = 1, 2, …, T, the Granger causality test involves estimating autoregressive models for both series and comparing the goodness of fit with and without including the lagged values of Xi as predictors of Xj (Foresti et al. 2006, Rodriguez-Caballero et al. 2014).

The null hypothesis of the Granger Causality test is that Xi does not Granger cause Xj, meaning that the past values of Xi do not provide any additional information for predicting Xj beyond what is already captured by the lagged values of Xj. The alternative hypothesis is that Xi does Granger-cause Xj, indicating that the past values of Xi provide significant additional information for predicting Xj.

To perform the Granger causality test, we estimate two autoregressive models: one with only the lagged values of Xj as predictors (restricted model) and another with the lagged values of both Xj and Xi as predictors (unrestricted model).

Let AR (Xj, p) represent the autoregressive model of Xj with lag p, and AR(Xj, Xi, p) represent the autoregressive model of Xj with lag p, including the lagged values of Xi as predictors.

The models can be expressed as follows:

$$AR(Xj,\ p):\ Xj(t)\ =\ cj\ +\ \Sigma\beta j,k\ *\ Xj(t-k)\ +\ \varepsilon j(t)\ (restricted\ model) \tag{5}$$

$$AR(Xj,\ Xi,\ p):\ Xj(t)\ =\ cj\ +\ \Sigma\beta j,k\ *\ Xj(t-k)\ +\ \Sigma\gamma j,k\ *\ Xi\ (t-k)\ +\ \varepsilon j(t)\ \ (unrestricted\ model) \tag{6}$$

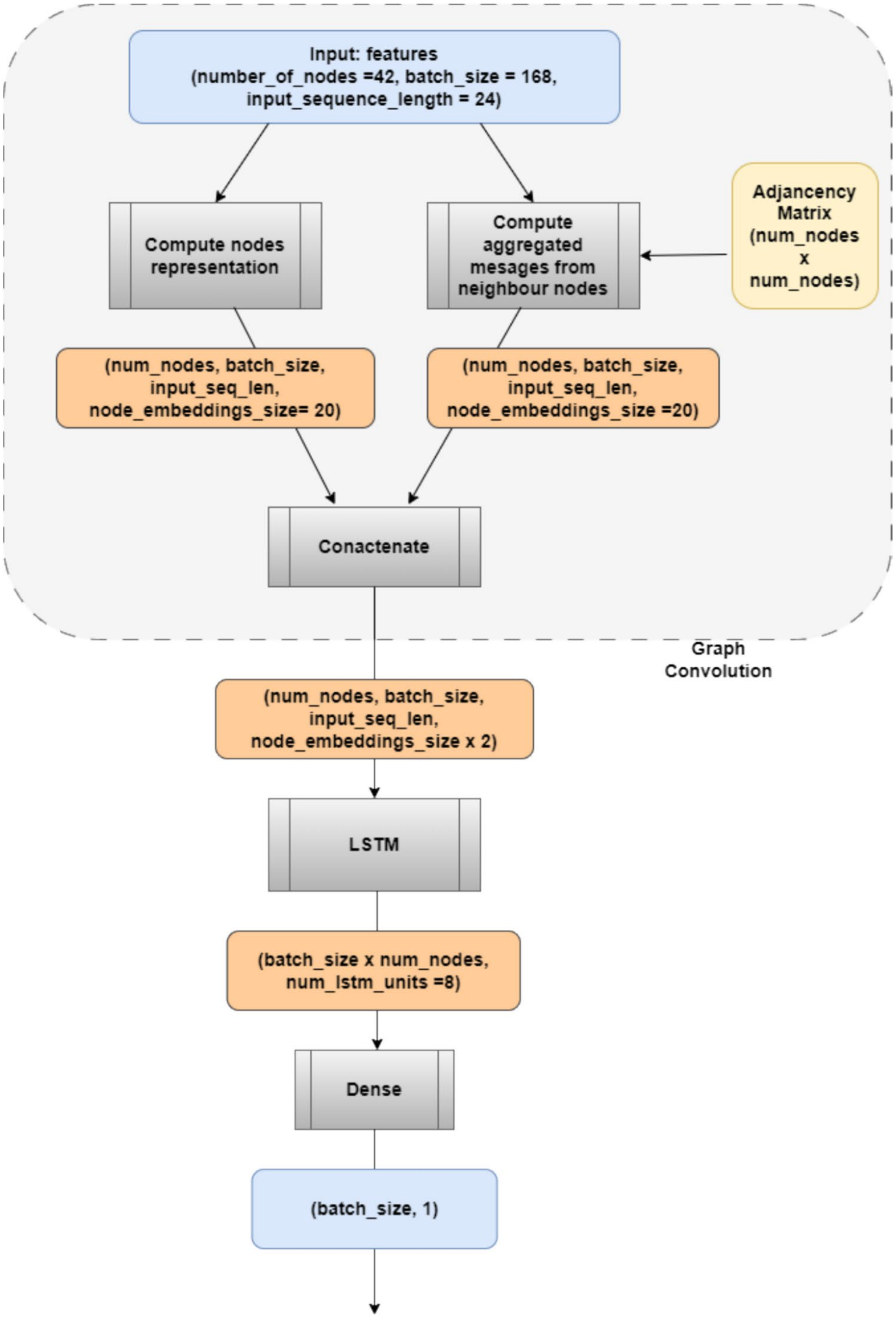
where c_j is the intercept, $\beta_{j,\,k}$ and $\gamma_{j,\,k}$ are the coefficients for the lagged values of Xj and Xi, respectively, and $\varepsilon_{j(t)}$ is the error term.

Let SSR_restricted represent the squared residual sum values for the restricted model, and SSR_unrestricted represent the squared residual sum values for the unrestricted model. Given a regression model with observed values y_i and predicted values $\bar{y}_i$ for i = 1, 2, …, n, the residual for each data point is defined as the difference between the observed value and the predicted value:

$$\varepsilon i\ =\ yi\ -\ \overline{yi} \tag{7}$$

The sums of squared residuals (SSR) are then calculated as:

Fig. 2 Hybrid GCN LSTM Model with annotations showing size of dataflows at different stages


Input: features
(number_of_nodes =42, batch_size = 168,
input_sequence_length = 24)
Compute nodes representation
Compute aggregated mesages from neighbour nodes
Adjancency Matrix (num_nodes x num_nodes)
(num_nodes, batch_size, input_seq_len, node_embeddings_size= 20)
(num_nodes, batch_size, input_seq_len, node_embeddings_size =20)
Conactenate
Graph Convolution
(num_nodes, batch_size, input_seq_len, node_embeddings_size x 2)
LSTM
(batch_size x num_nodes, num_lstm_units =8)
Dense
(batch_size, 1)

$$SSR = \Sigma(\varepsilon i^2) = \Sigma(yi - \overline{y}i)^2 \tag{8}$$

The distinction between the restricted and unrestricted models lies in the predictors used to calculate the predicted values, In the restricted model, only lagged values of X_j are used as predictors, while in the unrestricted model, both lagged values of X_j and X_i are used as predictors.

The F-statistic is computed as:

$$F = ((SSR_restricted - SSR_unrestricted)/p)/(SSR_unrestricted/(T - p - 1)) \tag{9}$$

where p is the number of lagged variables added to the unrestricted model compared to the restricted model and T is the total number of observations.

Determining the significance level to evaluate the probability of null hypothesis rejection involves comparing the F-statistic to the critical value derived from the F-distribution with the suitable degrees of freedom. Here null hypothesis rejection suggests the evidence of Granger causality, indicating that the lagged values of Xi provide additional information for predicting Xj. Hence the test assesses whether including the lagged values of Xi significantly improves the prediction of Xj.

By applying Granger causality to the multivariate time series, we expect to derive an adjacency matrix that represents the causal relationships between the variables in the time series. We believe the adjacency matrix derived in this manner will capture the prior and important spatial causal information of the dataset, especially in a low-data environment. We propose constructing the adjacency matrix using Granger causality with the following steps:

1. Set a lag value (the number of past values of the time series that is under analysis for usefulness in forecasting another) for the Granger Causality test. In our case, we tuned this as a hyperparameter, and it's set as 24 (Appendix 2).
2. Apply the Granger Causality test, which considers a pair of time series from the multivariate time series as input and performs a series of hypothesis tests (as shown above) to determine whether the set number of lag values of one of the time series Granger causes the other time series.
3. If the first time series Granger causes the second time series, the entry in the adjacency matrix denoting the link is set to 1, indicating a directed edge from the first time series to the second time series. Otherwise, the entry is set to 0.

Repeat steps 2–3 for all pairs of individual time series in the multivariate time series

Granger Causality test estimates vector autoregressive (VAR) models for each possible combination of variables and lags up to the maximum lag order. The algorithmic complexity involves estimating each VAR model by solving a set of k (k = maximum lag order) linear equations with n coefficients (number of time series variables). Hence, it's computationally efficient in quickly deriving prior causal relationships between the variables.

