Traffic Collision Prediction using Graph Neural Networks

Fatim Majumder

Motivation

- Los Angeles consistently **ranks among the most dangerous** cities in the United States to drive in.
- Notoriously **traffic congestion** at all hours of the day due to car dependency and poor public transport.
- **Gap** between ML research on traffic safety and civic engineers

Our goal is to identify the feasibility of using advanced data mining techniques to bridge this gap.

Related Works

Graph Neural Networks for Road Safety Modeling

(Nippani et al., NeurIPS, 2023) (arXiv)

- Creates **unified graph dataset** from accidents, traffic volume, and weather features over road networks of 7 states
- "...existing graph neural networks such as **GraphSAGE** [20] can predict the accident counts with **22% mean absolute error** (relative to actual counts) and whether an accident occurs on the road with over **87% AUROC**, averaged over eight states."

Research Questions

Can we...

1. **implement this paper on a novel dataset (city of LA?)**
2. **quantify the improvement over non-GNN methodology?**
3. **identify the features most correlated with accidents?**

Methodology

EDA and Accident Analysis

Data Mining and Pipeline

Data Pipeline Construction

GNN Model Selection

Graph Construction

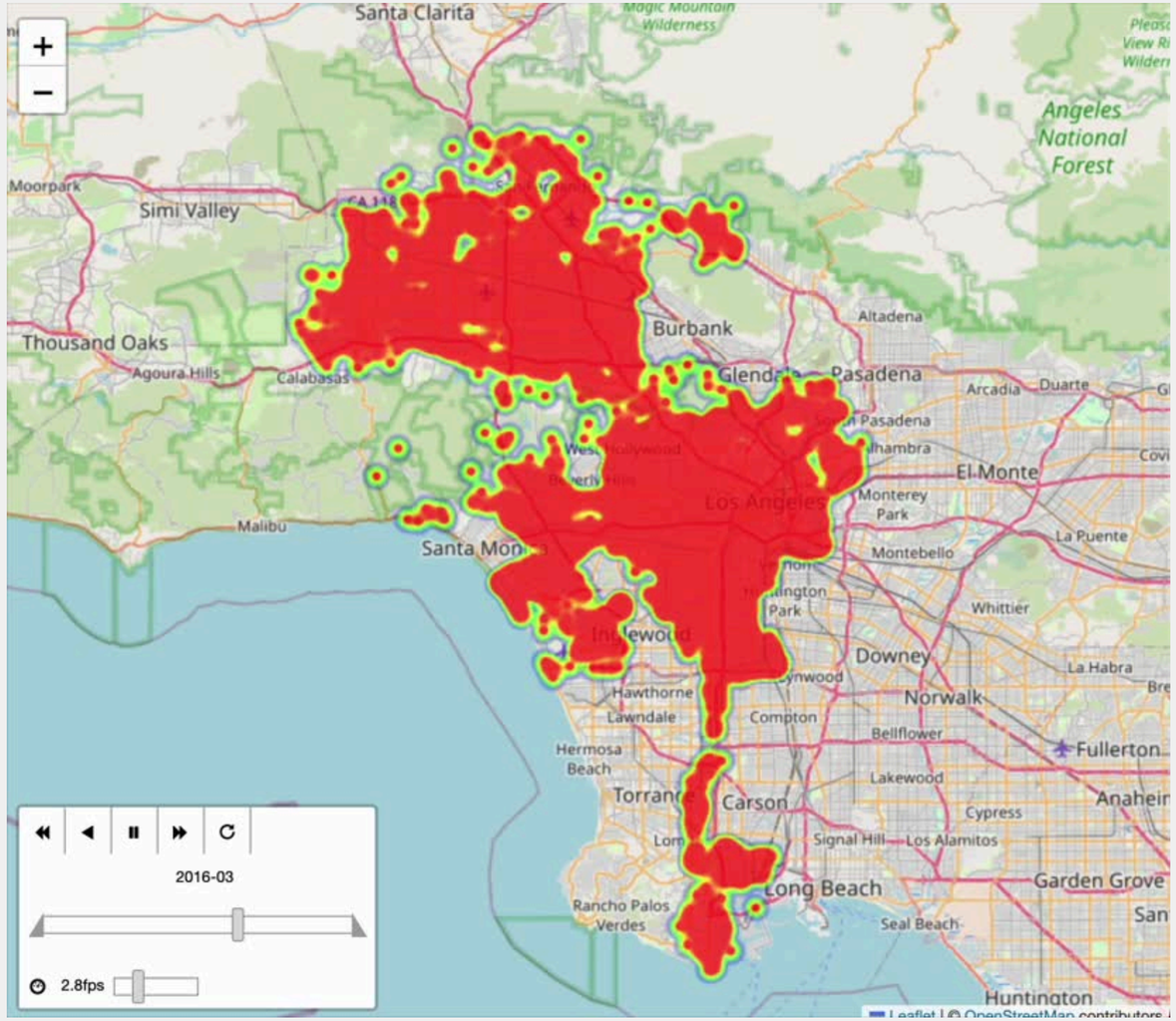
GNN Tuning

Evaluation

Initial EDA

Base accident dataset cleaning:

- 512k accident records
- **Remove** irrelevant features and missing values
- **Filter** by street collisions only
- Collision frequency drops off after 2022
    - lapse in police data tracking?
- Using only a heatmap of collisions offers **limited insights**.

