DS535 Recommender System and Graph Machine Learning

Group 4 Final Project Presentation 2023. 12. 05

Traffic Prediction

Forecasting **Real-time Traffic Information**

Based on **Historical Traffic Data**

Gwangwoo Kim 20235472

Joohyun Cho 20234431

Minkyung Choi 20234447

Content

1.1 Why Traffic Prediction?

- Traffic forecasting is crucial **to modern urban planning and transportation management**
- Efficient traffic flow is essential for **ensuring mobility, reducing congestion, and enhancing overall quality of life**
- To achieve these objectives, accurate and timely traffic predictions are very necessary

Staged Approaches for Digital Road Infrastructure Development

	Short-Term (2021-2025)	Medium-to-Long-Term (2026-2030)
Establishment of Digital Road Networks	Establishment of Major Road Communication Infrastructure Construction of Security Authentication System Development of IoT Communication Network	Expansion of Communication Infrastructure Construction Development of Universal Communication Platforms
Deployment of Smart Signals	Establishment of Smart Intersections and Responsive Signals	Expansion of Smart Intersection and Responsive Signal Construction
Traffic Flow Optimization	-	Development of Traffic Flow Optimization Technology Automated Variable Traffic Flow Separation Lane-Specific Convoy Driving Management Real-Time Signal Optimization for City-Wide Network
Assistance for Safe Driving	Establishment of ITS Service Roadmap Development of Edge-Type Peripheral Infrastructure Development of Guiding Technologies Supporting Negotiations and Coordination between Vehicles	Development of Guiding Technologies Supporting Negotiations and Coordination between Vehicles
Driving Implementer	Coordination with the Ministry of Land, Infrastructure and Transport, Local Infrastructure Management Agencies, Local Autonomous Bodies, and the Police Agency	

Source: Ministry of Land, Infrastructure and Transport

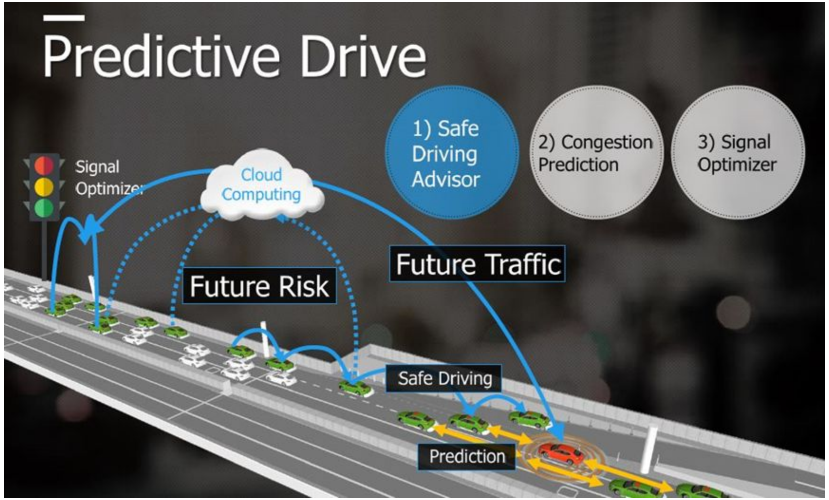
1.2 Related Work and Objective

Limitation of previous work

- Tree-based models are inadequate for applying regression and predicting continuous values
- RNNs suffer from time-consuming iterations and gradient vanishing or explosion problems with long sequences
- CNNs can only be applied to Euclidean data

Significance of GNN

- Capture complex relationships, long-range connections, and hierarchical structures within non-Euclidean data
- Effective in various node-level, edge-level, and graph-level prediction tasks



2.1 Data Description

- "Jeju Island Road Traffic Volume Prediction AI Competition"
- Organized by Jeju Technopark and the Jeju Special Self-Governing Province, managed by DACON

	id	base_date	day_of_week	base_hour	lane_count	road_rating	road_name	multi_linked	connect_code	maximum_speed_limit	vehicle_restricted
0	TRAIN_0000000	20220623	목	17	1	106	지방도1112호선	0	0	60	0
1	TRAIN_0000001	20220728	목	21	2	103	일반국도11호선	0	0	60	0
2	TRAIN_0000002	20211010	일	7	2	103	일반국도16호선	0	0	80	0
3	TRAIN_0000003	20220311	금	13	2	107	태평로	0	0	50	0
4	TRAIN_0000004	20211005	화	8	2	103	일반국도12호선	0	0	80	0
...	...	...	...	...	...	...	...	...	...	...	...
4701212	TRAIN_4701212	20211104	목	16	1	107	-	0	0	50	0
4701213	TRAIN_4701213	20220331	목	2	2	107	-	0	0	80	0
4701214	TRAIN_4701214	20220613	월	22	2	103	일반국도12호선	0	0	60	0
4701215	TRAIN_4701215	20211020	수	2	2	103	일반국도95호선	0	0	80	0
4701216	TRAIN_4701216	20211019	화	6	2	107	경찰로	0	0	60	0

	weight_restricted	height_restricted	road_type	start_node_name	start_latitude	start_longitude	start_turn_restricted	end_node_name	end_latitude	end_longitude	end_turn_restricted	target
0	32400	0	3	제3교래교	33.427747	126.662612	없음	제3교래교	33.427749	126.662335	없음	52
1	0	0	0	광양사거리	33.50073	126.529107	있음	KAL사거리	33.504811	126.52624	없음	30
2	0	0	0	창고천교	33.279145	126.368598	없음	상창육교	33.280072	126.362147	없음	61
3	0	0	0	남양리조트	33.246081	126.567204	없음	서현주택	33.245565	126.566228	없음	20
4	0	0	0	애월샷시	33.462214	126.326551	없음	애월입구	33.462677	126.330152	없음	38
...	...	...	...	...	...	...	...	...	...	...	...	...
4701212	0	0	0	대림사거리	33.422145	126.278125	없음	금덕해운	33.420955	126.27375	없음	20
4701213	43200	0	3	광삼교	33.472505	126.424368	없음	광삼교	33.472525	126.42489	없음	65
4701214	0	0	0	고성교차로	33.447183	126.912579	없음	성산교차로	33.444121	126.912948	없음	30
4701215	0	0	0	제6광령교	33.443596	126.431817	없음	관광대학입구	33.444996	126.433332	없음	73
4701216	0	0	0	서귀포경찰서	33.256785	126.50894	없음	시민공원	33.25713	126.510364	없음	35

2.2 Preprocessing for Graph Structure

The definition of edge relationships between nodes:

