POSTPRINT VERSION. The final version is published here :
Fall, T., Massicotte, D., Dessureault, J. S. et Ouameur, M. A. (2025, 25–27 June 2025). Urban Intersection Safety Risk Index: Machine Learning Methods for Real-Time Classification. Communication présentée 2025 25th International Conference on Digital Signal Processing (DSP), Pylos, Greece. doi: 10.1109/DSP65409.2025.11075008

Urban intersection safety risk index: Machine learning methods for real-time classification

Thierno Fall
Departement of Electrical and Computer Engineering
Université du Québec à Trois-Rivières
Trois-Rivières, Canada
Mame.Thierno.Mbacke.Fall@uqtr.ca

Daniel Massicotte
Departement of Electrical and Computer Engineering
Université du Québec à Trois-Rivières
Trois-Rivières, Canada
Daniel.Massicotte@uqtr.ca

Jean-Sébastien Dessureault
Departement of Mathematics and Computer science
Université du Québec à Trois-Rivières
Trois-Rivières, Canada
jean-sebastien.dessureault@uqtr.ca

Messaoud Ahmed Ouameur
Departement of Electrical and Computer Engineering
Université du Québec à Trois-Rivières
Trois-Rivières, Canada
messaoud.ahmed.ouameur@uqtr.ca

***Abstract*—The safety of urban intersections is a critical concern for city planners. Technological advancements, such as LiDAR sensors, enable better risk assessment for road users. This study proposes a hybrid model that combines Post-Encroachment Time (PET) data with unsupervised machine learning techniques, specifically DBSCAN clustering, to detect traffic anomalies. A generalized Pareto distribution (GPD) is then applied to estimate a risk index. Finally, categorical safety risk classification is performed using an optimizable neural network (ONN), support vector machine (OSVM), efficient logistic regression (ELR), and Gaussian Naïve Bayes (GNB). The impact of these methods is evaluated in real-time for urban traffic management in Trois-Rivières, Quebec, Canada. This work aims to assist decision-makers in urban traffic planning and accident prevention.**

***Index Terms*—Surrogate measures, Post-Encroachment Time (PET), Machine learning, DBSCAN, Pareto distribution, Classification.**

I. Introduction

Intelligent transport systems develop and integrate methods to overcome the high demand for economic concerns, the reliability and quality of infrastructures, and most importantly, the safety of road users. Each year, 1.35 million people are killed on the roads of the world, and another 20 to 50 million are seriously injured [1].

In road traffic, many users interact with each other, and the need for reliable measurement systems (on a week, day, and even hour basis) of interactions between users is growing. These interactions will likely generate conflicts (between vehicles, pedestrians, and motorcycles) as they cross paths in all directions. Thus, conflicts occur when traffic streams moving in different directions interfere. The number of possible conflict points at any intersection depends on the number of approaches, the turning movements, and the type of traffic control at the intersection [2].

PET is calculated as when the first road user leaves the conflict point and the second reaches the same point. It is defined as the time between moment t_1 when the first road user exits the conflict point and moment t_2 when the second enters the same conflict spot, as shown in Figure 1. The smaller the PET, the higher the risk of collision. Moreover, a PET value less than zero would indicate a crash occurrence [3].

Due to the rarity of dangerous events, such as accidents and near-misses, and the lack of timeliness, having a reliable model that gives a safety index of road traffic is challenging. Surrogate traffic measures can be used to determine how dangerous an intersection or road section is.

[4] proposes a conflict-based traffic safety assessment method by associating conflict frequency and severity with short-term traffic characteristics. [5] conducts a systematic review of conflict-based safety measures with a specific focus on the context of their applications. [6] uses conflict indicators, PET and Time to Collision (TTC) to identify pedestrian conflicts and predict pedestrian conflicts one cycle ahead, which can be 2-3 min, whereas [7] employed surrogate safety indicators to measure the safety level of pedestrian conflict with other road users to evaluate conflict risk.

