**Final Project Report**

**Manhattan Traffic Collision Map**

Chongyang Ren MSAUD 23’

Junling Zhuang MSCDP 23’

Yaoze Yu MSAUD 23

**Problem Definition**

Traffic collisions are a major concern in urban areas, and can lead to severe injuries or fatalities. Manhattan is known for its high-density road network and heavy traffic, making it important to understand the factors that contribute to traffic collisions in this area. In order to reduce the frequency and severity of collisions, it is necessary to identify the intersections that are most prone to accidents and determine the factors that contribute to their danger.

**Research Question**

What factors contribute to the frequency and severity of traffic collisions at intersections in Manhattan, and how can this information be used to predict the risk of collisions at different locations?

**Literature review**

1. Luis F. Miranda-Morenoa, Patrick Morencyb, Ahmed M. El-Geneidyc. (2011). The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections
2. Ayesha Shafique, Guo Cao, Zia Khan, Muhammad Asad, Muhammad Aslam. (2022). Deep Learning-Based Change Detection in Remote Sensing Images: A Review
3. Xiaojiang Li, Chuanrong Zhang, Weidong Li, Robert Ricard, Qingyan Meng, Weixing Zhang. (2015). Assessing street-level urban greenery using Google Street Viewand a modified green view index

**Data Sources**

1. Motor Vehicle Collisions - Crashes (<https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95>- From NYC Open Data)
2. Land Cover Raster Data (2017) – 6in Resolution

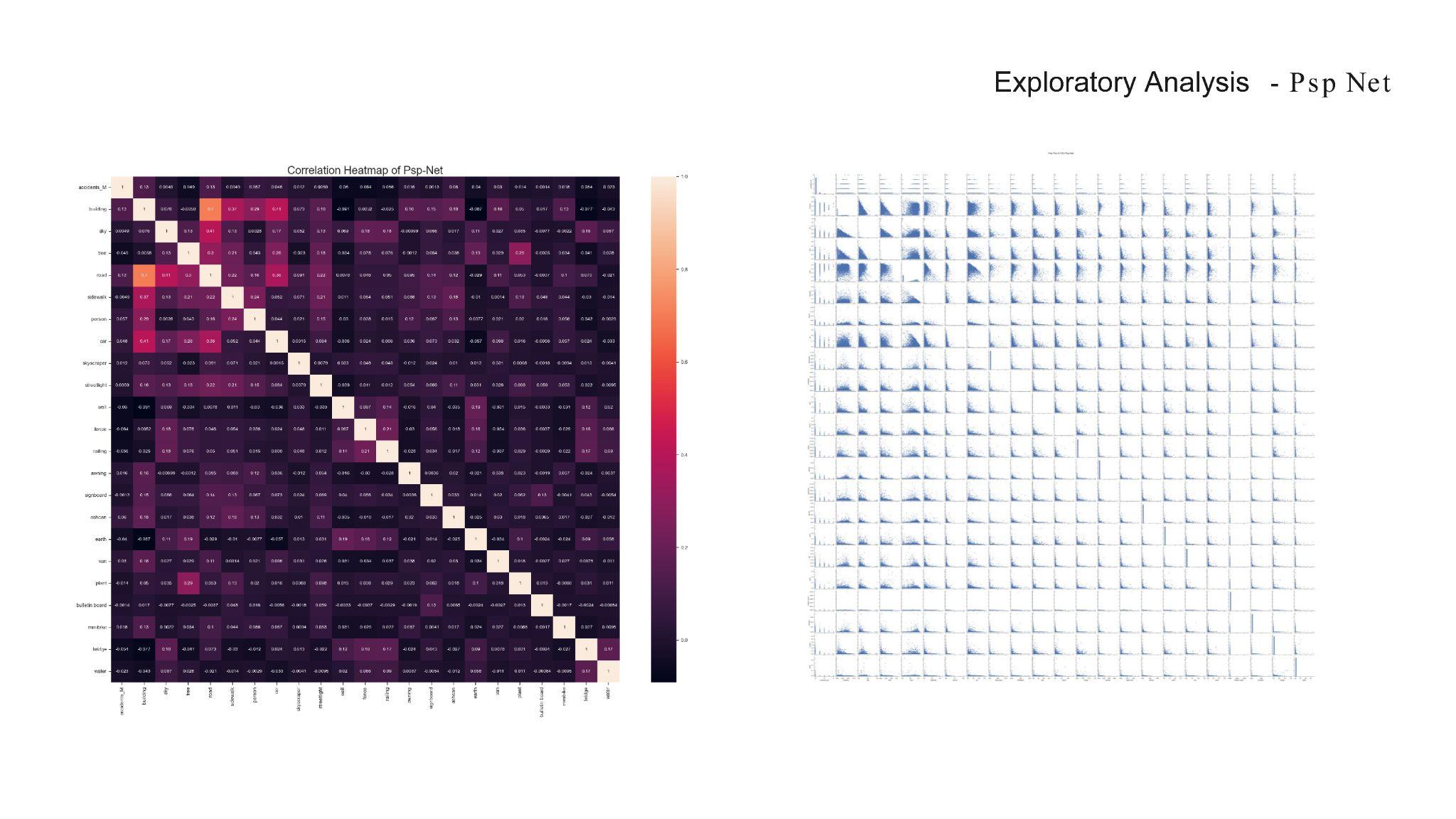
(<https://data.cityofnewyork.us/Environment/Land-Cover-Raster-Data-2017-6in-Resolution/he6d-2qns>- From NYC Open Data)