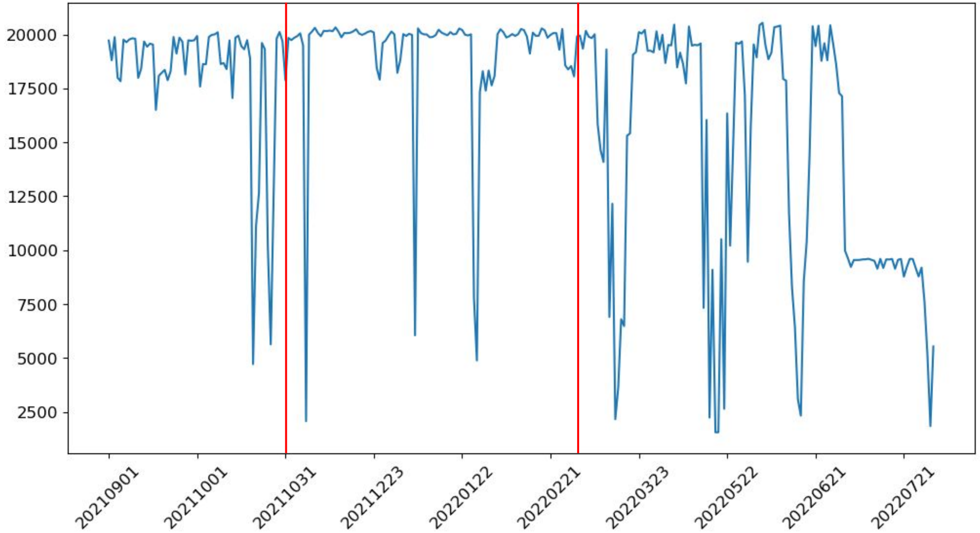
$$v_i = e^i_{start,end} \sim v_j = e^j_{start,end} \quad \text{if and only if}$$
$$e^i_{start} = e^j_{end} \quad \text{or} \quad e^i_{end} = e^j_{start},$$

The method described above has three remarks:

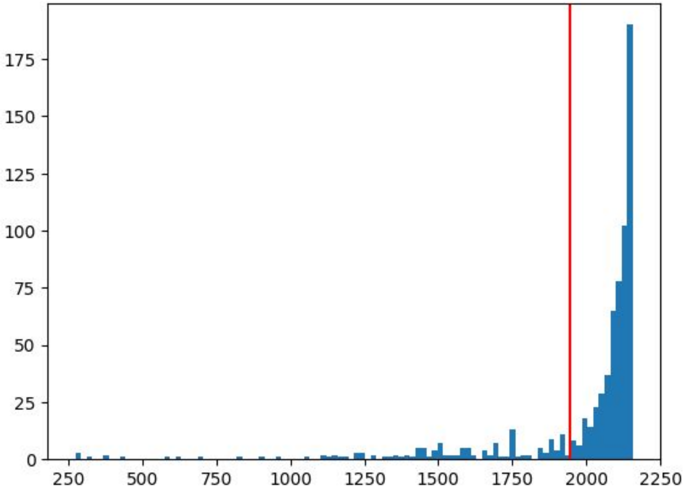
1. The resulting graph is directional in general
2. Self-loops or multiple edges could be created
3. Some nodes could be isolated, leading to a non-connected graph

Removed

2.2 Preprocessing for Graph Structure



Window from 2021.12 to 2022.02
to minimize missing values



The histogram of the number of the data within 3 months
The red vertical line is x = 24 * 90 * (9/10) = 1944

2.2 Preprocessing for Graph Structure

Statistics of DACON Jeju Traffic Graph

	Amount	Meaning
Nodes	564	Road
Edges	1426	Intersection
Node Feature Dimensions	10	Node features used
Targets	2160	Speeds per hour per day

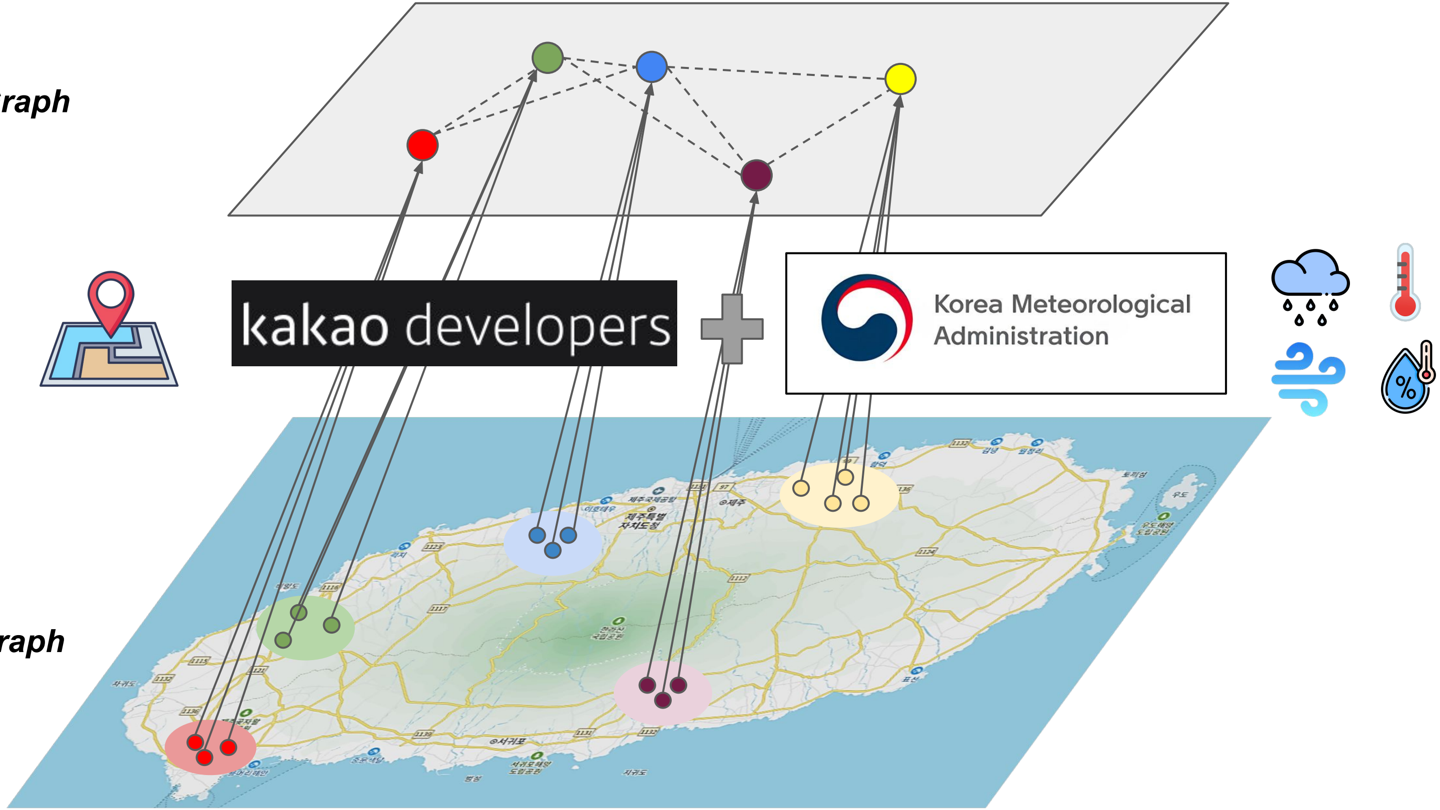
2.2 Preprocessing for Graph Structure

Administrative Area Graph

"Pyoseon-myeon"
"Pyoseon-myeon"

