Tabular Learning-Based Traffic Event Prediction for Intelligent Social Transportation System

Chen Sun, Shen Li, Dongpu Cao, Fei-Yue Wang, *Fellow, IEEE*, and Amir Khajepour, *Member, IEEE*

***Abstract*— Accurate forecasting of future traffic is a critical contemporary problem for transportation research. However, it is difficult to understand the feature patterns of traffic events due to the complexity of the traffic environment, heterogeneous factors, and lack of abnormal samples. This article proposes a framework to integrate the social traffic data and use the TabNet model to facilitate the representation learning task in traffic event prediction. With the tabular learning and model interpretability analysis, the importance of common traffic external factors toward traffic events is studied. The study has practical significance for regulating traffic planning and the development of the operational boundary for autonomous driving systems.**

***Index Terms*— Crowd sourcing, social transportation system, tabular learning, traffic prediction.**

I. Introduction

WITH rapid urbanization, the surge in motor vehicles has led to severe traffic accidents, resulting in injuries and huge economic losses. Prediction of traffic incidents is a crucial element of an intelligent transportation system (ITS). Traffic events prediction is vital for optimizing public transportation, making routes safer, and improving transportation infrastructure in a cost-effective manner, all to make roads safer [1]. It has become imperative to understand the patterns of traffic events in urban areas from the historical records and predict possible future hazards [2].

The existing studies can be divided into two categories: one interprets the traffic event prediction as a regression problem which to predict the future population of events given the historical number of accidents at a given time and location [1] and the other explores embedded features of traffic incidents and defines the traffic events prediction as a classification problem [3]. Yuan *et al.* [4] proposed a Hetero-ConvLSTM regression model, which incorporated spatial features to capture temporal trends and spatial heterogeneity of traffic data and integrate road, weather, and time factors for traffic accident prediction. This study used a grid partitioning approach to divide the region of interest with the first seven days of historical data as training to predict the number of accidents in the next week. The prediction accuracy improvement shows that the deep learning techniques can relieve the spatial heterogeneity problem. Following a similar route, researchers in [5] and [6] use the segment-based regression kriging (SRK) and the graph convolutional network to predict the traffic volume and region traffic dynamics. The regression model for traffic prediction can effectively generate an approximate model from historical data suitable for conventional traffic flow, road occupancy, and travel time prediction in a well-studied area. However, the regression strategy is highly dependent on geographic grid partitioning and local infrastructure data available [7] and often not able to consider the impact of nontime series and nonquantitative traffic-related factors for event prediction [8].

The classification methods categorize traffic states discretely and obtain the corresponding traffic event prediction based on the features extracted with neural networks [9], [10]. Classifiers can be widely applied to predict the traffic flow [11], traffic accidents [12], and accident severity [13]. Recently, Huang *et al.* [14] developed a deep dynamic fusion network to effectively transfer knowledge from external factors through the context-aware embedding modules and hierarchical fusion networks to achieve spatiotemporal pattern learning. The heterogeneous external factors are aggregated together to improve the forecasting performance, and the results are validated on extensive real data collected in New York City. It is further demonstrated in [15] that geo-spatiotemporal external factors can help to shape the traffic event prediction in the short and long term. Recognize the effectiveness of data mining for classifier training, Moosavi *et al.* [16] proposed a framework for collecting traffic data with an existing online database such as MapQuest [17]. They grouped the accident data in large U.S. cities and proposed an accident prediction model using long-short-term-memory (LSTM) components. Christ [18] showed that their best experimental results have an accuracy level of about 65%, which is not remarkably accurate for accident prediction since various human-related factors, such as driver negligence, that are not observable directly may lead to accidents.

Manuscript received November 28, 2021; revised April 10, 2022 and April 23, 2022; accepted April 25, 2022. This work was supported in part by the National Natural Science Foundation of China through the Major Project under Grant 52131201. *(Corresponding author: Chen Sun.)*

Chen Sun and Amir Khajepour are with the Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: c87sun@uwaterloo.ca; a.khajepour@uwaterloo.ca).

Shen Li is with the Department of Civil Engineering, Tsinghua University, Beijing 100084, China (e-mail: sli299@tsinghua.edu.cn).

Dongpu Cao is with the School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China (e-mail: oscar_cao2016@163.com).

Fei-Yue Wang is with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: feiyue.wang@ia.ac.cn).

Digital Object Identifier 10.1109/TCSS.2022.3170934

Overall, accurate traffic event prediction is still challenging, limited by the information we can obtain in the first place. Traffic accidents are influenced by a variety of heterogeneous external factors, such as traffic conditions, adverse weathers, or even random factors such as vehicle mechanical problems; drunk driving may also lead to traffic accidents. In addition, the spatial heterogeneity of accident occurrences and sparsity of accident data also challenge the traffic prediction models, making it hard to accurately predict individual accidents due to the lack of sufficient samples. In order to address these challenges, we propose to assemble available social data using existing infrastructure in [16]. Social data are cheaper and more accessible compared to those collected from roadside infrastructure under specific design [19]–[21]. The main contributions of this article are listed as follows. This article proposes a fusion framework for traffic events prediction, utilizing the heterogeneous factors parsed from social data. By leveraging the self-supervised learning step in TabNet, the proposed solution improved the prediction performance compared to [15] and [16] with better exploration on the edge cases. A variety of insights was gathered through experiment analyses of model effectiveness, expressiveness, and location correlations. The model interpretability analysis has been investigated, and the results may be applied to transportation research, urban planning, and driving risk analysis.

The remainder of this article is organized as follows. Section II discusses related works on traffic event prediction and learning with tabular data. The proposed methodology, data processing, and analysis tools are covered in Section III. Section IV explains the experiment design and corresponding ablation studies with model effectiveness, expressiveness, and interpretability analysis. Section V contains the discussion and suggestions toward the use case in the real-world implementation. Finally, Section VI concludes this article.

II. Related Works

A. Real-Time Traffic Event Prediction

Accurate traffic event prediction is an indispensable part of ITS and urban computing [22], which plays an essential role in regional traffic services such as route planning [23] and traffic mitigation [24]. Traffic event prediction focused on forecasting a combination of the states such as traffic flow [25], the occupancy rate of the road space [26], and potential accident severity [27], based on the historical traffic data [23].

The traffic data commonly consist of two parts: one is the ST sequence which can be a direct model of the traffic states [22] and the other is the external factors attached based on time and geographical sourcing [28]. The external factors can provide questions to enhance the prediction accuracy. Common external features include the weather condition, driver personal information, day-of-week, and time-of-day [29]. Although it is often hard to cover all the external factors, having some of the correlated features does effectively improve the prediction, which implicates the embedded causal relation between the driver-environment factors and the traffic events [30]. Due to the data-driven nature, the classical statistical approach and learning models are two major trends of traffic prediction research. The autoregressive integrated moving average (ARIMA) and its variants are commonly used strategies for time-series prediction based on classical statistical research and have been widely applied for urban computing problems [31]. The feature-based approach trains a regression model based on human-engineered traffic features [32] or utilizes the deep neural networks (DNNs) as an autofeature encoder [33]. The performance of the feature-based model depends heavily on feature selection. By assuming the internal traffic state transition following the Markovian property, the state-space models represent the uncertainty of the system and capture the implicit correlation structure of the data [34]. However, abnormal traffic events such as crashes or accidents suffer nonlinearity, whereas the state-space model is not optimal for modeling the complex traffic in large-scale problems [35].

B. Spatial–Temporal Pattern Mining

The real-time traffic events data differentiate it from other classical data studied in the literature with the dependencies in the spatial and temporal viewpoint [36], [37]. The ST data share structurally correlated features in space and time, which the traditional data mining strategies often miss and assume the data distribution to be independent and identically distributed (i.i.d) [38], [39].

Time-series data mining has been well explored in trajectory pattern mining and ST clustering in transportation study [40]. The spatial dependence is treated in the Euclidean space with convolutional neural network (CNN), recurrent neural network (RNN), and attention modules in much recent traffic prediction research [41], [42]. GCN is then proposed to model non-Euclidean spatial structure data based on spectral or spatial perspective [6]. The generic challenge of ST data is the autocorrelation due to the neighborhood effect both locationwise and timewise as mentioned in [43].

C. Learning With Tabular Data

Structured data, particularly the tabular data, is one of the most common data types in real world [44]. Thus, learning patterns from tabular data becomes a valuable and popular topic in the current machine learning field.