Despite the significant increase in interest in surrogate measures, many approaches lack real-time applicability. This study introduces a machine learning-based risk index capable of real-time classification of intersection safety levels. Our contribution uses PET as a surrogate measure and focuses on 1) machine learning techniques to evaluate the behavior of individual road intersections, 2) calculating the safety index for an hourly divided block of conflicts of all the data frames from ten intersections, and 3) using the generated safety risk level and conflict features (PET and speed) to train and test classification methods.

The paper is structured as follows: Section II presents the methodology, including data description, anomaly detection, the generalized Pareto Risk Index, and safety level classification. Section III discusses the results and their implications. Finally, conclusions are provided in Section IV.

II. Methodology

A. Data description

Traffic data are collected via Velodyne LiDAR sensors by third-party BlueCity. Thus, we assume the collected data completely satisfy the requirements of reliability, repeatability, and practicability [8]. The main fields from accessible conflict data are PET and speed (involved speed). The PET data range is from 0 to 10 seconds, whereas the speed can be up to 100 km/h. For the present contribution, PET and speed, more

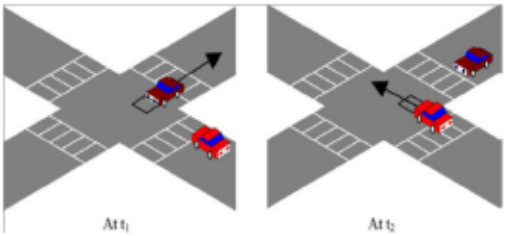


Fig. 1: Illustration of PET conflicts [3]

than 4 seconds and less than 20 km/h, respectively, are not considered. Concerning [3], we use the negated PET instead of the actual PET.

The conflicts are event-based and therefore remain stochastic. As depicted by Table I, the dataset comprises, in addition to PET and speed, date and time, types of users involved, and geographical locations to identify intersection hot spots.

B. Anomaly detection

The DBSCAN addresses the incompatibility of the classical cluster algorithm (i.e., K-means) to particular shaped data as stated in [9]. DBSCAN clustering is used to classify conflict points into normal and abnormal events. A density-based approach ensures the identification of rare, high-risk occurrences. After analyzing the architecture of our data, the intuitive notion of "clusters" and "noises" in a given database [9] is easily applicable. There is a large density of points around a range of PET and speed, which appear to be clustered and might be interpreted as normal or usual behavior in users' conflicts surrounded by rare and infrequent conflicts.

Serious conflicts leading to fatalities and general interactions between road users are less frequent than usual. It can be observed from the joint distribution shown in Figure 2 of PET and speed. Thus, these dangerous conflict data points appear to be noise with the DBSCAN technique according to the definition of the algorithm and the expected rarity of these events.

In this work, the threshold is not determined using the conventional parameter stability analysis method. Instead, it is defined as a percentile, computed from the ratio between the number of noisy points and the average number of points within the identified clusters. This ratio-based approach provides a data-driven and adaptive way to establish the threshold.

The parameters of the DBSCAN, clusters minimum points γ and maximal distance between points ϵ were determined empirically through a grid search to balance anomaly detection sensitivity and noise reduction. These values optimize the distinction between normal and anomalous traffic conflicts while maintaining high detection precision. That maneuver limits performance because intersection traffic activities are relatively different.

C. Generalized Pareto risk index

Exceedances of negated PET and speed are modeled using a GPD. Following the approach of [3], this study adheres closely to the formulations and theoretical foundations presented in [10]. The risk index is defined by Equation (1).

$$R_i x = Pr(X_i > \mu) = 1 - G(X_i) \tag{1}$$

Where $G(X_i)$ is the GPD function, μ is the threshold limit, and X_i is the random variable.