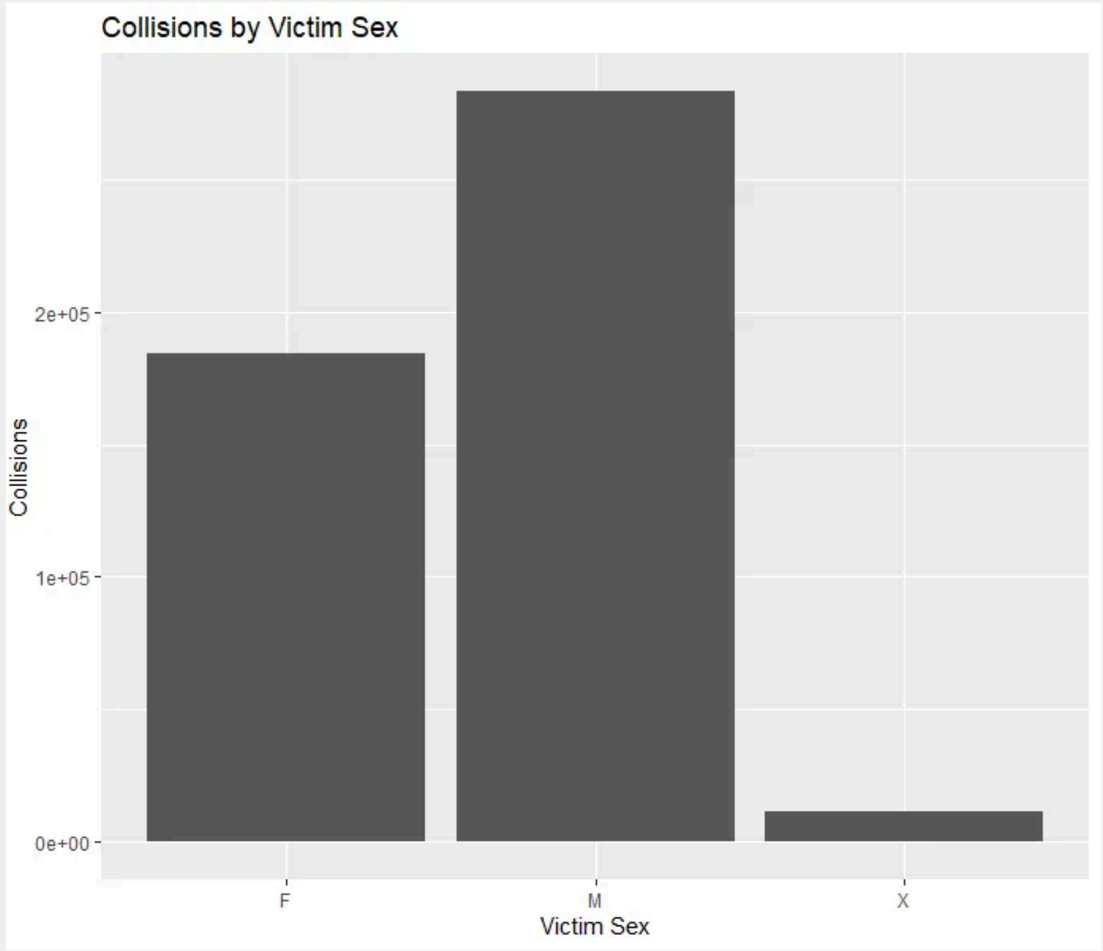


Gender Distribution

Higher exposure among males

- Men are **3x more likely** to commit traffic violations while driving and to be involved in accidents
- **Occupational exposure** (e.g., male-dominated industries involving driving) perhaps
- **Statistically** greater risk-taking behaviors? e.g speeding

How can we incorporate these **demographic patterns** into a model that captures the interactions involving this **driver behavior** to **predict high-risk collision** scenarios more accurately?