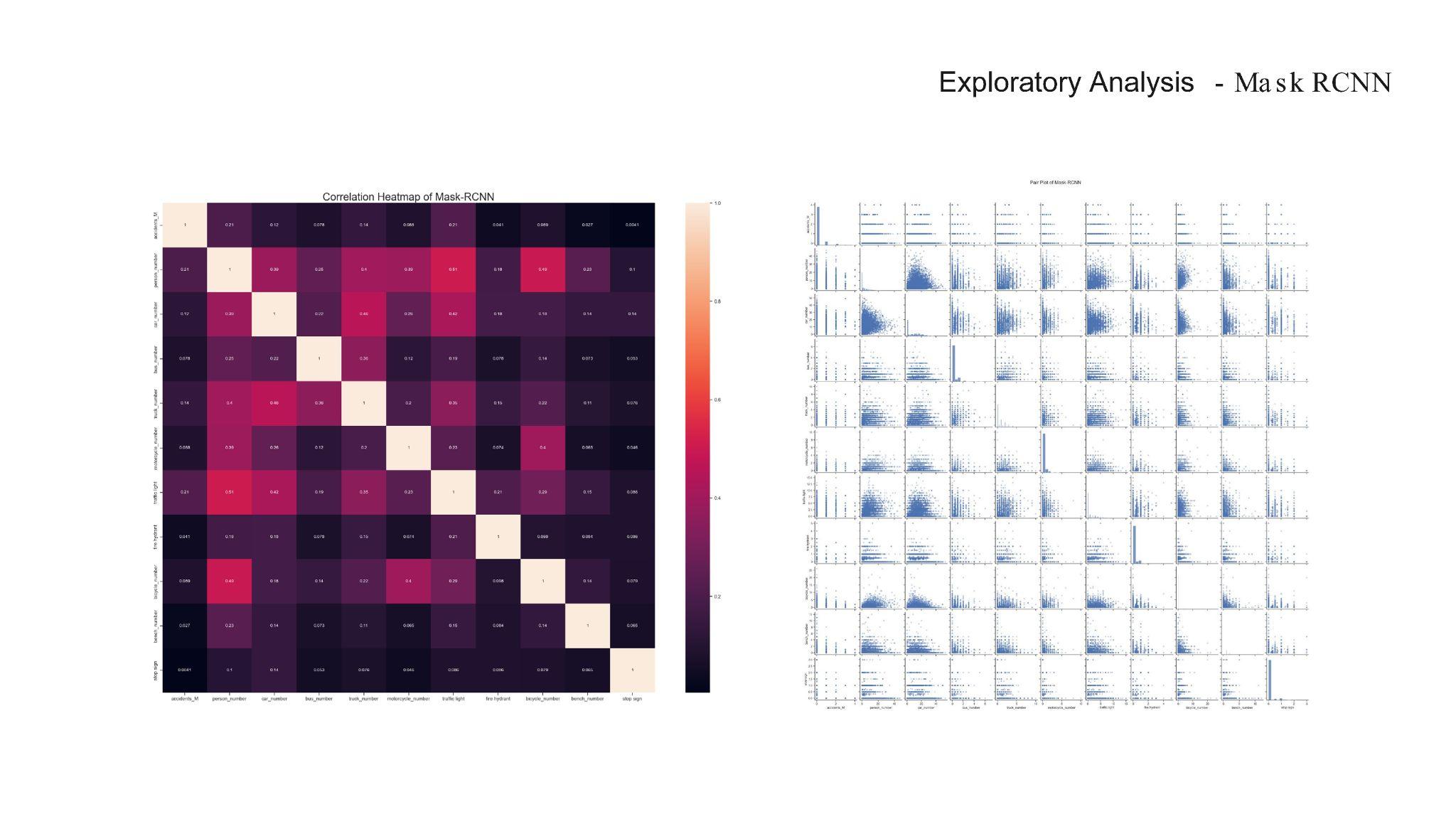
**Brief Description**

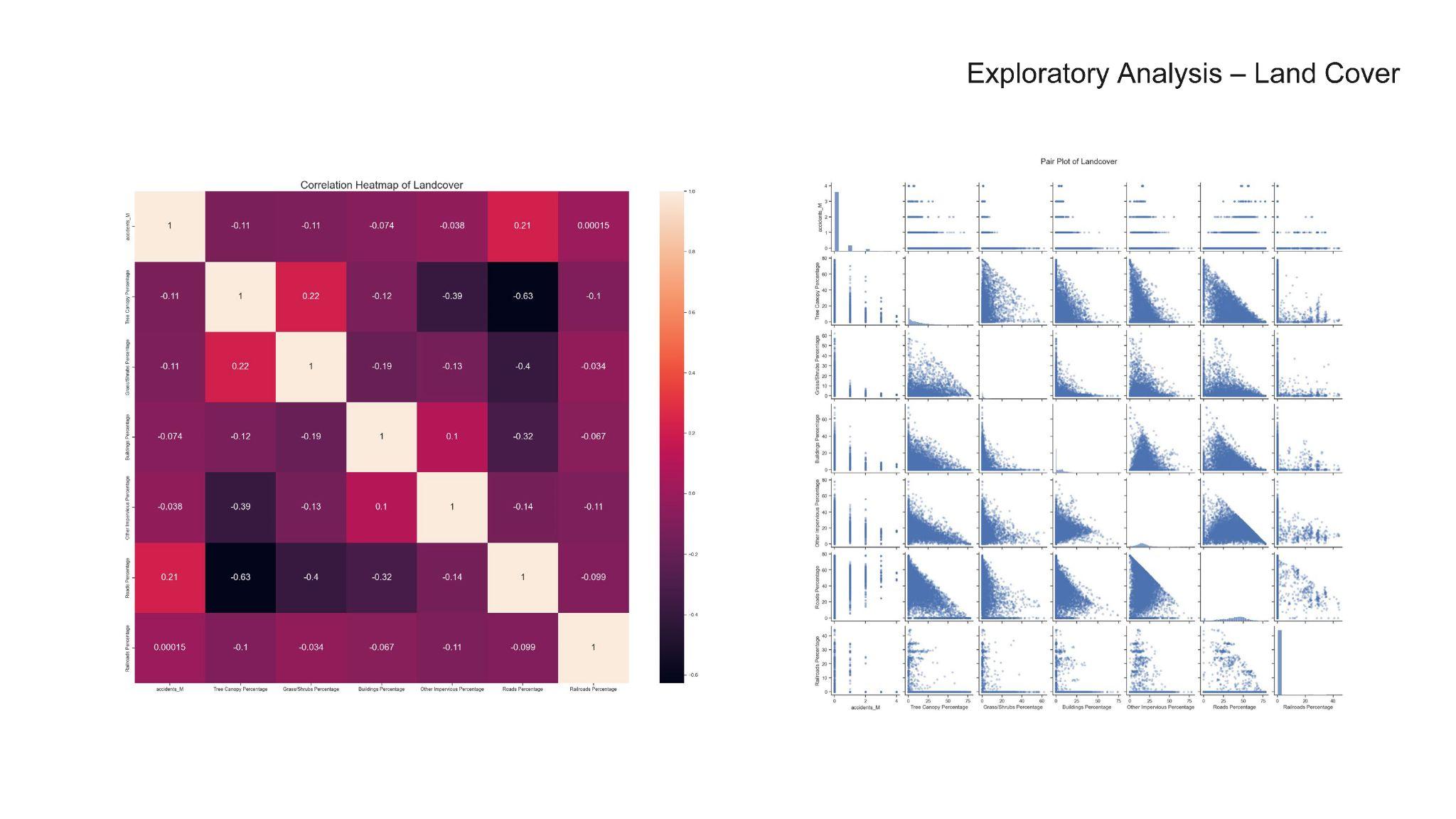
The Motor Vehicle Collisions crash table contains details on the crash event. Each row represents a crash event. The Motor Vehicle Collisions data tables contain information from all police reported motor vehicle collisions in NYC.