Granger causality tests assume that the input series are stationary to avoid spurious causality. Hence, a prerequisite for performing the Granger Causality test is that for any time series to have a predictive causality on another time series, both must be stationary. Before conducting Granger causality tests, we tested and confirmed stationarity in the variables between each time series considered using the Augmented Dickey-Fuller (ADF) procedure (Ventosa-Santaulària and Vera-Valdés 2008). Even if spurious causalities are derived by the Granger Causality test between a pair of series creating the wrong edge, the GCN part of the model has the ability to learn and ignore the neighbouring information from irrelevant edges. The following common strategies can be applied if the considered series for Granger causality fails the stationarity check (Hyndman and Athanasopoulos 2013) - If the series has trends or unit roots, differencing can be applied to make it stationary. Log transformation or Box-Cox transformation can be applied if the variance is non-constant. If differencing alone doesn't entirely remove the trends, detrending the series explicitly using linear regression or moving average can also help.

The new Adjacency Matrix for the Newport Road site is specifically derived using Granger Causality in this way to capture the prior causal information of the new location. Once a new adjacency matrix is calculated, it becomes the foundation for graph propagation and determines how the model combines the information from neighbouring nodes. When applying transfer learning as a weight initialisation scheme for GCN-LSTMs, the creation of node representations and computation of aggregated messages begin with the pre-trained weights and the new adjacency matrix. Here, the pre-trained weights of the GCN layers, which primarily drive feature transformation, retain information from the source graph. The new adjacency matrix (A′) defines each node's latest set of neighbours, altering the scope of information aggregation and influencing the aggregation of messages and how the embeddings evolve. The new adjacency matrix ensures the embeddings are adapted to the target graph structure. The aggregated messages M2 are computed as follows, with (A′ * R1) representing the element-wise multiplication between the adjacency matrix A′ and the node representations R1:

$$M2 = (A' * R1) * \hat{w}2 \tag{10}$$

Here, the pre-trained weights in $\hat{w}$ still retain general knowledge about how to aggregate and transform features.

Methodology

Dataset

For training the base models, the study utilised air quality data from the AURN (Automatic Urban and Rural Network) monitoring stations located in Port Talbot Margam, Swansea, Narberth, and Cardiff Centre in the United Kingdom. We used all the available pollutant values tracked by these monitoring stations along with 3 modelled metrological inputs –wind direction, wind speed and temperature captured by each of these monitoring stations for the continuous 11 years (spanning from 2011 to 2022), totalling 42 input features as part of the experimentation. Since these individual features have different ranges, we applied Standard Scaler as a pre-processing step.

NO_2 and PM_{10} are two of the pollutants for which achieving current objectives and limit values within the Air Quality Strategy is found to be the most challenging (DEFRA 2023). In Wales, out of the 44 declared AQMAs, 43 are based on consistently elevated NO_2 pollution (originating from road transport), with the exception of Port Talbot, which has been declared as AQMA due to PM_{10} emissions from local industry. As a result, this research focuses on reliable short-term forecasting of the pollutants - PM_{10} and NO_2, which are also responsible for the majority of AQMA declarations in the UK. The data from 2011 to 2021 is used for training the base models, and the data from 2022 is used as the test set. The specific pollutant values chosen in this study and their strong relationship with meteorological values have been extensively analysed in our previous study (Raj et al. 2022).

In order to smooth the data and make it relate better to what policy decision-makers are interested in, a rolling average of 24 h of PM_{10} and Nitrogen Dioxide (NO_2) for Port Talbot are considered as targets for training and prediction of the base model. It is modelled as a multivariate one-step regression problem to predict future value (12 h in advance) of one pollutant variable at a time (PM_{10} or NO_2) with all pollutant values and metrological values from all stations for the past 24-hour window provided as inputs.

In this study, the trained base models for predicting rolling average values of PM_{10} and NO_2 were further examined to determine their usefulness for transfer learning to a relatively new AURN monitoring station. Specifically, the study focused on the neighbouring and newly established Newport Road, Cardiff monitoring station, for which pollutant and modelled meteorological data were only available from 2019 onwards. The data from 2019 to 2021 of the Newport Road monitoring station is used to re-train the Transfer Learned models. Data from 2022 is used as the test data to estimate the Transfer learned model's performance, i.e., to forecast the 24-hour rolling average values of PM_{10} and NO_2 12 h in advance on unseen data.

The input variables used from the base model target station and the transfer learning model target station are plotted in Appendix 3 to visualise the generic patterns of pollutant values and meteorological values.

We have not used the conventional N-fold cross-validation as: (i) the dataset is reasonably large, (ii) there may be underlying annual and seasonal patterns in the time series data that N-fold cross-validation (even using different years for testing) would ignore (iii) for training and testing it is more realistic if the data represents a continuous sequence and (iv) repeating results for the same train/test split allows us to compare the reliability of different models better. Though one year of test data is reasonably large enough, it may not capture long-term downward trends of pollutants. For example, certain events or trends (e.g., pandemics, natural disasters, or policy changes) occurring in a specific year might skew the results and make the test set unrepresentative on exceptional occasions.

Baseline models

We compared the performance of the proposed hybrid GCN-LSTM model in predicting the future air quality values with three baseline models: - a stacked LSTM model, a Transformer model, a Transformer attention model and a hybrid 2DCNN-LSTM model. To prevent overfitting, the dropout layers are added to the models for regularization. The following base model topologies and the respective hyper parameters were optimised for the best prediction accuracy using Grid Search methodology with the early stopping criterion for the validation loss set as Mean Squared Error (MSE) with a patience of 5 epochs, where if the validation loss does not improve after five consecutive epochs, training ends. The hyperparameters fine-tuned for each model are explained in Appendix 2.

- Stacked LSTMs:

  LSTM networks are the most popular type of Recurrent Neural Networks (RNN) used for time series forecasting tasks and they effectively address issues of vanishing/exploding gradients in standard RNNs to learn long- and short-term dependencies in sequence data. They are successfully applied on air quality forecasting problems before (Reddy et al. 2018; Freeman et al. 2018; Li et al. 2017) and also to time series transfer learning (Fawaz et al. 2018; Fong et al. 2020) problems. We have considered the stacked LSTM model as a baseline for our comparison study.

  - Topology: Input layer with 42 features, 2 LSTM hidden layers with 16 nodes each and an output dense layer with 1 node predicting a single time step 12 hours in advance.
  - LSTM cell state and hidden state activation: Tanh, Recurrent Activation: Sigmoid.
  - Input window size: 24, Dropout: 0.2, Batch size: 168.

- Transformer:

  Transformers were originally devised for natural language processing tasks. Due to their sophisticated attention mechanisms, they stand out in managing sequential data. This feature also makes them apt for conducting time series tasks (Wen et al. 2022).

The transformer block constructs an input sequence's continuous representation, or embedding, via an encoder function, and it consists of:Layer Normalisation- which normalises the input layer, expediting training and stabilising the network. Multi-Head Attention- Facilitating simultaneous attention to diverse aspects of the input data from various representation subspaces at different positions. Feed Forward Network- A straightforward neural network applied individually to each position, enhancing the model's capacity to learn complex patterns. In our case a 1D convolutional layer with ReLU activation is applied to learn features and another 1D convolutional layer is applied to transform the features back to the original dimension.

- Layer Normalisation- which normalises the input layer, expediting training and stabilising the network. Multi-Head Attention- Facilitating simultaneous attention to diverse aspects of the input data from various representation subspaces at different positions. Feed Forward Network- A straightforward neural network applied individually to each position, enhancing the model's capacity to learn complex patterns. In our case a 1D convolutional layer with ReLU activation is applied to learn features and another 1D convolutional layer is applied to transform the features back to the original dimension.
- Topology: Our Transformer model is constructed by stacking 4 layers of the Transformer encoder, to deepen the network and enhance its learning capacity. Subsequent to the Transformer blocks, a Global Average Pooling layer is implemented to reduce output dimensionality, streamline the model and focus on pivotal features. Finally, the output layer is composed of a dense layer with a linear activation function, tailored to predict a single time step 12 hours in advance.
- Input layer is with 42 features, dimensionality of each attention head which determines the complexity and richness of the attention patterns that each head can learn is set to 128, number of attention heads which leads to better-represented vectors is set to 42, the size of the hidden layers in the feed-forward network inside the transformer is set to 84 and Number of units in the Dense networks after the encoder layer is set to 256 following hyper parameter tuning.

- Hybrid Graph Transformer Attention:

  The Graph Convolution layers from the proposed Hybrid GCN-LSTM model detailed in section 3 are combined with the aforementioned Transformer Attention model, replacing the LSTM layers to evaluate the performance of a Hybrid Transformer Attention model. The output of the graph convolution layers is reshaped into shape - (batch size, number of nodes, input sequence length, 2 x Number of node embedding size) and fed to the transformer model with attention head size set to 128, number of attention heads set to 42, the size of the hidden layers in the feed-forward network inside the transformer is set to 84 and the number of units in the dense networks after the encoder layer set to 256.

- Hybrid 2DCNN-LSTM:

  Since CNN and RNN can extract spatial and temporal features from data, respectively, combining these two methods for better prediction in time series has been successfully explored previously (Yin et al. 2020; Oehmcke et al. 2018; Kim et al. 2018). We have applied a hybrid Convolutional LSTM architecture (Huang et al. 2018; Zhang et al. 2020) as another baseline model for comparison and made some modifications on CNNs inspired by the paper (Hoseinzade et al. 2019). We applied 2D-CNN layers in the hybrid architecture for initial feature extraction. To extract spatial features, filters with kernel size (1 × number of nodes) are utilised and these filters cover all the features from a single time step and can combine them into a single higher-level feature. The subsequent 2D-CNN layer with kernel size (3,1) combines extracted features of different time steps to construct higher-level features for aggregating the available information in adjacent time periods. Finally, an LSTM layer is applied to extract any longer-term dependencies in the sequence data.

