HastGCN: A Hierarchical Attention-based Spatiotemporal Graph Convolution Network for Traffic Incident Impact Forecasting

ABSTRACT

Predicting the impact of the traffic incident is one of the essential research topics in the field of Intelligent Transportation Systems (ITS). However, the problem is very challenging to tackle since the traffic data usually is highly nonlinear and with implicit correlations with the impact of the traffic incidents. The existing traffic data-based studies on forecasting the impact of the traffic incidents are mostly incapable of modeling 1) the spatiotemporal correlations of the traffic data, and 2) the hierarchical topology of the road networks. In this paper, we propose the Hierarchical Attention-based Spatioemporal Graph Convolutional Network (HastGCN) model to solve the traffic incident impact forecasting problem. HastGCN has two major building blocks: the lower-level road encoder and the high-level predictor. At the lower level, we propose a spatiotemporal attention mechanism to model spatiotemporal correlations between the traffic sensors on the same corridor; we propose a ChebNet to infer the dynamic information cascade between the traffic sensors on the same corridor. At the higher level, a spatial attention mechanism is designed to learn the spatial relatedness between each road/corridor; a Graph Convolution Network is proposed to perform the prediction. With such model design, HastGCN is capable of considering the travel directions of the corridors by leveraging the spatial structures of the road networks, which is by far the one and only. Extensive experiments and comparisons to other baseline models on three real-work datasets from the CalTrans Performance Measurement System (PeMS) justify the efficacy of our model.

1 INTRODUCTION

The studies of early detecting the traffic incidents and estimating the impact of the non-recurrent congestions caused by traffic incidents have become increasingly important research topics due to the significant social and economic losses generated. A one-minute reduction on congestion duration produces a 65 US dollars gain per incident [1]. Although non-recurrent congestion is hard to predict due to its nature of randomness, the studies on impact and duration of the traffic incidents are still one of the major focuses for the traffic operators. The vast deployment of transportation traffic speed sensors and Traffic Incident Management Systems (TIMS) make the traffic speed data and traffic incident records ubiquitously accessible for the transportation operators. An efficient multi-task learning model can be implemented with an abundance of traffic data sources to provide an accurate prediction of incident duration.

The impact of a traffic incident in this paper is quantified by the time elapsed from the incident occurrence until all evidence of the incident has been removed from the incident scene. The life cycle of a traffic incident is split into five stages: Detection, Verification, Response, Clearance, and Recovery [4, 11]. Figure 2 shows the life cycle of a traffic incident. To accurately estimate the impact of a traffic incident in its early stages, there are several major weaknesses for the current research approaches in the research field's intelligent transportation systems. First, the existing methods ignore to extract high-level spatiotemporal features. Although the five-stage life cycle separation can serve as temporal features for the purposes of traffic management, such temporal features cannot be considered in traffic incident impact prediction tasks. It is critical to identify both spatial and temporal features in the early stages of the incident forming higher-level spatiotemporal features that can be used as a better indicator for predicting the incident impact. Second, it is difficult to predict the influence of the incident. In the research field of Traffic Incident Management (TIM), one of the most important tasks is to estimate the impact of the traffic incident in terms of its temporal duration the early stages. However, the performances of the conventional time series based methods are limited by their incapability of considering the spatiotemporal correlations between the traffic sensors, road corridors, and the temporal periods. Third, the current models ignore the spatial connectivity of the road networks, the traffic congestion cascades within the road network. As a consequence, the traffic patterns of incidents in their early stages are similar when the traffic incidents are topologically closer from the perspective of the road networks. Traffic incidents that are spatiotemporally closer should share more similar traffic speed patterns. Although some of the existing works tackling the traffic flow forecasting problems have proposed some graph-based method to model the spatial connectivity of the traffic sensors, most of their methods overly simplified the spatial correlation with naïve distance-based strategies. More complicated road network properties, such as the corridors' topologies and the travel directions of the corridors, are rarely considered [9].

The occurrence of a traffic incident will cause influences to nearby locations in both spatial and temporal dimensions. The records of the traffic sensors will reflect such influences. The traffic features at neighboring locations and time points are not isolated but dynamically correlated. Figure 1 demonstrates the spatiotemporal impact of traffic incidents. Therefore, the crux of current

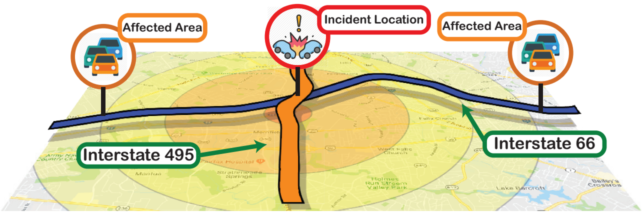

Figure 1: The spatial and temporal influence of a traffic incident. The orange and blue lines are Interstate highways. The traffic congestion of a traffic incident may be conveyed by the corridors linked to the location of the incident.

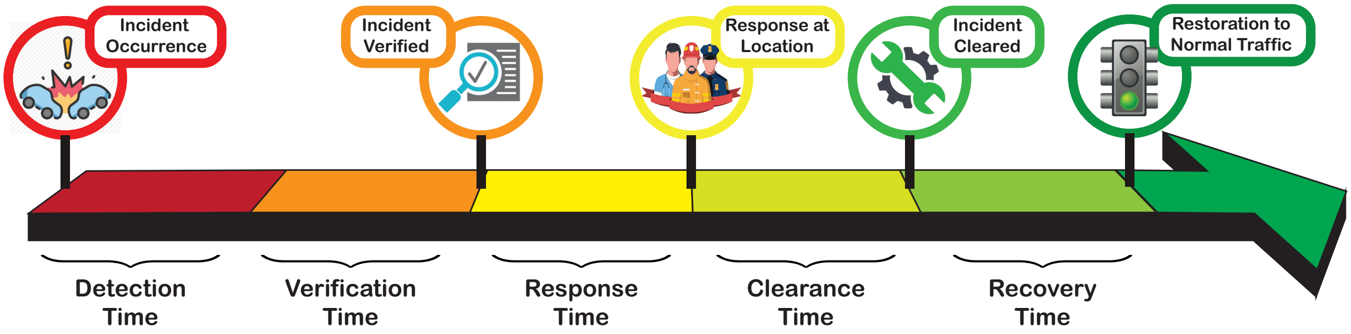

Figure 2: From the perspectives of traffic management and transportation operations, the life cycle of a traffic incident is separated into five stages: Detection, Verification, Response, Clearance, and Recovery

practice in predicting the impact of traffic incidents is to extract the spatiotemporal correlations effectively.