This dataset contains records of crash date, crash time and crash location and further detailed description. We extract **number of persons injured** and **number of persons killed** as the most important variable.

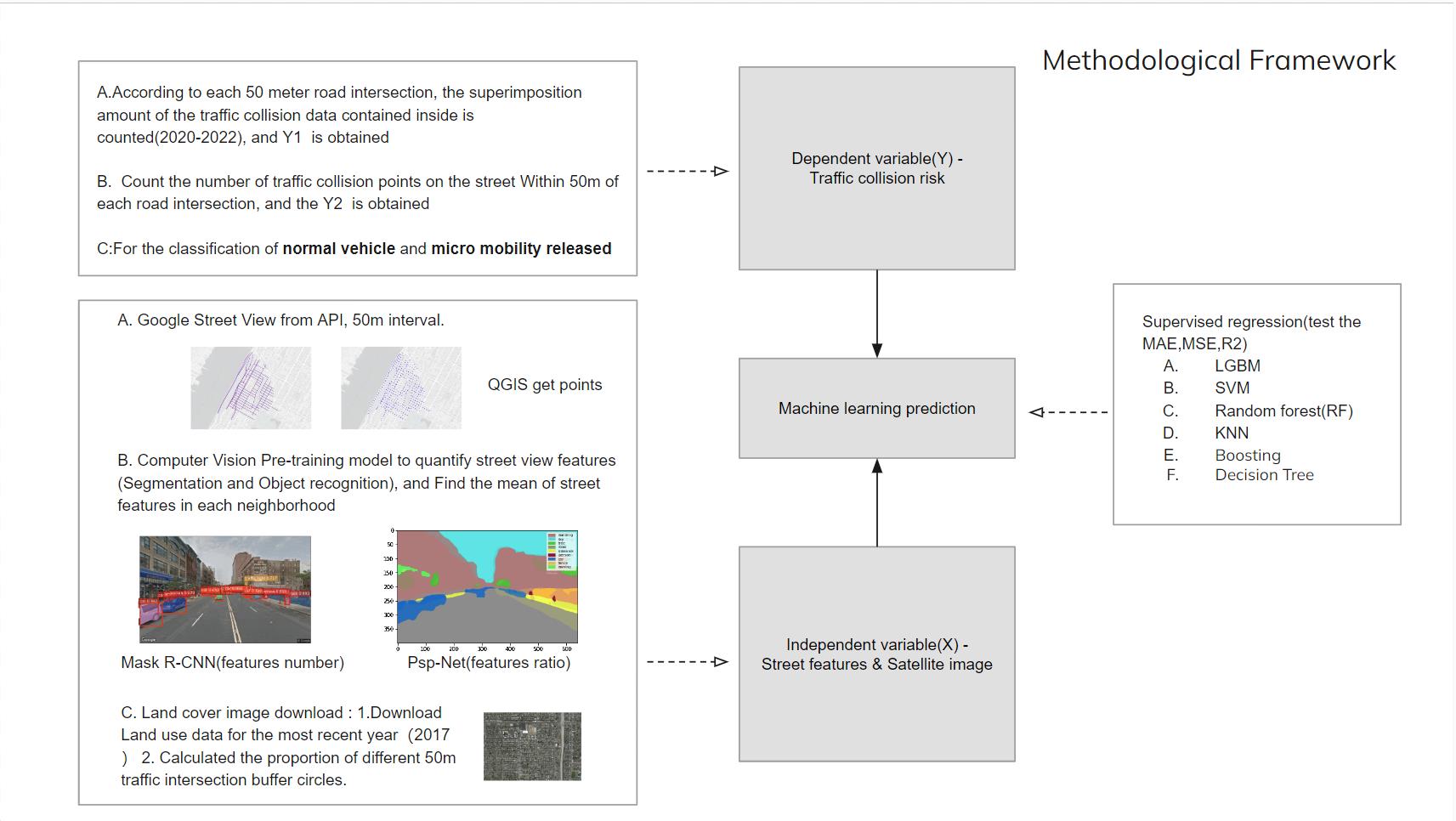
**Exploratory Analysis**

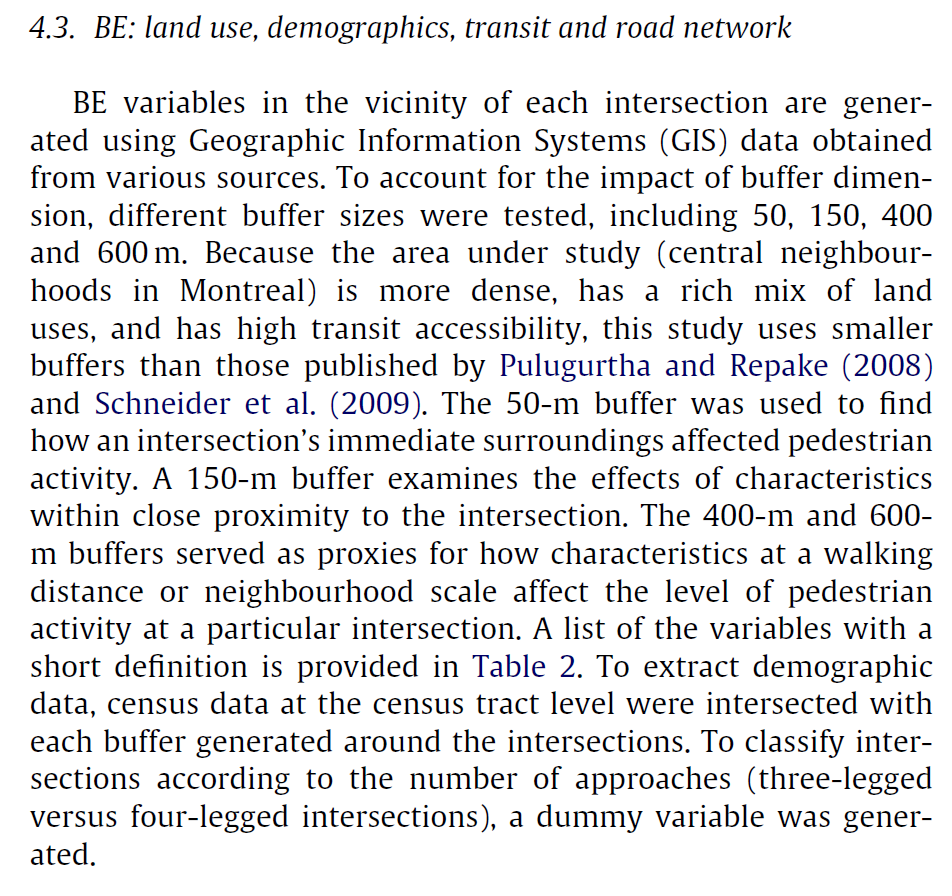
****

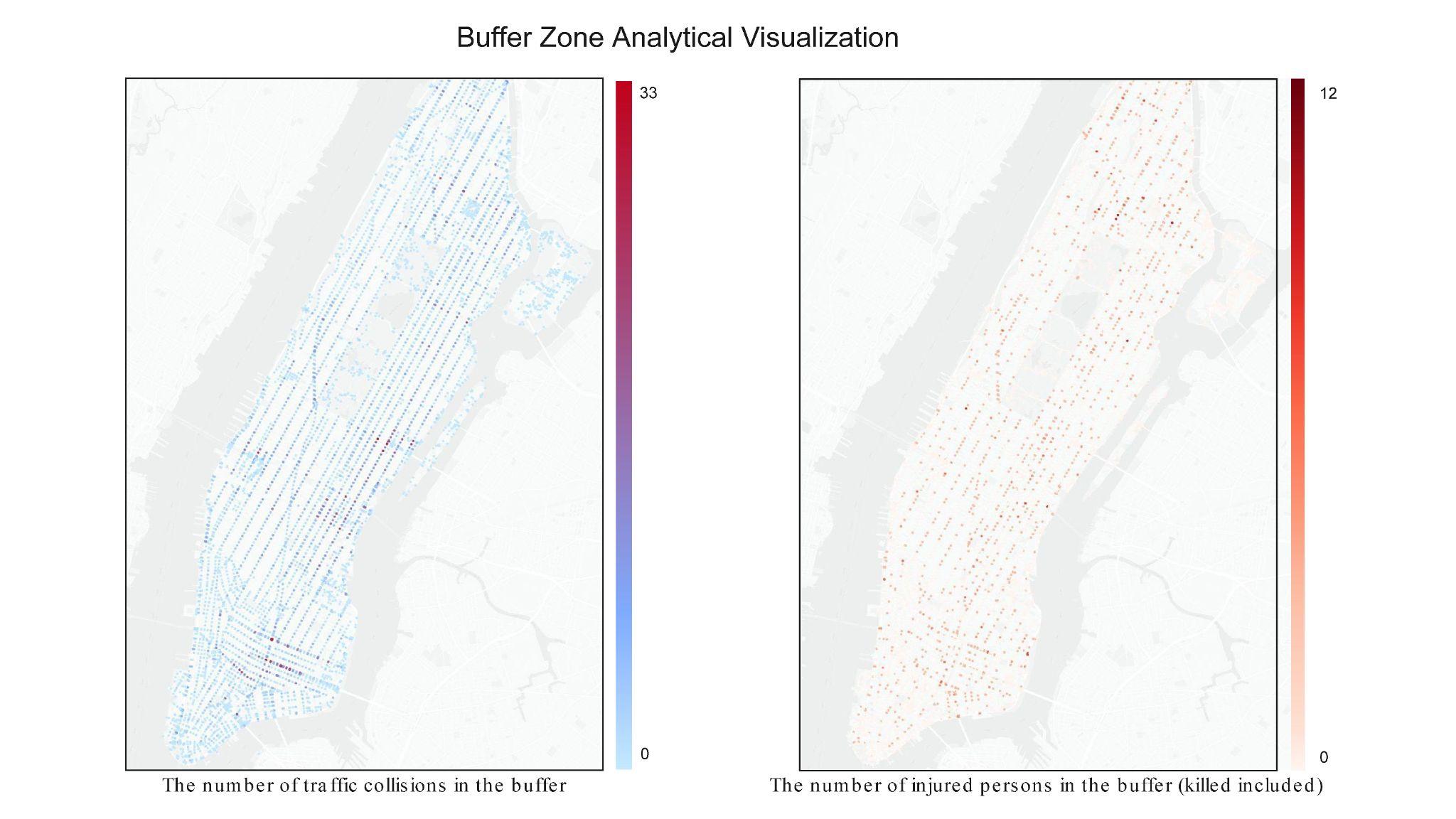




**Methodology Framework**

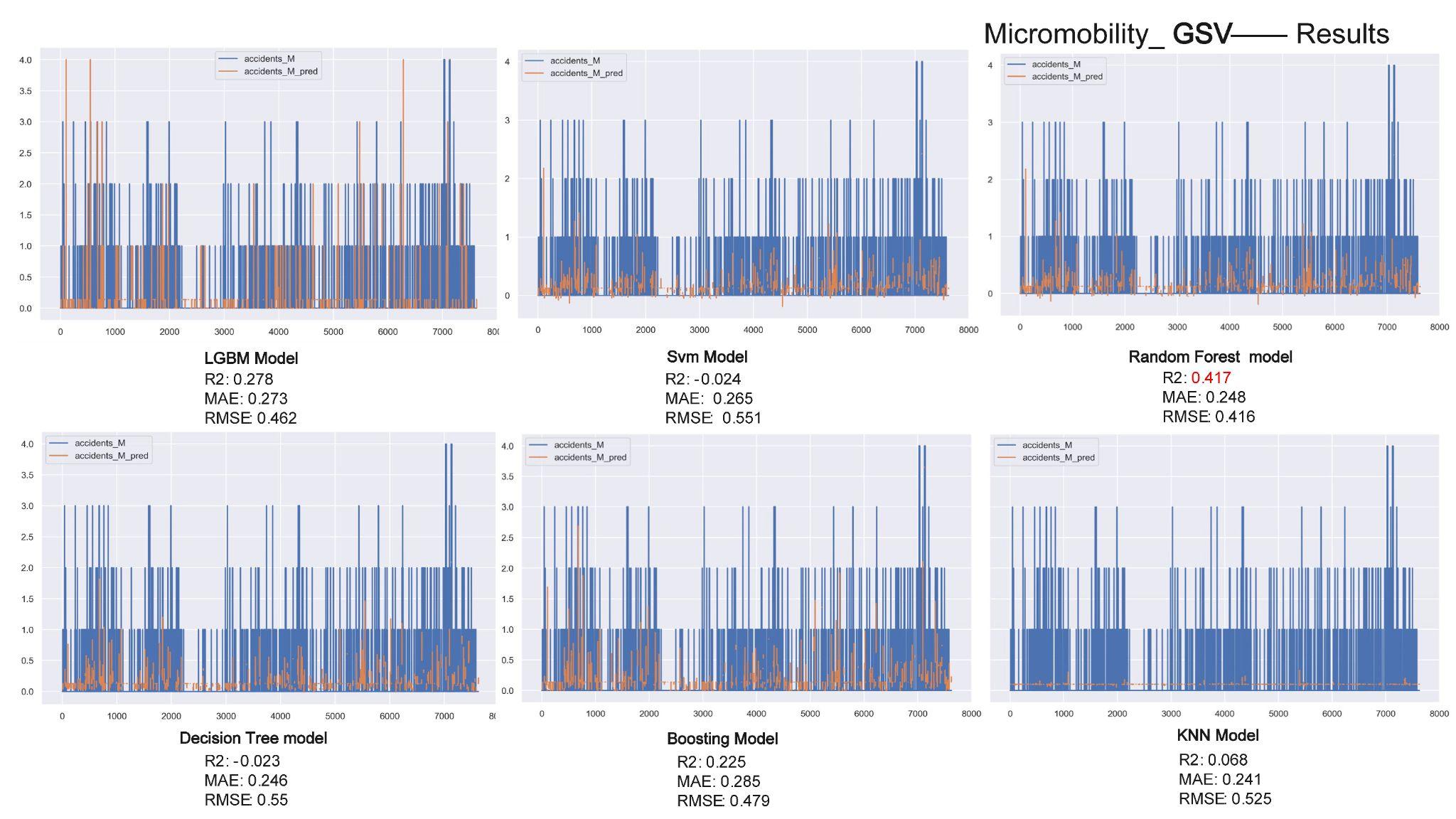
****

1. **Defining the Traffic Collision Risk (Dependent Variable Y)**
   1. 
   2. According to the paper“The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections”: Section 4.3 we referred to, **the 50-m buffer was used to find how an intersection’s immediate surroundings affected pedestrian activity, which fit our demand of researching how people get affected by traffic collisions happening around the intersections.**
   3. According to each 50 meter road intersection, the superimposition amount of the traffic collision data contained inside is counted(2020-2022), and the Y of the neighborhood scale is obtained.
   4. Count the number of traffic collision points on the street Within 50 meters of each road intersection.
   5. For the classification of normal vehicles and micro mobility released.

