Causal Analysis and Classification of Traffic Crash Injury Severity Using Machine Learning Algorithms

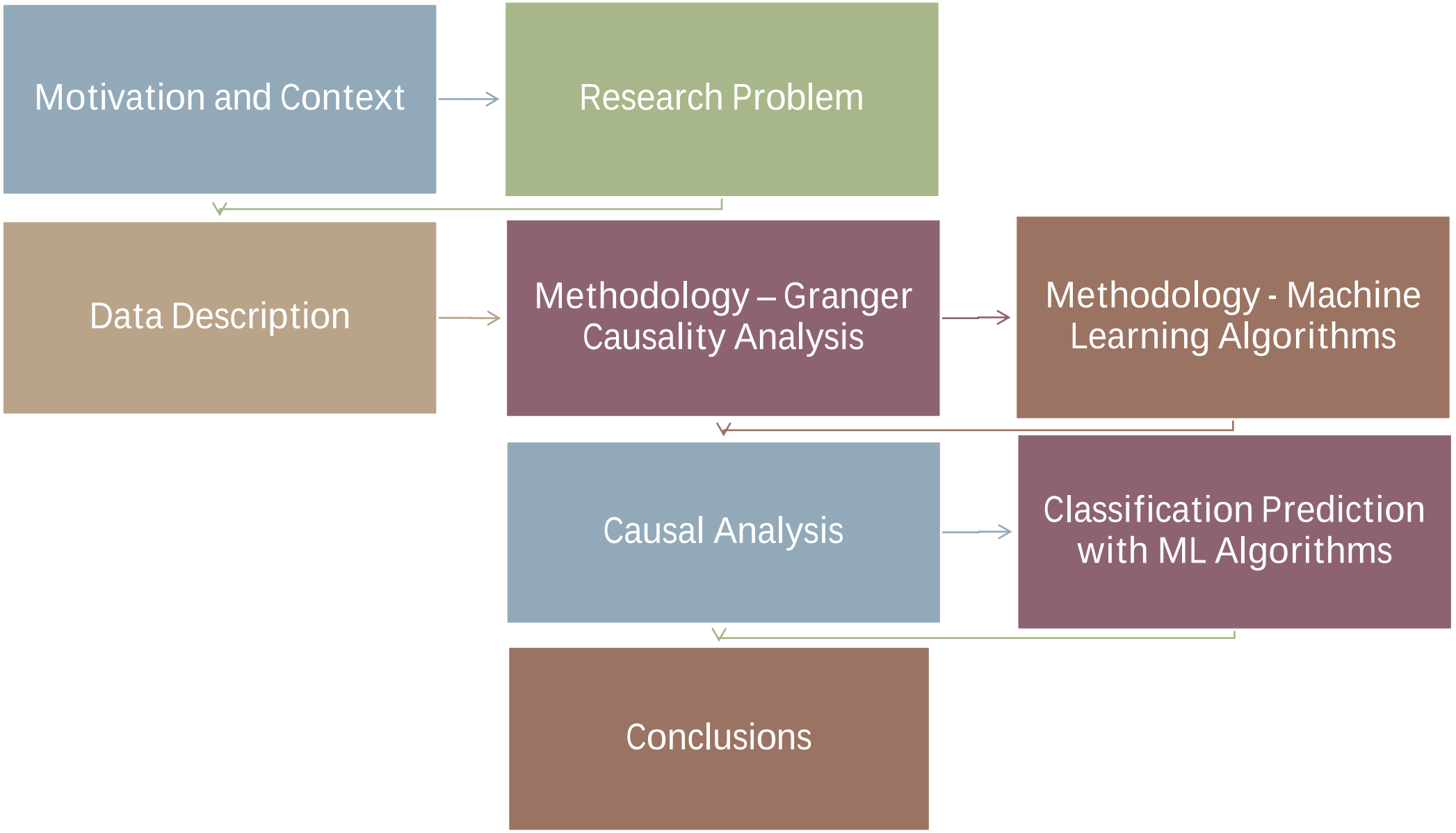
8th Road Safety and Simulation Conference
Athens, Greece
June 8 – 10, 2022