The decision tree (DT) approach is the most commonly used for tabular data learning recently. Meanwhile, ensemble learning is usually applied simultaneously with the standard DT algorithm to improve the overall performance. XGBoost [45] and LightGBM [46] are two most popular and powerful approaches of tree ensembles. Since the DNN approach has achieved great success in many areas such as images process [47], audio process [48], and translation [49], it shows great potential to solve many real-world problems. Hence, it seems feasible to solve the tabular learning problems by using the DNN approach. Over the years, an increasing number of researchers have worked in this field and have proposed many architectures such as CTGAN [50] and TabNN [51]; they tried different solutions and promoted the studies but did not achieve overwhelming performance in such field. Recently, a novel and state-of-the-art DNNs architecture called

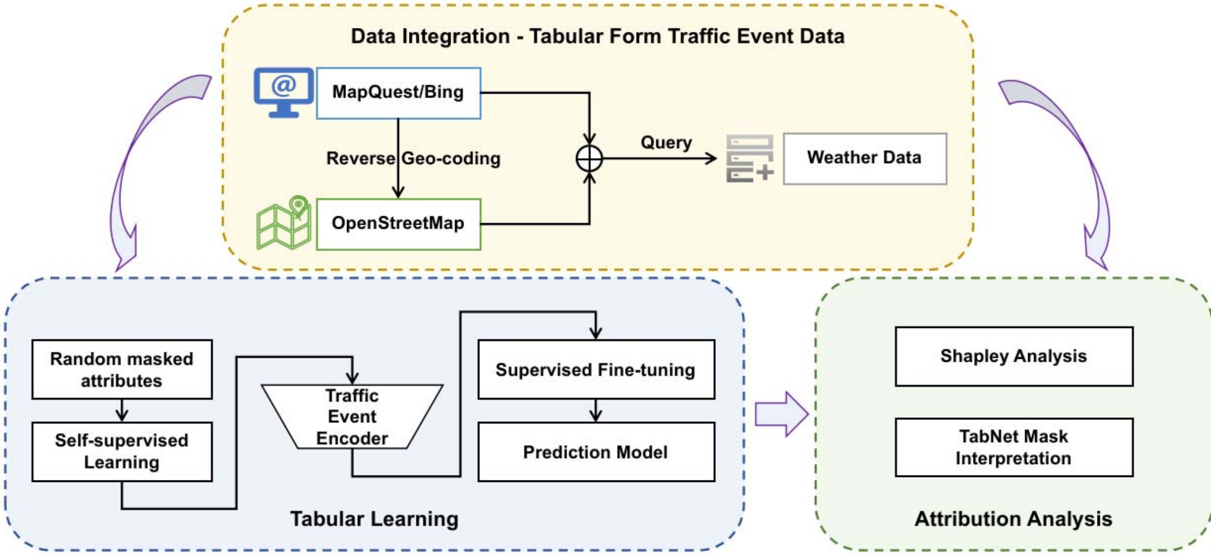


Fig. 1. Overall framework for traffic event prediction utilizing social available data and tabular learning.

TabNet [52] has been proposed by Google; it successfully achieved high performance and great interpretability in an end-to-end DNN model.

III. Methodology

In this section, the problem formulation and common definitions are presented first and then followed by the data augmentation and aggregation schemes developed in this article. The tabular learning strategy is also presented with the attribution analysis tool for the classification task as well as its interpretability. Fig. 1 shows an overview of the proposed framework integrating the social available data for traffic event prediction.

A. Problem Formulation

The spatial region centered at location $p = \langle lat, lon \rangle$ with radius r is denoted as p for simplification purpose. Radius r can be varied by application and the main assumption is to have an attribute-invariance property associated with given region on the external factors. The traffic data can be constructed by postevent recordings e_r, geographical properties g_p, and external environment representation e_e. The event recording commonly shares the structure $e_r = \langle p, t, desc, d_e \rangle$, where d_e concludes the event properties described by natural language $desc$.

In ITS social transportation setting [27], the set of traffic events $\mathcal{D}_{\rceil} = \{e_{r,1}, e_{r,2}, \ldots, e_{r,n}\}$ are often available through the postevent recording collection. The first goal is to first create a database of traffic events with external factors, e.g., $\mathcal{D} = \{(e_{r,1}, g_{p,1}, e_{e,1}), \ldots, (e_{r,n}, g_{p,n}, e_{e,n})\}$. Moreover, a proper model $\mathcal{M}$ is expected to predict the possibility as well as the severity of the event in the given region based on the historical recordings and available real-time measurements. Finally, the model should be "explainable" in the sense that able to automatically provide insights on the most related factors to the traffic events and their severity.

B. Data Integration

1) Traffic Data Collection: Urban traffic accident logs are often available in the form of postanalysis reports from police records. The National Highway Traffic Safety Administration (NHTSA) (URL: ftp://ftp.nhtsa.dot.gov/) provides comprehensive data on vehicle crashes such as Fatality Analysis Reporting System (FARS) [53]. FARS, created in 1975, contains data from the annual census of fatal motor vehicle crashes that took place on an open traffic way in USA. According to the database protocol, the data are collected, encoded, and then transmitted by designated FARS analysts who record and report over more than 125 data elements in standard forms. The 125 data elements consist of the crash scene details, vehicle information, driver states, and personnel [54]. The FARS data are complete; however, they cannot collect in real time and most information regarding the driver is not available ahead of crashes.

In recent years, the road traffic events have been collected and broadcast as social services provided by agencies such as MapQuest, Google, or Microsoft [55]. Moosavi *et al.* [16] have pulled data from the server and collected 2.27 million cases of traffic accidents between February 2016 and March 2019. The data collection process is continued and collected over 3 million traffic event records through the web API available.

2) Geoweather Property Extraction: The external factors related to the traffic events are collected through a query process based on (p, t) pairs in the traffic event recordings. The geographical properties are associated with annotations on the road map as junction, roundabout, stations, highway, and railway on the map. These road properties are obtained from OSM with the filtering method mentioned in [28].

Weather data are obtained similarly by querying (p, t), and the raw data come with attributes such as temperature, wind speed, precipitation, and pressure. In this work, the weather attached to the traffic event data is reorganized as close as

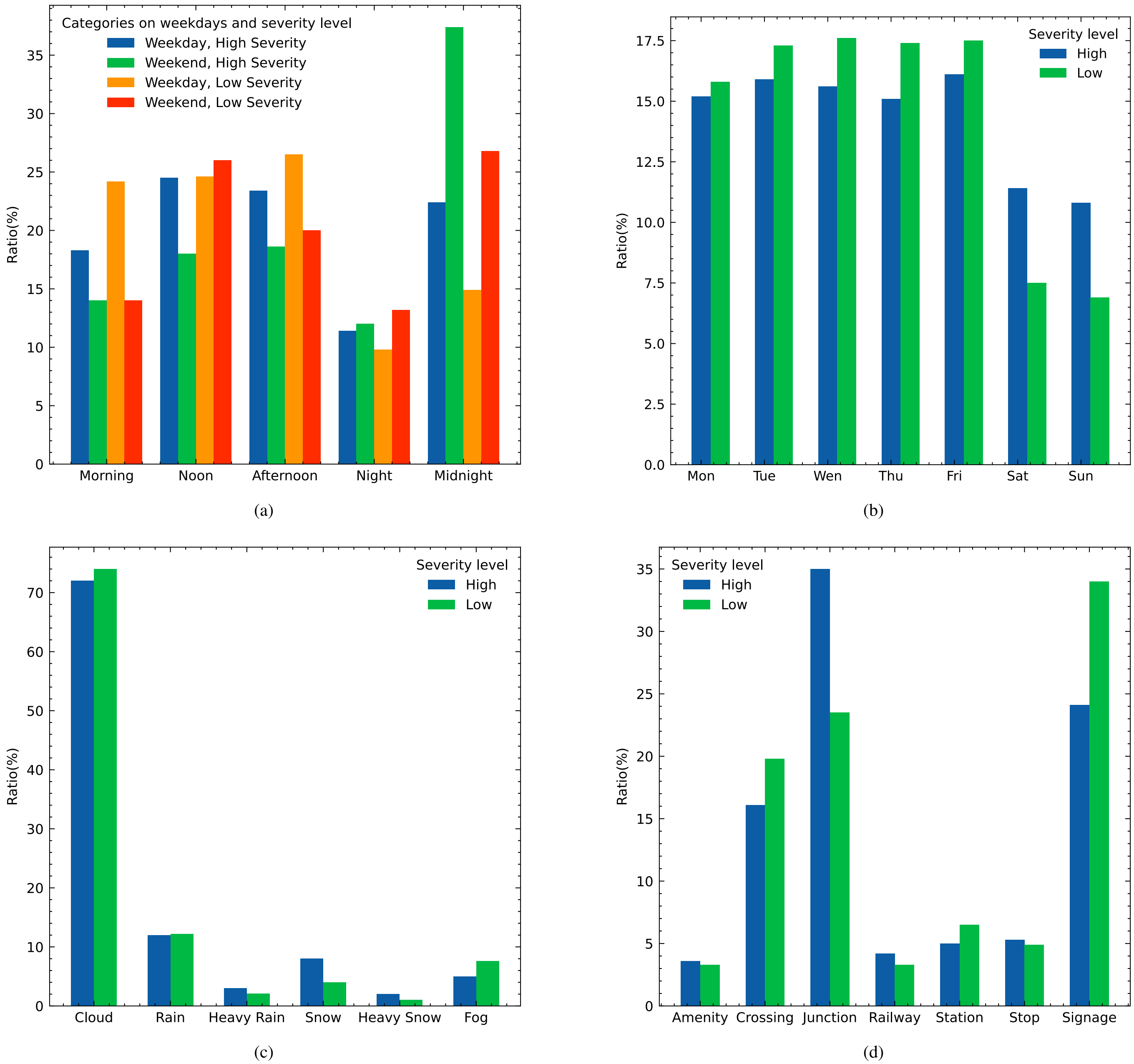


Fig. 2. Characteristics of traffic event dataset after data integration. The event severity distribution in terms of (a) and (b) time attributes, (c) weather and other external factors, and (d) geoproperties. The categories, such as bump and give-ways, are neglected in (d) due to the very small ratio in the dataset.