TABLE I: Conflicts fields informations

Feature	Description	Type	Range
PET	Time difference between two users crossing same point	Continuous	0 - 10 seconds
Speed	Speed difference between involved users	Continuous	$>$ 10 km/h
$1^{st} user$		Categorical	Car, Pedestrian, Bus, Bicycle, Motorcycle, Truck
$2^{nd} user$		Categorical	Car, Pedestrian, Bus, Bicycle, Motorcycle, Truck
Date and Time	Date and time of occurence	Date	
Position	GPS coordinates where the conflicts occurs	Decimal	Latitude and Longitude

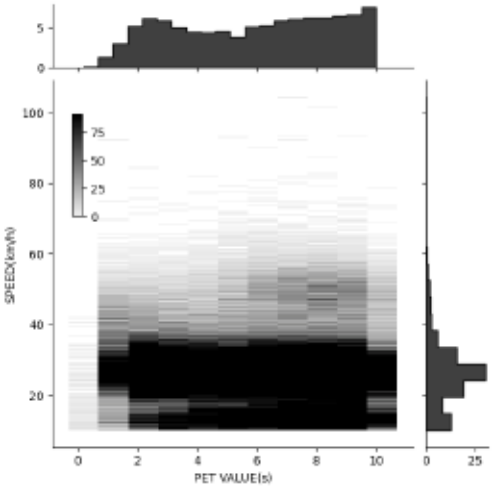


Fig. 2: Joint Marginal Histogram of PET and speed

As demonstrated by [10], the GPD function of (1) is formulated in (2).

$$G(x) = \exp\left(-\left[1 + \xi\frac{x-\mu}{\sigma}\right]^{\frac{1}{\xi}}\right) \tag{2}$$

where the parameters are the shape, the scale and the threshold namely $\xi \neq 0$, σ and μ, respectively.

The choice of the GPD is justified by the exponential decay observed in the tail of our data distribution [11], particularly for the negated PET. The parameters of the GPD distribution are estimated using maximum likelihood estimation (MLE), with 95% confidence intervals for parameter estimates to ensure an optimal fit to the data. Subsequently, Equation (1) is applied to the PET and speed exceedance data for the i^{th} hourly block, computing R_ip and R_is separately. The overall risk index R_i is then obtained as the mean of these two values. Once completed, a theoretical safety category is established based on the computed risk index.

$$R_i = \frac{R_ip + R_is}{2} \tag{3}$$

D. Classification of the safety risk level

All previously collected PET and speed data are available, making it straightforward to compute the risk index as de-

scribed in II-C. For classification, the input is conflict features PET and speed. We employ four machine learning models: neural networks, support vector machines (SVM), efficient logistic regression (ELR), and Gaussian naïve Bayes (GNB). Among these, the neural network and SVM undergo further optimization. The optimized parameters for the neural network include: the number of fully connected layers between 1 - 3, the activation function: ReLU, tanh, sigmoid or none, the regularization strength between 3.31 10 ^{10} - 3.31, the first and second layer sizes between 1-300 and standardized data or not. For the SVM model, the optimized parameters include: the kernel function: gaussian, linear, quadratic or cubic, the constraint level: 0.001 - 1000, the multiclass coding: One-vs-All or One-vs-One and standardized data or not.

The safety level classification is defined as follows: $green$ safety category if $R_t \leq 0.35$, $yellow$ if $0.35 < R_t \leq 0.65$ and if $R_t < 0.65$ then the category is red. Each conflict data point serves as an input for the classification model. The classification thresholds were empirically determined through sensitivity analysis to optimize model accuracy while ensuring meaningful risk differentiation.

III. Results and Discussion

The primary goal of this study is to accurately predict the safety level (category) of intersections in the city of Trois-Rivières. All simulations and implementations were carried out using MATLAB tools. For anomaly detection, the DBSCAN parameters are set to $\epsilon = 0.45$ and $\gamma = 220$. As shown in Figure 3, the identified abnormal data points align with those in the joint distribution presented in Figure 2.

In Figure 4, the blue histogram represents the probability density of PET exceedances, capturing values below the threshold (in the context of negative values). The bars indicate the frequency of these exceedances, normalized to reflect a probability density. The red line depicts the GPD fit applied to the exceedance data. The sharp increase near zero suggests that the distribution is heavily concentrated around low exceedance values, a characteristic commonly observed in distributions with Pareto-like tails.