Meghna Chakraborty
Timothy J. Gates
Jhelum Chakravorty

College of Engineering
MICHIGAN STATE UNIVERSITY

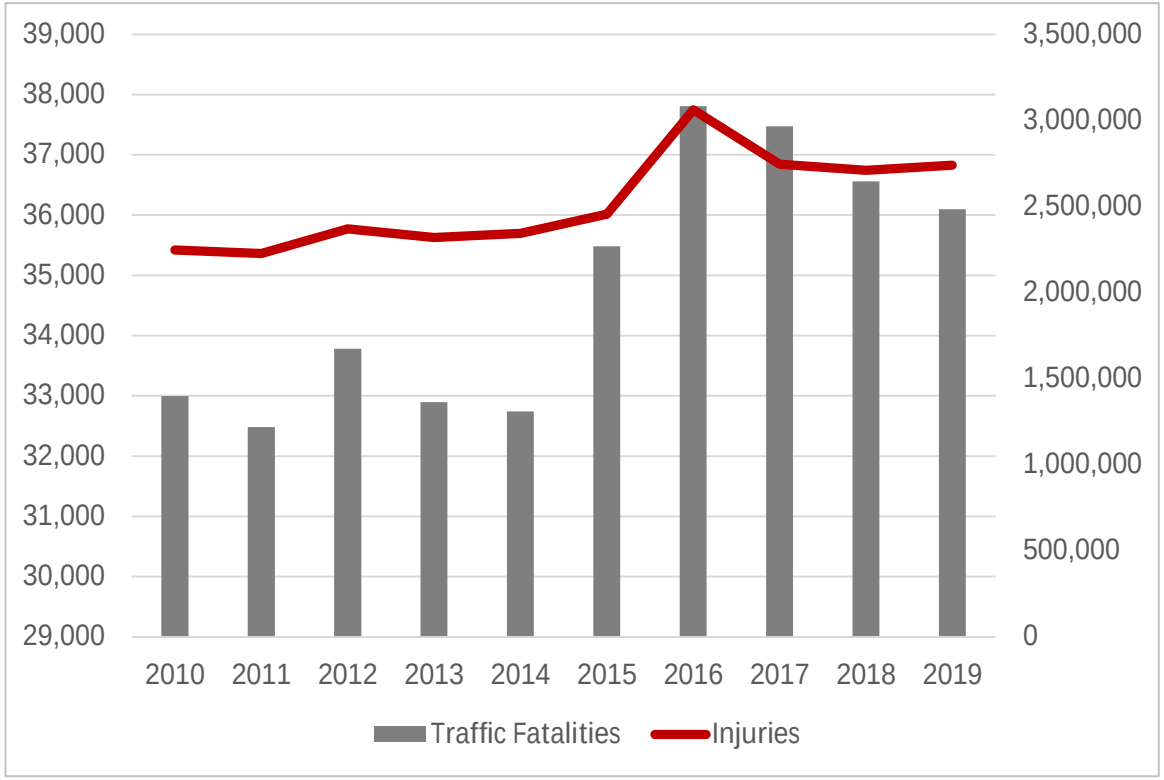
HITACHI
Inspire the Next
Hitachi Energy

Overview


Motivation and Context
Research Problem
Data Description
Methodology – Granger Causality Analysis
Methodology - Machine Learning Algorithms
Causal Analysis
Classification Prediction with ML Algorithms
Conclusions

Motivation and Context

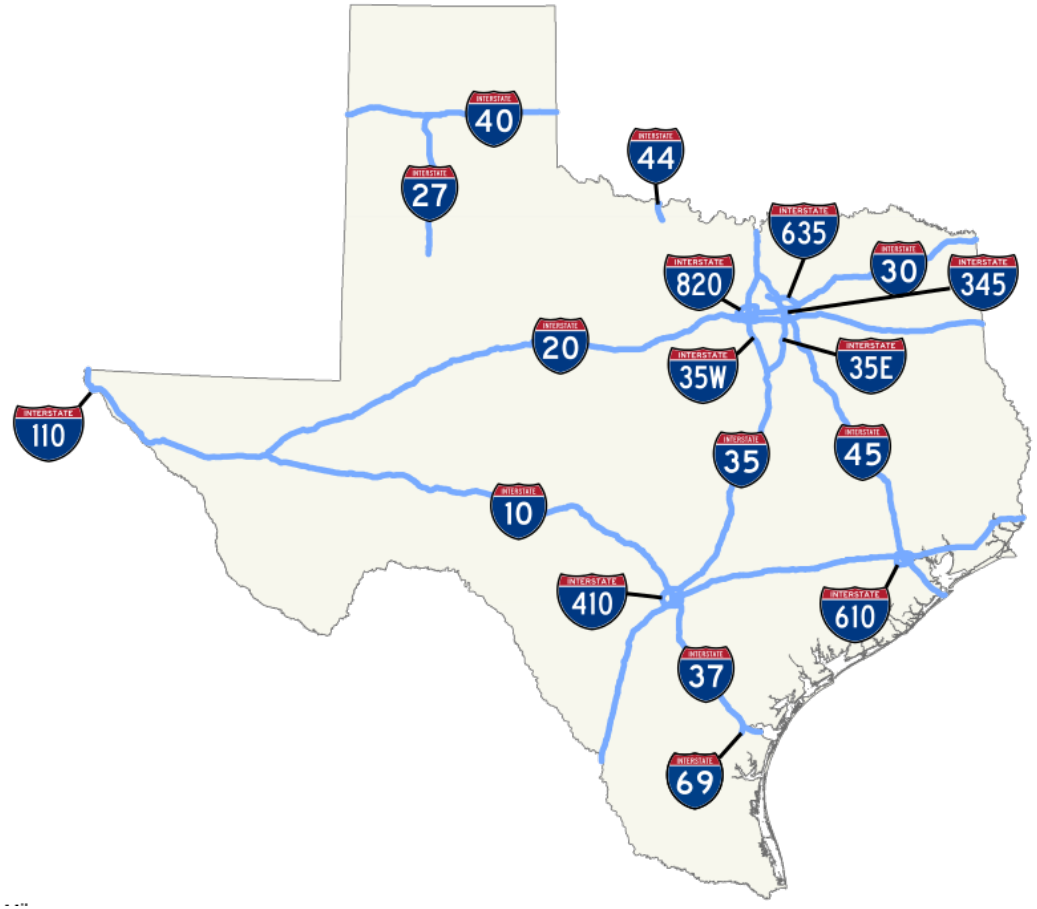
- Incredible economic and societal impact associated with traffic crashes (>36,000 fatalities and ~2.74 million injuries in US in 2019).
- Previous research largely utilized classical statistical techniques - methodological limitations not fully understood or accounted for.
- Emerging data mining techniques offer superior prediction and greater accuracy.
- Increasing availability of large-scale data.
- Highly imbalanced crash data – Need algorithms capable of dealing with mislabels in model training.
- Causal analysis with Granger causality – Widely used, especially for static data, and popular for identifying influential factors.