possible to the FARS encoding [53]. It is worth mentioning that quantitative ontology for the weather definition is still missing in academia. It is expected as a future work for explicitly specifying the operating conditions for the driving functions [29].

Notice that FARS has three severity levels (0: no injury, 1: injury, and 2: fatality) according to the worst-injured occupant in the crash. In comparison, the severity level in MS-Bing is defined with (1: low impact, 2: minor, 3: moderate, and 4: serious) with the event-type encoding considering accident, congestion, road hazard, construction, and miscellaneous. In this work, the severity level with injury or fatality with the severe level with accident and hazard is grouped as "severe accident." Fig. 2 provides details on the characteristics of event in the dataset.

C. Tabular Learning

In general, the constructed database often feeds into an ensemble DT model for tabular learning [56], [57]. The boosting tree models have become dominant in industry as the decision manifold of the model can be seen as hyperplane boundaries, which fit well for the tabular data and the inference share better interpretability due to the tree structure [58]. However, DTs often ignore the correlation between attributes and do not perform well for temporal data [59]. Thus, this

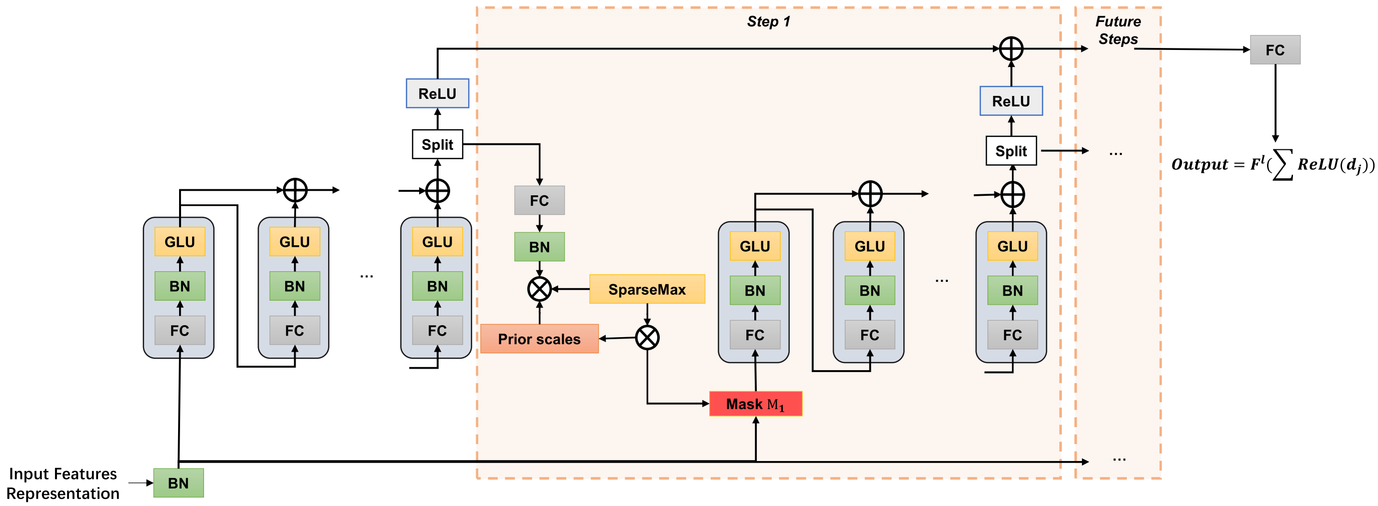


Fig. 3. TabNet encoder structure (changed from [52]).

article focuses on investigating the implementation of TabNet on the sparse and heterogeneous traffic event dataset and comparing the performance with the DT-based benchmark.

1) TabNet as Soft DT Additive Model: As shown in Fig. 3, the tabular data are encoded through multistep feature transforming. Instead of global feature normalization used in DTs, the batch normalization (BN) is applied for the input D-dimensional features $x \in \mathbb{B} \times \mathbb{D}$, with the batch size B.

The batch-normalized features $\bar{\mathbf{X}} \in \mathbb{R}^{B\times D}$ are then forwarded to the stacked fully connected (FC) layers, BN, and gated linear unit (GLU), which can be interpret by h^i in the following equation:

$$h^i(\bar{\mathbf{X}}) = (W^i * \bar{F}^i(\bar{\mathbf{X}}^i) + b^i) \otimes \sigma(V^i * \bar{F}^i(\bar{\mathbf{X}}^i) + c^i). \quad (1)$$

W^i and V^i are different convolution kernel at the ith stack with the $\bar{F}^i(\cdot)$ as the BN after FC operation. Similar to the definition proposed in [52], b^i and c^i are bias parameters, and $*$, $\otimes$, and $\sigma(\cdot)$ correspond to convolution, elementwise product, and the sigmoid function, respectively. BN is useful in adjusting the distribution of the output data of each layer so that it enters the activation function's "zone of action." The normalized residual with $(0.5)^{1/2}$ is proposed in [52] to stabilize the learning procedure to suppress the variance from changing dramatically.

The output of the stacked feature transformer is further split into two parts through

$$[\mathbf{d}_j, \mathbf{s}_j] = H_j(\mathbf{M}_j \cdot \tilde{\mathbf{X}}) \quad (2)$$

where $\mathbf{d}_j \in \mathbb{R}^{B\times N_d}$ is used for the final output of this computational model, whereas $\mathbf{s}_j \in \mathbb{R}^{B\times N_a}$ is used for the computation of next step mask layer for attentive transformer. The term H_j corresponds to the mapping combined by FC, BN, and GLU layers. The learnable mask $\mathbf{M}_j \in \mathbb{R}^{B\times D}$ is employed for salient feature selection in (2). The mask is obtained through sparsemax operation in (3) from the previous step as it encourages sparsity that directly map the high-dimensional vector to a simplex [60]

$$\mathbf{M}_j = \arg\min_{p\in\Delta} \|p - P_{j-1} \cdot H_j(\mathbf{s}_j)\|. \quad (3)$$

The term $P_j = \Pi_{k=1}^{j}(\gamma - M_j)$ is the prior scales term, which indicates the extent to which a feature has been used in previous steps in (3). According to human intuition, if a feature has been used many times in previous steps, it should not be selected by the model anymore, so the model reduces the weight of such features by this prior scales term. As reported in [52], the term γ controls the total allowed usage of previous features. If $\gamma = 1$, each feature will only be used once in one step. As γ increases, the restriction on feature reusing becomes softer. Finally, the prediction output is obtained through an FC layer after ReLU transform in the following form:

$$\text{output} = F^l(\sum_{j=1}^{N} \text{ReLU}(\mathbf{d}_j)). \quad (4)$$

Notice that the resulted mask in (3) stays in the exact shape of the input features and can be viewed as an attention distribution for the batch sample at the current step. Thus, different sample inputs will result in different attention distributions, guaranteeing the instancewise feature selection in favor of the multistep mask computation.

The regularization term as (5) is used to enhance the model ability to select features sparsely similar to [52]

$$\mathcal{L}_{\text{sparsity}} = \sum_{i=1}^{N_{\text{steps}}} \sum_{b=1}^{B} \sum_{j=1}^{D} \frac{-M_{b,j}[i]}{N_{\text{steps}} \cdot B} \log(M_{b,j}[i] + \epsilon). \quad (5)$$

The term L_{sparsity} is designed as entropy, and the optimization goal is to get the entropy as close to 0 as possible—the small number ϵ is used for numerical stability.