DACON Jeju Traffic Graph

(0.204418, 0.831966)
(0.379183, 0.779346)



2.2 Preprocessing for Graph Structure

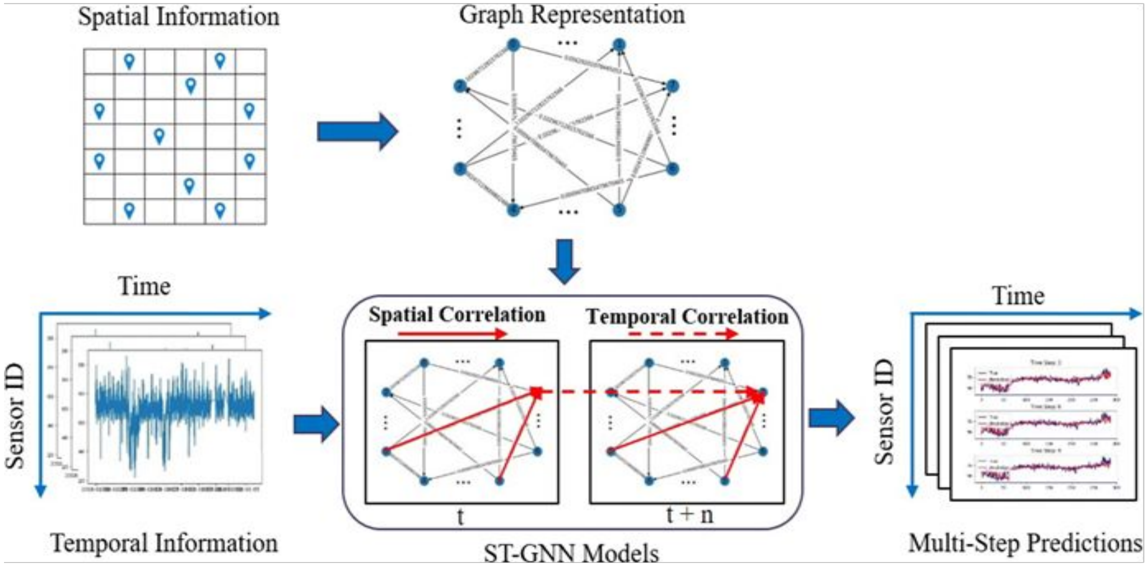
Statistics of Administrative Area Graph

	Amount	Meaning
Nodes	33	Administrative area
Edges	108	Road between two areas
Node Feature Dimensions	4	Node features used
Targets	2160	Made by the previous graph

3.1 Preliminary

Spatio-temporal Graph Neural Network(ST-GNN)

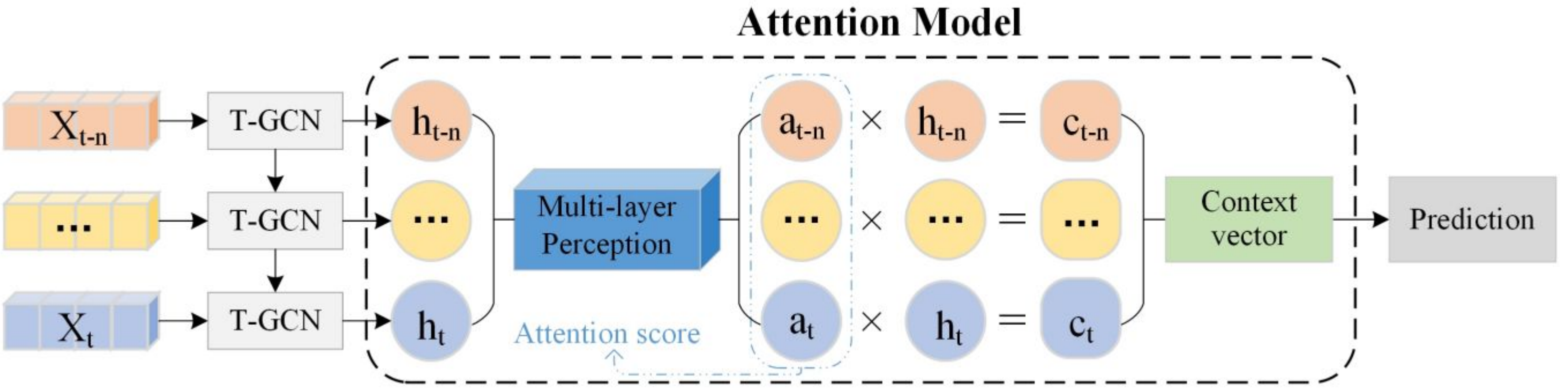
- Aims to model **dynamic node-level inputs** given the graph structure
- Can **integrate** GCN into RNN or into CNN, even attention method



3.1 Preliminary

Attention Temporal Graph Convolutional Network(A3T-GCN)

- Integrates **graph convolutional layers** tailored for handling **graph-structured data**
- Incorporates **temporal dependencies** to capture **sequential patterns** within the data
- Utilizes **attention mechanisms** to prioritize and **weigh important elements** in the data



3.2 Proposed Method

Challenge : Heterogeneous data, data quality

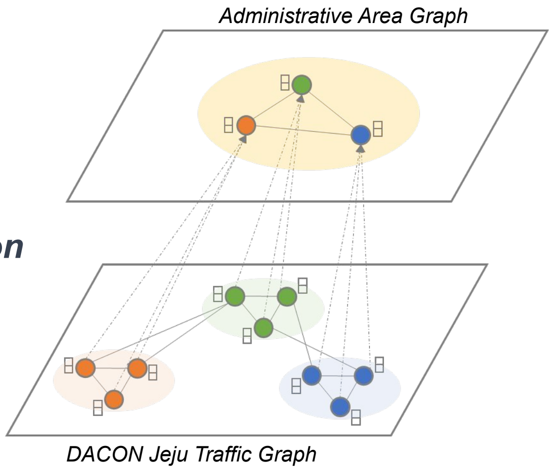
- Traffic prediction problems involve both **spatio-temporal data and external factors**, ***e.g. hierarchical administrative structure, weather, calendar information***

3.2 Proposed Method

Challenge : Heterogeneous data, data quality

- Traffic prediction problems involve both **spatio-temporal data and external factors**, ***e.g. hierarchical administrative structure, weather, calendar information***