Most existing methods are mostly infeasible to solve these challenges. Current feature learning methods such as ℓ_1-norm regularized methods such as Lasso [15] have properties in terms of feature selection. However, strong assumptions on the design matrix are required [20]. Zhan et al. [17] propose an M5P tree algorithm to predict traffic incidents' clearance time based on geometric and traffic features. Feature learning algorithms for biomarker identification [21] and social event indicators [18] are proved to be effective while finding higher-level features. However, most of them focus on learning important feature sets from attributes and does not apply to our encountered problem due to expensive computation. In these studies, they considered the duration of an incident to quantify the impact. However, their quantification strategies are designed to capture the one-time impact of the incident, instead of the time-varying nature of impact at different locations. In the recent decade, deep learning methods have been proven to be effective in a wide range of high-dimensional spatiotemporal datasets. For example, the application of convolutional neural networks (CNN) for computer vision tasks; and graph convolution networks (GCN) are utilized in spatial data inference. However, these methods cannot model the topology of the road networks and take the directions of the road corridors into account.

To address these challenges, we propose the **H**ierarchical **A**ttention-based **S**patio**t**emporal **G**raph **C**onvolution **N**etwork (HastGCN) based on spectral graph convolution with a multi-attention mechanism. A novel hierarchical structure of the Graph Convolution Network (GCN) is proposed to leverage the directions of each corridor. Such a design is intuitive because the traffic patterns only transmit along the corridors. For example, two traffic sensors deployed closely on a highway but on the opposite directions may record significantly different traffic patterns. The main contributions of this paper are summarized as follows:

- **Proposing a novel hierarchical structure for spatiotemporal graph convolution networks**. We leverage the topology of the road network to model the correlations of the traffic data on multiple levels. Specifically, the lower level is represented by the spatial relatedness between the traffic sensors on one corridor; the upper level is represented by the spatial connections between the corridors.
- **Modeling the directions of corridors with the hierarchical structure of the graph convolution network**. The hierarchy of the graph convolution network is constructed based on the separation of the corridors. With this design, only the traffic sensors on the same travel direction will be considered, which enables the proposed network to consider the travel directions.
- **Developing a spatiotemporal attention mechanism on multiple levels**. The HastGCN model is capable of capturing the dynamical dependencies between spatiotemporal features. The spatial and temporal attention mechanisms on the lower level of the hierarchy tackle the intervening dependency of the traffic sensors; the spatial attention on the upper level learns the dependencies between the corridors.
- **Extensive experiments and make comparisons to validate the effectiveness and efficiency of the proposed model**: We compare our proposed graph convolutional neural network with various baseline methods. We conduct experiments on three real-world traffic datasets. Both conventional methods and deep learning-based methods for traffic forecasting are selected for comparisons. Evaluations of various metrics study analyses are presented, illustrating the effectiveness of the proposed model.

The rest of our paper is structured as follows. In Section 2, we describe the problem setup of our work. In Sections 2 and 3, we present a detailed discussion of our proposed *HastGCN* model for predicting the impact of traffic incidents, and its solution for parameter learning. In Section 4, extensive experiment evaluations and comparisons are presented. In the last section, we discuss our conclusion.

2 PROBLEM STATEMENT

In this section, we first provide the mathematical definition of the traffic incident impact forecasting problem. Next, we describe two-step building blocks for the problem-solving process: 1) the hierarchical traffic connectivity network and 2) traffic incident impact forecasting. They collaborate to inference the future impact of the traffic incidents by modeling the multi-level spatial-temporal connectivity of the traffic network.

2.1 Hierarchical Traffic Connectivity Network

In this study, we consider a hierarchical structure for the road traffic network. This hierarchical structure is proposed because 1) the overall traffic network is constructed by a set of arterial roads such as interstates or major roads, 2) several traffic sensors are deployed on each arterial road.

Definition I: *Road level traffic network.* The original road level traffic network $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$ consists of a set of roads $\mathcal{R}$, where the $\mathcal{V}^*$ represents the set of road intersections, and $\mathcal{E}^*$ represents the set of road segments. The traffic network $\mathcal{G}^*$ represents the real topology of the road network connections. Based on the original road level traffic network $\mathcal{G}^*$, we construct a dual graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A}')$, where the $\mathcal{V}$ in the dual graph is a set of nodes $|\mathcal{V}| = N_r$, and the nodes set in the dual graph represents the roads in the road level traffic network $\mathcal{G}^*$; $\mathcal{E}$ is the set of edges in the dual graph, an edge $e \in \mathcal{E}$ linking nodes v_1 and v_2 is presented when the two roads r_1 and r_2 represented by v_1 and v_2 touches or intersects with each other; $A \in \mathbb{R}^{N_r \times N_r}$ denotes the adjacency matrix of the dual graph $\mathcal{G}$. Each node v_k on the dual traffic graph $\mathcal{G}$ has an attribute information that is generated from the lower-level road traffic network encoding.

Definition II: *Sensor level traffic network.* For each road $r \in \mathcal{R}$, a traffic sensor $v_{r_k} \in \mathcal{V}_r$ is deployed to record traffic features such as travel speed, volume and road occupancies, where $\mathcal{V}_r$ is the set of traffic sensors. Based on the connectivity of the road segments and the directions of the roads, we define a lower level traffic sensor connection network for a specific road $r \in \mathcal{R}$: $\mathcal{G}_r = (\mathcal{V}_r, \mathcal{E}_r, \mathbf{A}_r)$ where $\mathcal{V}_r = \{v_{r_1}, v_{r_2}, \dots, v_{r_n}\}$, and v_{r_k} is the k-th traffic sensor node deployed on road r, $\mathcal{E}_r$ is the set of edges links the traffic sensors, $\mathbf{A}_r \in \mathbb{R}^{n_r \times n_r}$ is the adjacency matrix representing the connectivity of the traffic sensors on road r.