2) Self-Supervised Feature Learning: The decoder in Fig. 4 is used to reconstruct representations from the encoded features. Mimic to the data augmentation process, the incomplete data with nulls as well as partially masked features are learned from the others. The encoder model trained through the self-supervised learning manner can effectively compress the features. The initial mask can be interpreted as $\mathbf{M_s} \in \{0, 1\}^{B\times D}$ and the encoder inputs with $(1 - \mathbf{M_s}) \cdot \mathbf{X}$. The goal of self-supervised learning is to reduce the difference between the

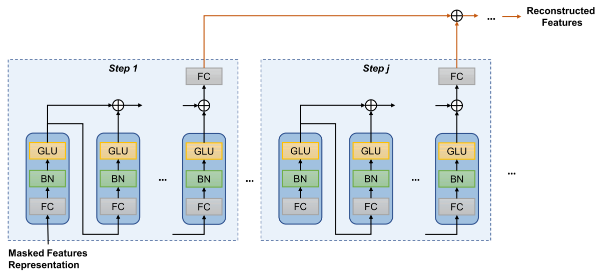


Fig. 4. TabNet decoder structure (changed from [52]).

real feature $\mathbf{M_s} \cdot \mathbf{X}$ and the reconstructed ones $\mathbf{M_s} \cdot \hat{\mathbf{M}}_\mathbf{s}$. The regularized mse is accepted in this article for the reconstruction loss in the following form:

$$\mathcal{L}_{\text{recon}} = \sum_{b=1}^{B}\sum_{j=1}^{D}\left|\frac{\left(\hat{\mathbf{X}}_{b,j} - \mathbf{X}_{b,j}\right)\cdot S_{b,j}}{\sqrt{\sum_{b=1}^{B}\left(f_{b,j} - 1/B\sum_{b=1}^{B} f_{b,j}\right)^2}}\right|^2. \quad (6)$$

Notice that the initial mask $\mathbf{M_s}$ is resampled in each round during the self-supervised learning process to learn the representation of the whole feature data instead of a set of local representations. Overall, self-supervised learning can use spatial–temporal data in the tabular form and enable the model to achieve better results even when there are imbalanced labels in the data.

D. Attribution Analysis

It is essential to understand the contributions of each factor (vehicle, environment, and traffic) toward the transportation events and their causality. In the proposed analysis, the attribution analysis is conducted with popular black-box-oriented Shapley strategy [61] and the in-built interpretation approach provided by TabNet [52].

1) Shapley Analysis: The Shapley method decomposes the regression equation and quantifies the contribution of each variable based on cooperative game [61]. The contribution to the prediction of a black-box model $\mathcal{M}$ is interpreted by

$$g(x) = \phi_0 + \sum_{j=1}^{M}\phi_j\theta_j' = \mathcal{M}(x) \quad (7)$$

where $\phi_j \in \mathbb{R}$ corresponds to the estimated contribution of the jth feature in the local instance x. The Shapley value ϕ_j can be then computed based on the following:

$$\phi_j(\mathcal{M}, x) = \sum_{S\subseteq\{\theta_1,\ldots,\theta_m\}\setminus\{\theta_j\}} \frac{|S|!(m-|S|-1)!}{m!} V\left(\mathcal{M}, \theta_j\right) \quad (8)$$

where the feature vector $x \in \mathbb{R}^m$ and the subset of feature combinations S with potentially 2^{m-1} combinations. The last term in (8) is the marginalized result over prediction in the subset $V = \mathcal{M}_x\left(S \cup \theta_j\right) - \mathcal{M}_x(S)$. The computational tools for Shapley are available in [61]–[63], which is accepted in this article for its computational efficiency.

2) TabNet Built-in Interpretability: The importance of each feature in different steps is represented in TabNet by learning the intrinsic mask. The final output is obtained through multiple ReLu activation layers (as shown in Fig. 3). If the sample's output at step j is negative, then features associated with the current step do not contribute to the final prediction. The contribution of each sample in batch b can be written as

$$\eta_{b,j} = \sum_{k=1}^{N_d}\text{ReLU}\left(d_{b,j}[k]\right). \quad (9)$$

The larger $\eta_b[i]$ is, the more contribution toward the final score of the output. Another interpretation mentioned in [52] is that (9) corresponds to the weight of each mask and finally reflects the feature's importance level as follows:

$$\psi_{b,j} = \sum_{k=1}^{N}\eta_{b,j}\cdot\mathbf{M}_{b,j}[k]. \quad (10)$$

IV. Experiment and Analysis

This section first compares different approaches in the traffic event data, followed by the interpretability analysis. Then, the ablation study and performance comparison are presented. Finally, the sparse feature problem and issues in instancewise feature selection are discussed. The experiments are implemented in Python using PyTorch [64] and scikit-learn [65] libraries on a Linux machine with GTX3080.

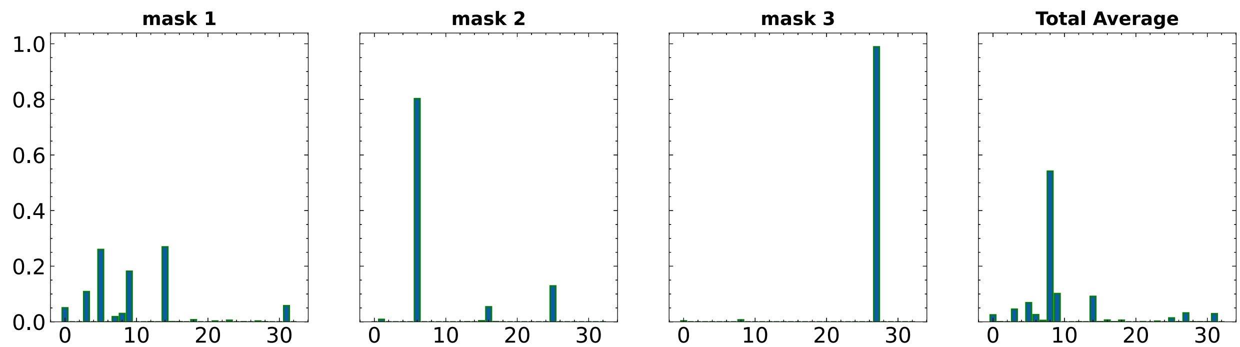
A. Model Exploration

The DNN [66], gradient boosting classifier (GBC) [67], and XGBoost model [68] are selected as baselines for comparison. The DNN is composed of four feedforward layers with three hidden layers with the same setting as in [15], with the Adam optimizer and initial learning rate at 0.01. GBC is a popular DT model, and the boosting stage is set as 100, with logistic loss. The learning rate of the GBC is set as 0.1. The XGBoost model shares the same 0.1 learning rate with the maximum boosting round as 100. The maximum depth is tuned between 5 and 15, where the best model performance is selected at 12. In our experiment, the TabNet model set the coefficient for feature reusage in the masks as 1.5. The momentum for BN is set as 0.3 as we found out that it provides better results than the default settings.

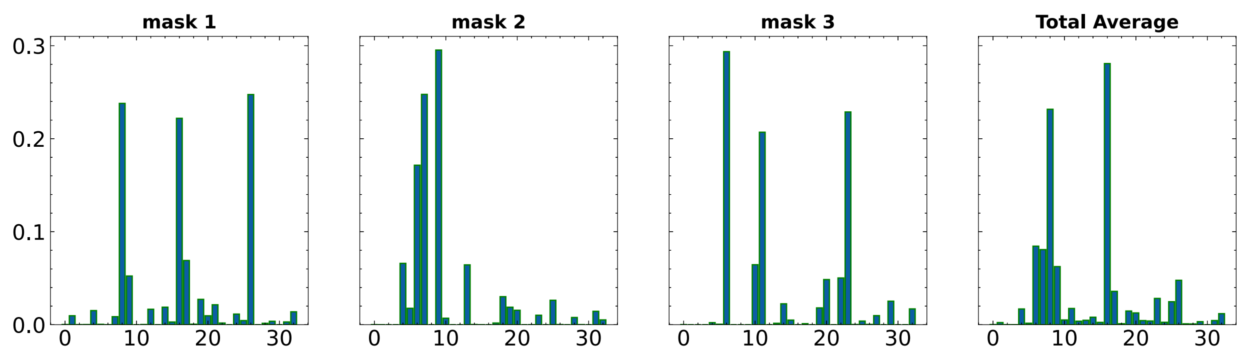
The area under curve (AUC), which corresponds to the area under the receiver operating characteristic (ROC) curve, is selected as the evaluation metric in this work as it can directly measure the goodness of the classifier. The AUC is selected to avoid using multiple F1-score to deal with measure the classifier's ability to identify the "rare cases" in [16]. The ROC is defined over the true positive rate (TPR) and false positive rate (FPR) hyperplane. The TPR in (11), also called sensitivity, corresponds to the percentage of all samples that were positive and correctly classified as positive as well, whereas the FPR is the percentage of all samples that were