A hierarchical structure by gathering administrative area information



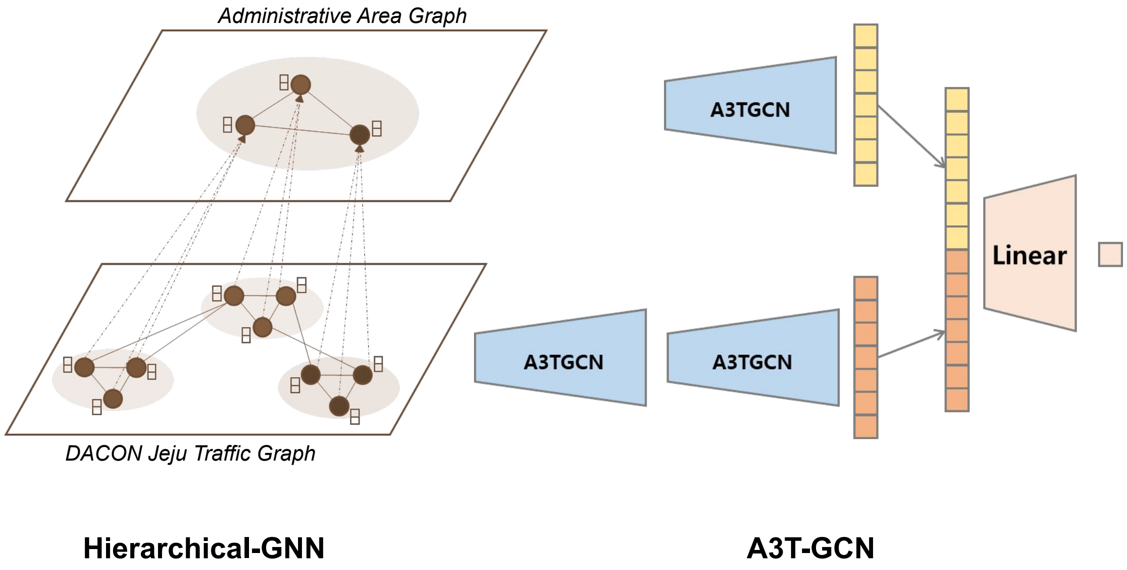
Hierarchical-GNN

3.2 Proposed Method

Challenge : Heterogeneous data, data quality

- Traffic prediction problems involve both **spatio-temporal data and external factors**, ***e.g. hierarchical administrative structure, weather, calendar information***

Each one was estimated using A3T-GCN and merged for the final inference



3.3 Formulation

Definition: Structure of `DACON Jeju Traffic Graph` and `Administrative Area Graph` at time t

$$G_t = (V_t, E_t, X_t, y_t)$$

- V denotes vertices(nodes)
- E denotes edges
- X denotes node features
- y denotes target speeds of the data

$$G_t^H = (V_t^H, E_t^H, X_t^H, y_t)$$

- H denotes hierarchical data

3.3 Formulation

Definition: Forward process of the proposed model

$$h_1 = \sigma(f_{\theta_1}(G_{t-n:t}))$$
$$h_2 = \sigma(f_{\theta_2}(h_1))$$
$$h_3 = \sigma(f_{\theta_3}(G^H_{t-n:t}))$$

- t-n:t denotes n graphs spanning from time t-n to t-1
  - ** \`$n$\` is a hyperparameter representing how many past time steps to consider*
  - ** \`$n$=24\` was chosen*
- f denotes A3T-GCN module
- θ: denotes neural network parameter
- σ denotes activation function

$$g_{\theta_4}([h_2, h_3])$$

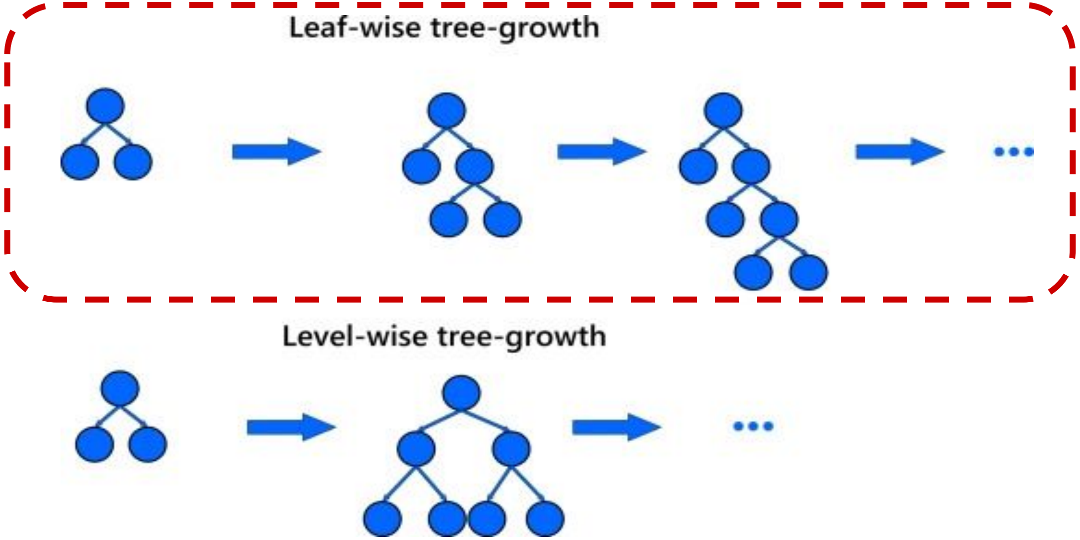
- g denotes fully-connected layer
- $[h_2, h_3]$ denotes concatenation
- The final prediction

$$L(\theta_1, \theta_2, \theta_3, \theta_4) = \|y_t - g(G_{t-n:t}, G^H_{t-n:t})\|^2$$

- Training by minimizing MSE loss with Adam optimaizer

4.1 Baseline: LightGBM

- **Light Gradient Boosting Machine(LightGBM)** is a **tree-based gradient boosting** framework
- **LightGBM** uses **leaf-wise growth** to make the tree **vertically**
- To expand, it selects the leaf **with the maximum delta loss,** reducing more loss than level-wise algorithms



4.2 Comparative Results

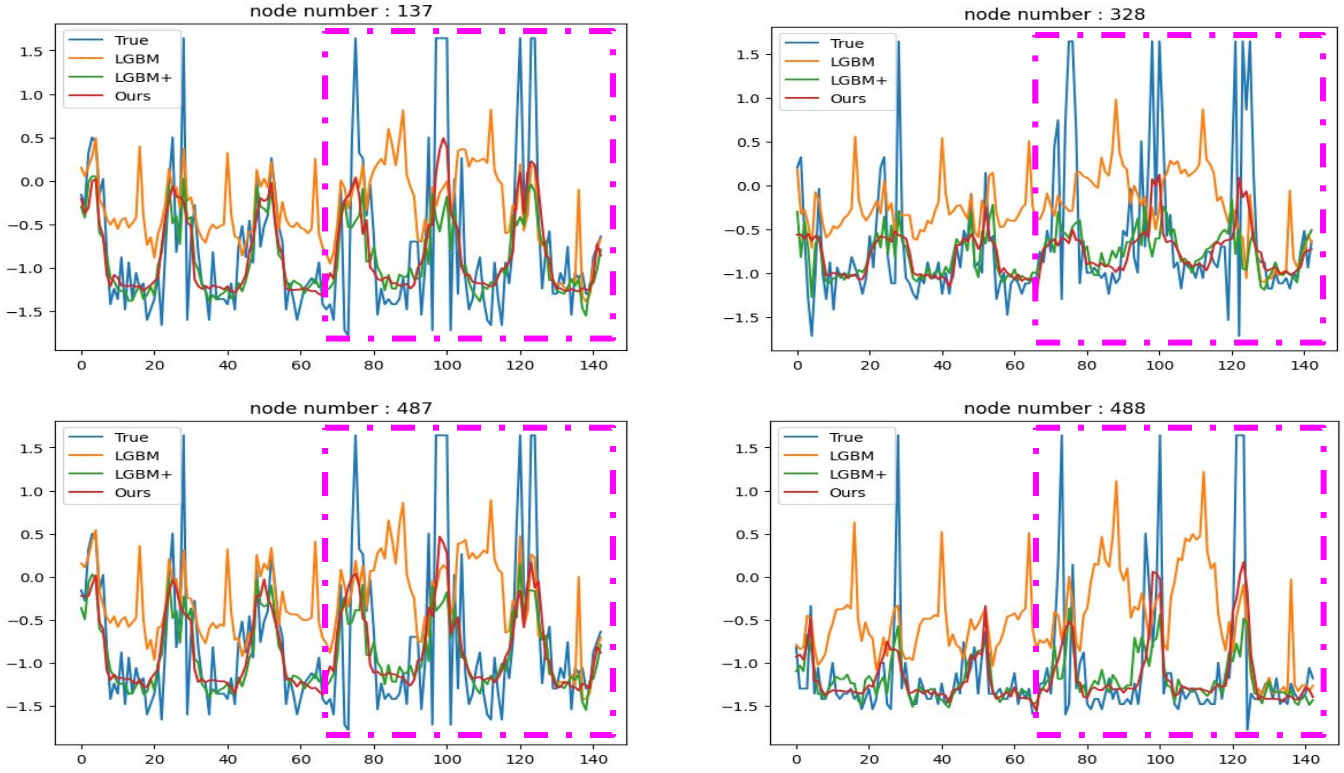
Comparison between LightGBM and Proposed Model