2.2 Traffic Incident Impact Forecasting

Suppose F is a the set of traffic features that can be recorded by the traffic sensors, we use $x_\tau^{f,n} \in \mathbb{R}$ to denote the value of the f-th feature of node n at time τ, and $\mathbf{x}_\tau^n \in \mathbb{R}^F$ is the traffic features for all features of node n at time τ. $\mathbf{X}_\tau = (\mathbf{x}_\tau^1, \mathbf{x}_\tau^2, ..., \mathbf{x}_\tau^N)^T \in \mathbb{R}^{N \times F}$ is the traffic features for all the traffic sensors at time τ.

Assume that we are given a collection of traffic incidents Φ from the traffic incident management system. For each traffic incident $\phi \in \Phi$ verified at time τ_v, we define the following temporal features in *detection time* and *early verification time.* Suppose the verification time of the traffic incident is τ_v, we define two important time periods **respond time** (*time between incident occurrence τ_o and incident verification time τ_v*) and **early verification time** (*a short period after the traffic incident verification time τ_v*) for feature construction. Both time periods temporal traffic features for the traffic incident ϕ are defined as:

Definition III: *Temporal features in respond time*: the previous p traffic feature readings before time τ_v: $\boldsymbol{\mathcal{X}}_\phi^- = \{\mathbf{x}_{\tau_v - p}, \mathbf{x}_{\tau_v - p+1}, ..., \mathbf{x}_{\tau_v - 1}\}$.

Definition IV: *Temporal features in early verification time*: the succeeding q traffic feature readings after time τ_v: $\boldsymbol{\mathcal{X}}_\phi^+ = \{\mathbf{x}_{\tau_v}, \mathbf{x}_{\tau_v+1}, ..., \mathbf{x}_{\tau_v+q}\}$, where $\mathbf{x}_\tau$ denotes traffic feature values at time τ for all traffic sensors, $\boldsymbol{\mathcal{X}}_\phi = \boldsymbol{\mathcal{X}}_\phi^- + \boldsymbol{\mathcal{X}}_\phi^+$, and $\boldsymbol{\mathcal{X}}_\phi \in \mathbb{R}^{(p+q) \times N \times F}$. $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{|\Phi| \times (p+q) \times N \times F}$ denotes the traffic features values of all the traffic incidents $\phi \in \Phi$. In addition, we define $y_\phi \in \mathbb{R}$ to be the target traffic incident duration in minutes.

Given the input data $\boldsymbol{\mathcal{X}}$, and $\mathrm{Y} = (y_1, y_2, ..., y_{|\Phi|})^T \in \mathbb{R}^{|\Phi|}$ as described above, our problem is now formatted as follows: for a target traffic incident ϕ, we take the corresponding traffic feature values $\boldsymbol{\mathcal{X}}_\phi$ as the observations, can we predict the future impact of the specific traffic incident y_ϕ in terms of it's duration in minutes? The problem is then formulated as solving the mapping:

$$\mathcal{F}(\boldsymbol{\mathcal{X}}) \rightarrow \mathrm{Y} \tag{1}$$

3 HASTGCN MODEL

In this section, we present our proposed model, HastGCN, which tackles the problem of forecasting the impact of traffic incidents. First, we discuss the construction process of the hierarchical graph for our proposed HastGCN model. We also show an overview of the design of the proposed framework. Then we detail the architecture of the graph convolution networks used on the two major building blocks of HastGCN: *Road Encoder*, and the *Predictor* of our model. These two building blocks work consecutively extracting both local (traffic sensor-level) and global (corridor-level) spatiotemporal features. Next, we introduce the attention mechanisms which play an important role in capturing the spatiotemporal correlations among traffic sensors, city corridors, and temporal time points for generating the predicted impact of traffic incidents. Finally, we describe the training procedure, which unifies the distinct components of the proposed framework.

3.1 Hierarchical Graph Construction

Two hierarchies of graph structures are constructed: 1) the sensor-level (lower) subgraphs, and 2) the road-level (higher) graph. The sensor-level subgraphs capture the spatiotemporal correlations between the traffic sensors on one specific corridor, which provide local traffic data perception. At the same time, the road-level graph is constructed based on the spatial connectivity between the corridors, which provides a global perspective.

3.1.1 *Road-Level Graph.*

According to Definition I, we describe the higher-level road graph with a dual graph representation $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A}')$, $\mathbf{A}'$ denotes the adjacency matrix of the dual graph, and $\mathcal{V}$ is the set of vertices in the dual graph which represents the set of roads $\mathcal{R}$. For road r_1 and r_2 which are represented by vertices v_1 and v_2 respectively in $\mathcal{G}$, the adjacency matrix $\mathbf{A}' \in \mathbb{R}^{N \times N}$ of the dual graph can be formulated as:

$$\mathbf{A}'_{i,j} = \begin{cases} 1, & \text{if condition } \zeta_s(r_1, r_2) \text{ holds.} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

where the ζ_s is a spatial condition that calculates the spatial relation between the input targets r_1 and r_2. In the HastGCN model, we select the ζ_s condition as "*ST_intersect*"[1] or "*ST_touches*"[2]. Such spatial condition will return True if r_1 and r_2 intersects or touches with each other.

3.1.2 *Sensor-Level Subgraphs.*

According to Definition II, we introduce the lower-level traffic sensor subgraph's representation $\mathcal{G}_r = (\mathcal{V}_r, \mathcal{E}_r, \mathbf{A}_r)$, where $\mathcal{G}_r$ denotes the subgraph of the traffic sensors $\mathcal{V}_r$ deployed on the road $r \in \mathcal{R}$, and $\mathbf{A}_r \in \mathbb{R}^{n \times n}$ is the

[1] https://postgis.net/docs/ST_Intersects.html

[2] https://postgis.net/docs/ST_Touches.html

adjacency matrix for the traffic sensors on the road r. $\mathbf{A}_r$ of the specific road r can be formulated as:

$$\mathbf{A}_{r_{i,j}} = \begin{cases} exp(-\frac{d_{ij}^2}{\sigma^2}), & i \neq j \text{ and } exp(-\frac{d_{ij}^2}{\sigma^2}) \geq \epsilon. \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

where $\mathbf{A}_{r_{i,j}}$ is represented by the weight of edge which is decided by d_{ij} (the distance between traffic sensor i and j). σ^2 and ϵ are thresholds to control the distribution and sparsity of matrix $\mathbf{A}_r$. This weighted distance calculation is applied in several previous traffic speed prediction works [2, 16]. Figure 3 shows the construction of the hierarchical graph.