In Figure 4a, the blue histogram represents the probability density of PET exceedances, capturing values below the threshold (in the context of negative values). The bars indicate the frequency of these exceedances, normalized to reflect a probability density. The red line depicts the GPD fit applied to the exceedance data. The strong alignment between the histogram and the red curve in Figure 4a demonstrates a good fit, validating the use of the GPD to model PET exceedance data. This confirms the suitability of GPD for capturing extreme PET values.

Table II provides a summary of the fit results for both PET and speed. The shape parameter (ξ = -0.16) indicates a bounded tail, implying a limited range of extreme values. However, slight discrepancies between the fitted GPD and the histogram in Figure 4a arise due to data sparsity in extreme-value regions.

In Figure 4b, the blue histogram represents the probability density of speed exceedances. The bars indicate the frequency of these exceedances, normalized to reflect a probability density. The red line shows the GPD fit applied to the speed exceedance data. The positive shape parameter (ξ = 0.19) indicates a heavy-tailed distribution, suggesting a higher probability of extreme values.

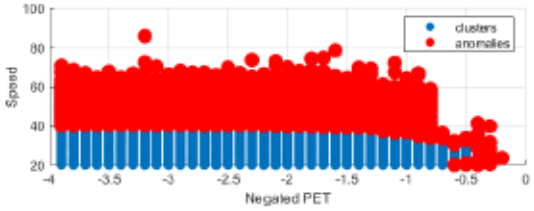
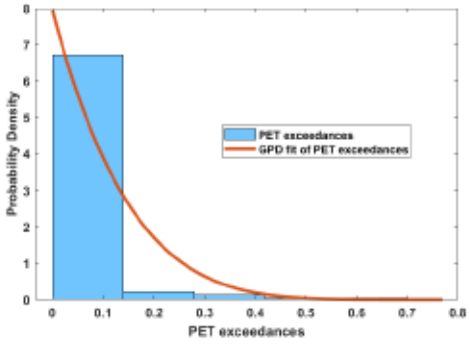
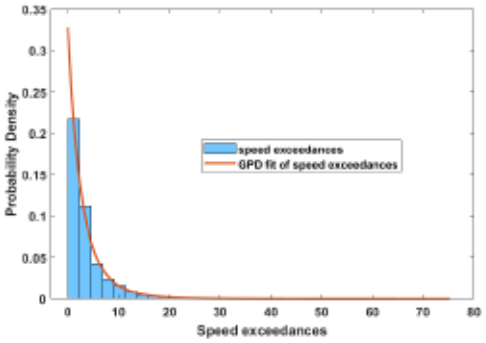


Fig. 3: DBSCAN anomaly detection



(a) Fit of Negated PET



(b) Fit of speed

Fig. 4: GPD Fit of PET and Speed Exceedances

For classification, 75% of the data is allocated for training and validation, with cross-validation applied to mitigate overfitting. The remaining 25% is reserved for testing. The training dataset consists of a total of 41079 samples, distributed across classes. The class distribution is as follows: Class $green$ – 5720 samples, Class $yellow$ – 28084 samples, and Class red - 7275 samples. As shown in Table III, overall accuracy is used as a key performance metric, where higher accuracy indicates better classification performance. Additionally, total cost represents the number of misclassified instances (lower is better), while the error rate serves as an indicator of the model's misclassifications tendency.

Further insights can be gained from metrics such as the true positive rate (TPR), which quantifies the proportion of correctly predicted instances within each class, and the false negative rate (FNR), which measures the proportion of misclassified negative instances. The confusion matrices presented in Tables IV detail the classification performance for validation and test datasets across different methods.

The comparative classification is conducted using four models: optimizable neural network (ONN), optimizable support vector machine (OSVM), efficient logistic regression (ELR), and Gaussian naïve Bayes (GNB). The evaluation metrics include accuracy (for validation and test datasets), TPR, FNR, and confusion matrices to assess model performance.