Summary statistics of the data prior to data balancing (FARS 2019)

Research Problem

- This study presents a methodological framework to model the severity of motor vehicle crashes on urban/suburban interstates.
- The analysis involves causal inference, using Granger causality tests and injury severity classification using different machine learning and deep learning algorithms including decision trees, random forests, extreme gradient boosting, and deep neural net.
- The output of the proposed crash severity classification approach includes three classes: fatal and severe injury (KA) crashes, non-severe and possible injury (BC) crashes, and property damage only (PDO) crashes.
- The background of this study is premised in Texas as it has historically been among the top states in terms of statewide fatalities in the U.S.
- Midblock crashes occurring on all urban/suburban interstates statewide evaluated.



Texas statewide interstate map

Data

- Crash data - Crash Records Information System (CRIS) by TxDOT.
- Crash location and period – All interstates in Texas (urban/suburban areas) between 2014 and 2019.
- A total of 156,166 crashes.
- Traffic volumes vastly vary across the freeways.
- Minimum speed limit for about 2% of crashes = 45 mph (crash locations include work zones).
- Over a quarter of observations have HOV lanes.
- Most crashes occurred on segments with more than 4 lanes (considering both traffic directions).

Parameter	Min.	Max.	Mean	S.D.
Fatal and Severe Injury Crashes (KA)	0	1	0.03	0.16
Non-severe and Possible Injury Crashes (BC)	0	1	028	0.45
Property Damage Only Crashes (O)	0	1	0.69	0.46
Annual Average Daily Traffic (AADT) (vpd)	4,606	330,096	144,961	63,631
Speed Limit (mph)	45	80	63.09	5.96
Proportion of Heavy Vehicle (%)	3.7	95.08	11.49	6.80
Single Vehicle Crashes (SV)	0	1	0.16	0.37
Work Zone Presence	0	1	0.13	0.34
Worker Prersent in the Work Zone	0	1	0.05	0.21
Vurnerable Road Usert Involved	0	1	0.004	0.06
Number of Lanes > 4	0	1	0.84	0.36
Commercial Vehicle Involved	0	1	0.15	0.36
Presence of High Occupancy Vehicle Lanes	0	1	0.27	0.44
Population > 50,000	0	1	0.93	0.16
Crash Time = Peak Hour	0	1	0.29	0.46
Dry Surface	0	1	0.83	0.37
Clear Weather	0	1	0.71	0.45
Daylight or Dark Unlighted	0	1	0.90	0.30
Curved Road Alignment	0	1	0.15	0.23
Left Turning Curved Road	0	1	0.10	0.30
Spiral Curved Road	0	1	0.01	0.08
Median Barrier Not Present	0	1	0.01	0.12
Median Width < 12 feet	0	1	0.44	0.50
No Left Shoulder Present	0	1	0.09	0.29
Left Shoulder Width < 6 feet	0	1	0.03	0.16
No Right Shoulder Present	0	1	0.01	0.03
Right Shoulder Width < 6 feet	0	1	0.001	0.01

Summary statistics of the data prior to data balancing

Causal Analysis

- Granger causality - One of the most commonly used notion of causality and has been used for causal inference from experimental data.
- Suppose X, Y and Z are three jointly distributed multivariate stochastic processes and consider the regression models.

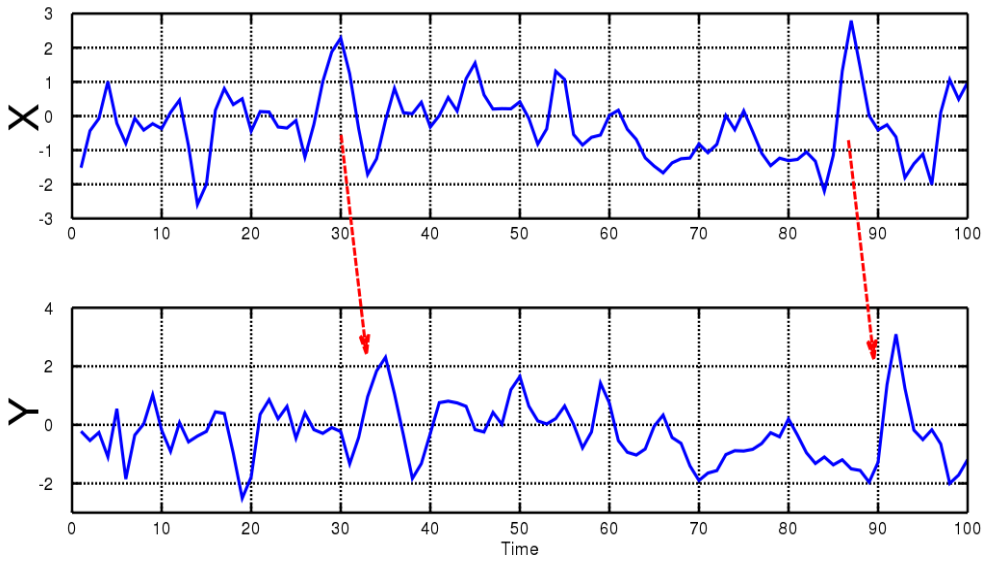
$$X_t = \alpha_t + (X_{t-1}^{(p)} + \bigoplus Z_{t-1}^{(r)}) \cdot A + \varepsilon_t$$

$$X_t = \alpha_t' + (X_{t-1}^{(p)} + \bigoplus Y_{t-1}^{(q)} \bigoplus Z_{t-1}^{(r)}) \cdot A' + \varepsilon_t'$$