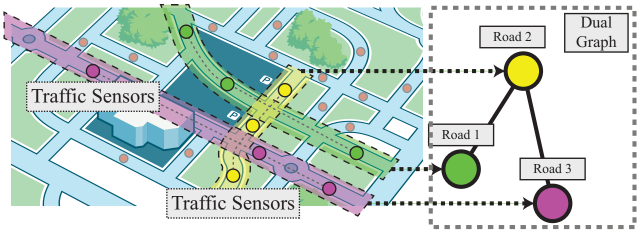

Figure 3: The Construction of the Hierarchical Graph. The road map on the left-hand side represents the traffic sensors. The green, yellow, and purple nodes represent the traffic sensors on the green, yellow, and purple roads, respectively. The traffic sensors on the same road construct a lower-level graph. The dual graph representation on the right-hand side represents the higher-level road connectivity graph.

3.2 Model Overview

Figure 4 presents the overall framework of the HastGCN model proposed in this paper. It consists of a two-level hierarchical structure of attention-based Graph Convolution Networks. The low-level GCNs are designed to capture the spatiotemporal correlations between the traffic sensors deployed on the same corridor, which provide the HastGCN model a local vision. The upper-level GCN is proposed to learn the spatial relationships between the city corridors, which provides a global perspective.

Assume that for the traffic incident $\phi \in \Phi$, the verification time for the incident is τ_v. According to the verification time of the traffic incident, we assign the early stage traffic features data as: $\boldsymbol{\mathcal{X}}_\phi = \boldsymbol{\mathcal{X}}_\phi^- + \boldsymbol{\mathcal{X}}_\phi^+$ where $\boldsymbol{\mathcal{X}}_\phi^-$ and $\boldsymbol{\mathcal{X}}_\phi^+$ are defined in Definitions III and IV respectively. Specifically, $\boldsymbol{\mathcal{X}}_\phi = \{\mathbf{x}_{\tau_v - p}, \mathbf{x}_{\tau_v - p+1}, ..., \mathbf{x}_{\tau_v + q}\} \in \mathbb{R}^{(p+q)\times N \times F}$. The overall goal of HastGCN is given the early stage traffic data $\boldsymbol{\mathcal{X}}_\phi$ to predict the impact y_ϕ of the traffic incident ϕ.

In the overview of proposed model depicted in Figure 4, the module blocks that handle input data tensor on the left-hand-side are called *Road Encoder* and the module block that makes predictions is called *Predictor*. For the Road Encoder module, the graph convolution is performed after the traffic sensor-level spatial attention layer, the output of the graph convolution is aggregated and then fed into the temporal attention layer. A linear transformation with ReLU activation is applied to generate the road encoder features:

$$F_s(\mathbf{h}_1) = w_{r_1}\text{ReLU}(w_{r_2}\mathbf{h} + b_{r_2}) + b_{r_1} \quad (4)$$

where $w_{r_1}, w_{r_2}, b_{r_1}, b_{r_2}$ are learnable parameters, $\mathbf{h}_1$ is the output from the temporal attention layer.

The input of the Predictor module is the encoded road features and the adjacency matrix of the roads, a road-level spatial attention layer is applied. A road-level graph convolution with kernel size equal to 1 is applied.

3.3 Spatial Attention

The HastGCN applies a spatial attention mechanism to model the spatial correlations of the traffic network dynamically. The spatial attention mechanisms are applied in both Road Encoder and Predictor modules of the proposed model.

3.3.1 Spatial Attention in Road Encoder. In order to capture the dynamic relations between the traffic sensor on a specific road, we implement the traffic sensor attention mechanism. This attention mechanism in the spatial dimension is proposed to model the traffic flow cascade travel along the target corridors. Based on the fact that the spatial connection relations of the traffic sensors and the road corridors are similar, we utilize the same spatial attention mechanism on both Road Encoder and Predictor modules of the model. The output of the attention layer is calculated as:

$$\boldsymbol{\alpha}_\phi = \boldsymbol{\lambda}_s \cdot \sigma((\boldsymbol{\mathcal{X}}_\phi \boldsymbol{\lambda}_1)\boldsymbol{\lambda}_2(\boldsymbol{\lambda}_3 \boldsymbol{\mathcal{X}}_\phi)^T + \mathbf{b}_s) \quad (5)$$

$$\boldsymbol{\alpha}'_{\phi_{ij}} = \frac{\exp(\boldsymbol{\alpha}_{ij})}{\sum_N^{j=1} \exp(\boldsymbol{\alpha}_{ij})} \quad (6)$$

$$\mathbf{c}^{\langle i \rangle} = \sum_j \boldsymbol{\alpha}'_{ij} \mathbf{o}^{\langle j \rangle} \quad (7)$$

where $\mathbf{c}^{\langle i \rangle}$ is the output of the attention layer in the Road Encoder module, $\boldsymbol{\alpha}'_{ij}$ is the attention weights, $\boldsymbol{\mathcal{X}}_\phi = (\mathbf{X}_{\phi_1}, \mathbf{X}_{\phi_2}, ..., \mathbf{X}_{\phi_{n_r}}) \in \mathbb{R}^{n_r \times F \times (p+q)}$ is the input of the Road Encoder module for the r-th road, $p + q$ is the temporal time range for the early stage of the traffic incident. $\boldsymbol{\lambda}_s, \mathbf{b}_s \in \mathbb{R}^{n_r}$, $\boldsymbol{\lambda}_1 \in \mathbb{R}^{(p+q)}$, $\boldsymbol{\lambda}_2 \in \mathbb{R}^{F\times(p+q)}$, $\boldsymbol{\lambda}_3 \in \mathbb{R}^F$ are learnable parameters, σ is the activation function. The calculated attention matrix $\boldsymbol{\alpha}$ stores the spatial correlations between the traffic sensors in $\mathcal{V}_r$, and the values in the attention matrix is updated dynamically according to the input data. Note that the Road Encoder models the traffic sensor locally within the scope of the roads, the HastGCN model considers a multiple road data input, a total number of $|\Phi|$ will be calculated, where Φ is the set of target roads. A softmax function is utilized to normalize the values in the attention matrices. Equaiton 7 calculates the output context of the spatial attention layer.