TABLE II: GPD Fit Parameters

Negated PET GPD fit			Speed GPD fit		
Shape:ξ	*Scale:*σ	*Thres.:*μ	*Shape:*ξ	*Scale:*σ	*Thres.:*μ
-0.16	0.12	-0.9	0.19	3.04	56.2

TABLE III: Test and validation results

	Validation			Test		
	Acc.	**T.Cost**	**E.Rate**	**Acc.**	**T.Cost**	**E.Rate**
ONN	*78.95%*	*6340*	*21.05%*	*79.37%*	*5084*	*20.63%*
OSVM	*77.69%*	*6720*	*22.31%*	*78.88%*	*5205*	*21.12%*
ELR	*76.82%*	*6983*	*23.18%*	*78.14%*	*5387*	*21.86%*
GNB	*75.82%*	*7285*	*24.18%*	*77.72%*	*5491*	*22.28%*

The majority of PET and speed data points fall into the *yellow* category, indicating a high classification accuracy for this class, as shown in the confusion matrix results in Tables IV. To address this imbalance, the dataset should be adjusted to ensure a more even distribution across all classes. Among the tested models, optimized neural network (ONN) achieved the highest accuracy, reaching 78.95% for validation and 79.37% for testing. It was followed by the optimized support vector machine (OSVM) with 77.69% and 78.88%, respectively, then ELR 76.82% and 78.14% and finally GNB 75.82% and 77.72%. The optimal parameters for the ONN model include: 2 fully connected layers, a tanh activation function, a regularization strength of $4.10e^{-10}$, a first layer size of 298, a second layer size of 3 and standardized data. For OSVM, the optimized parameters are: a quadratic kernel function, a constraint level of 39.76, One-vs-One multiclass coding, and standardized input data. ONN outperforms other models due to its ability to capture non-linear relationships, extract hierarchical features, and optimize hyperparameters for better adaptability. Its robustness in classifying both high-risk (red) and low-risk (green) scenarios makes it the most reliable choice for real-time risk assessment.

Some may argue that the GPD alone is sufficient for classifying conflicts. However, its limitations must be considered, particularly its reliance on complete data blocks over a set period before classification can be performed. While the classification methods in this study use instantaneous PET and speed as predictors (bi-variate input), they operate on individual data points rather than aggregated groups, enabling real-time processing. Moreover, their accuracy is still evaluated based on the GPD framework (theoretical classification of computed R_t).

IV. Conclusion

This study introduced a machine learning-based approach for real-time intersection risk assessment, leveraging DBSCAN anomaly detection and a GPD risk model to classify intersection safety levels. Unlike traditional methods based on historical accident data, our approach dynamically evaluates surrogate safety measures, such as PET and speed anomalies, to improve traffic management and accident prevention.

One limitation of our approach is the use of fixed DBSCAN parameters across all intersections, which may reduce its effectiveness in detecting anomalies in varying traffic conditions. To address this, we suggest a unified anomaly detection model that dynamically adjusts to different intersection profiles using a generalized bi-variate threshold. Additionally, to improve the detection of rare events, DBSCAN could be combined with an adaptive machine learning method that optimizes clustering parameters in real-time.