  - Topology – Input layer with 42 features, 2 x 2D convolutional layers with 8 filters each and kernel sizes (1,42) followed by (3,1). This is followed by an LSTM layer with 8 nodes and, finally, an output-dense layer with 1 node predicting a single time step 12 hours in advance.
  - Convolution layer activation: ReLU.
  - LSTM cell state and hidden state activation: Tanh, Recurrent Activation: Sigmoid.
  - Input window size: 24, Dropout: 0.2, Batch size: 168.

Comparison metrics and hypothesis tests

We evaluated the performance of regression models considered in this study using Root Mean Square Error (RMSE) and performed 10 runs of each combination with the same train /test split but with different seeds for the initial network weights. To compare the effectiveness of different algorithms considered in this study, we used mean differences of RMSE values, calculated over 10 runs for each model considered. We used the Mann-Whitney Wilcoxon test with the Bonferroni correction at a 95% level to assess the statistical significance of any observed differences in performance.

Results

We compared the prediction performance of 3 Models and Transfer Learned models under consideration on rolling average values of both the pollutants - NO_2 and PM_{10} separately.

Models considered

We initially compared the prediction performance of the baseline models considered. The rolling average values of PM_{10} and Nitrogen Dioxide (NO_2) from the past 24 h for Port Talbot monitoring station are considered as targets for training and prediction. For the comparison, the data from 2011 to 2021 is used for training the models and the data from 2022 is used as the test set.

The observations indicate (Fig. 3):

- Both GCN-LSTM and 2DCNN-LSTM models were found to be better than LSTM, Transformer Attention and Graph Transformer Attention models when forecasting both NO_2 and PM_{10} values.

The given observations suggest that:

- Incorporating spatial patterns, either through graph convolutional structures in the case of GCN-LSTM or through convolutional structures in the case of 2DCNN-LSTM, is more beneficial than solely relying on capturing temporal patterns using stacked LSTM or Transformer models. In contrast, both the LSTM model, and Transformer model primarily focus on capturing temporal dependencies in the data. While they are both capable of capturing some temporal patterns, it is found to be not fully exploiting the spatial information present in the data. It's can also be observed the proposed hybrid GCN-LSTM model is found to be performing better than the GCN-Transformer Attention model considered in this study. Similar results were observed with Transformer models on time series problems in previous studies as well (Zeng et al. 2023). Due to the poor baseline performance, we have omitted the stacked LSTM models, Transformer Attention and Graph Transformer Attention models from further analysis while executing Transfer Learning.

Effectiveness of transfer learned hybrid GCN-LSTM model

We have analysed the effectiveness of the proposed hybrid GCN-LSTM transfer learning model against CNN-LSTM. As explained in Section 6.2, the data from 2019 to 2021 of

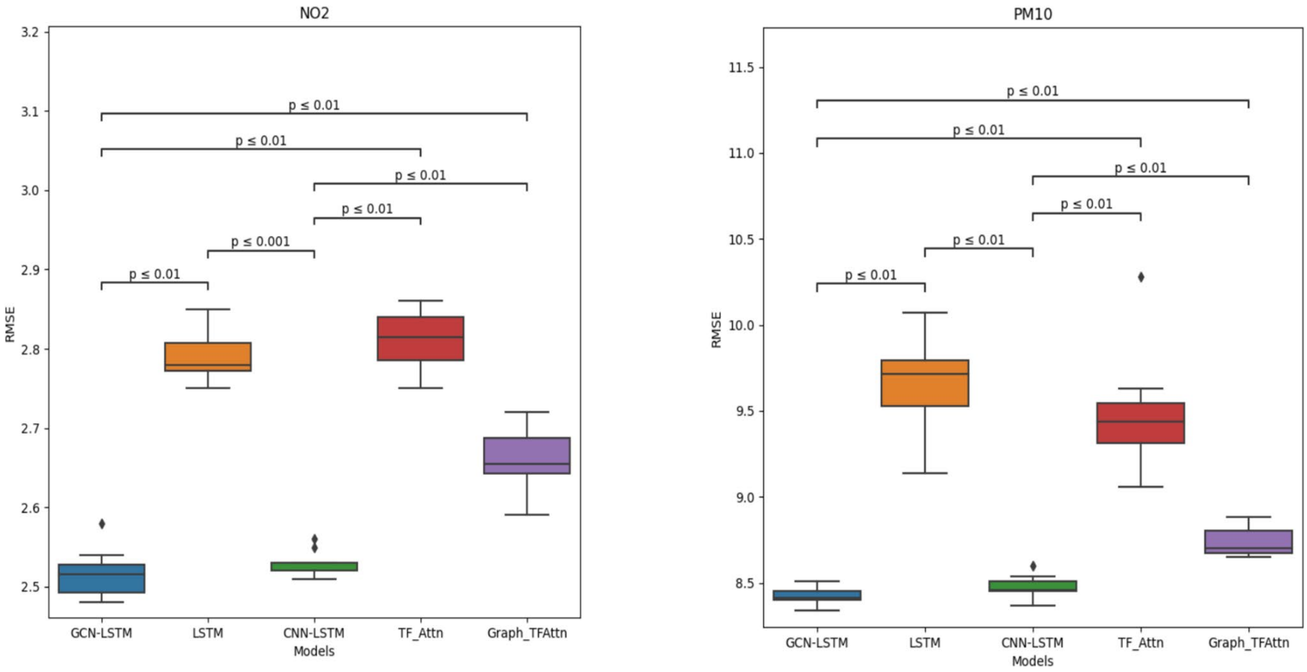


Fig. 3 Performance comparison of base models

the Cardiff Newport Road monitoring station is used for re-training the Transfer Learned models. Data from 2022 is used as the test data to estimate the Transfer learned model's performance on unseen data.

It is observed that (Fig. 4):

- The Transfer Learned hybrid GCN-LSTM models with the Adjacency Matrix specifically derived using Granger Causality for Cardiff-Newport Road site, are significantly more effective ($p < 0.001$) in forecasting both the pollutants than the other two Transfer Learned models compared in this study. The model has achieved a significant improvement in predictive accuracy, with an 8% increase in PM_{10} and a 7% increase in NO_2, as measured by RMSE values.

The given observation suggests that.

- Significant superiority of Granger Causality-based adjacency matrix: Among the Transfer Learned models, the one utilising the Adjacency Matrix derived using Granger Causality for the Cardiff-Newport Road site stood out as significantly more effective. The significance level of $p < 0.001$ indicates a strong statistical difference.
- Capturing prior causal information: The proposed hybrid GCN-LSTM model with the Granger Causality-based adjacency matrix is found to be effective in capturing prior causal information. In a low data environment, where the available historical data is limited for training (data from 2019 to 2021 in our case), leveraging this prior causal information is found to be particularly valuable comparing the base models.

Effectiveness of transfer learning over training de novo

To analyse the effectiveness of transfer learning strategy, we have compared the top performing models: GCN-LSTM and 2DCNN-LSTM starting from weights taken from models trained at the Port Talbot base station (transfer learning) with randomly initialised weights.

The observations indicate:

- The transfer learned models were more effective than their counterpart models trained from scratch. (Fig. 5)
- Transfer Learned models achieve convergence in only about half the number of epochs (15–20 epochs) required by models initialised using random weights (35–40 epochs).

The findings from these observations can be summarized as follows:

- Improved performance of transfer learned models: Fig. 6 demonstrates that the transfer learned models outperformed their counterparts trained from scratch. This improvement is statistically significant, indicating that the knowledge and generic feature extractors learned from the base dataset (Port Talbot) are still relevant and beneficial for forecasting pollutant values at the new site. The transfer learned models leverage the learned features, such as daily/weekly patterns and the effects of factors like wind, rain, wind direction, or temperature, to capture the complexities and disruptions in pollutant plumes.