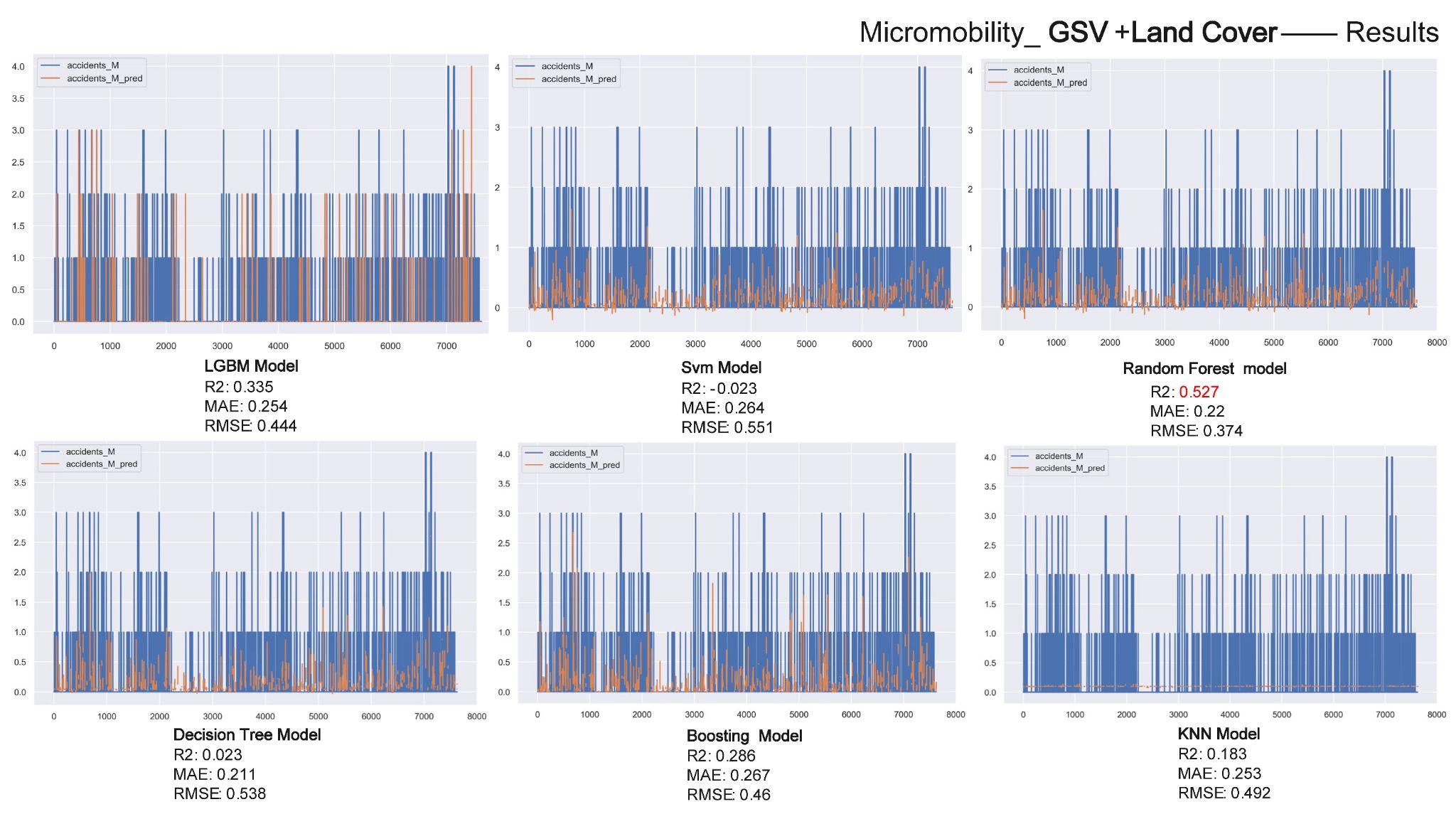


1. **Measuring the street views + Satellite Image (Independent Variable X）**
   1. GSV collection
      1. Download street pictures through google Api, 50m interval, 800x600 pixels.
      2. Use Mask R-CNN to get the type and number of features.
      3. Use Psp-Net to get the ratio of features.
      4. Calculate the mean of street features in the same buffer circle
   2. Land cover image download
      1. Download 2017 NYC land cover image.
      2. Calculated the proportion of different 50m traffic intersection buffer circles.
2. **Machine learning prediction**
   1. Use the training data of Manhattan 2020-2022 to predict the risk of collisions at different locations?
   2. Supervised regression(test the MAE,MSE,R2) to select a model of best performance.
      1. LGBM
      2. SVM
      3. Random forest(RF)
      4. KNN
      5. Boosting
      6. Decision Tree