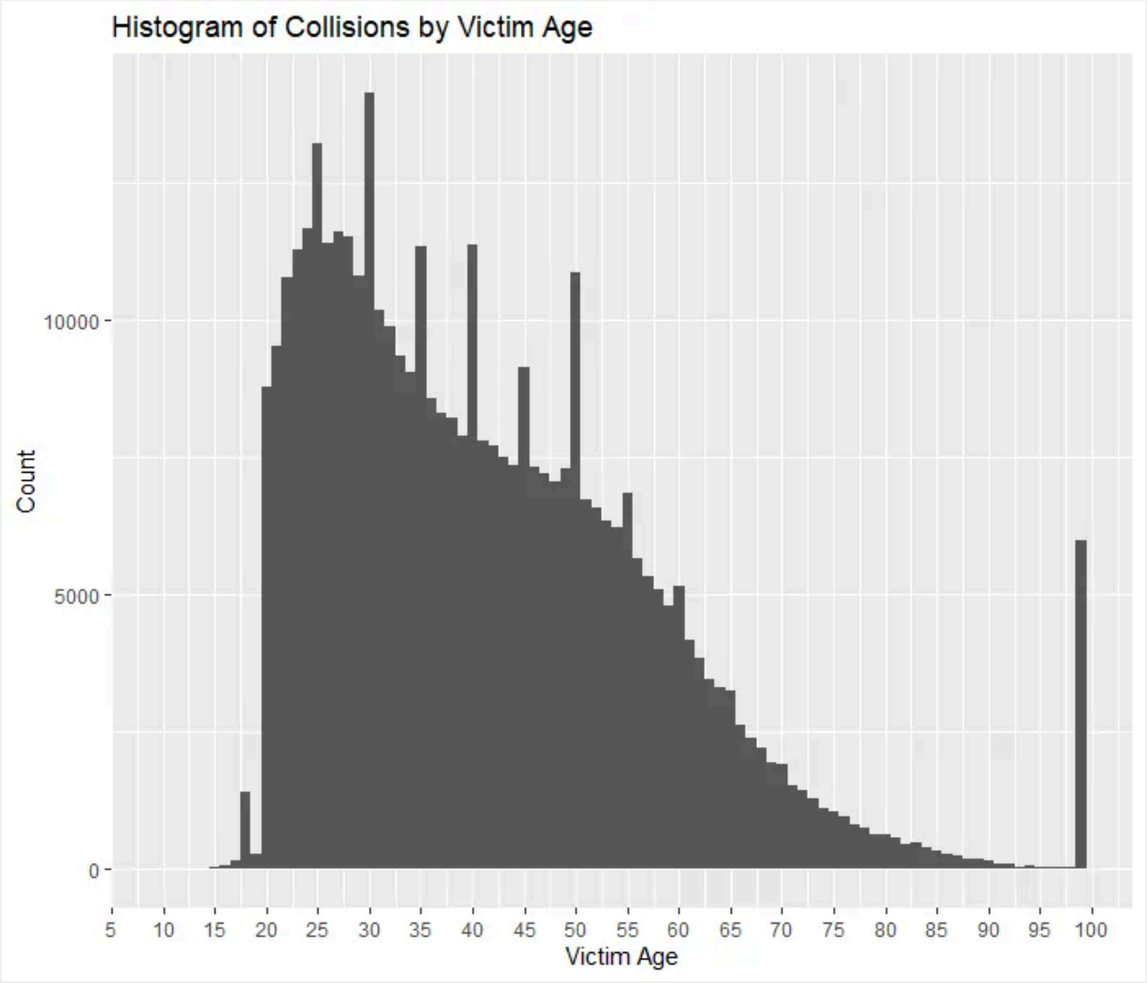


Victim Age

Children and Teens: fewer collisions involving younger individuals

Young Adults: higher exposure to collisions due to greater mobility and risk-taking behaviors typical of this age group.

Decline with Age: lower driving frequencies or risk tolerance among older populations.



Low-Collision Days

Why are certain holidays, like Presidents Day, consistently associated with **low-collision days**, while others, such as Independence Day, are not?

Could understanding the **intricate relationships** between holiday-specific travel patterns lead to developing more **effective predictive models** for collision risk?

date_occ	daily_count	day	hypothesis
2010-01-01	79	Saturday	New Year
2010-02-15	80	Monday	Presidents Day
2010-11-25	79	Thursday	Thanksgiving
2010-11-26	79	Friday	Thanksgiving
2010-12-28	82	Wednesday	Christmas / New Year
2012-02-20	84	Monday	Presidents Day
2012-12-31	82	Monday	New Years
2013-12-25	68	Tuesday	Christmas
2013-12-29	79	Sunday	Christmas / New Year
2014-01-21	81	Monday	MLK Day
2014-11-28	71	Thursday	Thanksgiving
2013-12-25	80	Wednesday	Christmas
2013-12-28	81	Saturday	Christmas / New Year
2013-12-29	86	Sunday	Christmas / New Year
2014-01-12	86	Sunday	Random Sunday
2019-01-21	89	Monday	MLK Day
2018-12-25	83	Tuesday	Christmas

Table 1: A table summarizing date occurrences, daily counts, day of the week, and hypothesis.

High-Collision Days

Increased traffic from end-of-week commutes, social outings, and early weekend travel.