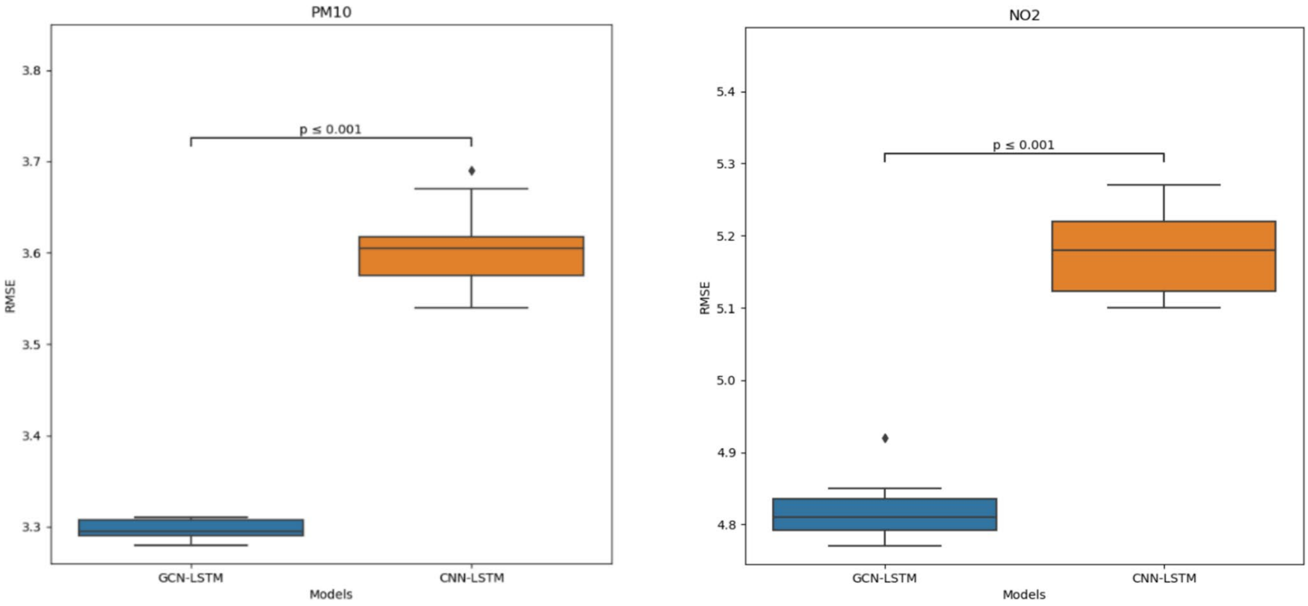


Fig. 4 Performance comparison of transfer learned models

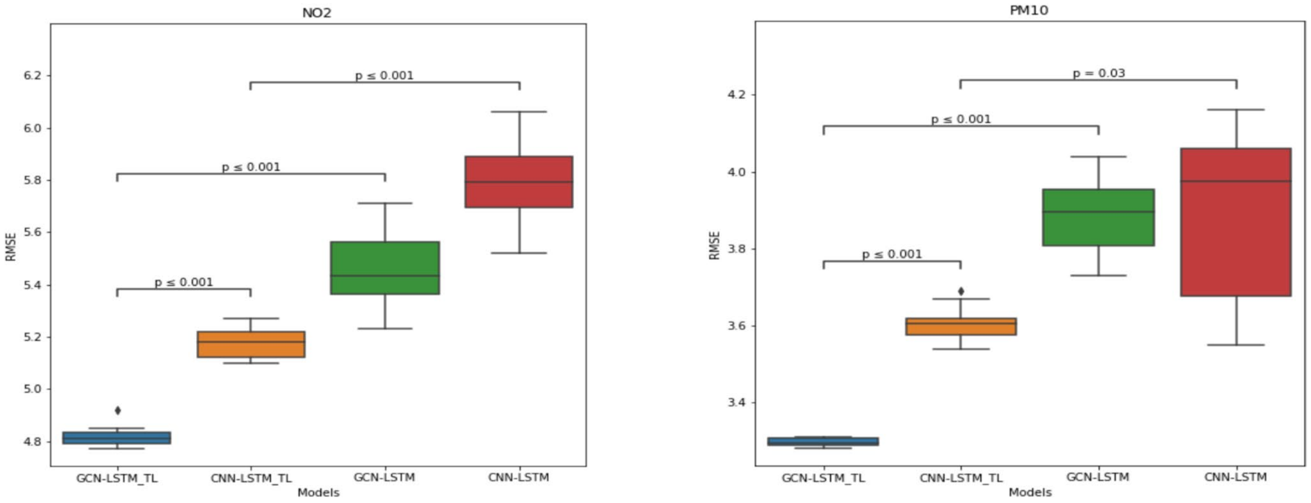


Fig. 5 Performance comparison of models trained using weights initialised via transfer learning and initialised to random values

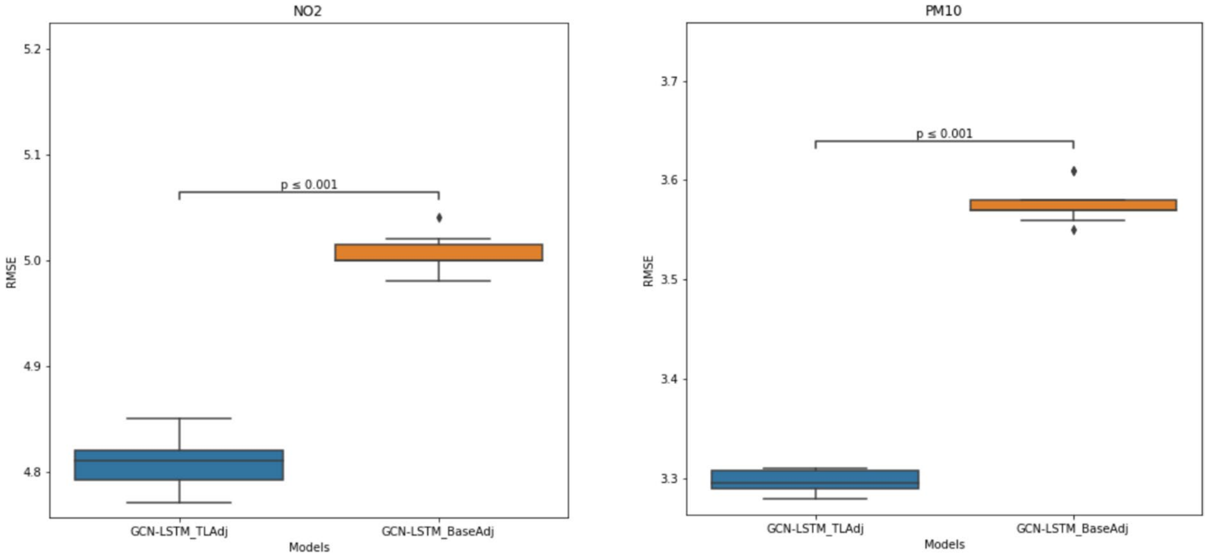


Fig. 6 Performance comparison when adjacency matrix is derived for new vs. using base data adjacency matrix

- Faster convergence with transfer learning: The transfer learned models achieved convergence in a significantly shorter number of epochs compared to models initialised with random weights. Typically, transfer learned models require approximately 15–20 epochs to converge, while models with random weights require 35–40 epochs. This faster convergence indicates that the pre-trained weights provide a good initialisation point, allowing the model to quickly adapt and fine-tune to the specific characteristics of the new site. As a result, transfer learning not only improves performance but also reduces training time, which can be advantageous in scenarios with limited computational resources or time constraints.

Effectiveness of (re)learned graph adjacency matrices

The additional analysis conducted as an ablation study compares the effectiveness of the proposed Granger Causality-based adjacency matrix derived specifically for the Cardiff-Newport Road site with the adjacency matrix derived for the base model.

The results indicate (Fig. 6).

– The newly derived adjacency matrix specific to the Cardiff-Newport Road site is significantly more effective on the Transfer Learned model than using the adjacency matrix derived for the base model.

For further visualisation, the directed edges towards NO_2 and PM_{10} values for base and newly derived adjacency matrix using Granger Causality where the causal relation is represented as a directed edge with value = 1 from row index parameter to column index parameter is shown in Fig. 7.

Out of the monitoring stations considered in this study, Port Talbot monitoring station tracks industrial background, Narberth tracks rural background, Cardiff tracks urban background, and Swansea and Newport Road track roadside background (Fig. 8 shows the geographical locations of the monitoring stations). It can be observed that the adjacency matrix infers some non-obvious causal relationships, and some typical examples of these causal relationships are shown in Figs. 9 and 10. For example, it can be observed the meteorological inputs from the Cardiff monitoring station - Temperature, Wind speed and Wind direction have a causal effect on NO_2 values of spatially neighbouring Newport Road site, but only Temperature and Wind speed from Cardiff have a causal effect on NO_2 values of Port Talbot site. The PM_{10} values of Newport Road have a causal effect from Ozone, CO and SO_2 values from neighbouring Cardiff, but PM_{10} moving average values in Port Talbot only have a causal effect from Ozone values of Cardiff. These typical examples show how site-specific prior causal information due to unique factors caused by local and spatial emission sources, weather patterns and geographical features could be incorporated into the adjacency matrix, hence contributing to improved forecasting performance by the GCN-LSTM model.

Transfer learning performance when more data is added

We have also checked what happens in each step when incorporating additional data for deriving an adjacency matrix by Granger Causality for GCN models and for re-training both the base models. We gradually included past data from 2019 to 2021 in a step-by-step manner with increments of 3 months for retraining the base models and deriving an adjacency matrix in the case of GCNs (for example using the data from 01 October 2021 for re-training the base models as a first step and then adding data in the increments of 3 months i.e., data from 01 July 2021 onwards for retraining in the next step). At each step, we evaluated the Transfer Learned models using data from 2022 as a test set to assess their performance on unseen data. We have also calculated the hamming distance of the adjacency matrix in each step from that of the adjacency matrix derived with just 3 months data.

It is observed that (Fig. 11):

- Both the Transfer Learned models had significantly better performance when more data was added for re-training the base models as expected.
- We have also done a paired t-test to find if the observed performance difference between the two models is statistically significant. It's observed the better performance of GCN-LSTM is statistically significant for both the pollutants - NO_2 (P-value: 0.011) and PM_{10} (*P*-value: 2.897e-07).
- It is observed as more and more data is added Hybrid 2DCNN-LSTM model's performance is found be improving faster than GCN-LSTM with both pollutant values. W.r.t to NO_2predictions, the performance difference between the models became insignificant when data from 2019–2021 is added for training.
- Hamming distance of Granger causality-based adjacency matrix at each step from the adjacency matrix derived with just 3 months of data shows the adjacency matrix starts converging from 12 months of data.