TABLE I
AUC OF EVENT PREDICTION FOR VARIOUS APPROACHES ON DIFFERENT REGION DATA

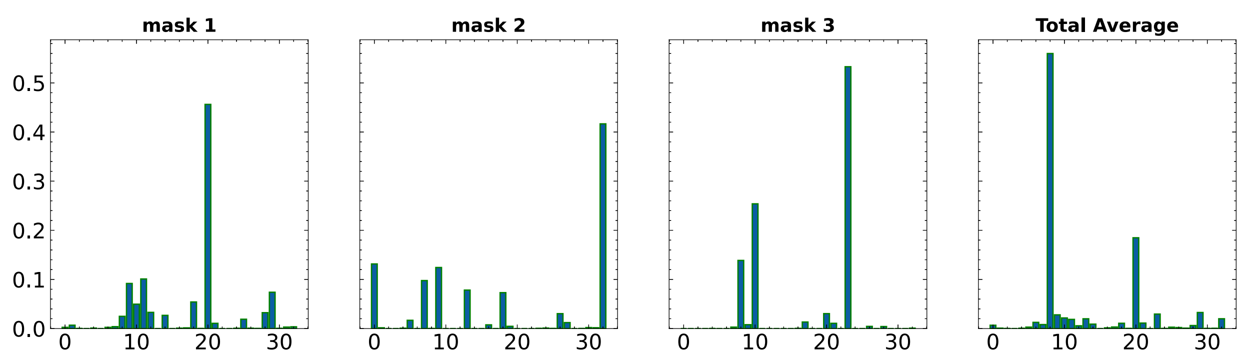
State \ Model	NN	GBC	XGBoost	TabNet w.o. self-supervise learning		TabNet w.t. self-supervise learning	
				Best Param	Mean + Variance	Best Param	Mean + Variance
MD	0.51582	0.81301	0.83703	0.86002	0.79956 ± 0.01128	**0.860635**	0.809999 ± 0.000898
OH	0.53511	0.84515	**0.93001**	0.87430	0.86148 ± 0.01089	0.914597	0.860758 ± 0.002728
GA	0.59253	0.77140	0.83447	0.83543	0.81148 ± 0.02577	**0.858092**	0.749283 ± 0.003610
NY	0.51253	0.67258	0.79099	0.83294	0.79982 ± 0.13414	**0.856614**	0.760518 ± 0.003687
CA	0.53581	0.64024	0.66574	0.80101	0.72146 ± 0.09498	**0.842342**	0.806890 ± 0.001239
All	0.56224	0.79185	0.75583	0.78979	0.73373 ± 0.10352	**0.895470**	0.819384 ± 0.008264



(a)



(b)



(c)

Fig. 5. Averaged feature importance (y-axis) in first three masks among the feature index (x-axis). Total average of (a) CA, (b) MD, and (c) OH samples.

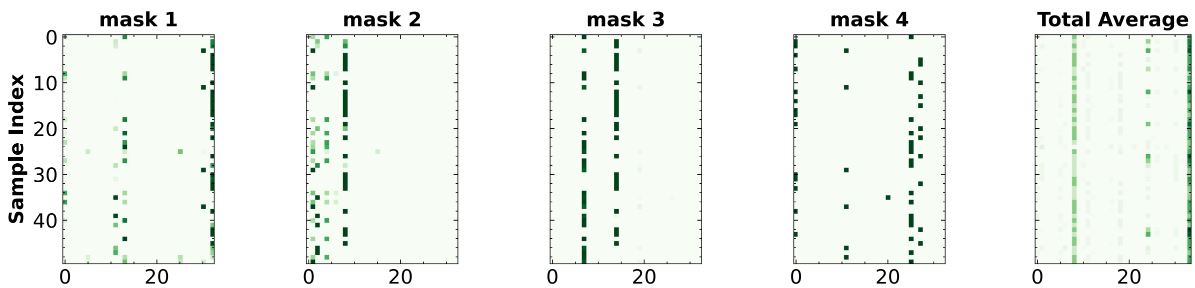
negative yet were wrongly classified as positive

$$\begin{aligned} \text{TPR} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{FPR} &= \frac{\text{FP}}{\text{FP} + \text{TN}}. \end{aligned} \tag{11}$$

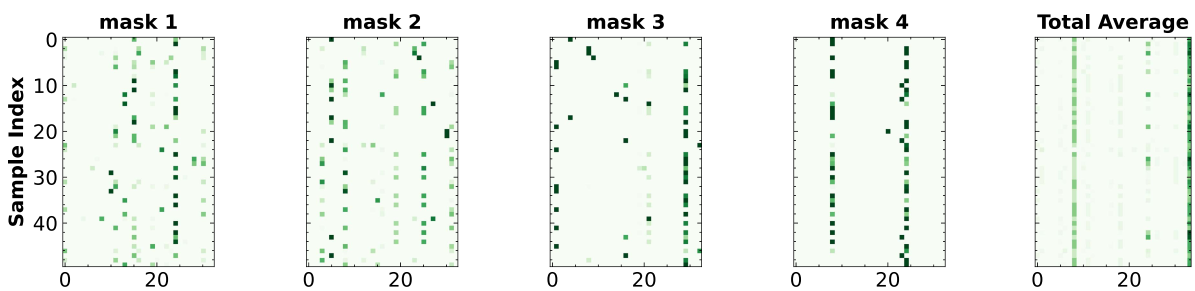
The models are deployed as regional service (in-state) cases and national cases. The regional service mask data from other states and the model are trained and tested over its regional data, which consider as a regional traffic service setting. On the other hand, the results on all states are trained and

TABLE II
TOP SIX FEATURES AND RELATIVE SIGNIFICANCE OF THE TRAFFIC EVENT PREDICTION IN VARIOUS STATES

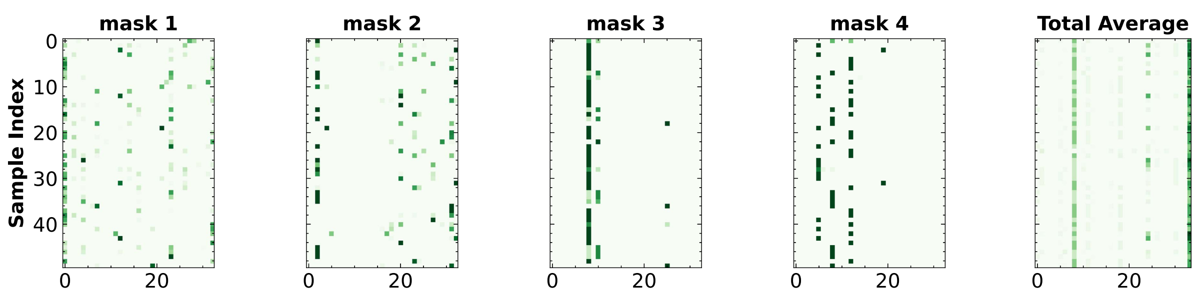
	MD		OH		GA		NY		CA	
XGBoost-Shapley	Laneblock	0.4712	Laneblock	0.4322	Laneblock	0.3197	Laneblock	0.2999	Laneblock	0.2419
	Flow-incident	0.3320	Flow-incident	0.2093	Flow-incident	0.2176	Flow-incident	0.2371	Flow-incident	0.2177
	Latitude	0.1103	Latitude	0.1012	Latitude	0.1008	Latitude	0.1079	Latitude	0.1270
	Longitude	0.1098	Longitude	0.0921	Longitude	0.1001	Longitude	0.0994	Longitude	0.1008
	Junction	0.0910	Construction	0.0727	Construction	0.0414	Construction	0.0428	Construction	0.0757
	Construction	0.0311	Junction	0.0474	Junction	0.0386	Weekend	0.0197	Bump	0.0609
TabNet	Laneblock	0.3033	Laneblock	0.5542	Laneblock	0.2969	Bump	0.2954	Laneblock	0.2782
	Construction	0.2555	Construction	0.1042	Rain	0.0854	Laneblock	0.2951	Stop	0.1106
	Flow-incident	0.1652	Turning Loop	0.0453	Bump	0.0661	Flow-incident	0.0511	Rain	0.1033
	Snow	0.0597	Humidity	0.0348	Period	0.0572	Weekend	0.0431	Bump	0.0975
	Railway	0.0455	Traffic Signal	0.0266	Wind speed	0.0440	Turning Loop	0.0398	Construction	0.0816
	Rain	0.0281	Flow-incident	0.0265	Amenity	0.0385	Construction	0.0218	Give-way	0.0702



(a)



(b)



(c)

Fig. 6. Feature significance mask $M[i]$ which indicates the selection at the ith step and the total average feature importance mask showing the global instancewise feature selection on various states samples in [16]. Brighter colors show a higher value. (a) CA. (b) MD. (c) OH.

validated over all the available data as a center traffic service setting. Both the studies follow the 60%, 20%, and 20% training–testing–validation split strategy. Five typical states are listed in Table I, namely, Maryland (MD), Ohio (OH), Georgia (GA), New York (NY), and California (CA). It can be seen that the XGBoost and TabNet outperform other baselines in all the cases listed. The TabNet and its self-supervised learning step significantly improve classification ability in most states, except at OH. The drop in performance is due to the differences in traffic complexity between OH and other listed states, which will be further investigated in Section IV-B.