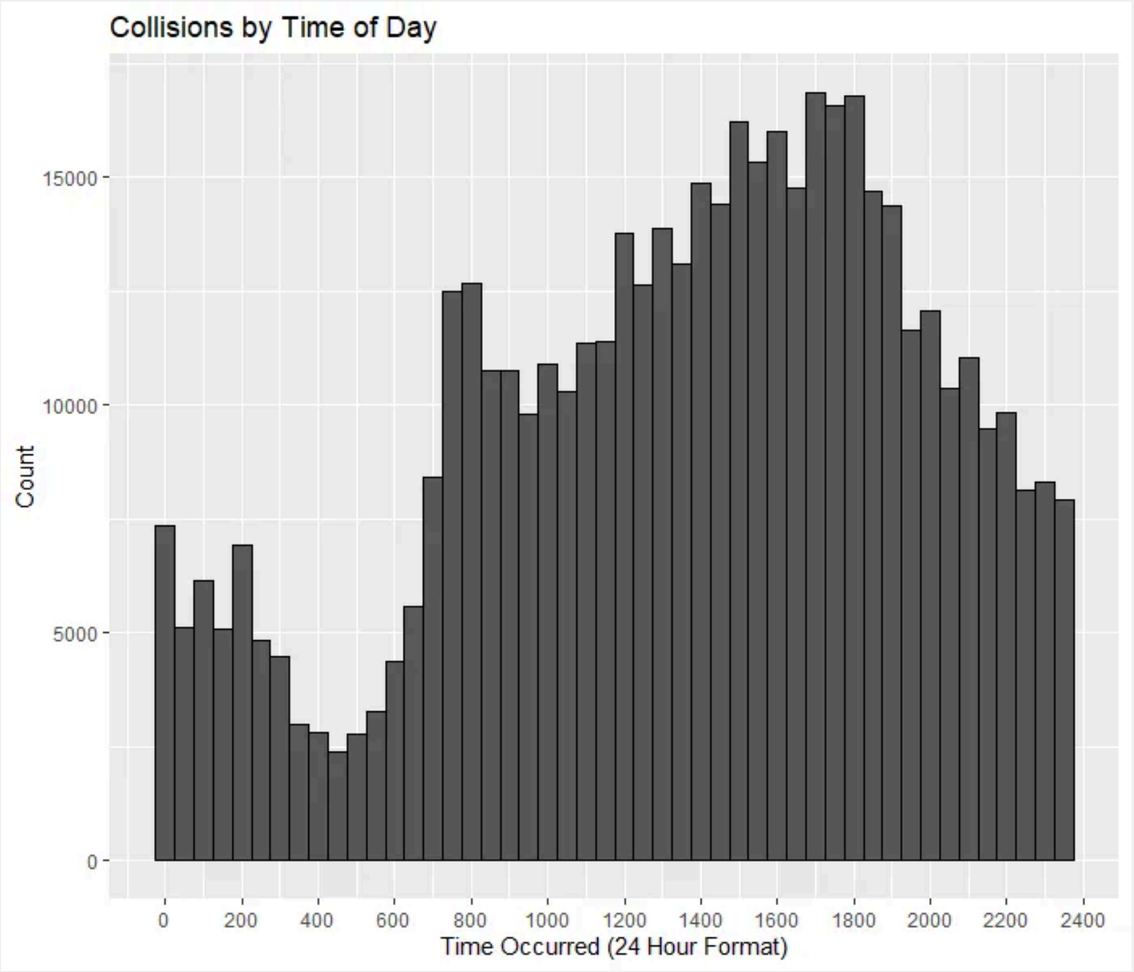
Notable high-collision day: December 15, 2017 (a Friday), which may reflect holiday shopping traffic or increased travel.

Would examining **weather conditions** on high-collision days also reveal underlying patterns, such as the **interplay** between weather events, and traffic density that contribute to **spikes in accidents**?

date_occ	daily_count	day
2010-02-05	202	Friday
2010-12-17	219	Friday
2015-10-09	213	Friday
2015-11-20	208	Friday
2016-02-17	218	Wednesday
2016-12-15	223	Thursday
2017-09-29	204	Friday
2017-10-20	202	Friday
2017-11-17	230	Friday
2018-01-08	212	Monday
2018-01-09	204	Tuesday
2018-02-16	217	Friday
2018-03-02	217	Friday
2018-10-12	216	Friday
2018-11-30	216	Friday
2018-12-07	207	Friday

Time of Collision

- **4 AM - 8 AM**
  - **Rapid spike** in traffic collisions as people head to work or school
  - Higher traffic density, hurried behavior, and potentially limited visibility during early morning hours
- **8 AM - 9:30 AM**
  - **Collisions dip** slightly as morning rush subsides and roads are less congested
- **9:30 AM - 6 PM**
  - **Steady increase** in collisions during the late morning and afternoon hours
  - **Peaks 5-6pm** (after-work rush hour)
- **6pm Onwards**
  - **Decline** after 6 PM and **sharply decrease** after 8 PM
  - Reduced traffic volume as commuters complete their evening travel