Our analysis also observed that there was no statistically significant prediction improvement while re-training the GCN-LSTM base model with the whole dataset (data from 2019 to 2021) using the adjacency matrix derived with the whole dataset and the adjacency matrix derived with just the previous year's data (data from 2021) with respect to both the pollutants considered here.

Execution time

We have compared the model execution speeds and the range of number of epochs required to converge a solution, with an early stopping criteria for the validation loss (with the set patience of 5 epochs where there is no improvement) with NVIDIA Tesla T4 GPU provided by Google Colab as the computational resource. It's observed hybrid GCN-LSTM models converge into a solution with half the execution required by 2DCNN-LSTMs (Table 1).

We have also checked computing time (without using any specific hardwarde accelerators like GPUs) required for GCN-LSTM to derive adjancency matrix using Granger Causlaity for the whole transfer learning dataset with that from 1 year of past data (as it's found the adjacency matrix derived using Granger Causality converges with just 1 year of past data as shown in Section 7.5) (Table 2).

Key findings

The key findings can be summarised as follows:

- Importance of spatial patterns: Extracting spatial patterns, either through graph convolutional structures in the case of GCN-LSTM or through convolutional structures in the case of 2DCNN-LSTM, is significantly more beneficial than solely relying on capturing temporal patterns using stacked LSTM or Transformer models.
- Capturing prior causal information: The proposed hybrid GCN-LSTM model with the Granger Causality-based adjacency matrix is found to be effective in capturing

Fig. 7 Edges of the adjacency matrix affecting the predicted pollutants of base model vs. transfer learned model (Abbreviations are shown in the Table 1, Appendix 1)

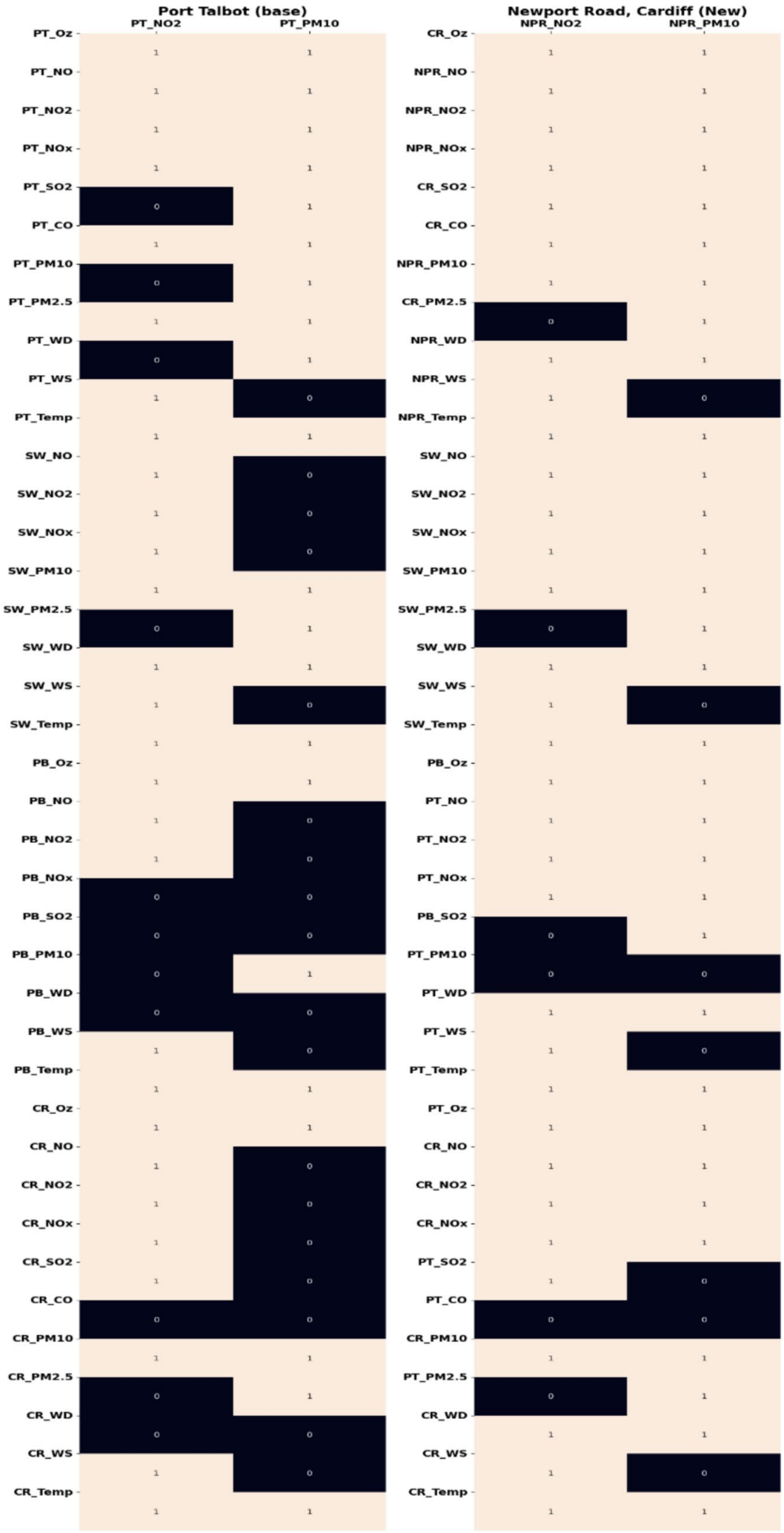
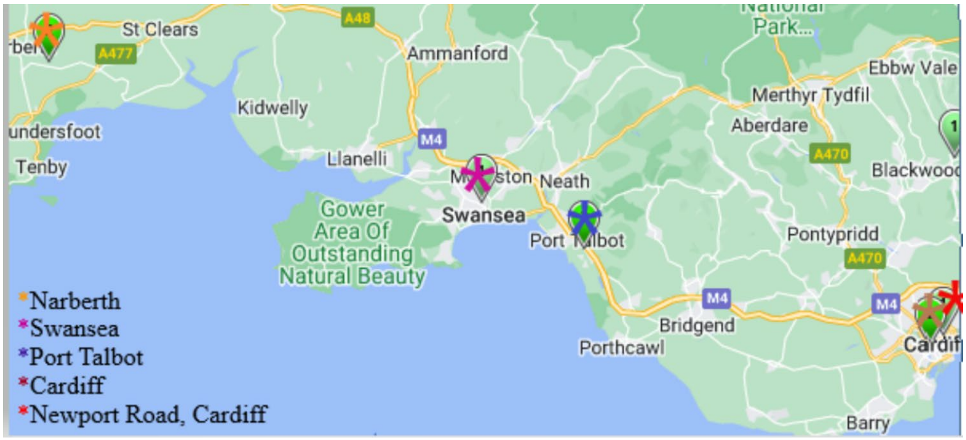