	LGBM	LGBM+	Ours
MAE	0.426	**0.224**	<u>0.270</u>
MSE	0.325	**0.124**	<u>0.157</u>
R^2 Score	0.649	**0.854**	<u>0.815</u>

- *`LGBM` denotes LightGBM vanilla model*
- *`LGBM+` denotes DACON competition winner's model*

4.2 Comparative Results

- Our model aimed to capture segments with peak blue, allowing for **better fitting of data** showing speed changes



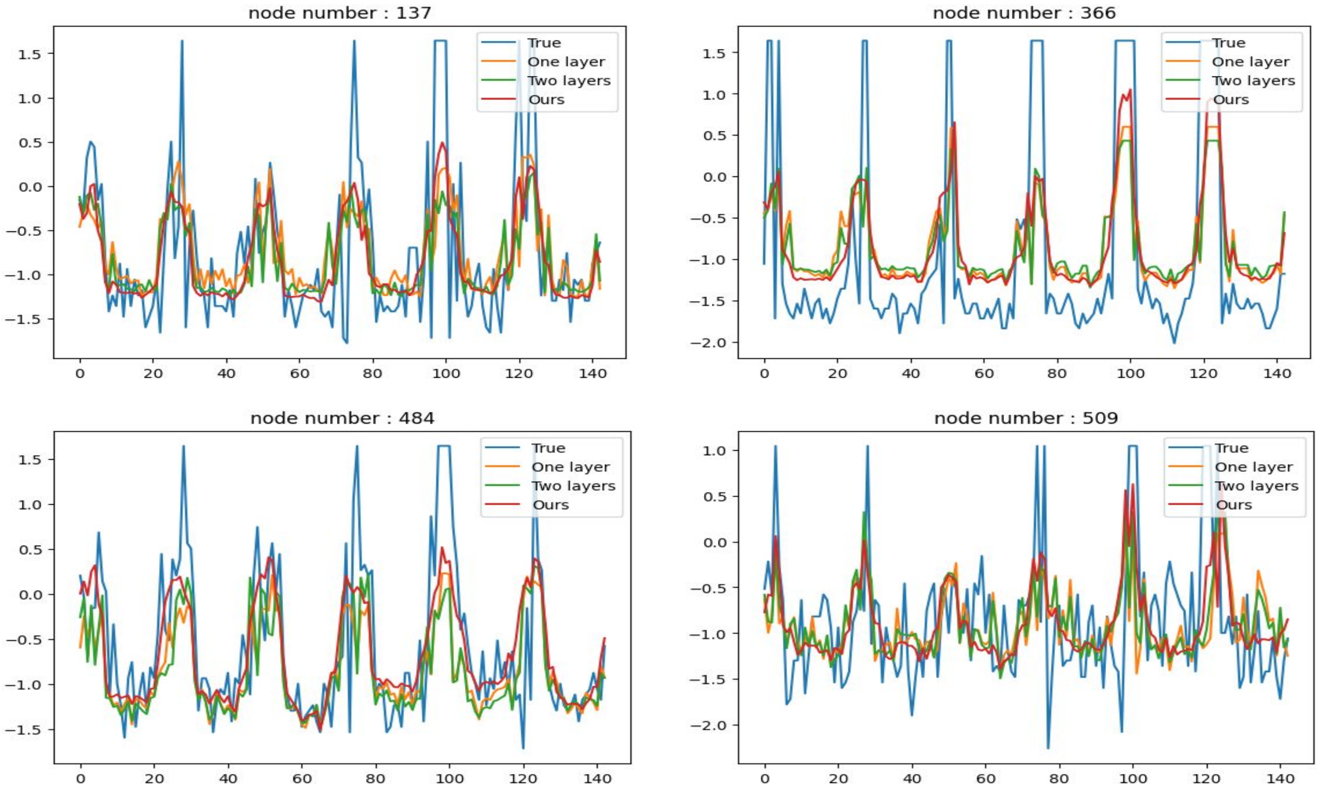
4.3 Ablation Study

	One Layer	Two Layers	Ours
MAE	0.314	0.277	**0.270**
MSE	0.199	0.169	**0.157**
R^2 Score	0.769	0.803	**0.815**

- *`One Layer` denotes structuring the A3T-GCN model with one layer*
- *`Two Layers` denotes structuring the A3T-GCN model with two layers*
- *`Ours` denotes structuring `Two Layers` with Hierarchical Information*