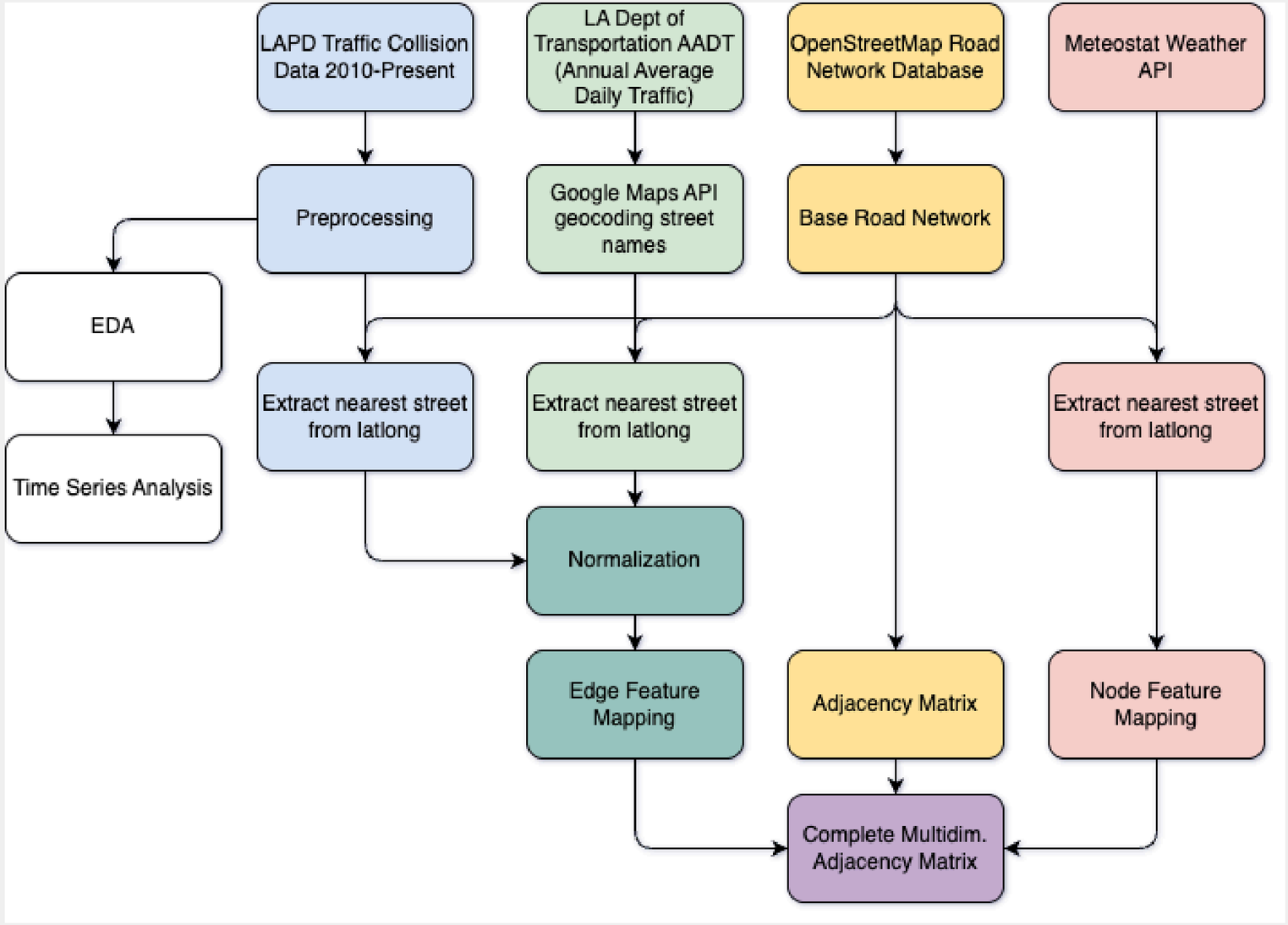


Data Sources and Pipeline

We created a **pipeline to extract and integrate** data from

- **OSM LA Road Network**
- **Collision Dataset**
- **LADOT Traffic Volume**
- **Meteostat Weather API**

to create a feature-rich for the GNN

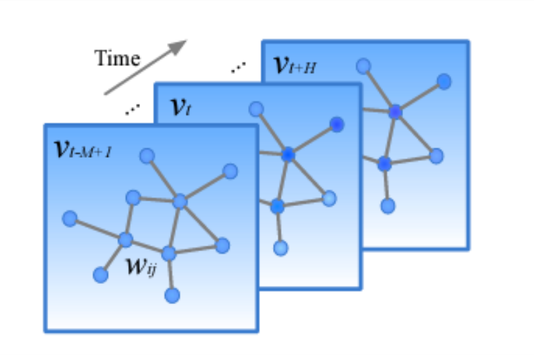


Why GNN?