Fig. 8 Geographical locations of monitoring stations considered



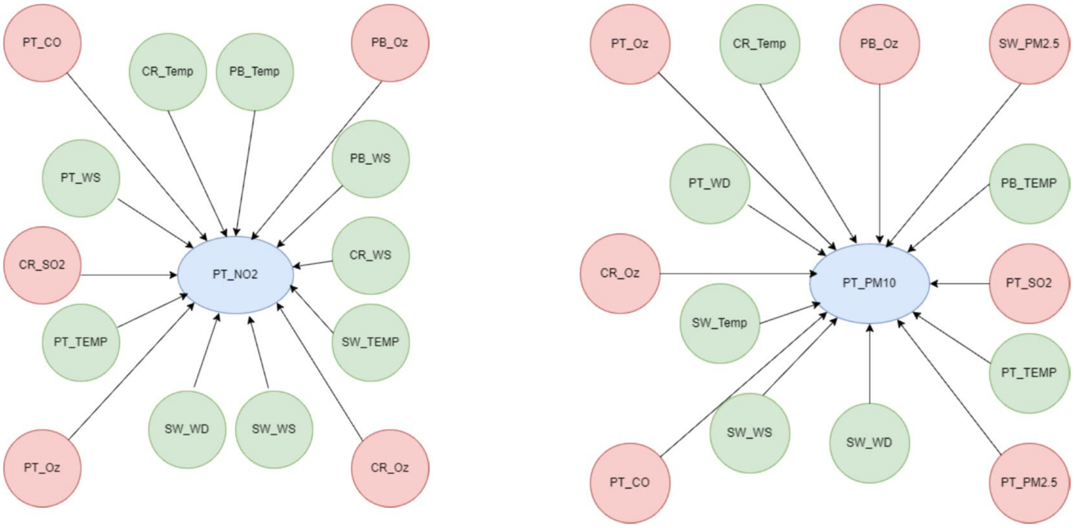


Fig. 9 Causal relations to NO_2 and PM_{10} in Port Talbot

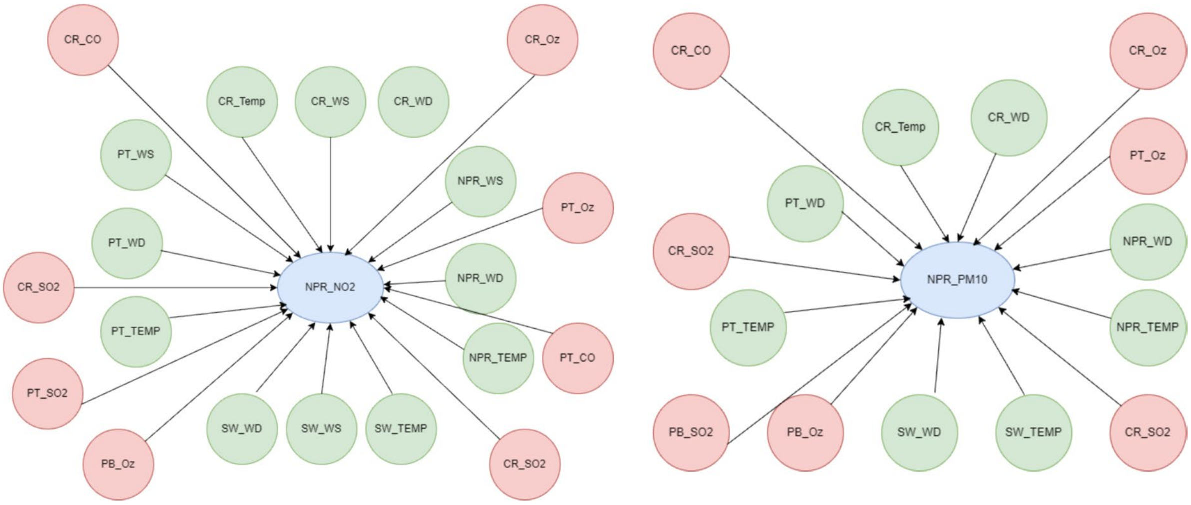


Fig. 10 Causal relations to NO_2 and PM_{10} in Newport Road, Cardiff

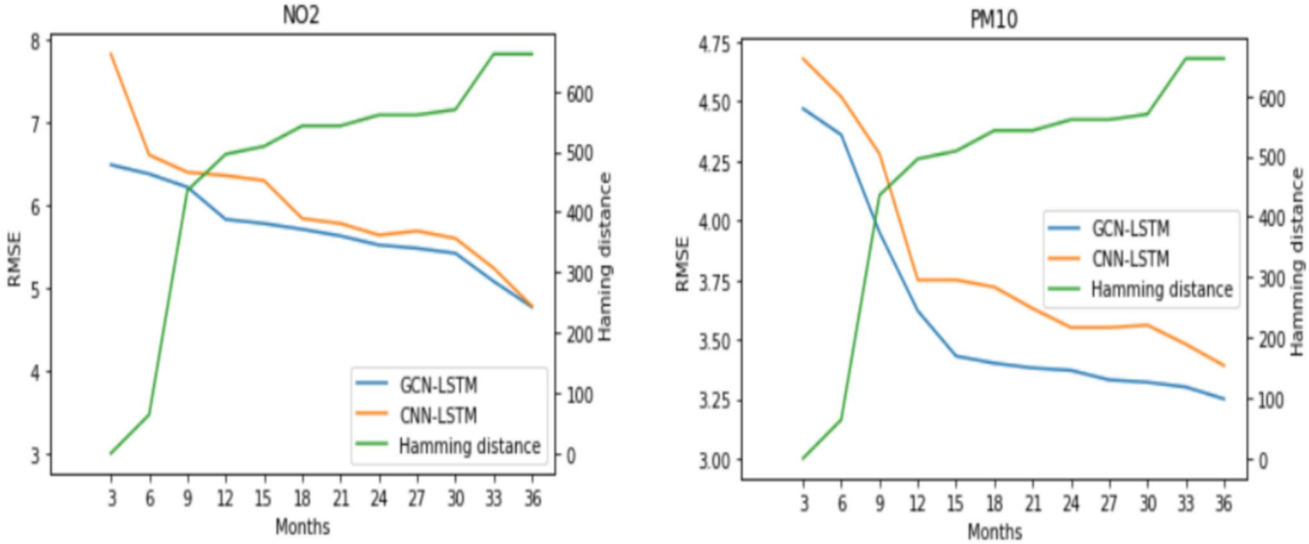


Fig. 11 Performance comparison of transfer learned models when more data is added step by step

Table 1 Execution time comparisons

Model	Training speed	Number of epochs for convergence
2DCNN-LSTM (Base model)	20ms	40–50
GCN-LSTM (Base model)	9ms	40–50
2DCNN-LSTM (Transfer Learning)	9ms	30–40
GCN-LSTM (Transfer Learning)	25ms	20–25

Table 2 Granger causality adjacency matrix computing time

Model	Training speed	Number of edges derived
Full past data (3 years data from 2019 to 2021)	20ms	1180
Past year data (data from 2021)	9ms	1106

prior causal information. GCNs are specifically designed to operate on graph-structured data and utilise the graph connectivity information. GCNs leverage the adjacency matrix to perform graph convolutions, enabling them to propagate information and capture relationships between nodes. This makes them suitable for incorporating prior causal information when multivariate time series data represented as graphs. In a low data environment, where the available historical data may be limited, leveraging this prior causal information is found to be particularly valuable since it makes the learning process faster and/or easier by automatically filtering out irrelevant links, whereas the CNN tends to learn this over time but it needs more data and training epochs for achieving that. The convolutional layers use a predefined kernel to scan over the input data and extract local features by aggregating information from nearby spatial locations but it doesn't have a mechanism to incorporate the prior causal dependencies encoded like in the adjacency matrix of GCN before training the model.

- Improved performance of transfer learned models: The transfer learned models leveraging the learned features, such as daily/weekly patterns and the effects of factors like wind, rain, wind direction, or temperature, to capture the complexities and disruptions in generic pollutant plumes, outperformed their counterpart models trained from the scratch.
- Significant superiority of Granger Causality-based adjacency matrix: Among the Transfer Learned models, the one utilizing the Adjacency Matrix derived using Granger Causality for the Cardiff-Newport Road site stood out as significantly more effective. The analysis suggests that this model outperformed the other two models in capturing the temporal and spatial causal relationships between different locations. The significance level of $p < 0.001$ indicates a strong statistical difference.

Discussion

Implementing deterministic predictive models in untested geographic areas poses significant challenges, especially in regions with distinctive pollutant characteristics and limited emission data. On the other hand, deep learning models with a data-centric approach are often more adaptable and capable of learning from historical data and adjusting quickly to new situations, provided reliable sensor data and robust real-time data collection are in place. The experimental outcomes demonstrated that spatio-temporal transfer learning implemented in this study by employing the GCN-LSTM model with a unique Granger Causality adjacency matrix was highly efficient. The results showed that these techniques can successfully transfer models trained in one area to another, even with limited additional data available. Significant differences in air quality data exist across regions due to variations in industrial activity, traffic patterns, population density, and vegetation. A model that has been trained or tested in one area may not be applicable to other regions. The model's effectiveness will be skewed towards identifying distinct pollution sources (such as a particular industrial type or prevailing transportation mode) if the area has distinctive pollution sources. Based on the current modelling experiments and results, though they are confined to the South Wales region in the UK, the proposed approach can be assessed in other geographical locations with diverse climatic conditions in future experiments to ensure the results are replicable in varied geographic and climatic conditions.

Conclusion

Combining precise short-term forecasting with episode-specific air quality management is expected to enable the proactive deployment of mitigation strategies to prevent peak episodes from happening. They can also help develop improved urban air quality information and forecasting systems, enhancing the capabilities of local authorities to successfully predict and describe air contamination episodes in advance on a day-to-day basis. New air quality monitoring stations are typically positioned in areas of concern to mitigate the effects of frequent peak episodes or to track overall background pollutant levels. In such cases, employing transfer learning techniques for air quality forecasting from nearby monitoring stations results in quick insights into future air quality readings with a relatively small amount of additional data from the target monitoring station. Experimental results from this study show that transfer learning effectively improved multi-variate time series prediction performance, with the transferred learned models outperforming models trained from scratch in a low data environment. This study compared the prediction performance of five base models (Stacked LSTM, Transformer, Hybrid 2DCNN-LSTM, GCN-Transformer and GCN-LSTM) on forecasting NO_2 and PM_{10} pollutant values. The GCN-LSTM, GCN-Transformer and 2DCNN-LSTM models were found to outperform the temporal models - LSTM and Transformer, highlighting the importance of capturing spatial patterns from different monitoring stations in addition to temporal trends when it comes to air quality prediction. Among the transfer learned models, the hybrid GCN-LSTM model with the Granger Causality-based adjacency matrix explicitly derived for the new site was found to be statistically more effective when forecasting with both the pollutants considered. GCNs are specifically designed to operate on graph-structured data and utilise the graph connectivity information. GCNs leverage the adjacency matrix to perform graph convolutions, enabling them to propagate information and capture relationships between nodes, making them suitable for incorporating prior causal information when multivariate time series data is represented as graphs. The results also suggest that prior causal information is essential in deploying a spatial transfer learning paradigm. Deriving an adjacency matrix using Granger Causality as proposed in this study, is found to be good at extracting this prior causal information from a spatio-temporal dataset. It is particularly useful in transferring knowledge in low-data environments, where Deep Learning models struggle to capture useful patterns and trends, and the predictive capabilities of these models are compromised due to inadequate training data.

