Supplemental Information

RF Method:

Santos et al. (2022) conducted a review of 56 studies on crash injury severity prediction published between 2001 to 2021 and concluded that RF algorithms usually offer the best results in injury severity prediction and have achieved the best performance 70% of the time whenever it was applied in the reviewed studies.

The RF method is based on the bagging principle and the random subspace method which involves developing a set of decision trees by selecting random predictors from the dataset (Breiman, 2001). The steps in RF classification using decision trees are described in Figure 1. Class of injury severity were used as the outcome variable in the RF model.

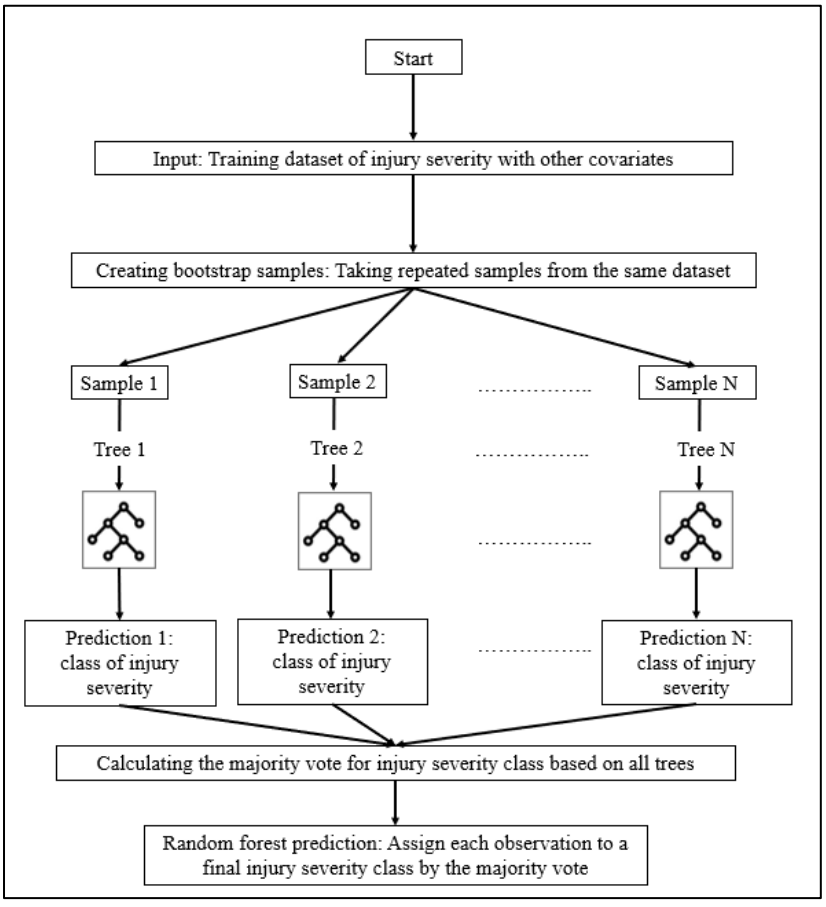

Figure 1: Steps followed by RF algorithm.

Resolving class imbalance issue in the dataset:

It should be noted that the dataset is highly imbalanced where 80% of the pedestrian injury severity is categorized as Major which can affect the prediction accuracy in the training and testing phase. With imbalanced datasets, RF algorithms do not obtain the necessary information about the minority class to make an accurate prediction which impacts the overall prediction accuracy of the model (Chen et al. n.d.). Also, in the testing dataset, the number of cases for minority classes becomes less than in the training dataset which impacts the prediction accuracy of the testing dataset. Therefore, multiple measures have been taken to resolve the data imbalance issue in this study. First, injury severity classes have been aggregated into two classes (Fatal or Major Injury: 93% and rest as Minor or None: 7%) to increase the number of minority classes. Then, the dataset were divided into training and testing dataset following stratified sampling which splits the data randomly, but keeps the same imbalanced class distribution for each subset. Below is the distribution of cases in training and testing dataset after stratified sampling.

Training data: Fatal or Major= 1505, Minor or None = 112

Testing data: Fatal or Major = 376, Minor or None = 28.

However, our datasets are still imbalanced. We applied the Random Over-Sampling Examples (ROSE) technique (Menardi and Torelli, 2014) to resolve the data imbalance issue in the training dataset. ROSE is a powerful algorithm that creates a sample of synthetic data by enlarging the feature space of minority class (here, Minor or None) cases. Furthermore, the majority class cases (here, Fatal or Major) are also under-sampled, leading to a more balanced dataset. The new cases are drawn from a conditional kernel density estimate of the two classes (Menardi and Torelli, 2014). The balanced training dataset has the following distribution of the classes.

Fatal or Major: 845 and Minor or None: 772.

We kept the testing dataset as it is as we wanted to test the trained model based on actual field data.