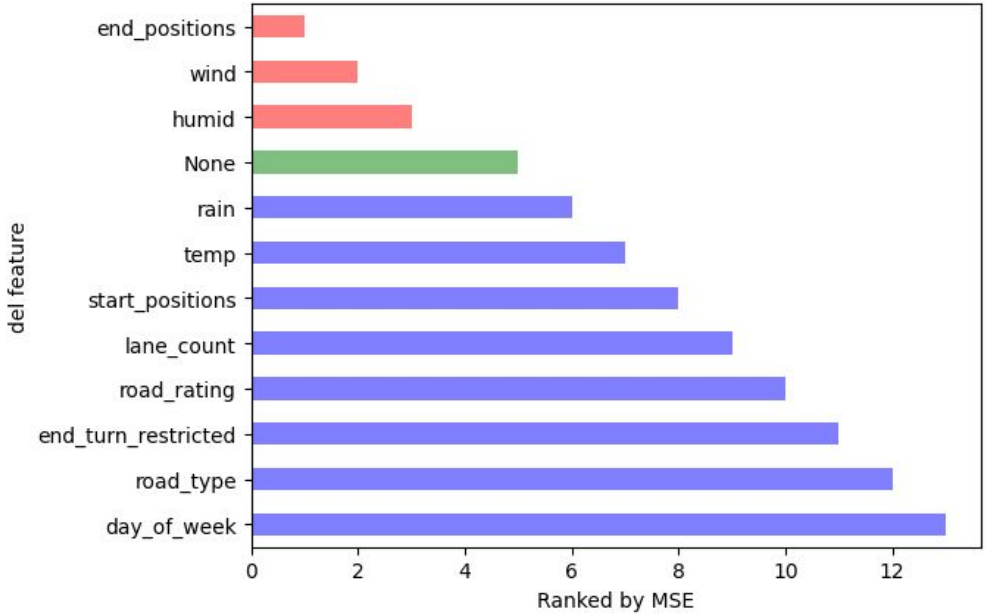
4.3 Ablation Study

- The adoption of **a hierarchical structure** led to **an improvement in performance**



4.4 Feature Importance Study

- We removed each feature at a time and sorted model performance **based on MSE**
- **Green 'None'** signifies the standard model **without omitting any feature**
- As the MSE is sorted in ascending order, if removing a certain feature resulted in a lower loss, it implies that the model performs better without this feature. **Thus, this feature is not significant (the red ones)**



5.1 Conclusion and Future Work

Conclusion:

1. Proposed a model combining temporal GNN and hierarchical information
2. Highlighted the significance of utilizing hierarchical information through ablation and feature importance studies
3. Identified robustness to abrupt changes compared to tree-based models

Future Work:

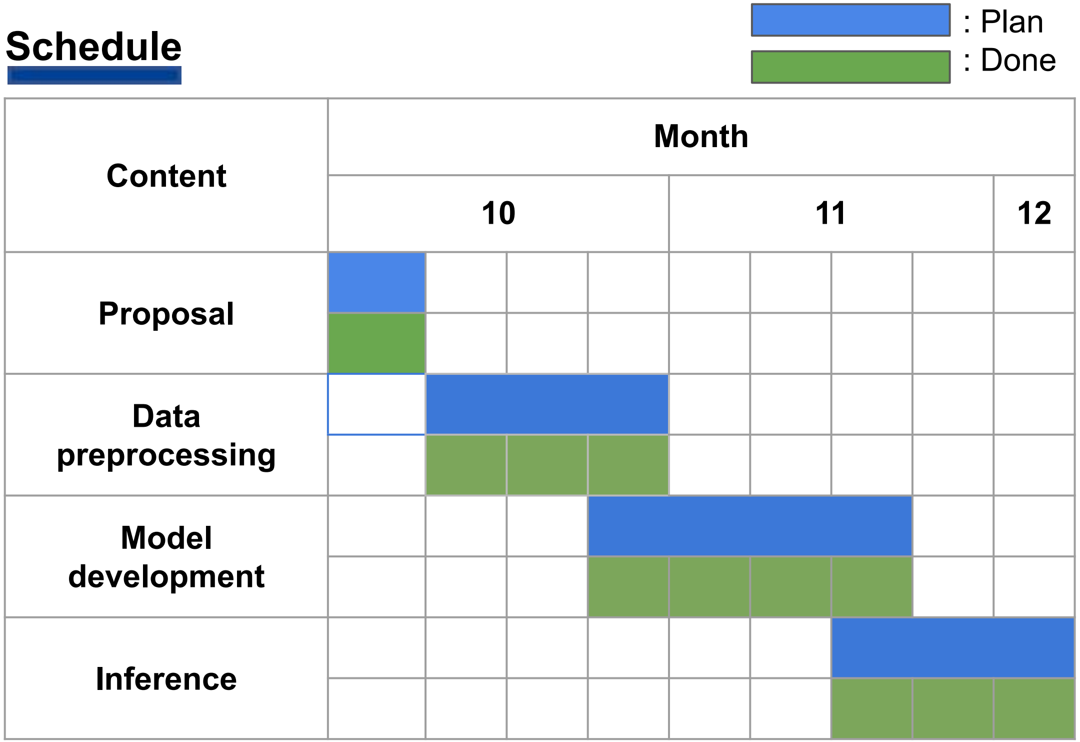
1. Solely relied on a basic MSE loss function

   → ***Some research indicates the suitability of Dynamic Time Warping Loss or DILATE(DIstortion Loss including shApe and TimE) for deep learning-based time series forecasting***
2. The interference from the end position, as revealed in experiments, is counterintuitive.

   → ***Attempts could involve creating surrogate variables using the difference between start and end positions***
3. Further investigation is required to determine if the model structure is well-suited for the data.

Schedule & Role

Schedule



Role

Task	Person
Background Research	Joohyun Cho
Data Processing & Graph Generation	Joohyun Cho, Minkyung Choi
Model Development	Gwangwoo Kim
Inference	Gwangwoo Kim, Joohyun Cho
Presentation	Minkyung Choi

Reference

[1] 제주도 도로 교통량 예측 AI 경진대회, DACON, https://dacon.io/competitions/official/235985/

[2] “초단기실황", 기상자료개방포털, https://data.kma.go.kr/data/rmt/rmtList.do?code=400&pgmNo=570

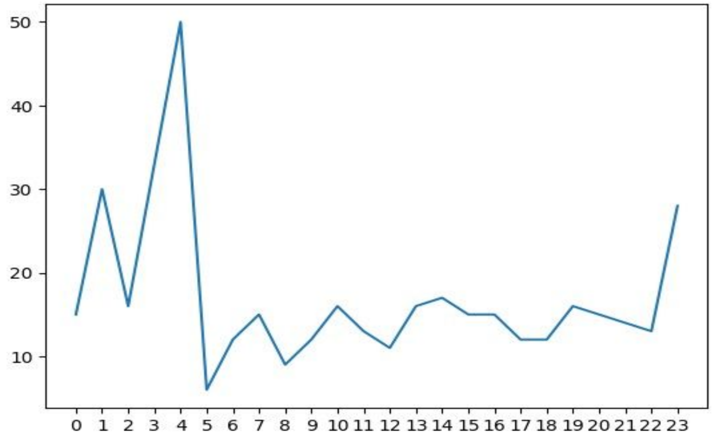
[3] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

[4] Bai, J., Zhu, J., Song, Y., Zhao, L., Hou, Z., Du, R., & Li, H. (2021). A3t-gcn: Attention temporal graph convolutional network for traffic forecasting. ISPRS International Journal of Geo-Information, 10(7), 485.