- **Traditional** data structures cannot capture both **spatial and temporal patterns** in data
- Number of accidents is **dependent** on both **static and dynamic** node and edge features

STCGN (Spatio-Temporal Graph Convolutional Network) - Yu et. al 2017

- **Convolution layers** capture spatial relationships by applying **filters** to neighboring road segments and identifying patterns (e.g connectivity)
- **Temporal convolutional components** process **sequential data** to uncover time-based patterns (e.g seasonal trends)

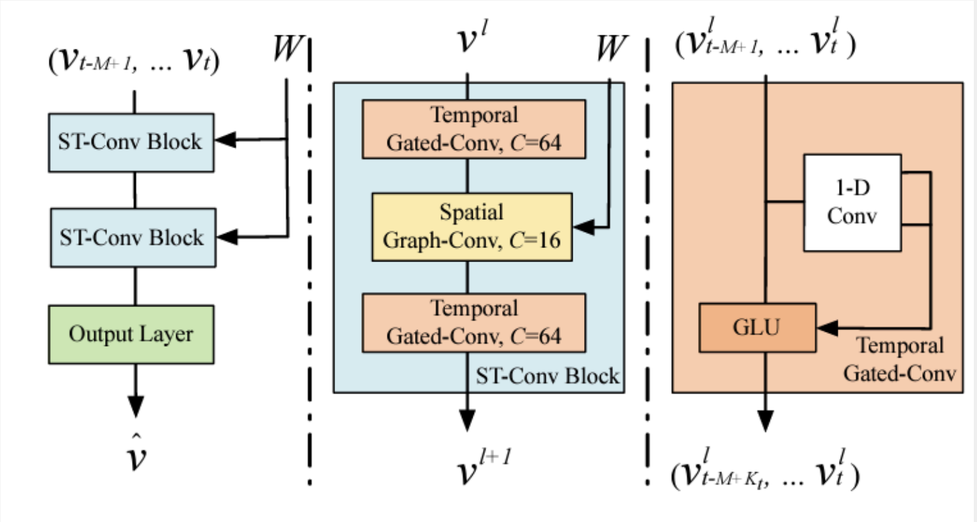


2 **ST-Conv** (Spatio-Temporal Convolution) blocks