The Granger causality of Y on X, given Z, is given by

$$G_{Y \to X|Z} = \ln \frac{var(\varepsilon_t)}{var(\varepsilon_t')}$$

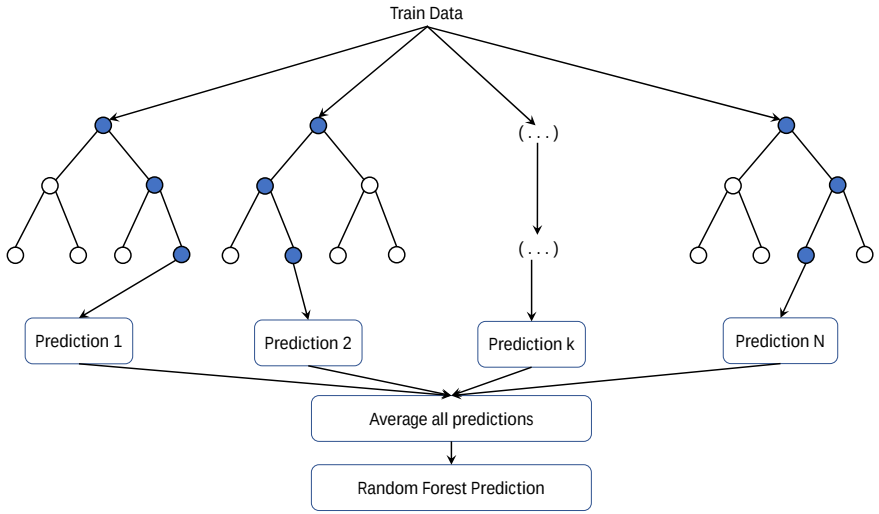
where,
A and A' are the regression coefficients,
α and α' are constants, and
ϵ_t and ϵ_t' are the residuals