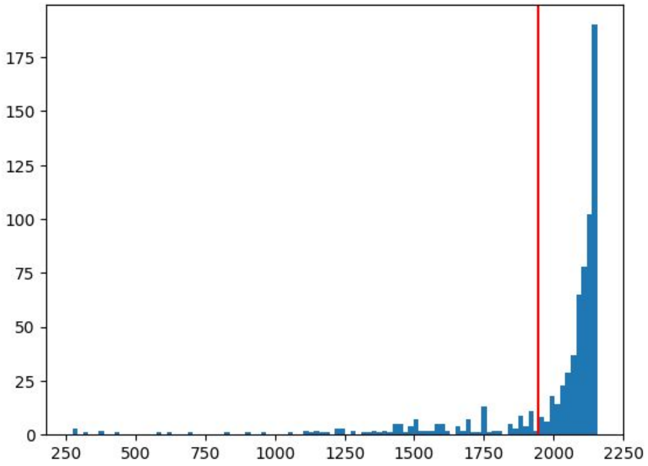
Thank You!

Appendix 1. Data Interpolation

- **Nodes with timestamp records exceeding 90% are selected, followed by interpolation**
- **Most missing values occur at dawn, where traffic volume is low, indicating the potential for faster driving, these gaps are interpolated using the `maximum_speed_limit`**



Target values of a region within 24 hours



The histogram of the number of the data within 3 months
The red vertical line is x = 24 * 90 * (9/10) = 1944

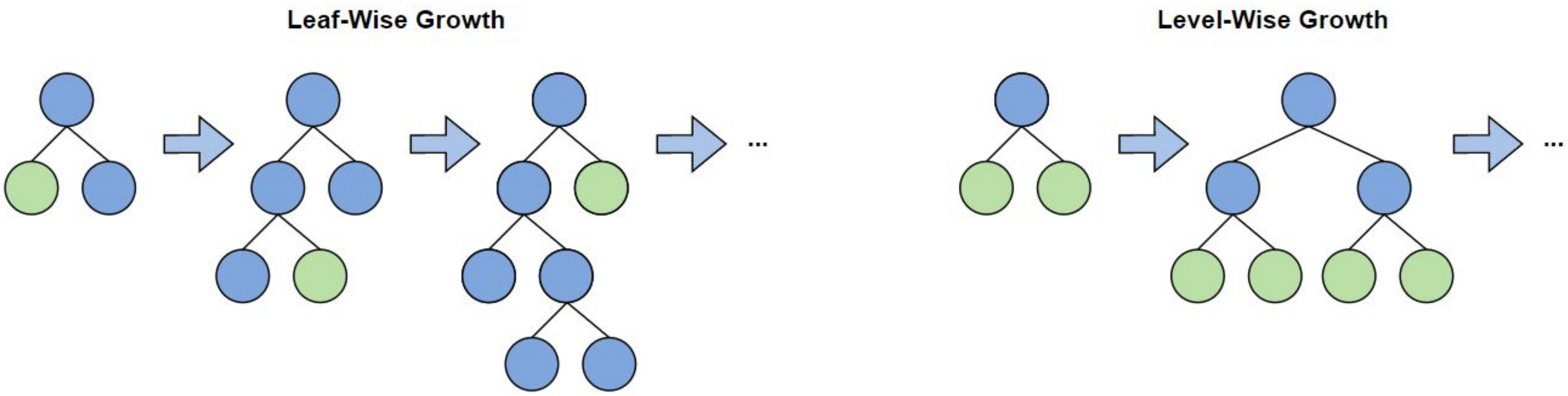
Appendix 2. Data Description

Korea Meteorological Administration

- "Jeju Island Road Traffic Volume Prediction AI Competition"
- Organized by Jeju Technopark and the Jeju Special Self-Governing Province, managed by DACON

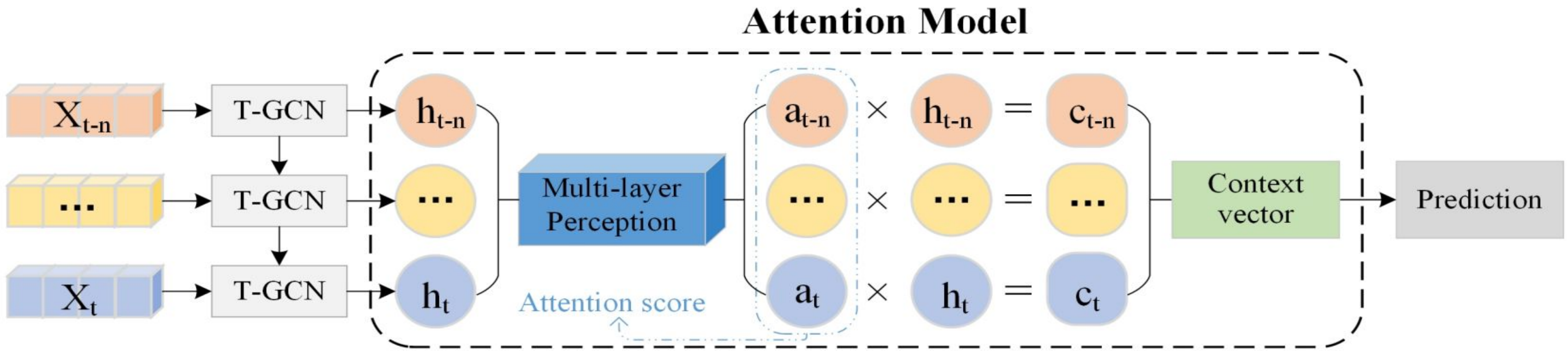
Spatial Info	Temporal Info	Not used
• lane_count: The number of lanes on the road • road_rating: The quality or rating of the road • road_type: The classification or type of the road • start_latitude: The latitude of the start node • start_longitude: The longitude of the start node • end_latitude: The latitude of the end node • end_longitude: The longitude of the end node • end_turn_restricted: Indicates if there are any restrictions on turning at the end node • start_node_name: The name of the start node • end_node_name: The name of the end node • start_address: Address based on start_node latitude and longitude • end_address: Address based on end_node latitude and longitude	• base_date: The date when the data was recorded • day_of_week: The day of the week corresponding to the base date • base_hour: The hour of the day when the data was recorded • target: Average speed (km/h) • rain: Rainfall recorded at the corresponding geographic location and date • temp: Temperature recorded at the corresponding geographic location and date • humid: Humidity level recorded at the corresponding geographic location and date • wind: Wind speed recorded at the corresponding geographic location and date	• maximum_speed_limit: The highest legal vehicle speed allowed on the road • id: An identification number for each record in the dataset • multi_linked: Indicates whether this segment of the road is shared by two or more routes • connect_code: A code that represents the road connection • weight_restricted: The maximum legal vehicle load allowed on the road • height_restricted: The maximum legal vehicle height allowed on the road • start_turn_restricted: Indicates if there are any restrictions on turning at the start node • road_name: The name of the road • vehicle_restricted: The type of vehicles restricted on the road

Appendix 3. LightGBM



- Light GBM expands trees in a vertical way, while other algorithms expand trees horizontally
- It selects a leaf with the maximum delta loss to expand
- The leaf-wise algorithm can reduce more loss than the level-wise algorithm

Appendix 4. A3T-GCN



1. T-GCN

   With n historical data, perform GCN(within graph) and GRU(within time series) to get spatiotemporal characteristic

2. Attention module

   Hidden states were inputted into the attention model to get the context vector.

3. Context vector

   Context vector is calculated by the weighted sum.