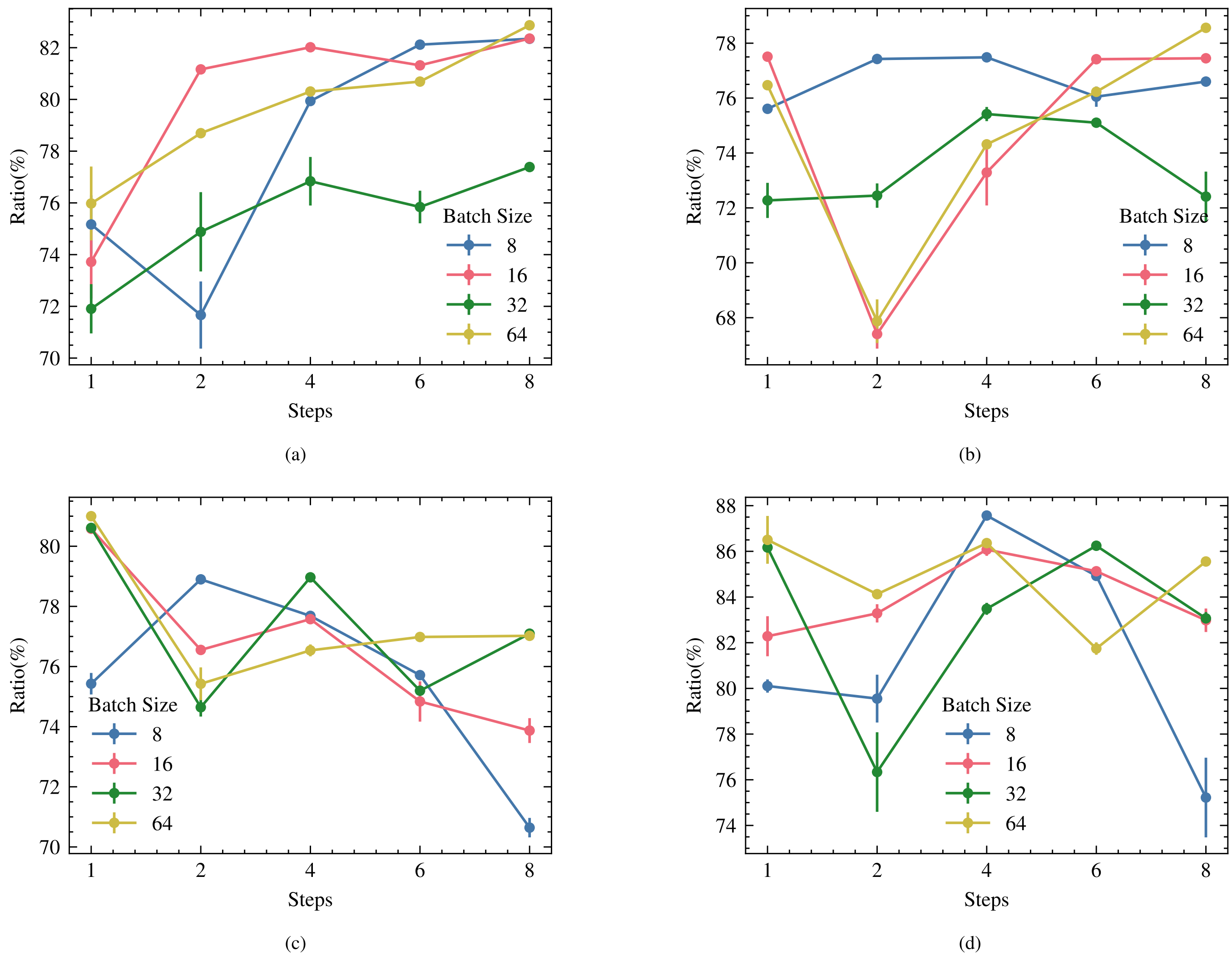


Fig. 7. Model performance on AUC over different batches and step selections on various statewise data. (a) CA. (b) NY. (c) MD. (d) OH.

The other observation from Table I is that the self-supervised learning process can significantly help to stabilize the learning procedure, which can be reflected from the AUC variance of the output model performances. Self-supervised learning provides a judicious choice of the encoded features that help to improve the supervised learning when the labeled dataset is limited. The advantage of the fast convergence is shown in Table I, which is helpful in the continual learning [69] as the transportation service map out as local services.

B. Model Interpretability

The TabNet model provides the instancewise feature selection and self-interpretability through mask weight of each step mentioned in (10). The average feature significance of each model mask step toward the given samples in various states is shown in Fig. 5. The x-axis in Fig. 5 are the feature index, where the top eight features and its weight are listed in Table II. It can be seen that the feature significance of the event prediction model has strong local characteristics. The prediction model in CA region data shares strong dependence in a small set of features at each step. In contrast, the feature significance is much more alike in MD and OH due to the west and east coast traffic difference.

Table II lists the top six significant features extracted from XGBoost and the studied TabNet model. The top importance features are extracted based on the Shapley strategy mentioned in (8) with the average absolute relative Shapley values scaled with $\tilde{\phi}_j = \|\phi_j\|/\sum\|\phi\|$. On the other hand, the aggregate decision contribution is calculated based on (9). Significant features extracted from the Shapley method for the XGBoost model are consistent across different geographical regions. Notice that the "Laneblock" feature contributes most toward the decision for both models due to the high correlation between the blocked lane with a traffic event in the U.S. accident dataset [16]. The DT model is more overfitting than the TabNet model because the GNSS location is included in the most significant features. Although the geographic coordinates correlate with the event's occurrence and severity in terms of the data, there are no causal relations based on the human interpretation. On the contrary, the TabNet model reveals better in terms of causality interpretation. For example, the inference of traffic events between each state of the TabNet model has relatively more consistent with the characteristics of the distribution of traffic facilities in those corresponding states shown in Table II.

Fig. 6 shows the aggregate feature importance for the samples from Table II. It can be observed that the feature importance is almost all zero for those irrelevant features based on the TabNet model. At each decision step, the TabNet model merely focuses on the relevant ones. For the CA-mask step 3, feature-7 and feature-14 are the most significant ones for most of the samples examined. At the same time, both the MD and OH samples showed great total average feature importance on feature-8 and feature-34. Fig. 6 also shows that the TabNet model yields good instancewise feature selection. The salient features are chosen corresponding to different input samples, which can overcome the overfitting problem and improve the utilization of data samples.

C. Ablation Study

The impact of ablation cases is shown in Fig. 7 on the expressiveness and the mask step configurations over different regions. For all cases, the number of training iterations is optimized on the validation set. The performance of the choice of batch size and the encoder steps shares strong locality, where the geographical traffic characteristics shape the required expressiveness in feature representations. Typically, in the CA state, the larger the batch size and the wider step choices improve the model performance in this region.

On the contrary, the model prediction performance decreases when we increase the choice of steps for a state like MD since the traffic complexity is much simpler than CA and the factors of traffic event occurrence are relatively more sparse. The number of decision steps can be interpreted as the number of split nodes in the DT. The introduction of more steps in the TabNet model to ensemble multiple trees improves the model's expressiveness. Fig. 7 also shows complex models (step=8), and a larger batch size tends to give better training results. The larger the batch size is, the more accurate the direction of descent it determines, and the less training oscillation it causes within a specific range. It can be noticed that the traffic complexity of a region and the degree of indistinguishability of the features associated with traffic events correspond to the complexity of the model it requires.

V. Conclusion and Discussion

For traffic event prediction at the ITS service level, it is unlikely to expect optimal prediction performance with a uniform model for different geographic and traffic environments. A more practical approach would be to design models with suitable parameters for geographical characteristics and data sources. The model hyperparameters indeed have a significant impact on the model performance. The TabNet model investigated in this article inherits the advantages of the DT-based approach (interpretability and sparse feature selection) and the advantages of DNN (representation learning). Due to its instance interpretation capability and unsupervised pretraining, the TabNet model performs better in traffic event prediction as a local ITS service.