- **3 layers**: temporal, spatial, temporal

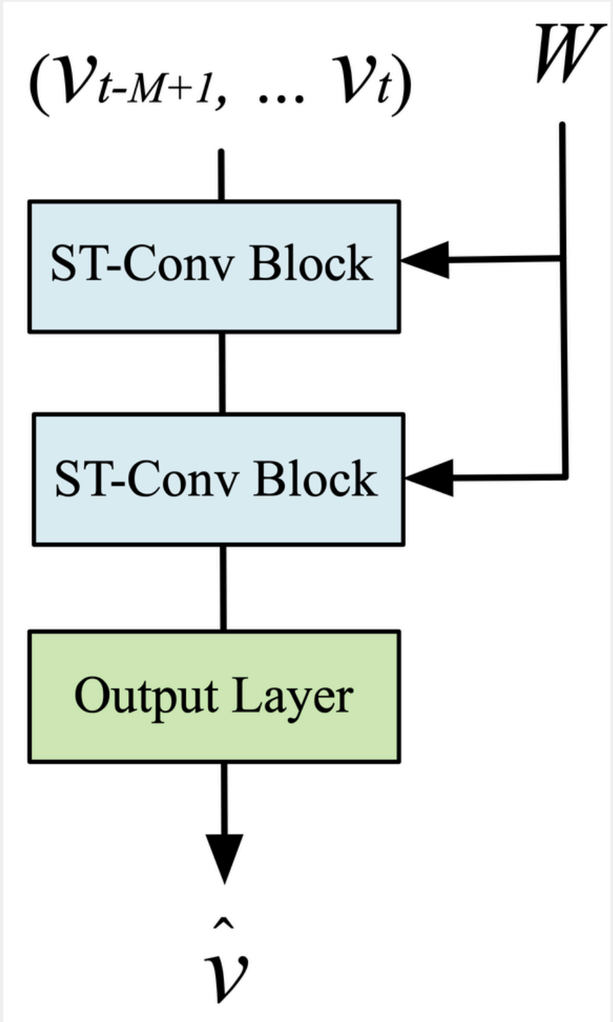
Output Layer

- Final temporal convolution **+** Fully connected layer



ST-Conv Block

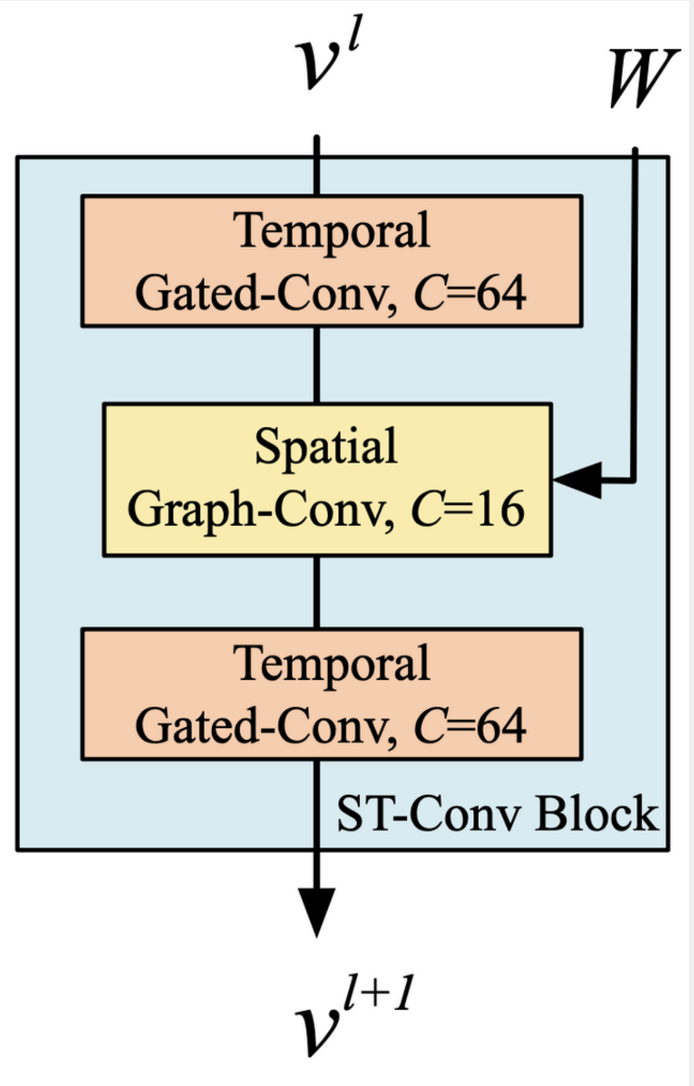
Captures **spatial dependencies** between nodes (e.g., road segments) and **temporal dependencies** (e.g., traffic trends over time)



Spatial Graph Convolution

Models **spatial relationships** by performing graph convolutions.

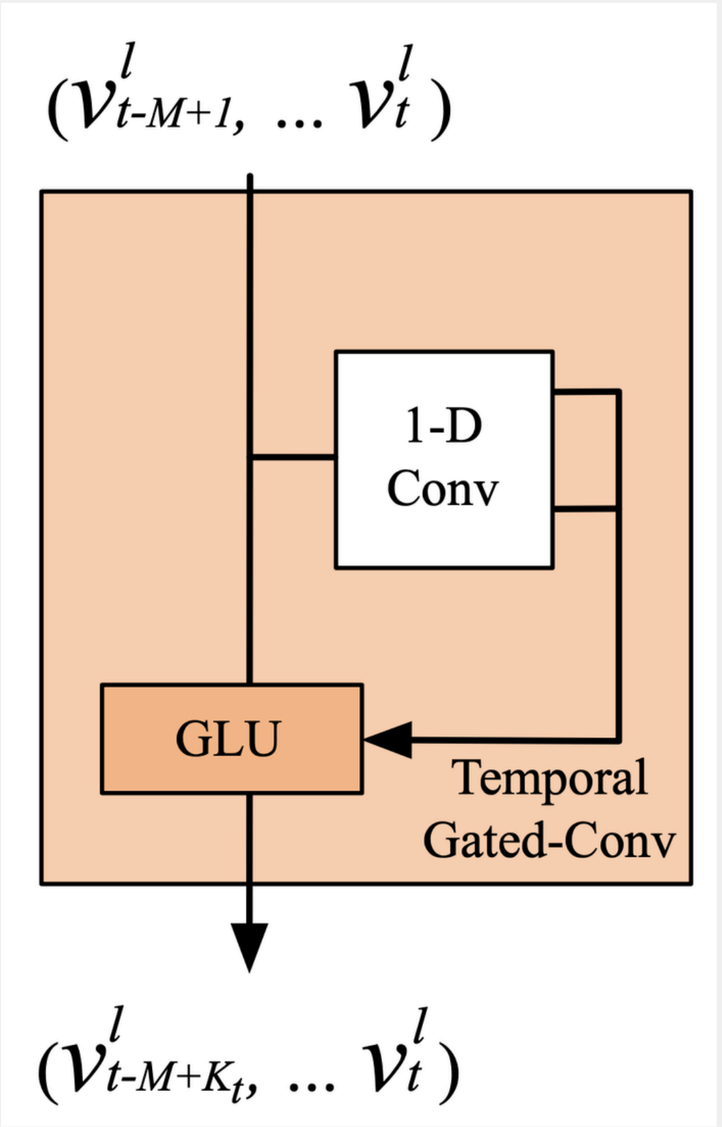
- Uses a **graph adjacency matrix** to compute the **influence of neighboring nodes** (e.g., how congestion at one road impacts connected roads).