Appendix 1

Table 3 Abbreviations used in the adjacency matrix binary heatmap

PT_NO	PortTalbot_Nitric oxide
PT_NO2	PortTalbot_Nitrogen dioxide
PT_NOx	PortTalbot_Nitrogen oxides as nitrogen dioxide
PT_PM10	PortTalbot_PM10 particulate matter
PT_PM2.5	PortTalbot_PM2.5 particulate matter
PT_WD	PortTalbot_Modelled Wind Direction
PT_WS	PortTalbot_Modelled Wind Speed
PT_Temp	PortTalbot_Modelled Temperature
SW_NO	Swansea_Nitric oxide
SW_NO2	Swansea_Nitrogen dioxide
SW_NOx	Swansea_Nitrogen oxides as nitrogen dioxide
SW_PM10	Swansea_PM10 particulate matter
SW_PM2.5	Swansea_PM2.5 particulate matter
SW_WD	Swansea_Modelled Wind Direction
SW_WS	Swansea_Modelled Wind Speed
SW_Temp	Swansea_Modelled Temperature
PB_Oz	Pembroke_Ozone
PB_NO	Pembroke_Nitric oxide
PB_NO2	Pembroke_Nitrogen dioxide
PB_NOx	Pembroke_Nitrogen oxides as nitrogen dioxide
PB_SO2	Pembroke_Sulphur dioxide
PB_PM10	Pembroke_PM10 particulate matter
PB_WD	Pembroke_Modelled Wind Direction
PB_WS	Pembroke_Modelled Wind Speed
PB_Temp	Pembroke_Modelled Temperature
CR_NO	Cardiff_Nitric oxide
CR_NO2	Cardiff_Nitrogen dioxide
CR_NOx	Cardiff_Nitrogen oxides as Nitrogen Dioxide
CR_SO2	Cardiff_Sulphur dioxide
CR_CO	Cardiff_Carbon monoxide

Table 3 (continued)

CR_PM10	Cardiff_PM10 particulate matter
CR_WD	Cardiff_Modelled Wind Direction
CR_WS	Cardiff_Modelled Wind Speed
CR_Temp	Cardiff_Modelled Temperature
NPR_NO	Newport Road_Nitric oxide
NPR_NO2	Newport Road_Nitrogen dioxide
NPR_NOx	Newport Road_Nitrogen oxides as nitrogen dioxide
NPR_PM10	Newport Road_PM10 Particulate Matter
NPR_WD	Newport Road_Modelled Wind Direction
NPR_WS	Newport Road_Modelled Wind Speed
NPR_Temp	Newport Road_Modelled Temperature

Appendix 2

Table 4 Hyperparameter tuning

Model	Hyperparameters	Tuning Range	Chosen optimum value	
Stacked LSTMs	Dropout		[0.2,0.3,0.4]	0.2
	Batch Size		[128,168,336]	168
	Input window size		[24,72]	24
Transformer Attention	Head size		[64,128]	128
	Number of attention heads		[42,84]	42
	Batch Size		[168,336]	168
	Input window size		[24,72]	24
	Dropout		[0.2,0.3,0.4]	0.3
GCN-Transformer Attention	Input window size		[24,72]	24
	Node embedding size		[5,10,20]	10
	Head size		[64,128]	128
	Number of attention heads		[42,84]	42
	Batch Size		[168,336]	168
	Dropout		[0.2,0.3,0.4]	0.3
2DCNN-LSTM (Spatial model)	Dropout		[0.2,0.3,0.4]	0.3
	Batch Size		[168,336]	336
	Input window size		[24,72]	24
GCN-LSTM (Spatial model)	Dropout		[0.2,0.3,0.4]	0.3
	Batch Size		[168,336]	168
	Input window size		[24,72]	24
	Node embedding size		[10,20]	20
	Granger Causality Lag value (K)		[1,6,12,24]	24

Appendix 3 – Input variable plots

Input variables of the base model target monitoring station and transfer learning target monitoring station are plotted below. For brevity only 4 years ranging from 2019 to 2022 is plotted here (Fig. 12 and 13).