   Forecasting results were outputted using the fully connected layer.

Appendix 5. Hierarchical A3T-GCN

$$h_1 = \sigma(f_{\theta_1}(G_{t-n:t}))$$

$$h_2 = \sigma(f_{\theta_2}(h_1))$$

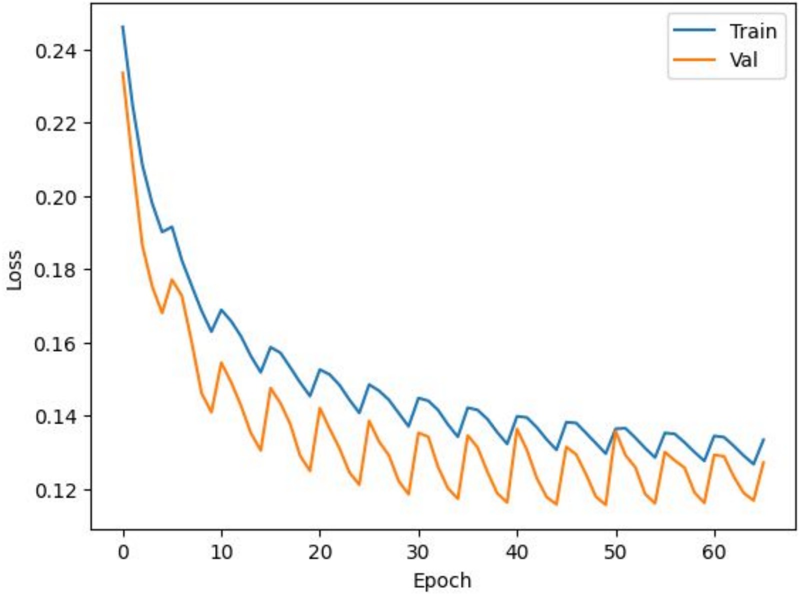
$$h_3 = \sigma(\boxed{f_{\theta_3}(G^H_{t-n:t})}) \Rightarrow \boxed{f_{\theta_3}([G^H_{t-n:t}, h_2])}$$

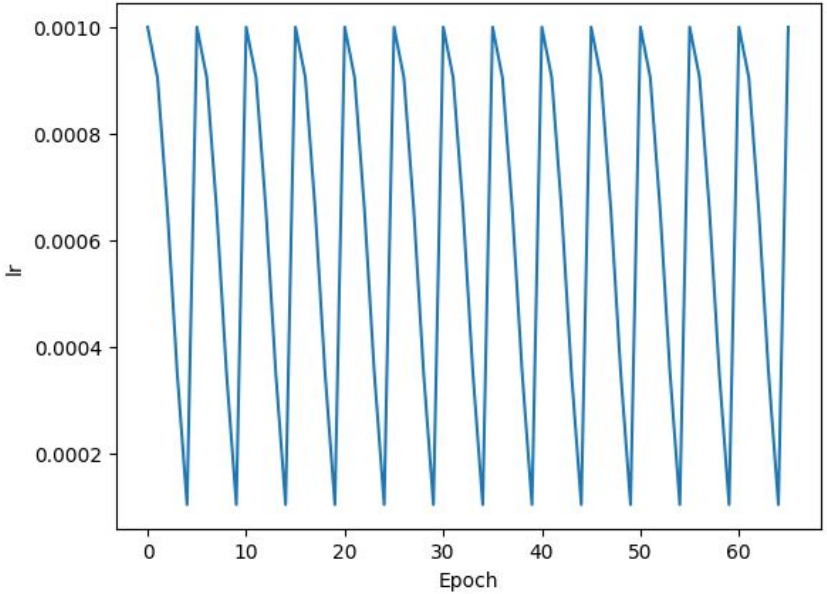
$G^H_{t-n:t}$ (33, 5, 24)

h_2 (564, 32) -> (33, 32)

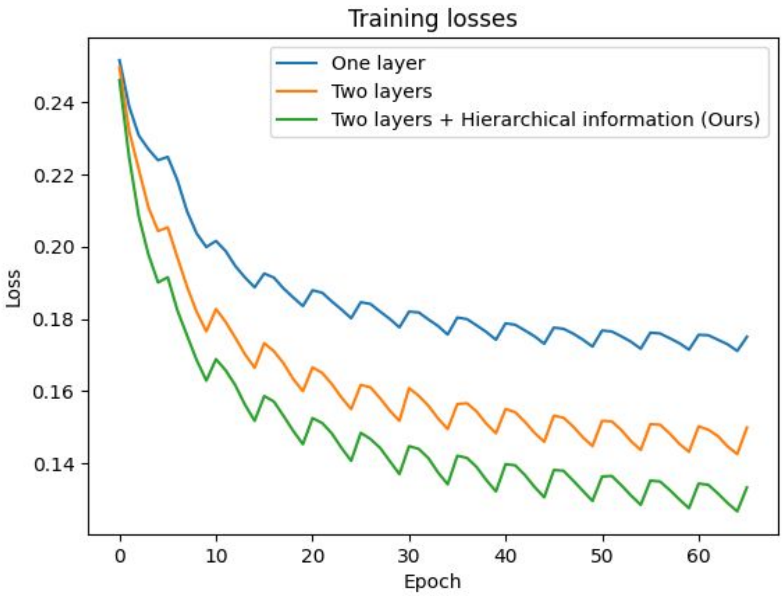
$\Rightarrow$ temporal graph (33, 37, 24)

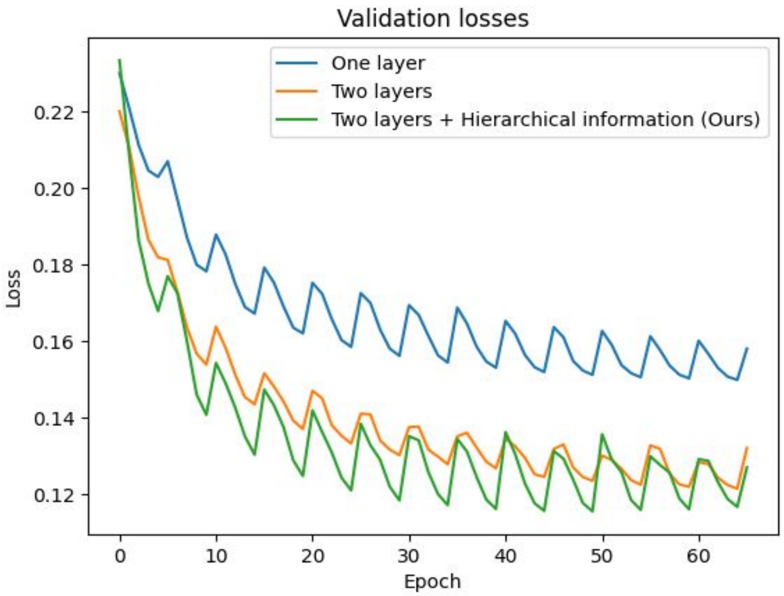
Appendix 6. Training Log





Appendix 7. Ablation Loss





Appendix 8. Feature Importance Study