We also found key factors that affect the differentiation of traffic event prediction on the causal analysis of fixed-time-based traffic event prediction. The traffic flow through which the road infrastructure can pass and the congestion level of its road segments impact traffic times in almost all locations. This aspect confirms the critical influence of traffic infrastructure layout on traffic safety on-road sections, as mentioned in [70]. On the other hand, our analysis starts from the social traffic data and is based on the posterior analysis. We get how much of the influence of abnormal events such as lane blocks and road construction on potential traffic events. Traffic facilities and weather conditions are essential in defining ODD for autonomous driving safety. The interpretability analysis in this article of these external factors on traffic events can help define the ODD boundaries for autonomous driving systems in the future. It is worth noticing that the driver-related factors [54] are not considered in this work due to the information privacy issue. In the future, integrating driver characteristics for complete traffic event prediction can be considered a potential research direction when the methods and platforms for collecting driver data are available.

References

[1] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 4, pp. 606–616, Aug. 2011.

[2] F. Zhu, Y. Lv, Y. Chen, X. Wang, and F. Wang, "Parallel transportation systems: Toward IoT-enabled smart urban traffic control and management," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4063–4071, Oct. 2020.

[3] H. Yao, X. Tang, H. Wei, G. Zheng, and Z. Li, "Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction," in *Proc. Conf. Artif. Intell.*, vol. 33, 2019, pp. 5668–5675.

[4] Z. Yuan, X. Zhou, and T. Yang, "Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 984–992.

[5] Y. Song, X. Wang, G. Wright, D. Thatcher, P. Wu, and P. Felix, "Traffic volume prediction with segment-based regression Kriging and its implementation in assessing the impact of heavy vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 232–243, Jan. 2019.

[6] L. Zhao *et al.*, "T-GCN: A temporal graph convolutional network for traffic prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3848–3858, Sep. 2019.

[7] Y. Ma, M. Chowdhury, A. Sadek, and M. Jeihani, "Integrated traffic and communication performance evaluation of an intelligent vehicle infrastructure integration (VII) system for online travel-time prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1369–1382, Sep. 2012.

[8] I. Lana, J. Del Ser, M. Velez, and E. I. Vlahogianni, "Road traffic forecasting: Recent advances and new challenges," *IEEE Intell. Transp. Syst. Mag.*, vol. 10, no. 2, pp. 93–109, Apr. 2018.

[9] D. A. Tedjopurnomo, Z. Bao, B. Zheng, F. Choudhury, and A. K. Qin, "A survey on modern deep neural network for traffic prediction: Trends, methods and challenges," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 4, pp. 1544–1561, Apr. 2022.

[10] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, "Toward effective mobile encrypted traffic classification through deep learning," *Neurocomputing*, vol. 409, pp. 306–315, Oct. 2020.

[11] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Sep. 2015.

[12] L. G. Cuenca, E. Puertas, N. Aliane, and J. F. Andres, "Traffic accidents classification and injury severity prediction," in *Proc. 3rd IEEE Int. Conf. Intell. Transp. Eng. (ICITE)*, Sep. 2018, pp. 52–57.

[13] R. E. AlMamlook, K. M. Kwayu, M. R. Alkasisbeh, and A. A. Frefer, "Comparison of machine learning algorithms for predicting traffic accident severity," in *Proc. IEEE Jordan Int. Joint Conf. Electr. Eng. Inf. Technol. (JEEIT)*, Apr. 2019, pp. 272–276.

[14] C. Huang, C. Zhang, P. Dai, and L. Bo, "Deep dynamic fusion network for traffic accident forecasting," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 2673–2681.

[15] S. Moosavi, M. H. Samavatian, A. Nandi, S. Parthasarathy, and R. Ramnath, "Short and long-term pattern discovery over large-scale Geo-spatiotemporal data," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2905–2913.

[16] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. Ramnath, "Accident risk prediction based on heterogeneous sparse data: New dataset and insights," in *Proc. 27th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2019, pp. 33–42.

[17] M. P. Peterson, "Mapquest and the beginnings of web cartography," *Int. J. Cartogr.*, vol. 7, no. 1, pp. 1–7, May 2021.

[18] D. Christ, "Simulating the relative influence of tire, vehicle and driver factors on forward collision accident rates," *J. Saf. Res.*, vol. 73, pp. 253–262, Jun. 2020.

[19] Z. Zhang, Y. Li, and H. Dong, "Multiple-feature-based vehicle supply–demand difference prediction method for social transportation," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 4, pp. 1095–1103, Aug. 2020.

[20] X. Xue, F. Chen, D. Zhou, X. Wang, M. Lu, and F.-Y. Wang, "Computational experiments for complex social systems—Part I: The customization of computational model," *IEEE Trans. Computat. Social Syst.*, early access, Nov. 15, 2021, doi: 10.1109/TCSS.2021.3125287.

[21] H. Lu *et al.*, "Social signal-driven knowledge automation: A focus on social transportation," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 3, pp. 737–753, Jun. 2021.

[22] Z. Yu, C. Licia, W. Ouri, and Y. Hai, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1–55, Sep. 2014.

[23] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, "Big data analytics in intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 383–398, Jan. 2019.

[24] M. Di Vaio, G. Fiengo, A. Petrillo, A. Salvi, S. Santini, and M. Tufo, "Cooperative shock waves mitigation in mixed traffic flow environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 12, pp. 4339–4353, Dec. 2019.

[25] Y. Wu, H. Tan, L. Qin, B. Ran, and Z. Jiang, "A hybrid deep learning based traffic flow prediction method and its understanding," *Transp. Res. C, Emerg. Technol.*, vol. 90, pp. 166–180, May 2018.

[26] R. Tian, J. Bi, Q. Zhang, and Y. Liu, "Research on lane occupancy rate forecasting based on the capsule network," *IEEE Access*, vol. 8, pp. 38776–38785, 2020.

[27] C. Sun *et al.*, "Accident prediction in mesoscopic view: A cpss-based social transportation approach," in *Proc. IEEE 1st Int. Conf. Digital Twins Parallel Intell. (DTPI)*, Jul./Aug. 2021, pp. 306–311.

[28] C. Sun *et al.*, "Proximity based automatic data annotation for autonomous driving," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 2, pp. 395–404, Mar. 2020.

[29] C. Sun, Z. Deng, W. Chu, S. Li, and D. Cao, "Acclimatizing the operational design domain for autonomous driving systems," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 2, pp. 2–16, Mar./Apr. 2021.

[30] S. Xie, S. Chen, N. Zheng, and J. Wang, "Modeling methodology of driver-vehicle-environment system dynamics in mixed driving situation," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct./Nov. 2020, pp. 1984–1991.

[31] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results," *J. Transp. Eng.*, vol. 129, no. 6, pp. 664–672, Nov. 2003.

[32] W. Li *et al.*, "A general framework for unmet demand prediction in on-demand transport services," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2820–2830, Aug. 2019.

[33] J. Zhang, Y. Zheng, J. Sun, and D. Qi, "Flow prediction in spatio-temporal networks based on multitask deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 3, pp. 468–478, Mar. 2019.

[34] N. Polson and V. Sokolov, "Bayesian particle tracking of traffic flows," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 2, pp. 345–356, Feb. 2018.

[35] D. Zang, J. Ling, Z. Wei, K. Tang, and J. Cheng, "Long-term traffic speed prediction based on multiscale spatio-temporal feature learning network," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3700–3709, Oct. 2019.

[36] S. Shekhar *et al.*, "Spatiotemporal data mining: A computational perspective," *ISPRS Int. J. Geo-Inf.*, vol. 4, no. 4, pp. 2306–2338, Oct. 2015.

[37] S. Shekhar, R. R. Vatsavai, and M. Celik, "Spatial and spatiotemporal data mining: Recent advances," in *Next Generation Data Mining*. Boca Raton, FL, USA: CRC, Dec. 22, 2008, pp. 573–608, ISBN: 9780429147562, doi: 10.1201/9781420085877.

[38] E. Dittrich *et al.*, "A spatio-temporal latent atlas for semi-supervised learning of fetal brain segmentations and morphological age estimation," *Med. Image Anal.*, vol. 18, no. 1, pp. 9–21, Jan. 2014.

[39] G. Atluri, A. Karpatne, and V. Kumar, "Spatio-temporal data mining: A survey of problems and methods," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–41, Jul. 2019.

[40] G. Andrienko, N. Andrienko, P. Bak, D. Keim, S. Kisilevich, and S. Wrobel, "A conceptual framework and taxonomy of techniques for analyzing movement," *J. Vis. Lang. Comput.*, vol. 22, no. 3, pp. 213–232, Jun. 2011.

[41] K. Niu, C. Cheng, J. Chang, H. Zhang, and T. Zhou, "Real-time taxi-passenger prediction with L-CNN," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4122–4129, May 2019.

[42] S. Yang, W. Ma, X. Pi, and S. Qian, "A deep learning approach to real-time parking occupancy prediction in transportation networks incorporating multiple spatio-temporal data sources," *Transp. Res. C, Emerg. Technol.*, vol. 107, pp. 248–265, Oct. 2019.