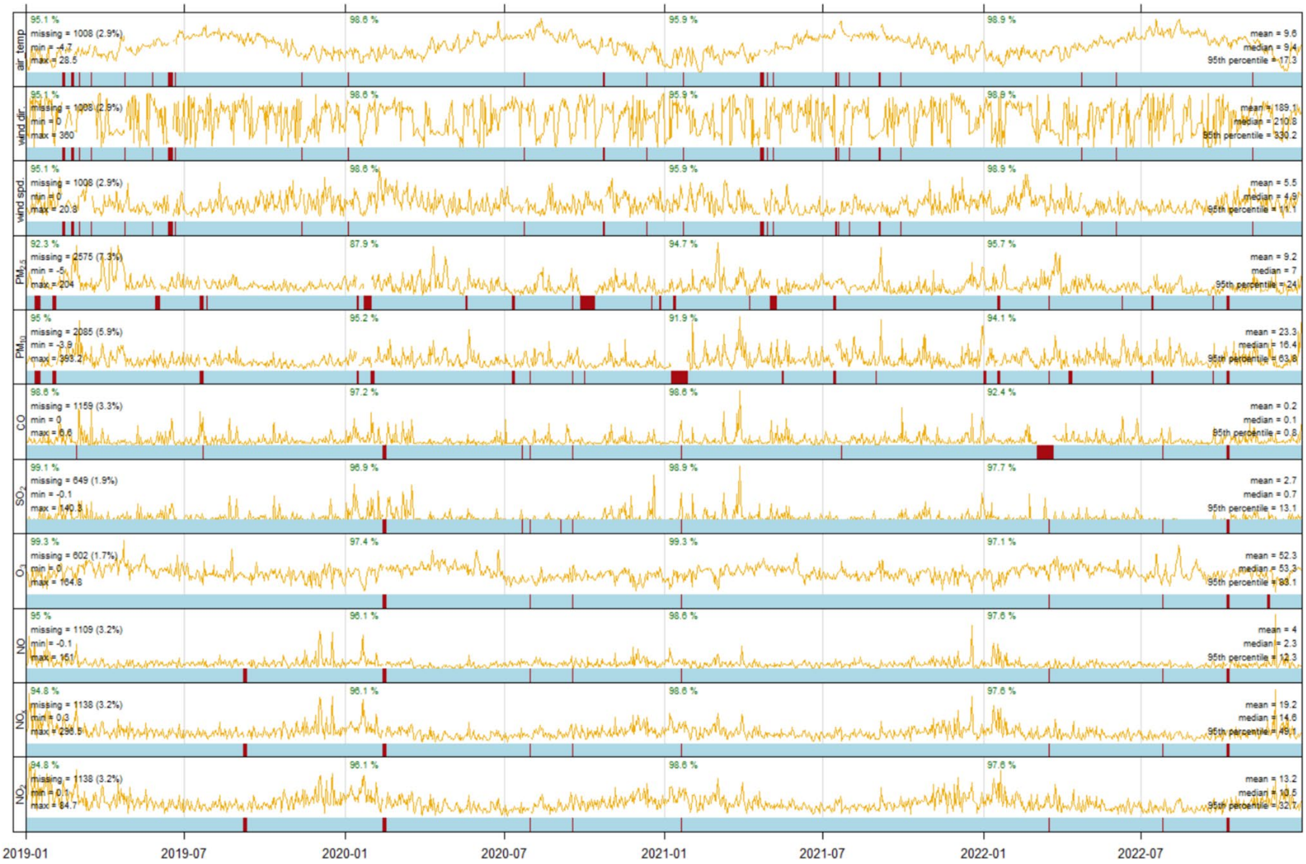


Fig. 12 Input variables from Port Talbot monitoring station

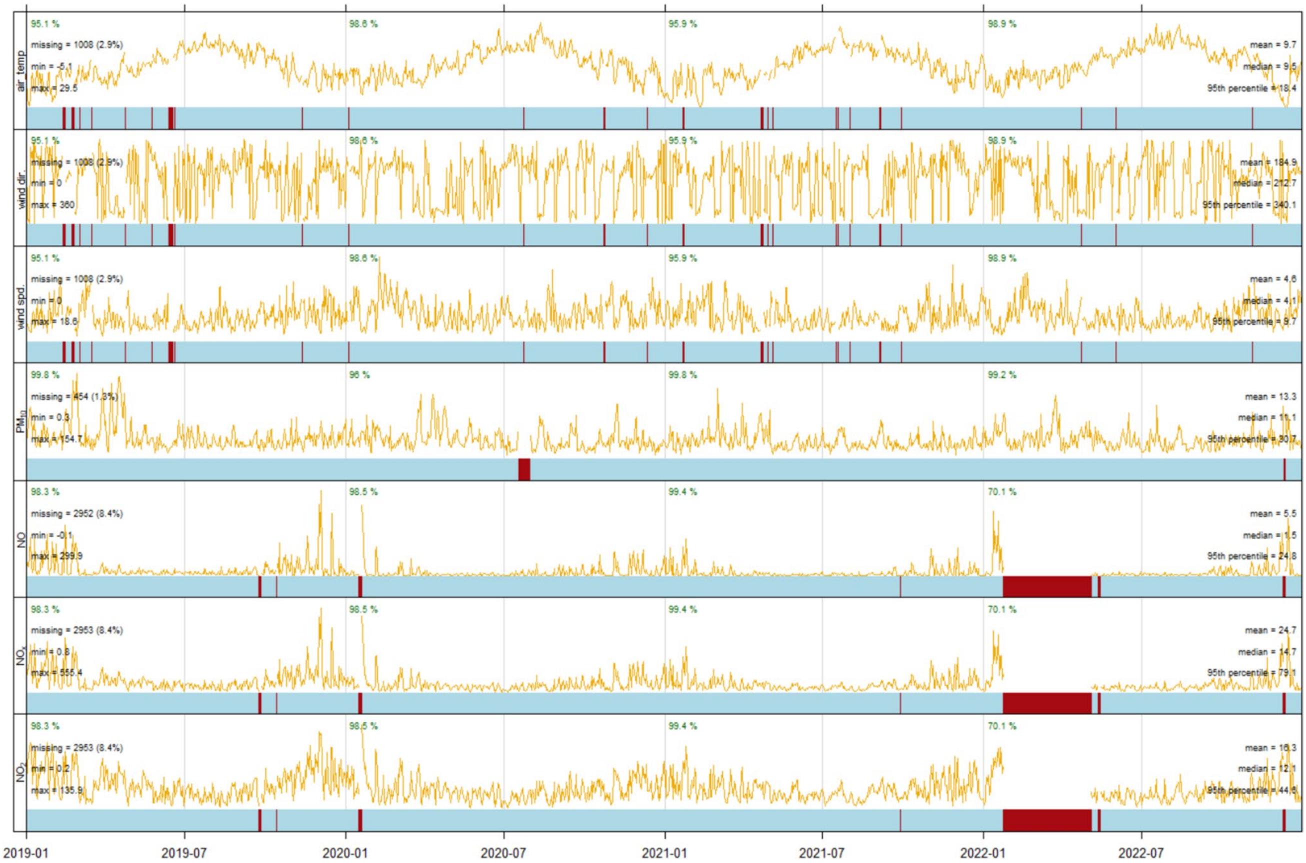


Fig. 13 Input variables from Newport Road, Cardiff monitoring station

Author contributions Sooraj Raj; Responsible for extracting the data running the experiments and capturing the results. Responsible for preservation of the original data on which the paper is based, verification that the figures and conclusions accurately reflect the data collected and that manipulations to images are in accordance with Nature journal guidelines, and minimisation of obstacles to sharing materials, data and algorithms through appropriate planning. Prof. Jim Smith; Responsible for supervising and reviewing the modelling experiments and the contents of the paper. Prof. Enda Hayes; Responsible for guidance on the air quality forecasting background specific to the monitoring stations considered for this study and setting the end-user objectives of the research experiments considered in this study.

Funding No funding has been received for this research from any parties.

Data availability The datasets generated during and/or analysed during the current study are available for public access in the UK Government repository - https://uk-air.defra.gov.uk/data/.

Declarations

Ethics approval and consent to participate Not applicable as the study involves analysis and modelling using public domain data and no individual data is used.

Consent for publication Authors consent to publish this material and no individual data is used in this research for getting separate consent for publication.

Competing interests Not applicable.

References

Bird JJ, Kobylarz J, Faria DR, Ekárt A, Ribeiro EP (2020) Cross-domain MLP and CNN transfer learning for biological signal processing: EEG and EMG. IEEE Access 8:54789–54801

DEFRA UK (2023) Air Pollution in the UK 2022 [online]. Available from: https://uk-air.defra.gov.uk/assets/documents/annualreport/air_pollution_uk_2022_issue_1.pdf

Dhole A, Ambekar I, Gunjan G, Sonawani S (2021) An ensemble approach to multi-source transfer learning for air quality prediction. In 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (pp. 70–77). IEEE. https://doi.org/10.1109/ICCCIS51004.2021.9397138

Duan Z, Xu H, Huang Y, Feng J, Wang Y (2022) Multivariate time series forecasting with transfer entropy graph. Tsinghua Sci Technol 28(1):141–149

Fawaz HI, Forestier G, Weber J, Idoumghar L, Muller PA (2018) Transfer learning for time series classification. In 2018 IEEE international conference on big data (Big Data) (pp. 1367–1376). IEEE. https://doi.org/10.1109/BigData.2018.8621990

Fong IH, Li T, Fong S, Wong RK, Tallon-Ballesteros AJ (2020) Predicting concentration levels of air pollutants by transfer learning and recurrent neural network. Knowl Based Syst 192:105622

Foresti, P., 2006. Testing for Granger Causality between Stock prices and Economic Growth. MPRA Paper No. 2692, online at https://mpra.ub.uni-muenchen.de/2962/1/MPRA_paper_2962.pdf. Accessed 25 Feb 2018

Freeman BS, Taylor G, Gharabaghi B, Thé J (2018) Forecasting air quality time series using deep learning. J Air Waste Manag Assoc 68(8):866–886

Ge L, Wu K, Zeng Y, Chang F, Wang Y, Li S (2021) Multi-scale spatiotemporal graph convolution network for air quality prediction. Appl Intell 51:3491–3505

Hoseinzade E, Haratizadeh S (2019) CNNpred: CNN-based stock market prediction using a diverse set of variables. Expert Syst Appl 129:273–285

Huang CJ, Kuo PH (2018) A deep CNN-LSTM model for particulate matter (PM2. 5) forecasting in smart cities. Sensors 18(7):2220

Hyndman RJ, Athanasopoulos G (2013) 8.1 Stationarity and differencing. Forecasting: Principles and practices, Melbourne, Australia, OTexts.

Jmour N, Zayen S, Abdelkrim A (2018) Convolutional neural networks for image classification. In: 2018 international conference on advanced systems and electric technologies (IC_ASET) (pp. 397–402). IEEE. https://doi.org/10.1109/ASET.2018.8379889

Kim TY, Cho SB (2018) Web traffic anomaly detection using C-LSTM neural networks. Expert Syst Appl 106:66–76

Laptev N, Yu J, Rajagopal R (2018) Reconstruction and regression loss for time-series transfer learning. In: Proceedings of the Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) and the 4th Workshop on the Mining and LEarning from Time Series (MiLeTS), London, UK vol. 20, pp. 1–8

Li X, Peng L, Yao X, Cui S, Hu Y, You C, Chi T (2017) Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environ Pollut 231:997–1004

Liao Q, Zhu M, Wu L, Pan X, Tang X, Wang Z (2020) Deep learning for air quality forecasts: a review. Curr Pollution Rep 6:399–409

Lin CH, Lin YC, Wu YJ, Chung WH, Lee TS (2021) A survey on deep learning-based vehicular communication applications. J Signal Process Syst 93:369–388

Oehmcke S, Zielinski O, Kramer O (2018) Input quality aware convolutional LSTM networks for virtual marine sensors. Neurocomputing 275:2603–2615

Otović E, Njirjak M, Jozinović D, Mauša G, Michelini A, Stajduhar I (2022) Intra-domain and cross-domain transfer learning for time series data—how transferable are the features? Knowl Based Syst 239:107976

Pinto G, Wang Z, Roy A, Hong T, Capozzoli A (2022) Transfer learning for smart buildings: a critical review of algorithms, applications, and future perspectives. Adv Appl Energy 5:100084

Qi Y, Li Q, Karimian H, Liu D (2019) A hybrid model for spatiotemporal forecasting of PM2. 5 based on graph convolutional neural network and long short-term memory. Sci Total Environ 664:1–10

Raj S, Smith J, Hayes E (2022) Exploring deep learning architectures for localised hourly air quality prediction. In: International Conference on Artificial Neural Networks. Cham: Springer International Publishing. pp. 133–144

Reddy V, Yedavalli P, Mohanty S, Nakhat U (2018) Deep air: forecasting air pollution in Beijing, China. Environ Sci, 1564.

Ribani R, Marengoni M (2019) A survey of transfer learning for convolutional neural networks. In: 2019 32nd SIBGRAPI conference on graphics, patterns and images tutorials (SIBGRAPI-T) pp. 47–57. IEEE. https://doi.org/10.1109/SIBGRAPI-T.2019.00010

Rodriguez-Caballero CV, Ventosa-Santaulària D (2014) Granger causality and unit roots. J Stat Econometric Methods 3(1):97–114

Shojaie A, Fox EB (2022) Granger causality: a review and recent advances. Annual Rev Stat Its Application 9:289–319

Soekhoe D, Van Der Putten P, Plaat A (2016) On the impact of data set size in transfer learning using deep neural networks. In: Advances in Intelligent Data Analysis XV: 15th International Symposium, IDA 2016, Stockholm, Sweden, October 13–15, 2016, Proceedings 15 (pp. 50–60). Springer International Publishing. https://doi.org/10.1007/978-3-319-46349-0_5

Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J (2016) Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging 35(5):1299–1312

Ventosa-Santaulària D, Vera-Valdés JE (2008) Granger-causality in the presence of structural breaks. Econ Bull 3(61)

Wen Q, Zhou T, Zhang C, Chen W, Ma Z, Yan J, Sun L (2022) Transformers in time series: a survey. arXiv preprint. https://doi.org/10.48550/arXiv.2202.07125

Yin C, Zhang S, Wang J, Xiong NN (2020) Anomaly detection based on convolutional recurrent autoencoder for IoT time series. IEEE Trans Syst Man Cybernetics: Syst 52(1):112–122

Zeng A, Chen M, Zhang L, Xu Q (2023) Are transformers effective for time series forecasting? In Proceedings of the AAAI conference on artificial intelligence 37(9):11121–11128

Zhang Q, Lam JC, Li VO, Han Y (2020) Deep-AIR: a hybrid CNN-LSTM framework forfine-grained air pollution forecast. arXiv preprint arXiv:2001.11957. https://doi.org/10.1109/ACCESS.2022.3174853

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.