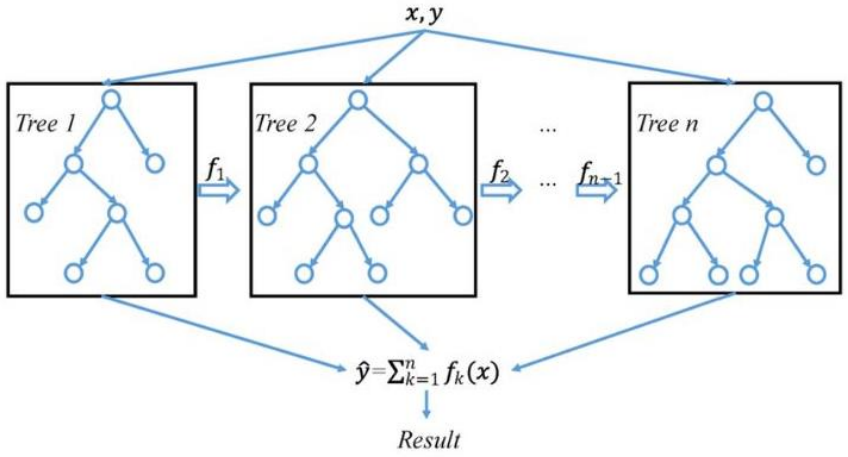


Granger Causality; Source – Wikipedia

Machine Learning Algorithms for Classification

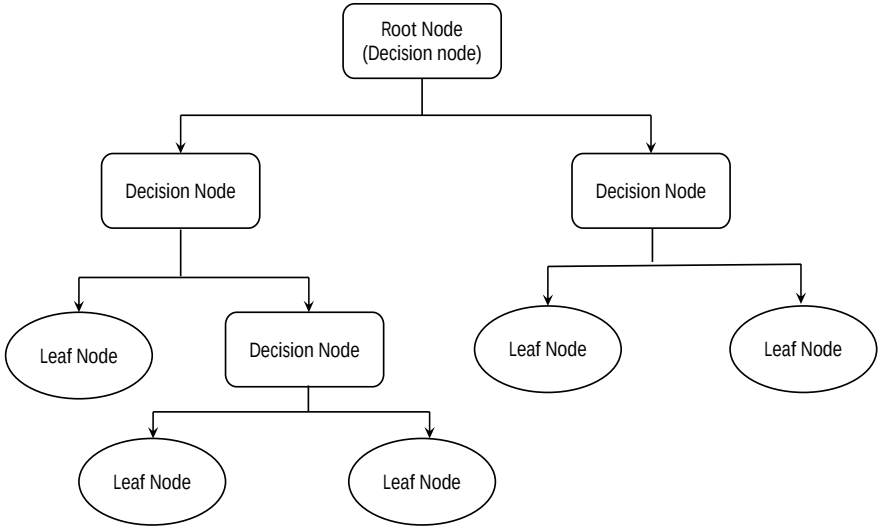
Decision Trees

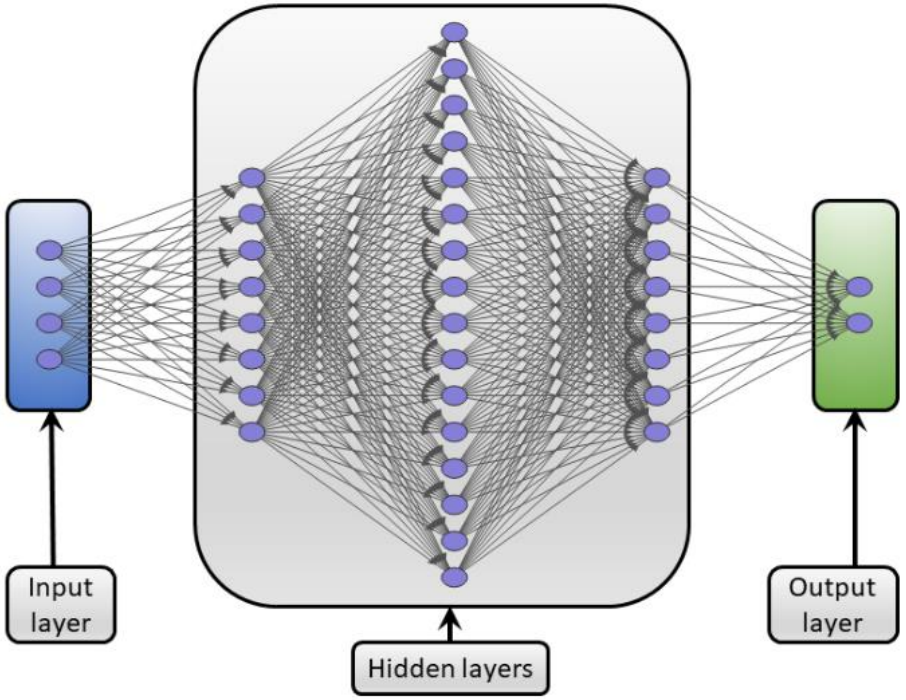




Extreme Gradient Boosting

Random Forests

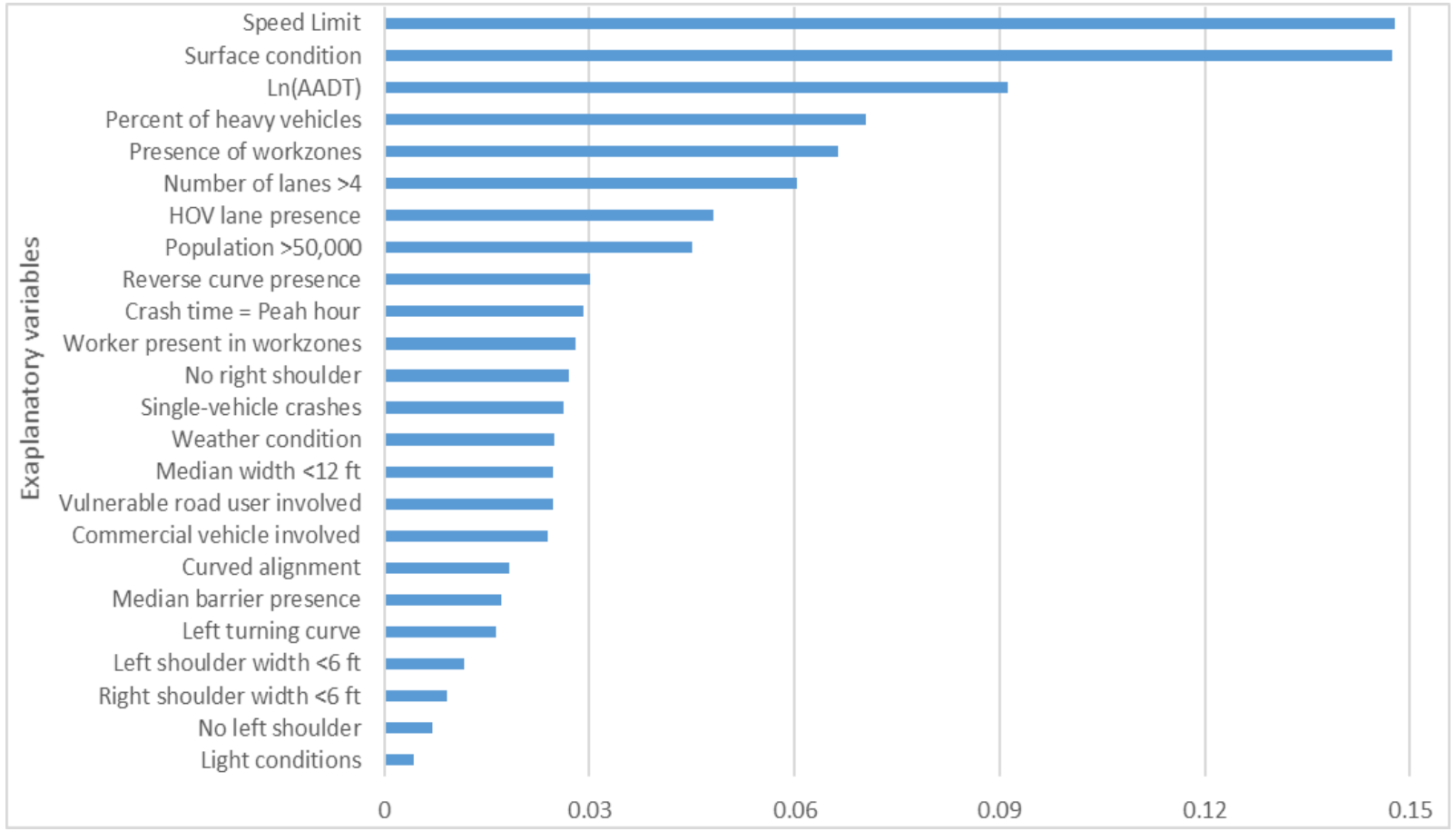




Deep Neural Net

Granger Causality

- Evaluated a total of 24 variables – also used for full classifiers.
- Identified and rank-ordered a set of most influencing predictors based on causality scores.