3.3.2 Spatial Attention in Predictor. The Predictor module models the spatial correlations at a higher level, it captures the traffic pattern dynamic between the corridors. Unlike the attention layers in the Road Encoders, there is only one attention layer in the Predictor module. To calculate the attention matrix, We apply the same logic as we used to calculate the attention matrices in the Road Encoders. The difference is that the output $\boldsymbol{\alpha}$ is singular. Here the calculation is omitted due to the similarity with the spatial attention of the Road Encoders.

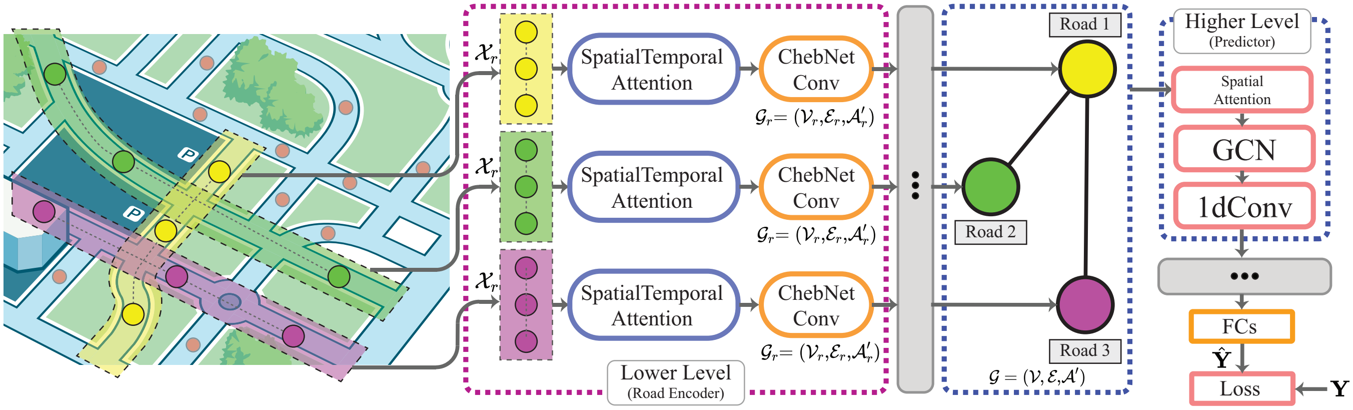
$\mathcal{X}_r$
$\mathcal{X}_r$
$\mathcal{X}_r$
SpatialTemporal Attention
SpatialTemporal Attention
SpatialTemporal Attention
ChebNet Conv
ChebNet Conv
ChebNet Conv
$\mathcal{G}_r = (\mathcal{V}_r, \mathcal{E}_r, \mathcal{A}'_r)$
$\mathcal{G}_r = (\mathcal{V}_r, \mathcal{E}_r, \mathcal{A}'_r)$
$\mathcal{G}_r = (\mathcal{V}_r, \mathcal{E}_r, \mathcal{A}'_r)$
Lower Level (Road Encoder)
Road 1
Road 2
Road 3
$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}')$
Higher Level (Predictor)
Spatial Attention
GCN
1dConv
•••
FCs
$\hat{\mathbf{Y}}$
Loss
$\mathbf{Y}$

Figure 4: The framework of the Hierarchical Attention-based Spatiotemporal Graph Convolution Network (HastGCN). The $\mathcal{X}_r$ in the figure represents the road sensor data tensor for the specific road. The middle section of this figure depicts the graph convolution block for the lower-level graph. The SpatialTemporal Attention sub-block is constructed sequentially. This block of neural networks can be stacked multiple times. The right section of this figure depicts the higher-level graph convolution for the road connectivity network. This block of neural networks can be stacked multiple times.

3.4 Temporal Attention for Road Encoders

In the Road Encoder module, the temporal attention layer is applied after the spectral convolution layer for temporal correlation caption on one of the dimensions of the input data tensor. The temporal attention mechanism is only applied in the Road Encoder. In the Predictor, the temporal attention is not applied because the time dimension of the data tensor is convoluted in the Road Encoder. Similar to the spatial attention, we calculate to highlight different importances in the time dimension:

$$\boldsymbol{\beta}_{\phi} = \boldsymbol{\eta}_t \cdot \sigma((\boldsymbol{\mathcal{X}}_{\phi}\boldsymbol{\eta}_1)\boldsymbol{\eta}_2(\boldsymbol{\eta}_3\boldsymbol{\mathcal{X}}_{\phi}) + \mathbf{b}_t) \tag{8}$$

$$\boldsymbol{\beta}'_{\phi_{ij}} = \frac{\exp(\boldsymbol{\beta}_{ij})}{\sum_{N}^{j=1}\exp(\boldsymbol{\beta}_{ij})} \tag{9}$$

$$\mathbf{c}^{\langle i \rangle} = \sum_{j} \boldsymbol{\beta}'_{ij}\mathbf{o}^{\langle j \rangle} \tag{10}$$

where $\mathbf{o}^{\langle j \rangle}$ is the output of the graph convolution layer in the Road Encoder module, $\boldsymbol{\beta}'_{ij}$ is the attention weights, $\boldsymbol{\mathcal{X}}_{\phi} = (\mathbf{X}_{\phi_1}, \mathbf{X}_{\phi_2}, ..., \mathbf{X}_{\phi_{n_r}}) \in \mathbb{R}^{n_r \times F \times (p+q)}$ is the input of the Road Encoder module for the r-th road, $p+q$ is the temporal time range for the early stage of the traffic incident. $\boldsymbol{\eta}_t$, $\mathbf{b}_t \in \mathbb{R}^{(p+q)\times(p+q)}$, $\boldsymbol{\eta}_1 \in \mathbb{R}^{n_r}$, $\boldsymbol{\eta}_2 \in \mathbb{R}^{F \times n_r}$, $\boldsymbol{\eta}_3 \in \mathbb{R}^{F}$ are learnable parameters, σ is the activation function. The calculated attention matrix $\boldsymbol{\beta}$ represents the temporal dependencies between the time point i and j. Equaiton 10 calculates the output context of the spatial attention layer.