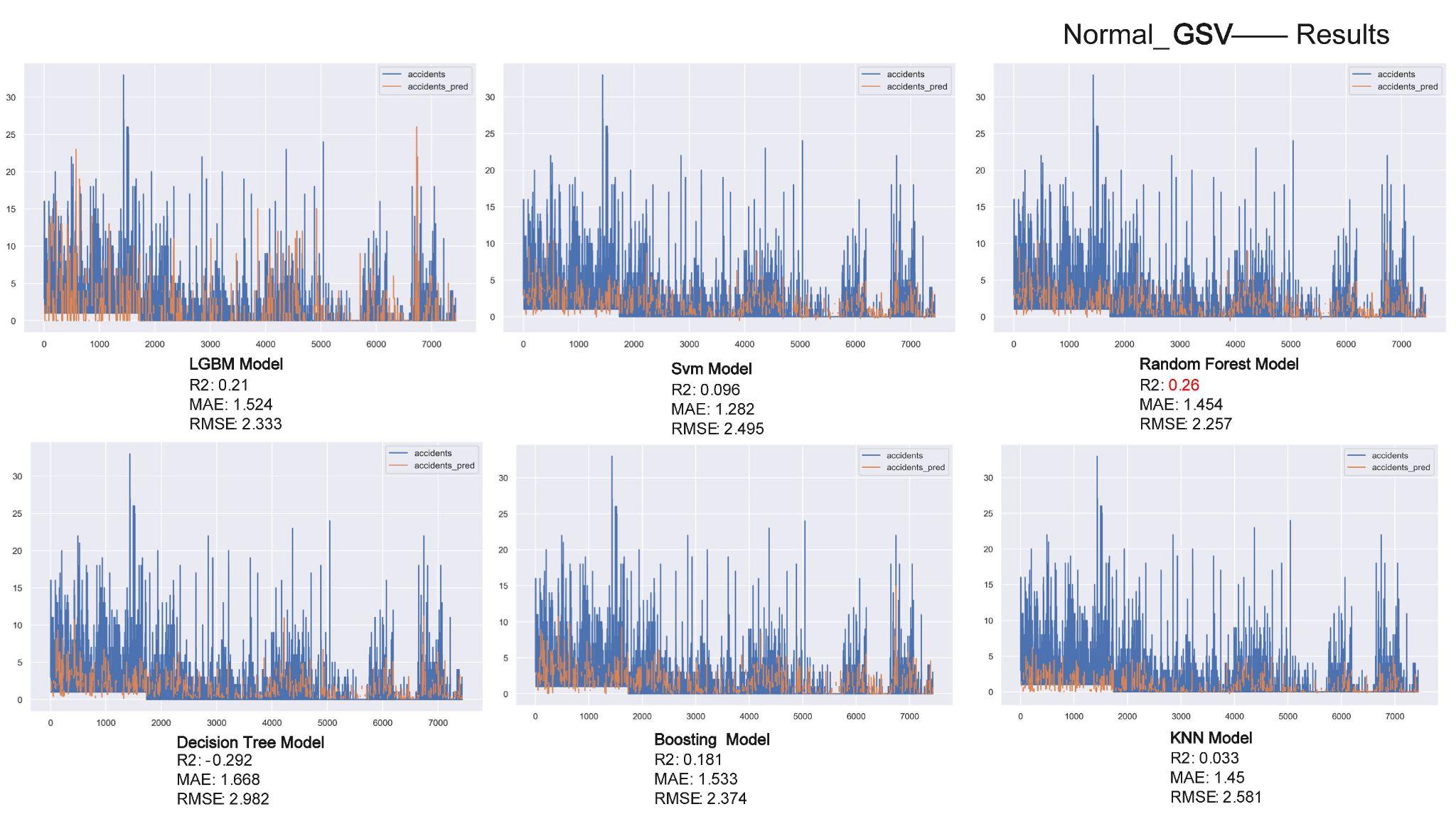
**Results**

****

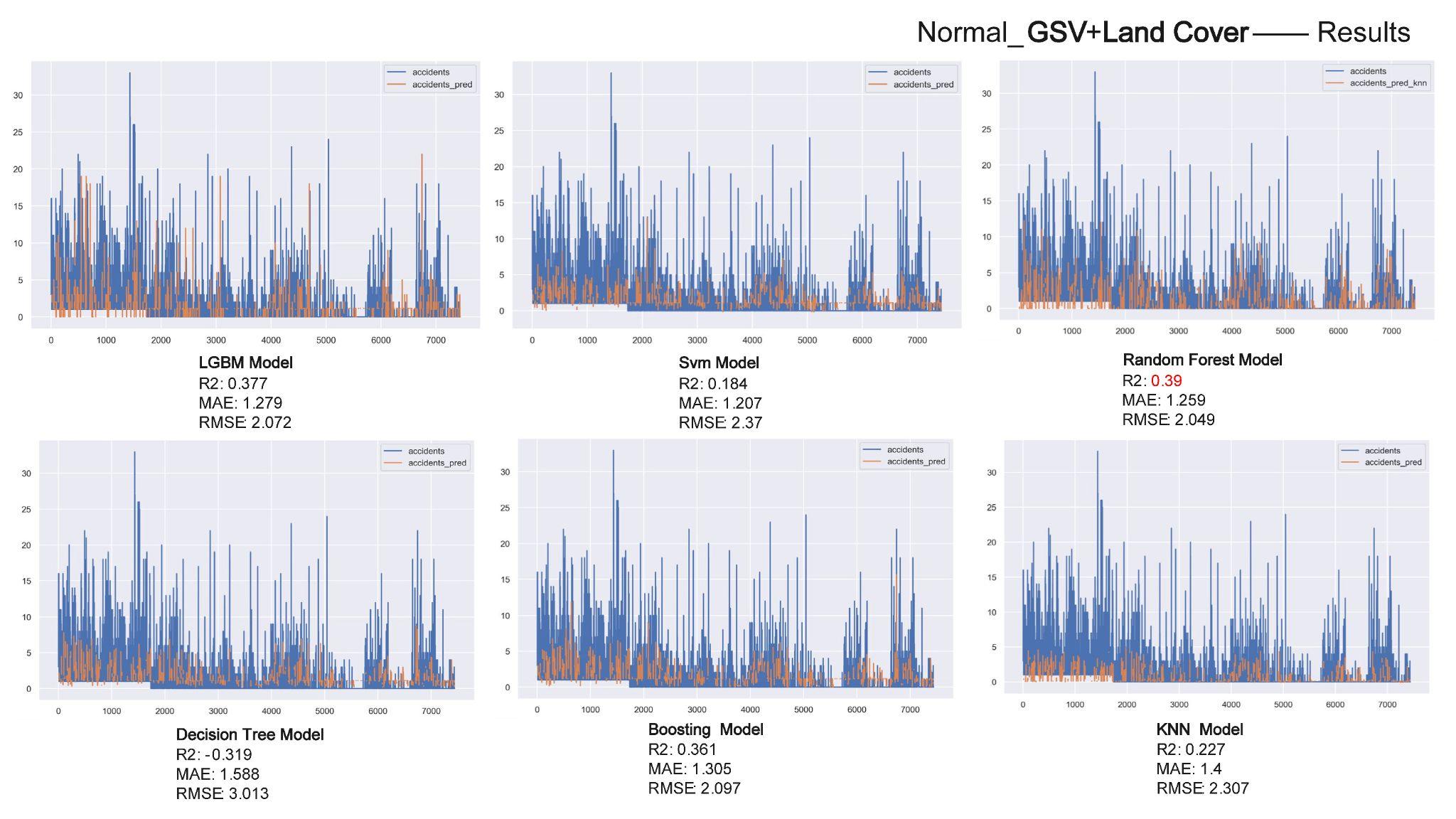
1. Firstly, for Micromobility, we conducted a regression analysis between the number of accidents events within the traffic intersection buffer, considering only GSV data, and the street-level built environment features segmented by Mask R-CNN and PSPNet. The LightGBM, Gradient Boosting, and Random Forest models all achieved an R2 score greater than 