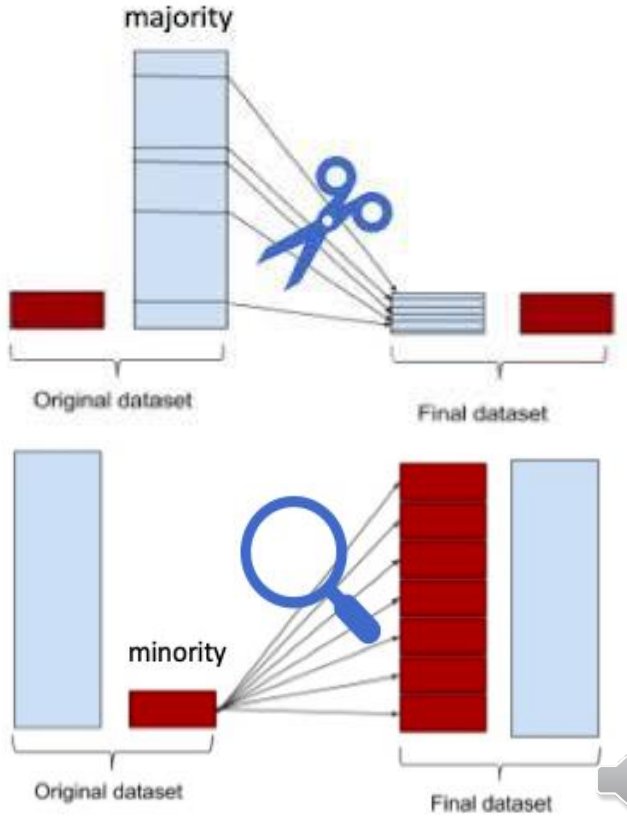


Rank ordering of the predictors based on causality scores

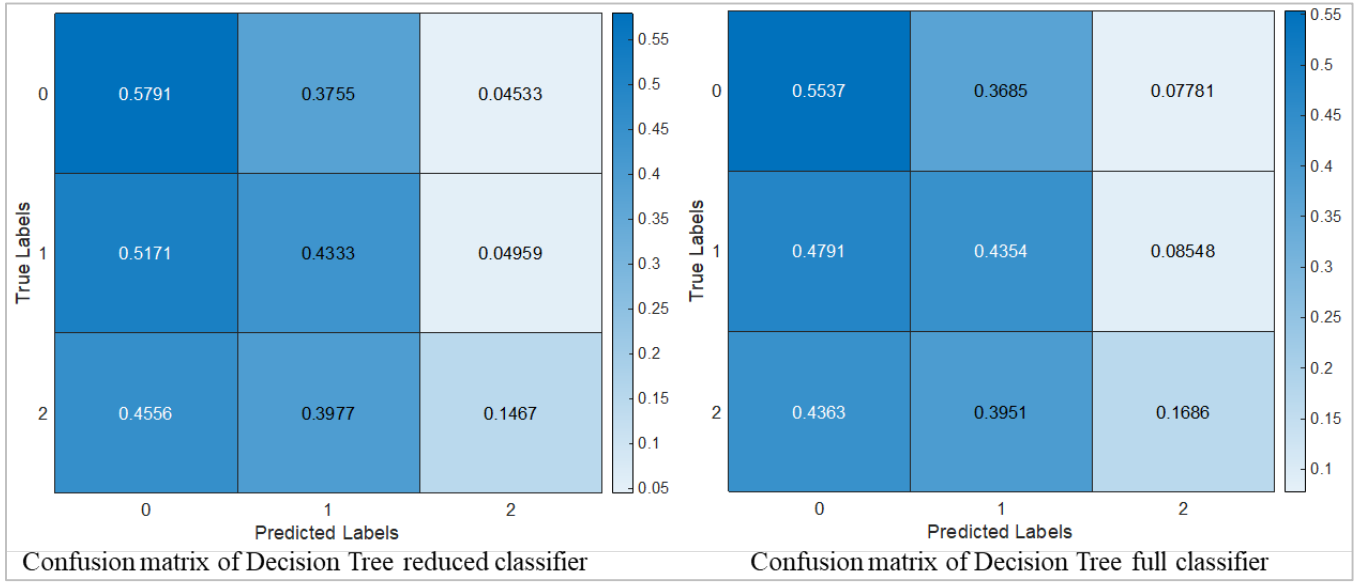
- The optimal lag for the VAR model was 4 (using AIC).
- Ultimately, a total of 17 predictors with the highest scores were selected – used for reduced classifiers.