3.5 Graph Convolution for Road Encoder and Predictor

Given the graph representation of traffic sensor network on a given road $\phi \in \Phi$, $\mathcal{G}_r = (\mathcal{V}_r, \mathcal{E}_r, \mathbf{A}_r)$, the spectral graph convolution is operated in the Fourier domain. An essential operator for spectral graph analysis is the Laplacian matrix $\mathbf{L}$, which is defined by $\mathbf{L} = \mathbf{D} - \mathbf{A}_r$, where $\mathbf{D} \in \mathbb{R}^{N \times N}$ is the degree matrix, and $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{r_{ij}}$. With the definitions of the degree matrix and the Laplacian matrix, we further calculate the normalized Laplacian matrix by $\mathbf{L}_n = I_N - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}_r\mathbf{D}^{-\frac{1}{2}} \in \mathbb{R}^{N \times N}$, where I_N is the identity matrix. After the normalization process, the Laplacian matrix $\mathbf{L}_n$ is now symmetric positive semidefinite, then its spectral decomposition is represented as $\mathbf{L}_n = \mathbf{U}\Lambda\mathbf{U}^T$. $\mathbf{U}$ is comprised of orthogonal and normalized eigenvectors $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_N] \in \mathbb{R}^{N \times N}$ and $\Lambda = diag([\lambda_1, ..., \lambda_N])$ is the combination of eigenvalues $\lambda \in \mathbb{R}^N$. Then the spectral convolution can be defined in the Fourier domain as:

$$y = \sigma(\mathbf{U}g_{\theta}(\Lambda)\mathbf{U}^T x) \tag{11}$$

where x, y are convolution input and output respectively, g_{θ} is the filter of the convolution process, and σ is the activation function. This formation is feasible for spectral convolution process, however, the requires expensive computation complexity for large scale graph structures. To reduce the computation complexity, approximation should be applied to the filter $g_{\theta}(\Theta)$. We apply m^{th} order Chebyshev polynomials $T_m(x)$ to the filter $g_{\theta}(\Theta)$:

$$g_{\theta}(\Lambda) \approx \sum_{m=0}^{K} \theta_m T_m(\widetilde{\Lambda}) \tag{12}$$

$$\widetilde{\Lambda} = \frac{2}{max(\lambda)}\Lambda - I_N \tag{13}$$

The approximation was first proposed by Hammond et al. [6]. The Chebyshev polynomials are recursively defined as $T_m(x) = 2xT_{m-1}(x) - T_{m-2}(x)$, with $T_0(x) = 1$ and $T_1(x) = x$. Kipf et al. [8] further limit the number of order m to be 1, along with the max eigen value to be 2. The Graph Convolution Network is now represented as:

$$\mathbf{Y} = (\mathbf{D} + I_N)^{-\frac{1}{2}}(\mathbf{A} + I)(\mathbf{D} + I_N)^{-\frac{1}{2}}X\Theta \tag{14}$$

For the Predictor component, target is the corridor graph $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$, the input $X \in \mathbb{R}^{N \times C}$ is the hidden encoded road features

Table 1: Dataset Summary for Heterogeneous Sources

		Districts		
		PeMS04	PeMS08	PeMS10
Traffic Data (CalTrans)	#Traffic Sensors	187	145	109
	Time Spans	2019/09-2019/10	2019/09-2019/10	2019/09-2019/10
Incident Records (RITIS)	#Traffic Incidents	2395	2276	2093
Road Network (TIGER/Line)	#Roads/Corridors	48	40	34

generated by the Road Encoder, the output $Y \in \mathbb{R}^{N \times F}$ is the learned hidden features based on the connectivity of the corridors. With the graph structure of G, weight cascade between the connected nodes can be modeled by the filter g_θ. Thus HastGCN can be utilized as an appropriate model for spatially connected corridors.

4 EXPERIMENT

This section evaluates our proposed model through extensive experiments. We first summarize the details of the experiment environment including datasets, the data pre-processing methods, and the experiment configuration settings. We select several previously proposed baseline methods for comparison purposes. Extensive experiments between our proposed model and the baseline methods are presented and analyzed.

4.1 Datasets

Our experiments are conducted on three heterogeneous data sources: real-world traffic data, topological road network data for graph adjacency matrix generation, and real-world traffic incident records dataset. The summaries for the heterogeneous data sources are shown in Table 1.

4.1.1 Traffic data. We evaluate the performance of our proposed model via three real-world large-scale datasets PeMS04, PeMS08, and PeMS10 collected from the Caltrans Performance Measurement System (PeMS) [3]. The data describes the occupancy rate, between 0 and 1; average vehicle travel speed of different car lanes of Bay area/Sacramento, San Bernardino/Riverside, and Stockton areas in California respectively.

4.1.2 Road network data. We utilize the TIGER/Line shapefiles and related database files of the state of California[3] to represent the topological connections and relations between the roads/corridors. The road shapefiles are an extract of selected geographic and cartographic information from the U.S. Census Bureau's Master Address File / Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) Database.

4.1.3 Traffic incident records data. Traffic incident records data: We collect traffic incident records from the Regional Integrated Transportation Information System (RITIS)[4]. The critical information such as incidents verification time, incident location in latitude and longitude, and incident duration in minutes.

4.2 Comparison Methods

To evaluate the performance of the traffic incident duration prediction, five conventional baseline methods are considered in our experiment: ℓ_2 regulized linear regression (ridge regression), ℓ_1 regulized linear regression (LASSO), support vector regression (SVR), Naïve multi-task learning model (nMTL), and TITAN model. Two state-of-the-art deep learning methods for traffic flow forecasting are also selected for comparison: ASTGCN and GeoMAN. Due to the different goals of prediction, we change the output to be 1-step prediction in the implementations of these methods.

