2023 TRANSPORTATION TECHNOLOGY TOURNAMENT

Crash Data Reporting

Using A Machine Learning (ML) Model to Automatically Assign Proper Collision Types based on Qualitative Crash Data

2023 Transportation Technology Tournament

GMU & UVA

In Collaboration with DDOT

1. Report Number	TTT-0000	
2. Designated Agency	Washington District of Columbia Department of Transportation (DDOT)	
3. Sponsoring Agency	Transportation Technology Tournament, sponsored by the National Operations Center of Excellence (NOCoE) and the U.S. DOT ITS JPO PCB program	
4. Type of Report	Final – Concept of Operations	
5. Project Title	Crash Data Reporting: Using A Machine Learning (ML) Model to Automatically Assign Proper Collisions Type based on Qualitative Crash Data.	
6. Performing Organization Name and Address	· George Mason University, 4400 University Dr, Fairfax, VA 22030 · University of Virginia, 351 McCormick Rd, Charlottesville, VA 22904	
7. Student's Contact:	Eun Young Kim, ekim46@gmu.edu	PhD student, George Mason University
	Yan Ai, yai@gmu.edu	PhD student, George Mason University
	David Lin, dl9bbk@virginia.edu	Undergraduate student, University of Virginia
8. Faculty Advisor's Contact:	Duminda Wijesekera, dwijesek@gmu.edu	Professor, George Mason University
	Cing-Dao Kan, cdkan@gmu.edu	Professor, George Mason University
	Chung-Kyu Park, ckpark@gmu.edu	Research Professor, George Mason University
9. Industry Advisor's Contact:	Derek Voight, derek.voight@dc.gov	District Department of Transportation
10. Submission Date	May 26, 2023	
11. Abstract	This project intends to recommend solutions with a high-level concept-of-operations for an automatic process of crash data reporting. New proper collision types considering Vulnerable Road Users (VRUs) were introduced and used for validating the proposed Machine Learning (ML) model with qualitative crash data provided by the District Department of Transportation (DDOT). This project is expected to improve issues of the current data processing of crash data reporting with an automatic process focused on data quality and completeness and collision types considering crashes involving in VRUs.	
12. Original "Problem" Statement from DDOT	The project team would provide values by automating the process of assigning a proper collision type based on the narrative descriptions using artificial intelligence approaches.	
13. Number of Pages	Concept-of-Operation: 12 Pages | Additional Resource: 3 Pages	

Contents

List of Figures

List of Tables

List of Acronyms

DDOT	District Department of Transportation
MPD	Metropolitan Police Department
VRU	Vulnerable Road User
NHTSA	National Highway Traffic Safety Administration
FARS	Fatality Analysis Reporting System
TARAS	Traffic Accident Analysis and Research System
GIS	Geographic Information System
OCTO	Office of the Chief Technology Officer
ML	Machine Learning (ML)
NCAP	New Car Assessment Program
SVM	Support Vector Machine
CNN	Convolutional Neural Network
ETL	Extract, Transform, and Load

1. Overview

This report demonstrates a high-level concept-of-operations describing an automatic process of the crash data reporting for proposed collision types based on qualitative data by using an artificial intelligence technology. In this report, the crash data processing through the proposed Machine Learning (ML) model proved for automatically assigning the proposed collision types considering crashes involving Vulnerable Road Users (VRUs). The District Department of Transportation (DDOT) provided samples of crash datasets and clearly identified their expectations to improve the current crash data reporting system with a consideration of vulnerable road users and an artificial intelligence technology. Two main proposed solutions are presented here: 1) to introduce proper collision types considering crashes involving VRUs, and 2) to use a Machine Learning (ML) model to automatically assign the collision type based on unstructured and qualitative data.

1.1. Motivation

Crash data have been collected and stored in various ways for consideration of road traffic safety. Using artificial intelligence technologies for crash data has been popular to improve existing safety management systems or utilize more information from existing safety databases. An automatic process of crash data reporting would be efficient to manage crash data and to provide crash information for road user safety.