Crash Severity Classification

- Classification predictions using
  - Decision Trees
  - Random Forests
  - Extreme Gradient Boosting (XGBoost)
  - Deep Neural Net
- Prediction performances between full and reduced classifiers compared.
- 3 injury classes:
  - Fatal and severe injury (KA),
  - Non-severe and possible injury (BC), and
  - No injury or property damage only (O or PDO).
- Train vs test: 80% of data for training. 20% of data for testing
- Data Balancing: Training data was balanced by
  - Random under-sampling of no injury class, and
  - Over-sampling of the data of the remaining classes using the SMOTE
- Prediction Performance Metrics:
  - F1 score and AUC (Area Under the ROC Curve) score, and
  - Confusion matrix for both full and reduced classifiers.



Classification Predictions



Confusion matrix of Decision Tree reduced classifier

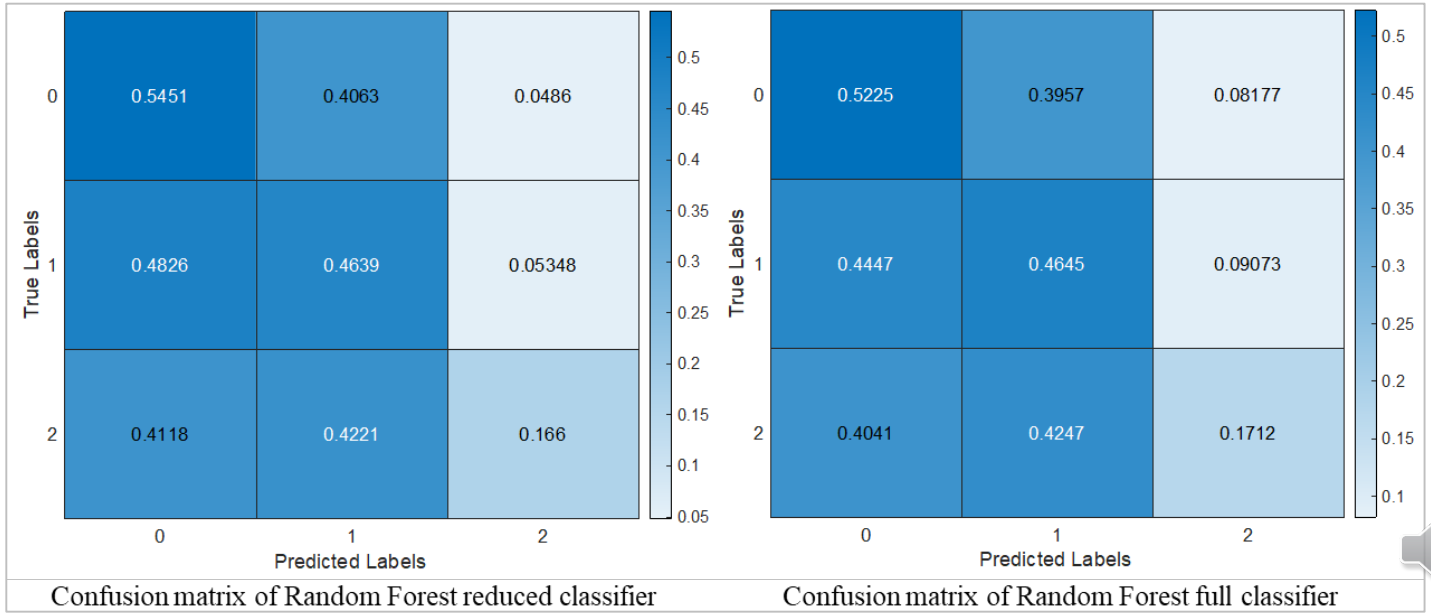
Confusion matrix of Decision Tree full classifier

Decision Trees

- For PDO (Label 0) class performance improves with reduced classifier.
- True positives for both KA (Label 2) and BC (Label 1) classes decrease with reduced classifier,

Random Forests

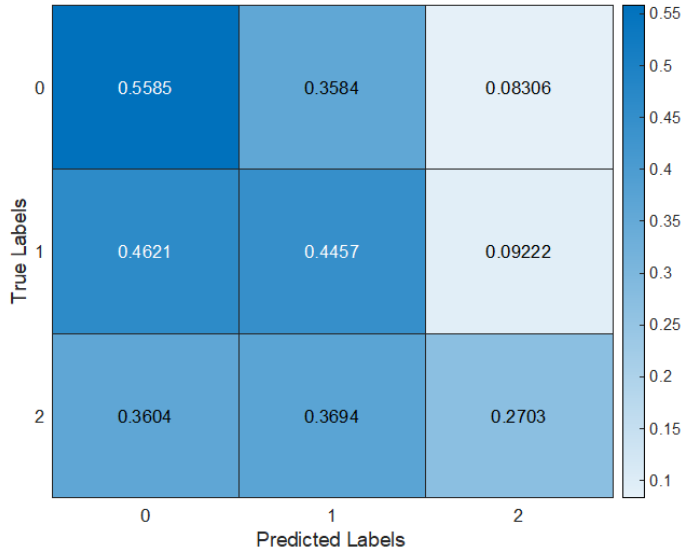
- The classification performance for the PDO crashes improves for the reduced classifier.
- For the BC class, it remains almost same.
- For KA class, the performance of the reduced classifier degrades.