- **ℓ_2 Regulized Linear Regression (Ridge)** [13]. Only temporal features are considered in this method. In our implementation, the λ parameter of Ridge is searched from $\{10, 100\}$. This model only considers the temporal features on duration prediction. No multi-task for arterial road connectivity and grouped temporal features are considered.
- **ℓ_1 Regulized Linear Regression (LASSO)** [14]. In our experiment, the λ parameter of LASSO is searched from $\{1, 10, 100\}$. The feature configurations applied by this model is the same as the ridge regression model.
- **Support Vector Regression (SVR)** [14]. Only temporal features are considered in this method. The model parameters are selected with $c = 1$ and $\epsilon = 0.1$. This model considers similar temporal features with ridge regression and LASSO methods, no multi-task features for connectivity is considered.
- **Naïve Multi-task Learning Model (nMTL)** [19]. Spatial and temporal features are considered in this method. The correlations between tasks are intuitively constrained by ℓ_2 norm, and within each task, the importance of the features are constrained by ℓ_1 norm. The penalty parameter λ is searched from $\{1, 10, 100\}$.
- **TITAN** [4]. Spatial and temporal features are considered in this method. In our implementation of this method, we apply this method on the complete set of traffic incidents, and the target feature is selected to be the temporal features. The penalty parameter λ is searched from $\{1, 10, 100\}$.
- **GeoMAN** [10]. Spatial and temporal features are considered in this method. The GeoMAN model is a multi-level attention-based recurrent neural network model proposed for the geo-sensory time series prediction problem. In our implementation of this method, we apply this the model structure, while modifying the output prediction to be 1-step.
- **ASTGCN** [5]. Spatial and temporal features are considered in this method. This model can process the traffic data directly on the original graph-based traffic network and effectively capture the dynamic spatial-temporal features. In our implementation of this

[3] https://catalog.data.gov/dataset/tiger-line-shapefile-2015-state-california-primary-and-secondary-roads-state-based-shapefile

[4] https://ritis.org/archive/incident

Table 2: Traffic Incident Impact Forecasting Comparisons (RMSE (Min), MAE (Min), MAPE (%))

Method	PeMS04			PeMS08			PeMS10		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Ridge	90.7262	73.2653	92.6015	86.2420	66.3947	84.4614	80.8172	61.3363	81.4579
LASSO	89.0958	71.5649	88.4577	74.5746	54.9547	68.4185	70.2032	52.4042	66.1313
SVR	84.5132	68.2331	87.1200	70.6437	52.2577	66.4771	65.8972	46.4696	61.6294
nMTL	67.7107	58.3578	78.6850	54.2898	39.2072	54.4242	53.1043	39.6155	40.4327
TITAN	72.7224	57.2332	80.8274	42.0785	33.0092	48.6796	45.7338	34.3774	41.8743
ASTGCN	66.3035	49.8112	73.5991	37.1646	27.7668	44.4699	37.5648	27.5014	37.0826
GeoMAN	64.9024	53.1784	74.1148	37.6363	**26.9859**	47.8037	37.5347	33.6329	36.8926
HastGCN	**64.1561**	**52.5912**	**71.9804**	**35.9933**	27.1678	**44.1087**	**36.5005**	**32.2848**	**36.7703**

method, we apply this the model structure, while modify the output prediction to be 1-step.

4.3 Evaluation Metrics

To justify the performance of the proposed model on traffic incident duration prediction, we adopt root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). These metrics are widely utilized in the field of traffic duration prediction studies [7, 9, 12, 22], it reflects the predictive performance of the proposed model. Equations 15, 16, and 17 represent the calculations of the selected evaluation metrics:

$$RMSE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2} \tag{15}$$

$$MAE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i}^{N}|y_i - \widehat{y}_i| \tag{16}$$

$$MAPE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i}^{N}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \tag{17}$$

where N is the total number of records; $\boldsymbol{y}$ is the predicted traffic incident durations represented in vector; $\hat{\boldsymbol{y}}$ is the ground truth value of the corresponding record, which is also represented in vector. y_i and $\widehat{y}_i$ are the i^{th} predicted result and the i^{th} ground truth value respectively.

4.4 Results and Discussion

The motivation of our model is to predict the impact of the traffic incident with the traffic sensors readings in its early stages. The impact of the traffic incident is quantified with the duration time, namely the time between the occurrence and the clear time. In order to accurately perform the prediction, our HastGCN model models three major aspects of the traffic sensor data: the spatial connections, the temporal time sequence, and the spatiotemporal dependencies. The locations of the sensors represent the spatial aspect of the traffic sensor data, and the location information is further refined by considering the topology of the road network. The time series of traffic records represent the temporal aspect of the traffic data. The attention mechanisms model the dynamical dependencies of the spatiotemporal features.

According to these three major modeling aspects, we group the comparison methods into three groups:

- Conventional methods with temporal features only.
- Conventional methods with temporal features and spatial constraints.
- Deep learning methods with spatiotemporal features and attention mechanisms.

All the experiments are conducted with the test dataset, which is separate from the training set, we apply 60% for training, 20% for validation, and 20% for testing on the experiment dataset. The results are reported in Table 2. We consider Ridge, LASSO, and SVR to be the group of "temporal features only" because these methods consider only the traffic sensor readings in the early stages to be the feature. The methods nMTL and TITAN fall into the group of "temporal features with spatial constraints," this is because these methods apply model the spatial connections with a multi-task framework that provides parameters sharing. The methods ASTGCN, GeoMAN, and our proposed HastGCN fall into the last group of "deep learning methods with spatiotemporal features and attention mechanism." Because these methods apply graph convolution networks (GCN) or recurrent neural networks (RNN) to model the spatiotemporal correlations. And these methods all utilize the attention mechanism to model the dynamics of the feature dependencies.

4.4.1 *Comparisons between Temporal-Only Models.*

Our proposed model HastGCN consistently and significantly outperforms the conventional methods (Ridge, LASSO, and SVR) that only consider the temporal features. This suggests that these three models struggle to modulate the impact of traffic incidents with the singular input information. Similar results about the weak prediction performance using these two models on the same task are reported in the previous work [4]. While comparing the results between the different groups separated by the aspects considered, we can find that the group with only temporal features do not outperform the group that considers temporal features with spatial constraints. This observation shows that considering the spatial aspects in the task of incident impact forecasting is necessary. As shown in Table 1, for different districts (PeMS04, PeMS08, and PeMS10), the number of corridors is different. And the number of corridors in the dataset are is equal to the number of nodes in the Predictor module of our proposed model. According to this property of dataset, for the first group of methods, we observe that with the number of corridors in the dataset increasing, the performance of the incident impact prediction drops. In contrast, the performance of the prediction decreases when the number of corridors is smaller. As a result, we conclude that including the spatial information on a dataset with a smaller number of corridors in our model helps to improve its sensitivity to the incident impact predictions.