0.2. The MAE and RMSE values were centered around 0.2 and 0.4, respectively.



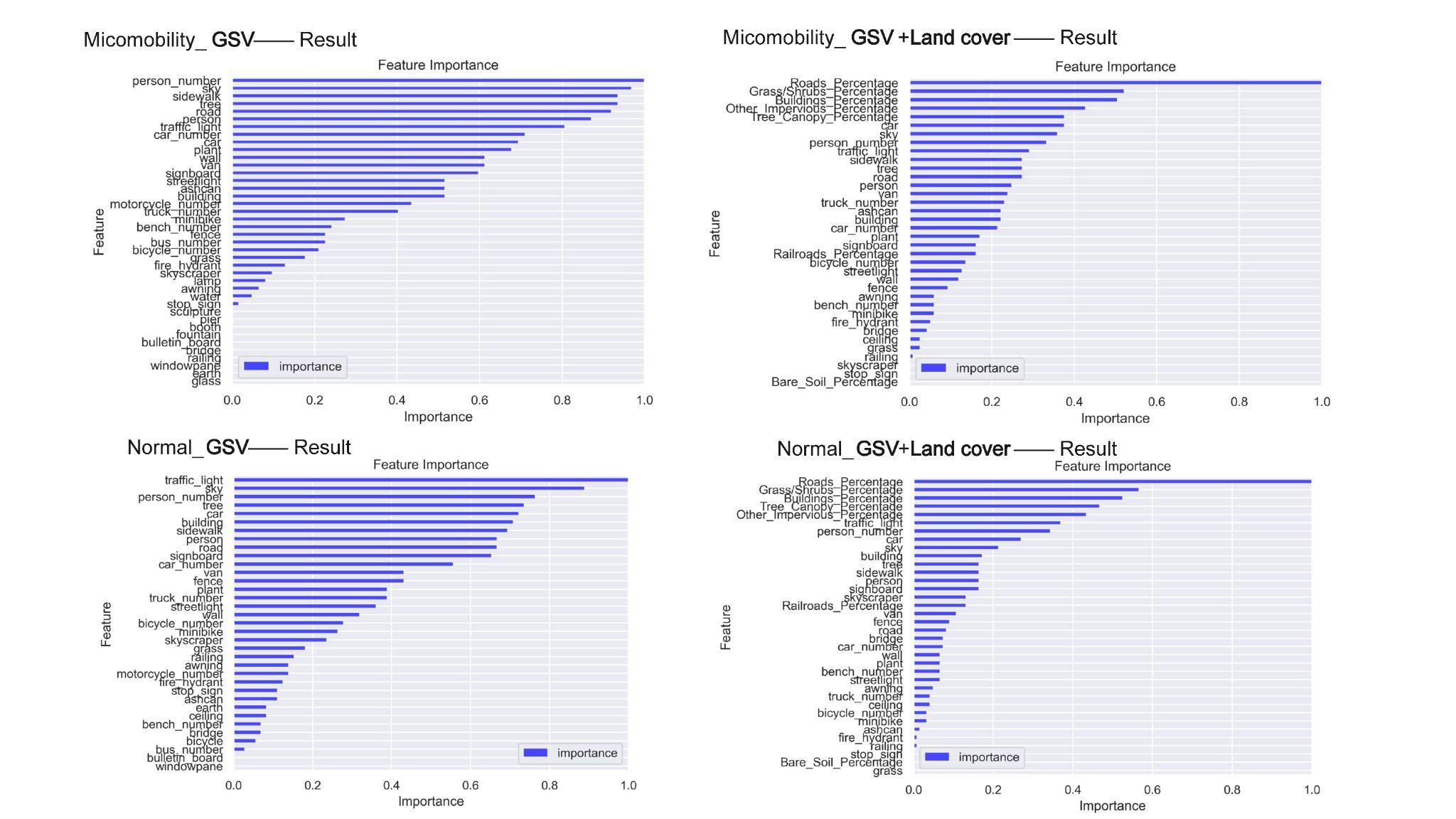
1. However, when we incorporated Land Cover into our analysis, the overall R2 scores experienced a significant improvement, particularly for the Random Forest model, which increased from 0.41 to 0.52. The RMSE values for each model also decreased by approximately 0.5 to 0.1 when compared to the cases considering only GSV data.