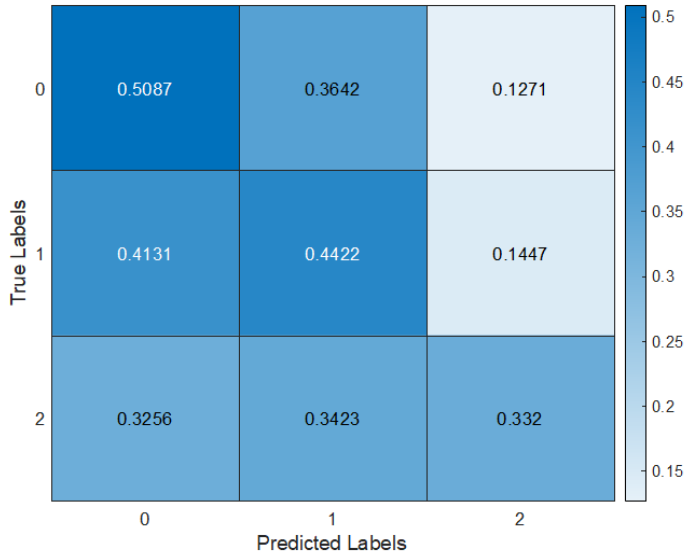
Confusion matrix of Random Forest reduced classifier

Confusion matrix of Random Forest full classifier

Classification Predictions



Confusion matrix of XGBoost reduced classifier



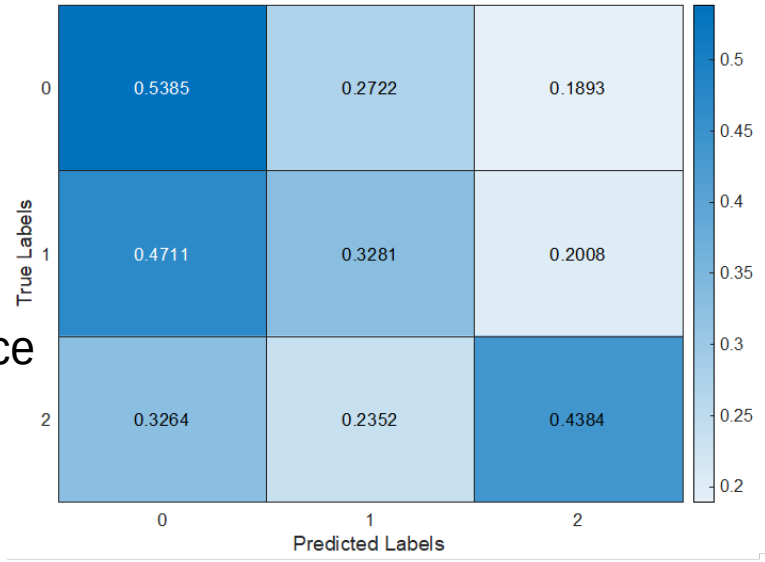
Confusion matrix of XGBoost full classifier

Extreme Gradient Boosting (XGBoost)

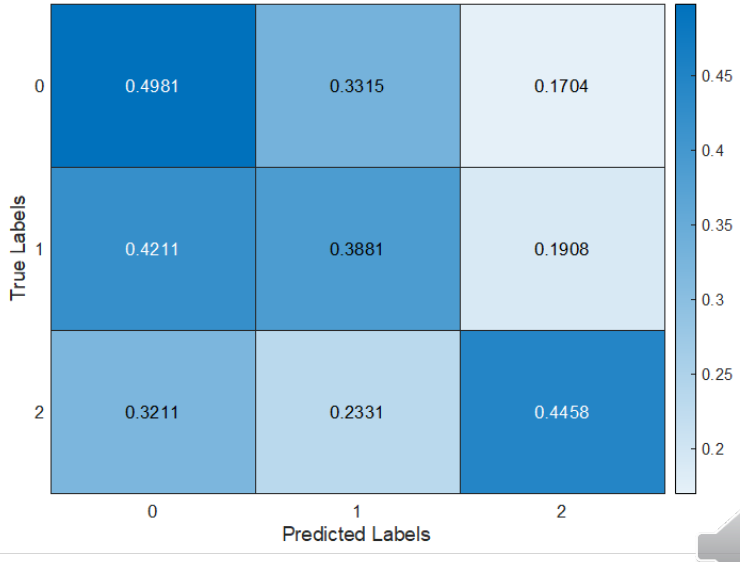
- Considerable improvement in prediction performance for KA class compared to DT and RF classifiers.
- The predictive performance of the reduced classifier improves substantially for PDO class.
- For BC class, it remains almost the same.

Deep Neural Net

- Considerable improvement in prediction performance for KA class compared to all other classifiers.
- For KA and BC classes, performance of reduced classifier improves.
- Prediction performance degrades for PDO crashes with reduced classifier.



Confusion matrix of Deep Neural Net reduced classifier



Confusion matrix of Deep Neural Net full classifier

Conclusions

- The study presents a methodological framework involving the development of causal inference and injury severity classification for freeway traffic crashes.
  - Granger causality test to identify and rank-order the influential features.
  - Four different classifiers, including Decision Tree, Random Forest, Extreme Gradient Boosting (XGBoost), and Deep Neural Net.
  - Output classes: KA crashes, BC crashes, and PDO crashes.
  - Most influencing factors: speed limit, surface/weather conditions, traffic volume, presence of workers in workzones, HOV lanes etc.
  - Efficacy of the Granger causality was demonstrated by achieving improved or comparable results between reduced order and full order models.
  - Decision tree and random forest classifiers provided the greatest performance for PDO and BC crash severities, respectively.
  - For the KA class, the rarest class in the data, deep neural net classifier performed most superior.

Thank you!

Contact:
Meghna Chakraborty

chakra43@msu.edu
https://www.linkedin.com/in/meghnach/