4.4.2 Comparisons between Temporal Models with Spatial Constraints. Our model also consistently outperforms the group of methods that considers the temporal features with spatial constraints. In particular, for the task of predicting the impact of traffic incidents, our proposed HastGCN model achieves 7% to 10% less Root Mean Squared Error than the TITAN method; HastGCN also achieves 3% to 19% less Root Mean Squared Error than the nMTL method. We argue that this boost of performance is attributed to both spatial and temporal attention mechanisms. In our mechanism, the lower-level spatial attention layer allows the traffic sensors to apply spatial dependencies among all other sensors, and dynamically attend to only the important sensors in its neighborhood. The temporal attention layer allows the Road Encoder module to look back at all the hidden states for each encoder at each step and dynamically attend to only the important states in each sequence. As a result, we conclude that by considering the spatiotemporal attention mechanisms in a graph convolution network model helps to improve the performance of impact prediction.

4.4.3 Comparisons between Deep Learning Methods. Although our proposed HastGCN model outperforms the deep learning-based methods in general, it achieves the least advantages during the comparisons. In particular, for the task of predicting the impact of traffic incidents, our proposed HastGCN model achieves less Root Mean Squared Error than the GeoMAN method; HastGCN also achieves 1% to 2% less Root Mean Squared Error than the ASTGCN model. This achievement of performance may be generated by considering the hierarchical structure of the graph convolution networks. With the hierarchical graph structure, the traffic sensors on different corridors will not be convoluted, even if the two traffic sensors are closely located. This setting not only increases the robustness of the proposed model but also enables the HastGCN model to consider the directions of the corridors.

5 CONCLUSION

In this paper, we present a hierarchical attention-based spatiotemporal graph convolution network (HastGCN) to predict the impact of the traffic incidents. This model expands the traditional graph convolution network by considering a hierarchical structure that is capable of modeling graphs with sub-graphs. With this extension, the model considers not only the distances between the traffic sensors but also the directions of the corridors where the traffic sensors are located. We employ spatial and temporal attention mechanisms that encourage the traffic sensors and the corridors to build dependencies between each other and group by the sequences that leverage dynamically learned weights. Our model is evaluated on three real-world traffic data collected from the Caltrans Performance Measurement System (PeMS). And the experimental results demonstrate that our model can consistently outperform the existing temporal features-based and deep learning-based methods at predicting the future impact of the traffic incident at its early stages of development.

REFERENCES

[1] Martin W Adler, Jos van Ommeren, and Piet Rietveld. 2013. Road congestion and incident duration. *Economics of transportation* 2, 4 (2013), 109–118.

[2] Cen Chen, Kenli Li, Sin G Teo, Xiaofeng Zou, Kang Wang, Jie Wang, and Zeng Zeng. 2019. Gated Residual Recurrent Graph Neural Networks for Traffic Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 485–492.

[3] Chao Chen, Karl Petty, Alexander Skabardonis, Pravin Varaiya, and Zhanfeng Jia. 2001. Freeway performance measurement system: mining loop detector data. *Transportation Research Record* 1748, 1 (2001), 96–102.

[4] Kaiqun Fu, Taoran Ji, Liang Zhao, and Chang-Tien Lu. 2019. Titan: A spatiotemporal feature learning framework for traffic incident duration prediction. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 329–338.

[5] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 922–929.

[6] David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. 2011. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis* 30, 2 (2011), 129–150.

[7] Asad J Khattak, Jun Liu, Behram Wali, Xiaobing Li, and ManWo Ng. 2016. Modeling traffic incident duration using quantile regression. *Transportation Research Record* 2554, 1 (2016), 139–148.

[8] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[9] Ruimin Li, Francisco C Pereira, and Moshe E Ben-Akiva. 2018. Overview of traffic incident duration analysis and prediction. *European Transport Research Review* 10, 2 (2018), 22.

[10] Yuxuan Liang, Songyu Ke, Junbo Zhang, Xiuwen Yi, and Yu Zheng. 2018. Geoman: Multi-level attention networks for geo-sensory time series prediction.. In *IJCAI*. 3428–3434.

[11] Kaan Ozbay and Pushkin Kachroo. 1999. Incident management in intelligent transportation systems. (1999).

[12] Hyoshin Park, Ali Haghani, and Xin Zhang. 2016. Interpretation of Bayesian neural networks for predicting the duration of detected incidents. *Journal of Intelligent Transportation Systems* 20, 4 (2016), 385–400.

[13] Srinivas Peeta, Jorge L Ramos, and Shyam Gedela. 2000. Providing real-time traffic advisory and route guidance to manage Borman incidents on-line using the hoosier helper program. (2000).

[14] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.

[15] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 1 (2005), 91–108.

[16] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2018. Spatio-temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*.

[17] Chengjun Zhan, Albert Gan, and Mohammed Hadi. 2011. Prediction of lane clearance time of freeway incidents using the M5P tree algorithm. *IEEE Transactions on Intelligent Transportation Systems* 12, 4 (2011), 1549–1557.

[18] Liang Zhao, Junxiang Wang, and Xiaojie Guo. 2018. Distant-supervision of heterogeneous multitask learning for social event forecasting with multilingual indicators. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[19] Liang Zhao, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2016. Hierarchical incomplete multi-source feature learning for spatiotemporal event forecasting. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2085–2094.

[20] Jiayu Zhou, Jun Liu, Vaibhav A Narayan, Jieping Ye, Alzheimer's Disease Neuroimaging Initiative, et al. 2013. Modeling disease progression via multi-task learning. *NeuroImage* 78 (2013), 233–248.

[21] Jiayu Zhou, Zhaosong Lu, Jimeng Sun, Lei Yuan, Fei Wang, and Jieping Ye. 2013. Feafiner: biomarker identification from medical data through feature generalization and selection. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1034–1042.

[22] Yajie Zou, Kristian Henrickson, Dominique Lord, Yinhai Wang, and Kun Xu. 2016. Application of finite mixture models for analysing freeway incident clearance time. *Transportmetrica A: Transport Science* 12, 2 (2016), 99–115.