Eigenvector spatial filtering approach:

Eigenvector Spatial Filter (ESF) approach is used to account for spatial autocorrelation and uses geographical coordinates to specify eigenvectors across geographic distance (Griffith, 2003; Murakami and Griffith, 2019). In this study, coordinates of the traffic collision locations were used to obtain the eigenvectors. ESF creates a spatial connectivity weight matrix of the data points. In this study, it was a 2021 x 2021 matrix for 2021 collision locations. However, it should be noted that only a subset of the generated eigenvectors demonstrates spatial autocorrelations (Griffith and Chum, 2013; Murakami and Griffith, 2019). For the injury severity data, total 232 pairs of eigenvectors were found to be sufficient to show spatial dependence in the data. As the size of our dataset is comparatively low and have only 38 non-spatial predictors, instead of using all these synthetic eigenvectors in the RF model as predictor variables, we used the vector of all

approximated eigenvectors (EV)[1] as a proxy variable to capture the spatial autocorrelation in the dataset. Readers are referred to Murakami and Griffith (2019), Griffith and Chun (2013), McCord et al. (2020), and Islam et al. (2022) for a detailed understanding of the methodology used and its application.

The role of non-spatial predictors in the RF models:

Here, we briefly discussed the ten non-spatial predictors of highest importance in the RF models. Results suggest that pedestrian conditions (e.g., normal, inattentive) and driver's conditions (e.g., inattentive, slowly stopping, driving properly, failing to yield right of way) are important predictors of pedestrian injury severity. Studies have found that pedestrian's inattentiveness and driver's conditions such as being inattentive, failing to yield right of way are more likely to cause higher pedestrian injury severity (Zhai et al., 2019; Khan and Habib, 2022) whereas lower severity is expected when driver's condition is normal (Pour-Rouholamin and Zhou, 2016). Also, aggressive driving and major arterial as the collision location are important predictors for injury severity in Toronto. Studies have also found that aggressive driving and collisions at the arterials increase the likelihood of higher injury severity of pedestrians (Lin et al. 2019; Khan and Habib, 2022). Previous studies also found that dark lighting conditions during the collision time at the collision location cause higher injury severity among pedestrians (Haleem et al. 2015; Hossain et al. 2022). The RF model of injury severity of pedestrians in Toronto also identifies dark lighting conditions as an important predictor of injury severity. In terms of demography, individuals between 0-19 years of age have been identified as vulnerable road users and are at-risk of getting severely injured in traffic collisions (Cloutier et al. 2021). The developed RF model also indicates this as an important predictor of injury severity.

References:

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Chen, C. Liaw, A. & Breiman, L. (n.d.). Using Random Forest to Learn Imbalanced Data. https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf

Cloutier, M. S., Beaulieu, E., Fridman, L., Macpherson, A. K., Hagel, B. E., Howard, A. W., ... & Rothman, L. (2021). State-of-the-art review: preventing child and youth pedestrian motor vehicle collisions: critical issues and future directions. *Injury prevention*, *27*(1), 77-84.

Griffith, D. A. (2003). *Spatial autocorrelation and spatial filtering: Gaining understanding through theory and scientific visualization*. Berlin, Germany: Springer.

[1] Calculated based on Griffith and Chum (2013) and Murakami and Griffith (2018) with *R* package "*spmoran*".

Griffith, D. A., & Chun, Y. (2013). Spatial autocorrelation and spatial filtering. In M. M. Fischer & P. Nijkamp (Eds.), *Handbook of regional science* (pp. 1477–1507). Berlin, Germany: Springer. https://doi.org/10.1007/978-3- 642-23430-9_72

Haleem, K., Alluri, P., & Gan, A. (2015). Analyzing pedestrian crash injury severity at signalized and non-signalized locations. *Accident Analysis & Prevention*, *81*, 14-23.

Hossain, A., Sun, X., Thapa, R., & Codjoe, J. (2022). Applying association rules mining to investigate pedestrian fatal and injury crash patterns under different lighting conditions. *Transportation research record*, *2676*(6), 659-672.

Khan, N. A., & Habib, M. A. (2022). Exploring the impacts of built environment on pedestrian injury severity involving distracted driving. *Journal of safety research*, *80*, 97-108.

Lin, P. S., Guo, R., Bialkowska-Jelinska, E., Kourtellis, A., & Zhang, Y. (2019). Development of countermeasures to effectively improve pedestrian safety in low-income areas. *Journal of traffic and transportation engineering (English edition)*, *6*(2), 162-174.

McCord, M. J., McCord, J., Davis, P. T., Haran, M., & Bidanset, P. (2020). House price estimation using an eigenvector spatial filtering approach. International Journal of Housing *Markets and Analysis, 13*(5), 845-867.

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery, 28*, 92-122.

Murakami, D. and Griffith, D.A. (2019) Eigenvector spatial filtering for large data sets: fixed and random effects approaches. *Geographical Analysis, 51* (1), 23-49.

Pour-Rouholamin, M., & Zhou, H. (2016). Investigating the risk factors associated with pedestrian injury severity in Illinois. *Journal of safety research*, *57*, 9-17.

Santos, K., Dias, J. P., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. *Journal of Safety Research, 80*, 254-269.

Zhai, X., Huang, H., Sze, N. N., Song, Z., & Hon, K. K. (2019). Diagnostic analysis of the effects of weather condition on pedestrian crash severity. *Accident Analysis & Prevention*, *122*, 318-324.