A consideration of VRUs has been getting more attention because VRUs accounted for a growing share of all United States roadway fatalities in recent years. According to the National Highway Traffic Safety Administration's (NHTSA) Fatality Analysis Reporting System (FARS), VRU fatalities have decreased overall from 14,308 in 1980 to 12,352 in 2019. However, the proportion of traffic fatalities for VRUs was 28 percent in 1980 and increased to 34 percent by 2019. In a 2020 Pedestrians Traffic Safety Fact Sheet, the total number of pedestrian fatalities in the District of Columbia is listed as very low but the percentage of pedestrian fatalities is the second highest with the highest being in New Jersey.

1.2. Operational Objectives

DDOT has provided a public service named "Crashes in DC" on an application "OpenData DC" in order to deliver crash information on the Washington D.C. map. The purpose of Crashes in DC is to provide crash location data (within 24 hours) to users interested in exploring the crash data in the Washington D.C. However, DDOT has been aware that the crash data contain significant gaps and quality issues. Therefore, DDOT is interested in some novel approaches for crass=h data reporting based on unstructured and qualitative data such as narratives and descriptions. DDOT identified two objectives for improving the current data processing for the crash data reporting: 1) to define proper collision types considering crashes involving VRUs based on qualitative crash data, and 2) to Introduce an artificial intelligence technology for automatically assigning the proper collision types.

1.3. Approach

Two main approaches were defined to fulfill the two aforementioned objectives:

(1) Proper Collision Types: A need for new collision categorizations was significant to consider crashes involving VRUs because a current typology of collision types does not reflect the consideration of VRUs. Thus, the new collision types needed to be proposed for an automatic process of the crash data reporting. An assumption was presented that the proposed collision types are identified based on unstructured and qualitative data such as narratives or descriptions. The proposed collision types are anticipated to be validated by a Machine Learning (ML) model.

(2) A Machine Learning (ML) Model: A Machine Learning (ML) model is used to automate the process of assigning collision types. The ML model can output the collision types automatically, which can greatly enhance the efficiency of crash data management.

An analytics process overview including the two approaches was delineated in Figure 1: the below flow chart describes details of each step of the analytics process.