TABLE IV: Test confusion matrix results

ONN		**Predicted**			
		Green	**Yellow**	**Red**	**FNR**
True	**Green**	*2945(**71.8%**)*	*1072(26.1%)*	*83(2%)*	***28.2%***
	Yellow	*980(7%)*	*12229(86.8%)*	*886(6.3%)*	*13.2%*
	Red	*204(3.2%)*	*1859(28.8%)*	*4388(**68%**)*	***32%***
OSVM		**Predicted**			
		Green	**Yellow**	**Red**	**FNR**
True	**Green**	*2444(59.6%)*	*1605(39.1%)*	*51(1.2%)*	*40.4%*
	Yellow	*501(3.6%)*	*13051(92.6%)*	*543(3.9%)*	*7.4%*
	Red	*145(2.2%)*	*2360(36.6%)*	*3946(61.2%)*	*38.8%*
ELR		**Predicted**			
		Green	**Yellow**	**Red**	**FNR**
True	**Green**	*2643(64.5%)*	*1438(35.1%)*	*19(0.5%)*	*35.5%*
	Yellow	*648(4.6%)*	*12994(92.2%)*	*453(3.2%)*	*7.8%*
	Red	*199(3.1%)*	*2630(40.8%)*	*3622(56.1%)*	*43.9%*
GNB		**Predicted**			
		Green	**Yellow**	**Red**	**FNR**
True	**Green**	*2321(56.6%)*	*1740(42.4%)*	*39(1%)*	*43.4%*
	Yellow	*435(3.1%)*	*13226(**93.8%**)*	*434(3.1%)*	***6.2%***
	Red	*174(2.7%)*	*2669(41.4%)*	*3608(55.9%)*	*44.1%*

The GPD effectively models the overall distribution of PET and speed exceedances, offering valuable insights for risk assessment. However, its performance may be less reliable when applied to a single intersection.

In terms of classification performance, the optimized neural network (ONN) outperformed other models, particularly in identifying high-risk (red) and low-risk (green) scenarios. Gaussian Naïve Bayes (GNB) demonstrated fair accuracy for moderate-risk (yellow) classification, but ONN remained the most effective overall, achieving the highest global accuracy. Further refinements, such as data balancing techniques and feature engineering, could enhance classification robustness and generalization to real-world conditions.

These findings highlight the potential of machine learning in traffic safety analysis and provide a valuable framework for integrating real-time risk assessment into urban traffic management systems. ONN is the best among the four to classify *green* and *red*, GNB is fair to classify *yellow* according to validation and test, but ONN is the best according to global accuracy.

Acknowledgment

This work has been funded by the Natural Sciences and Engineering Research Council of Canada, Canada Foundation for Innovation, Ville de Trois-Rivières, and the Research Chair in Signals and Intelligence of High-Performance Systems.

REFERENCES

[1] M. Eskandari Torbaghan, M. Sasidharan, L. Reardon, and L. C. Muchanga-Hvelplund, "Understanding the potential of emerging digital technologies for improving road safety," *Accident Analysis Prevention*, vol. 166, p. 106543, 2022.

[2] A. Boyle and C. O'Flaherty, *Highways, Fourth Edition*. Taylor & Francis, 2002.

[3] P. Songchitruksa and A. P. Tarko, "The extreme value theory approach to safety estimation," *Accident Analysis Prevention*, vol. 38, no. 4, pp. 811–822, 2006.

[4] Y. Hu, Y. Li, C. Yuan, and H. Huang, "Modeling conflict risk with real-time traffic data for road safety assessment: a copula-based joint approach," *Transportation Safety and Environment*, vol. 4, no. 3, p. tdac017, 08 2022. [Online]. Available: https://doi.org/10.1093/tse/tdac017

[5] A. Arun, M. M. Haque, S. Washington, T. Sayed, and F. Mannering, "A systematic review of traffic conflict-based safety measures with a focus on application context," *Analytic Methods in Accident Research*, vol. 32, p. 100185, 2021.

[6] S. Zhang and M. Abdel-Aty, "Real-time pedestrian conflict prediction model at the signal cycle level using machine learning models," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 3, pp. 176–186, 2022.

[7] R. Ezzati Amini, K. Yang, and C. Antoniou, "Development of a conflict risk evaluation model to assess pedestrian safety in interaction with vehicles," *Accident Analysis Prevention*, vol. 175, p. 106773, 2022.

[8] W. D. Glauz and D. J. Migletz, "Application of traffic conflict analysis at intersections," *NCHRP Report*, 1980.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.

[10] S. Coles, "An introduction to stat. modeling of extreme values," *Journal of the American Statistical Association*, vol. 97, 2001.

[11] T. M. Inc., "Design time series narx feedback neural networks," Natick, Massachusetts, United States, 2024.