[43] S. Wang, J. Cao, and P. Yu, "Deep learning for spatio-temporal data mining: A survey," *IEEE Trans. Knowl. Data Eng.*, early access, Sep. 22, 2020, doi: 10.1109/TKDE.2020.3025580.

[44] J. Bughin, J. Seong, J. Manyika, M. Chui, and R. Joshi, "Notes from the AI frontier: Modeling the impact of AI on the world economy," *McKinsey Global Inst.*, pp. 1–64, Apr. 2018.

[45] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[46] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3146–3154.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[48] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 173–182.

[49] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[50] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," 2019, *arXiv:1907.00503*.

[51] O. S. Arık and T. Pfister, "Tabnet: Attentive Interpretable Tabular Learning," *AAAI*, vol. 35, pp. 6679–6687, 2021.

[52] S. O. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," 2019, *arXiv:1908.07442*.

[53] N. C. Briggs, R. S. Levine, W. P. Haliburton, D. G. Schlundt, I. Goldzweig, and R. C. Warren, "The fatality analysis reporting system as a tool for investigating racial and ethnic determinants of motor vehicle crash fatalities," *Accident Anal. Prevention*, vol. 37, no. 4, pp. 641–649, Jul. 2005.

[54] L. Yan, Y. He, L. Qin, C. Wu, D. Zhu, and B. Ran, "A novel feature extraction model for traffic injury severity and its application to fatality analysis reporting system data analysis," *Sci. Prog.*, vol. 103, no. 1, pp. 1–23, Mar. 2020.

[55] L. Zhu and J. D. Gonder, "A driving cycle detection approach using map service API," *Transp. Res. C, Emerg. Technol.*, vol. 92, pp. 349–363, Jul. 2018.

[56] M. Gashler, C. Giraud-Carrier, and T. Martinez, "Decision tree ensemble: Small heterogeneous is better than large homogeneous," in *Proc. 7th Int. Conf. Mach. Learn. Appl.*, Dec. 2008, pp. 900–905.

[57] D. Che, Q. Liu, K. Rasheed, and X. Tao, "Decision tree and ensemble learning algorithms with their applications in bioinformatics," *Softw. Tools Algorithms Biol. Syst.*, vol. 696, pp. 191–199, 2011.

[58] N. Manwani and P. S. Sastry, "Geometric decision tree," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 42, no. 1, pp. 181–192, Feb. 2012.

[59] W. Li and A. W. Moore, "A machine learning approach for efficient traffic classification," in *Proc. 15th Int. Symp. Modeling, Anal., Simulation Comput. Telecommun. Syst.*, Oct. 2007, pp. 310–317.

[60] A. Martins and R. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 1614–1623.

[61] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon *et al.*, Eds. Curran Associates, 2017, pp. 4765–4774.

[62] M. Sundararajan and A. Najmi, "The many Shapley values for model explanation," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2020, pp. 9269–9278.

[63] A. Ghorbani and J. Zou, "Data shapley: Equitable valuation of data for machine learning," in *Proc. Int. Conf. Mach. Learn.*, May 2019, pp. 2242–2251.

[64] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BS, Canada, vol. 32, Dec. 2019, pp. 8026–8037.

[65] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 10, pp. 2825–2830, Jul. 2017.

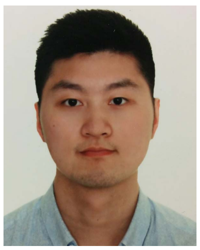
[66] D. Andreoletti, S. Troia, F. Musumeci, S. Giordano, G. Maier, and M. Tornatore, "Network traffic prediction based on diffusion convolutional recurrent neural networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2019, pp. 246–251.

[67] X. Ma, C. Ding, S. Luan, Y. Wang, and Y. Wang, "Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 9, pp. 2303–2310, Sep. 2017.

[68] B. Sun, T. Sun, and P. Jiao, "Spatio-temporal segmented traffic flow prediction with ANPRS data based on improved XGBoost," *J. Adv. Transp.*, vol. 2021, pp. 1–24, May 2021.

[69] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6467–6476.

[70] L. C. Bento, R. Parafita, and U. Nunes, "Intelligent traffic management at intersections supported by V2V and V2I communications," in *Proc. 15th Int. IEEE Conf. Intell. Transp. Syst.*, Sep. 2012, pp. 1495–1502.



Chen Sun received the B.Eng. degree in automation from the University of Electronic Science and Technology of China, Chengdu, China, in 2014, and the M.A.Sc. degree in electrical and computer engineering from the University of Toronto, Toronto, ON, Canada, in 2017. He is currently pursuing the Ph.D. degree in mechanical and mechatronics engineering with the University of Waterloo, Waterloo, ON.

He is a member of the Cognitive Autonomous Driving Laboratory (CogDrive), University of Waterloo, and supervised by Prof. Dongpu Cao and Prof. Amir Khajepour. His research interests include end-to-end autonomous driving, safety validation for cyber-physical systems, and planning and control for robots.

Shen Li received the Ph.D. degree from the University of Wisconsin–Madison, Madison, WI, USA, in 2018. He is currently a Research Associate with Tsinghua University, Beijing, China. His research interests include intelligent transportation systems (ITSs), architecture design of connected automated vehicle highway (CAVH) systems, vehicle-infrastructure cooperative planning and decision method, traffic data mining based on cellular data, and traffic operations and management.

Dongpu Cao received the Ph.D. degree in mechanical engineering from Concordia University, Montreal, QC, Canada, in 2008.

He is the Canada Research Chair in Driver Cognition and Automated Driving. He is currently an Associate Professor and the Director of the Waterloo Cognitive Autonomous Driving (CogDrive) Laboratory, University of Waterloo, Waterloo, ON, Canada. He has contributed more than 200 articles and three books. His current research focuses on driver cognition, automated driving, and cognitive autonomous driving.

Dr. Cao received the SAE Arch T. Colwell Merit Award in 2012, IEEE VTS 2020 Best Vehicular Electronics Paper Award, and three best paper awards from the American Society of Mechanical Engineers (ASME) and IEEE conferences. He is an IEEE VTS Distinguished Lecturer.

Fei-Yue Wang (Fellow, IEEE) received the Ph.D. degree in computer and systems engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990.

In 1990, he joined The University of Arizona, Tucson, AZ, USA, where he became a Professor and the Director of the Robotics and Automation Laboratory and the Program in Advanced Research for Complex Systems. In 1999, he founded the Intelligent Control and Systems Engineering Center, Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Overseas Chinese Talents Program from the State Planning Council and the 100 Talent Program from CAS. In 2002, he joined the Laboratory of Complex Systems and Intelligence Science, CAS, as the Director, where he was the Vice President for Research, Education, and Academic Exchanges with the Institute of Automation from 2006 to 2010. In 2011, he was named the State Specially Appointed Expert and the Director of the State Key Laboratory for Management and Control of Complex Systems, Beijing. His current research interests include methods and applications for parallel systems, social computing, parallel intelligence, and knowledge automation.

Dr. Wang was elected as a fellow of International Council on Systems Engineering (INCOSE), International Federation of Automatic Control (IFAC), American Society of Mechanical Engineers (ASME) and American Association for the Advancement of Science (AAAS). In 2007, he was a recipient of the National Prize in Natural Sciences of China and the Outstanding Scientist by ACM for his research contributions in intelligent control and social computing. He was a recipient of the IEEE Intelligent Transportation Systems (ITS) Outstanding Application and Research Awards in 2009, 2011, and 2015 and the IEEE SMC Norbert Wiener Award in 2014. He was the Chair of the IFAC TC on Economic and Social Systems from 2008 to 2011. He was the President of the IEEE ITS Society from 2005 to 2007, the Chinese Association for Science and Technology, USA, in 2005, and the American Zhu Kezhen Education Foundation from 2007 to 2008. He was the Vice President of the ACM China Council from 2010 to 2011. He is currently the President of the IEEE Council on Radio Frequency Identification. Since 2008, he has been the Vice President and the Secretary-General of the Chinese Association of Automation.

Amir Khajepour (Member, IEEE) received the B.S. degree from Ferdowsi University, Mashhad, Iran, in 1990, the M.S. degree from the Sharif University of Technology, Tehran, Iran, in 1992, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 1996.

He is currently a Professor of mechanical and mechatronics engineering with the University of Waterloo, where he is also the Canada Research Chair in Mechatronic Vehicle Systems. He has developed an extensive research program that applies his expertise in several key multidisciplinary areas. He has authored over 350 journal and conference publications and five books. His research interests include system modeling and control of dynamic systems. His research has resulted in several patents and technology transfers.

Prof. Khajepour is a fellow of The Engineering Institute of Canada, The American Society of Mechanical Engineers, and The Canadian Society of Mechanical Engineering.