Temporal Gated Convolution

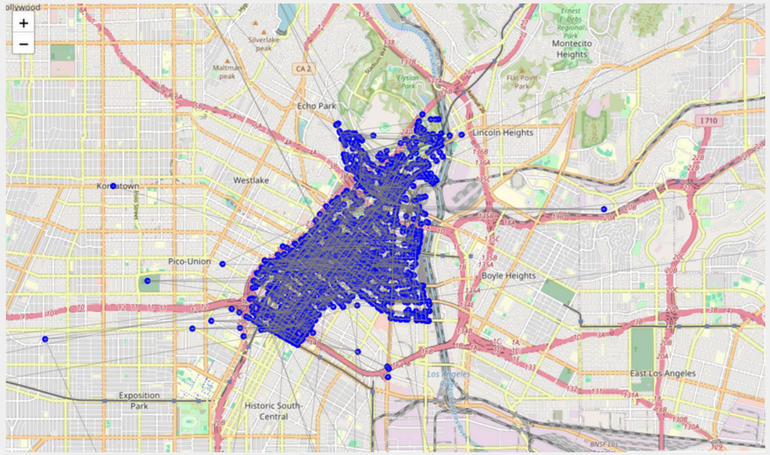
Captures time-based patterns using **gated convolutional layers**, which apply a **learned filter** across the sequence of inputs.

- **Gated Linear Unit** (GLU) selectively passes or suppresses information from the input
- allows the network to focus on the most relevant features while **ignoring noise** or less important details



Graph Structure

Given graph G(V,E),
V = Street intersections
E = Streets



Node Features

Static
- latitude
- longitude

Dynamic
- average surface temp
- total precipitation
- avg wind speed

Edge Features

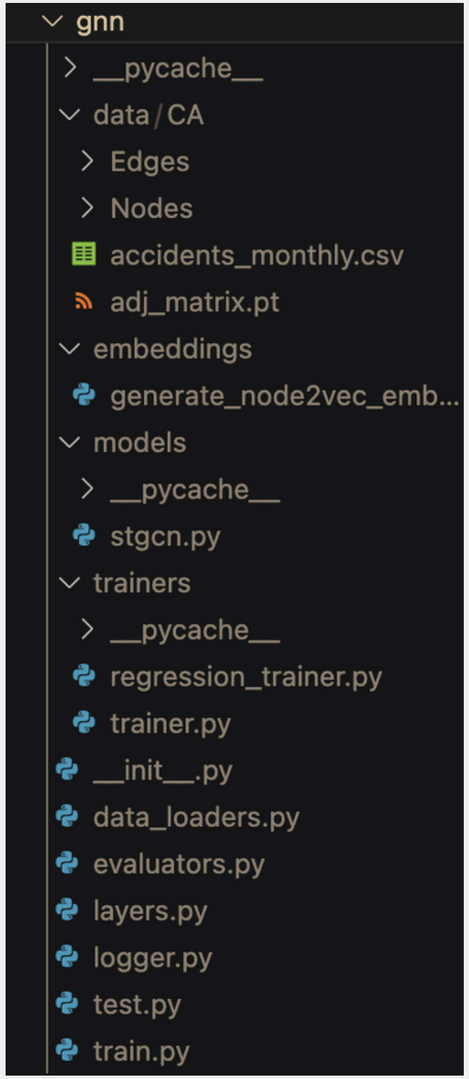
Static
- One way/two way
- Highway, residential
- Road length
- Annual average daily traffic (AADT)

Dynamic
- Monthly aggregated Accident Counts

GNN Tuning

- Learning rate: 0.001
- Training epochs: 100
- Runs: 3
- Batch size: 8124
- Data split:
  - Train: 2010-2014
  - Validate: 2015-2018
  - Test: 2019-2022

Only able to **manual search** for hyperparameters due to large network size and limited access to compute. We guided our hyperparameter tuning with research and ultimately chose these based on their balance of performance and compute cost.



Model Evaluation

Classification

Did at least one accident occur on this road this month?

Table 1:

Metric	Train (%)	Valid (%)	Test (%)
ROC-AUC	89.41 ± 1.59	67.89 ± 0.96	88.32 ± 1.40
F1	21.47 ± 5.10	9.15 ± 1.54	20.05 ± 5.22
AP	29.11 ± 7.21	12.31 ± 2.19	27.69 ± 5.48
Recall	36.24 ± 13.33	34.31 ± 29.64	38.32 ± 12.85
Precision	35.71 ± 8.63	16.46 ± 3.43	31.99 ± 7.76

- ROC-AUC disproportionately high compared to other evals - good at
- Possibly due to class imbalance - far less positive edges than negative

Regression

Predicting monthly accident counts

Table 1: Regression

Metric	Train	Valid	Test
MAE	1.25 ± 0.05	1.35 ± 0.07	1.30 ± 0.06
MSE	2.10 ± 0.10	2.35 ± 0.12	2.25 ± 0.11

- Many road segments have 1-3 accidents per month
- Poor across the board, but stable - with more tuning, this could be a better application

Limitations

Current data pipeline is highly complex and **resource-intensive**

- Long runtimes as we could only train with CPUs instead of NVIDIA GPUs
- Model requires further refinement to handle data complexity

Future Work

- Examine correlation between graph features (node indegree/outdegree, betweenness centrality) and accident counts
- Conduct ablation studies to analyze impact of traffic volume, weather, and other features
- Visualization of predicted collision hotspots for public policy analysis

Thank you