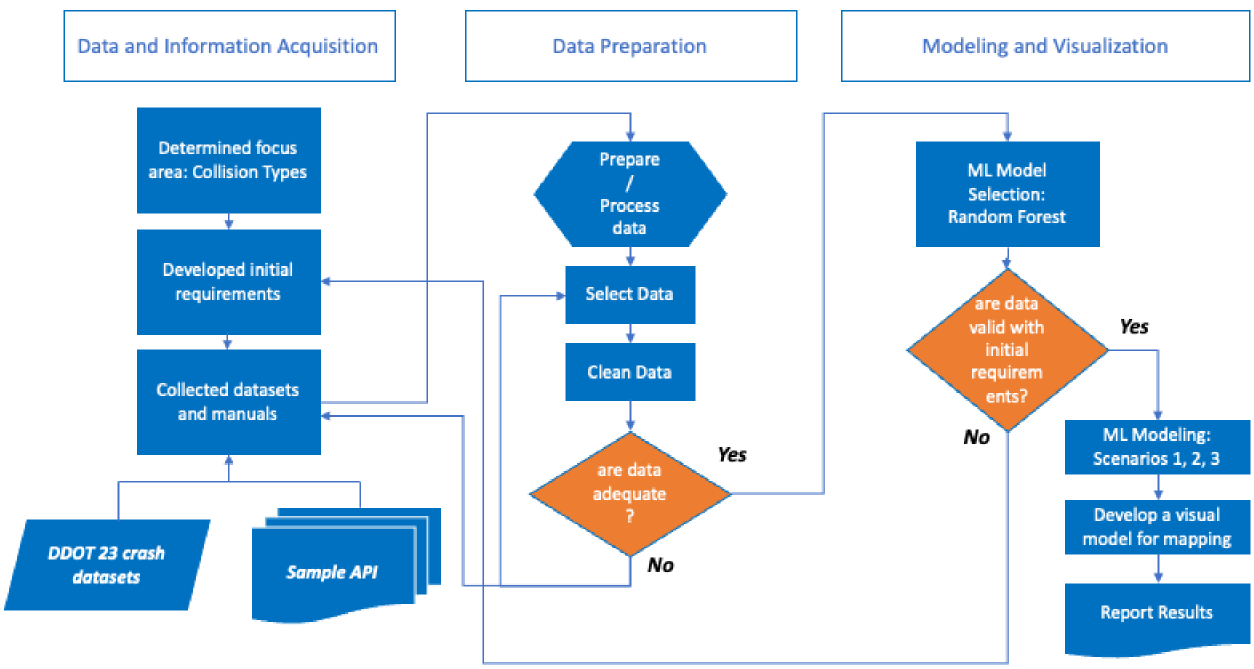

Figure 1 An analytics process overview

1.4. Stakeholders

An understanding of stakeholders for the crash data reporting was fundamental to ensure that all needs of the stakeholders can be mirrored with the proposed solutions. In addition, interactions of among stakeholders were delineated to learn and clarify the data processing for the crash data reporting. The stakeholders were grouped into three domains described in Figure 2: the below diagram portrays interactions of the stakeholders.