1. We employed the same approach for normal vehicles as well. When considering only GSV data, the Decision Tree Model, LightGBM Model, and Random Forest Model all exhibited relatively high R2 scores, though the Decision Tree Model displayed a negative correlation overall. At the same time, the MAE and RMSE values, in comparison to those for micro mobility, increased significantly, reaching 1.5 and 2.3, respectively.



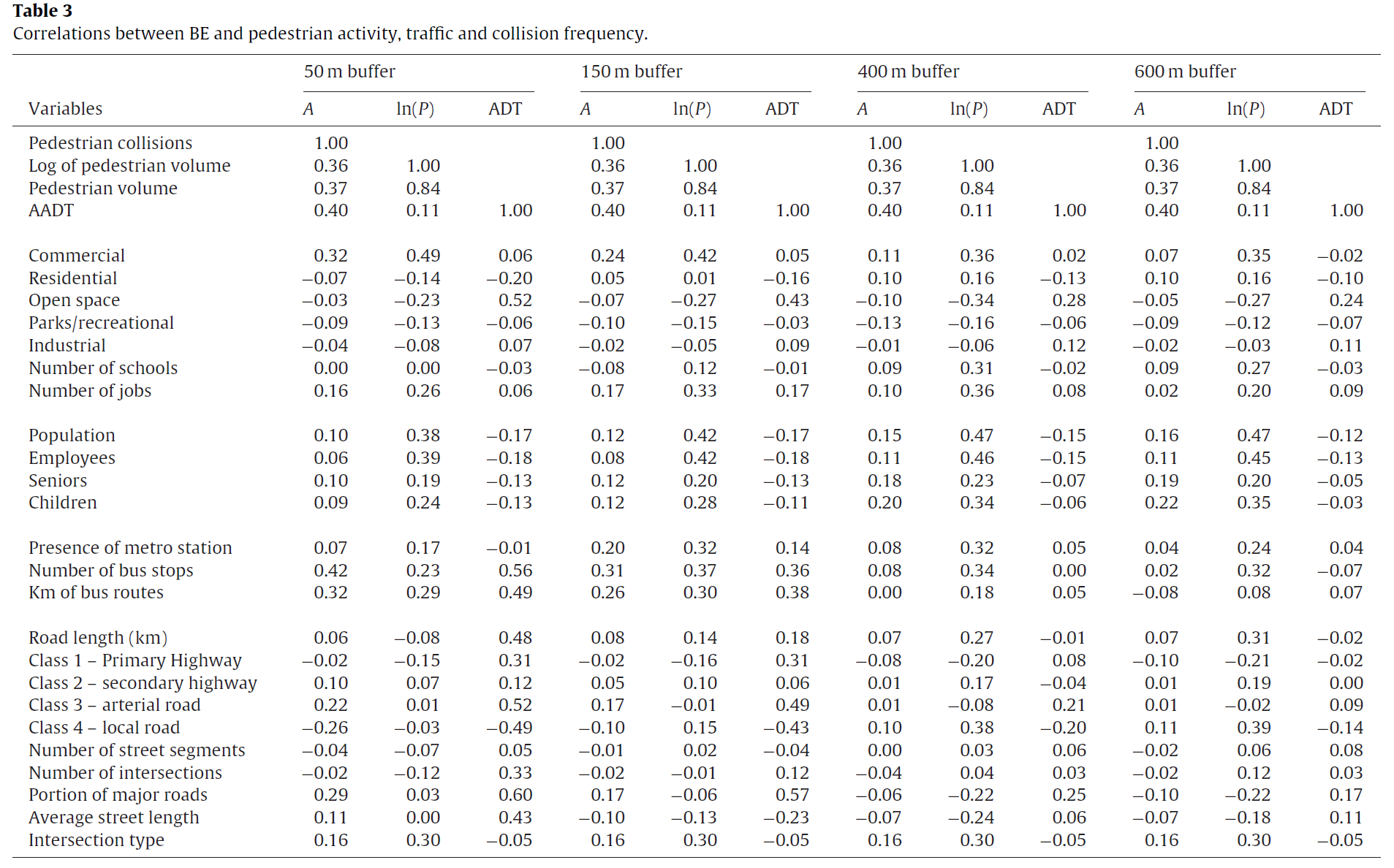
1. Upon incorporating Land Cover into our analysis, the overall performance of the models improved considerably. Among the six models, the Random Forest model demonstrated the best performance.



1. We then ranked the features' importance during the regression of the Random Forest model using GINI importance. We found that, without considering Land Cover, parameters such as traffic light, tree, car, buildings, sky, and person had a relatively high impact. However, after including Land Cover, the road percentage became the most influential factor.

**Conclusions & Implications & Limitations**

* Random Forest (RF) model takes the slowest time but has the highest R² and the best performance among all models.
* According to the paper “The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections”: Table 3 we referred to, A correlation matrix for each buffer size is generated, including 50, 150, 400 and 600m. Therefore, The diameter size of the buffer zone should also be adjusted to 150, 400, 600 meters for comparative analysis.

****

* After debugging the model with relatively optimal performance again, it is hoped that in the future, the risk of traffic collisions at road intersections in other low data areas can be predicted.