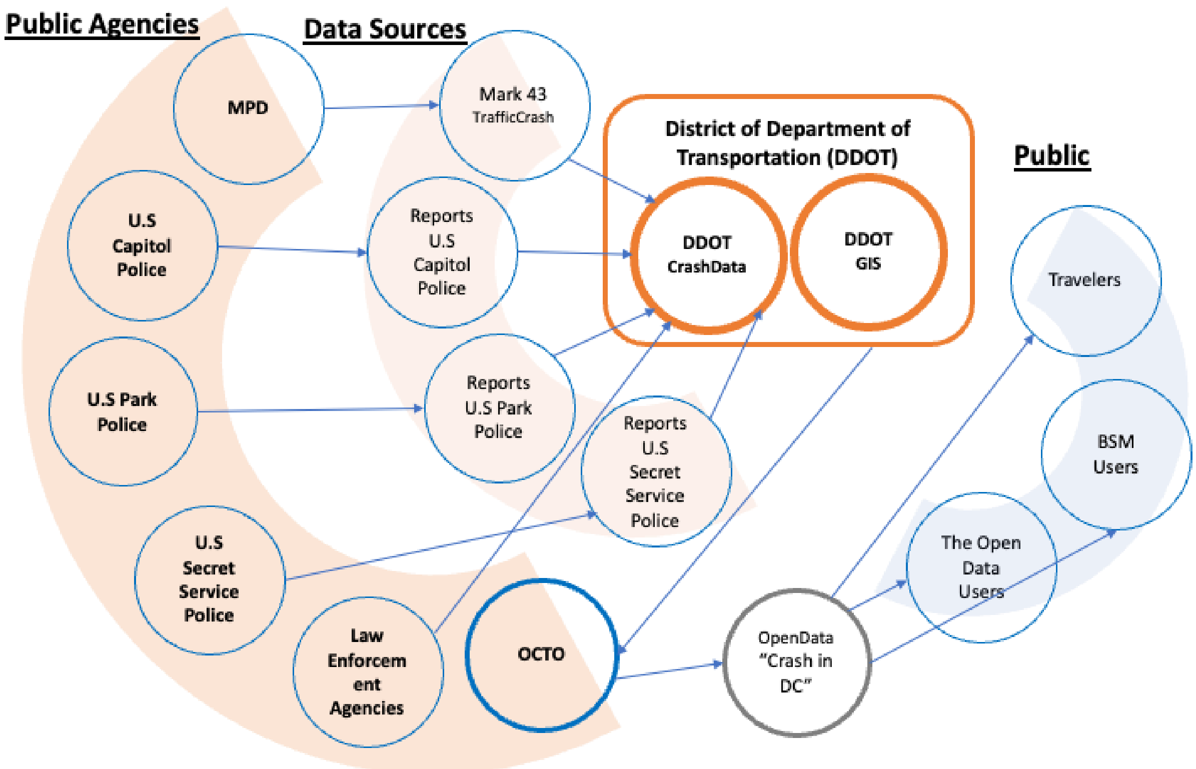

Figure 2 Block diagrams of interacting stakeholders

2. Definition

2.1. Problem Definition and Scope

DDOT has been aware of issues within the current crash data reporting on account of data quality and completeness. DDOT addressed two main expectations for improving the current crash data reporting: 1) to consider crashes involving VRUs to define proper collision types, and 2) to improve the current data processing for the crash data reporting with unstructured and qualitative data to automatically assign the collision types by using an artificial intelligence technology.

In order to satisfy these expectations, defining existing problems was carefully conducted focusing on the current data processing for the crash data reporting and collision types considering VRUs based on unstructured and qualitative data. The first problem was an issue of not matching crash information with the MPD's crash reports for the crash data reporting. There are several variations caused the issue by changing post-crash information such as injury levels, delayed reporting at the crash data processing, data entry error, and different data sources between DDOT and MPD. Another problem was that data attributes for identifying collision types do not consider VRU's crashes to classify collision types.

The scope of the problems was outlined with three main tasks:

1. Ensure that defining proper collision types considers crashes involving VRUs.
2. Establish an approach on how to build an automatic process of the crash data reporting to assign the proper collision types based on unstructured and qualitative data.

2.2. User Needs

User needs were specified with considering the aforementioned stakeholders:

UN 1.1: An automatic process of assigning proper collision types

DDOT needs a novel automatic process of the crash data reporting based on narrative descriptions to assign proper collision types.

UN 1.2: Proper collision types including crashes involving VRUs

DDOT needs to consider crashes involving VRUs for defining collision types reflected by both a traditional collision typology and additional collision types with VRUs.

UN 1.3: Estimate cost and time to evaluate benefits and limitations

DDOT needs to understand values for implementing the automatic process of the crash data reporting.

3. Proposed Solutions

Two main approaches were proposed to find out suitable solutions for the defined problems. The following two sections: 3.1 and 3.2 describes the details of the proposed solutions:

3.1. New Collision Types

3.1.1. Data Source and Attributes

DDOT provided 23 revised crash datasets during the calendar years of 2001 to 2023. All attributes were reviewed to select data attributes having qualitative data values such as Narrative or Description as well as a data attribute (Manner of Crash) directly describing collision types. A strategy was established to decide on data attributes for the next step; building a Machine Learning (ML) (ML) model by defining common data attributes and comparative data attributes. The different comparative data attributes were introduced to develop analysis scenarios for the ML model. Details of datasets and data attributes provided by DDOT are in Appendix A.

3.1.2. Existing Collision Types

A categorical data attribute “Manner of crash value” directly refers to collision types with categories such as Front-to-Rear, Front-to-Front, Angle, Sideswipe, Rear-to-Side, and Rear-to-Rear. It is regarded as a traditional collision typology for vehicle crashes. Within each typology, crashes involving VRUs were not considered any categories and left blanks within the data values.

The NHTSA FARS has large VRU collision types defined by VRUs pre-events. Whereas, the consumer vehicle safety rating programs well-known as the New Car Assessment Program

(NCAP) for USA and Euro developed the VRU collision types based on simple prior actions and locations of VRUs such as crossing to a vehicle, longitudinal to a vehicle, and turn across path to a vehicle. The comparative data attributes were selected following the NCAP VRU collision types considering VRUs prior actions and location for the abovementioned analysis scenarios.

3.1.3. Proposed Collision Types

A proposed solution for defining proper collision types was drawn with the simple concept of adding two categories: VRU-pedestrians and VRU-bicyclists at the traditional collision typology. Based on analyzing data attributes associated with VRUs, VRU crashes could roughly be distinguished by two crash types: pedestrian crashes and bicyclist crashes. The classification of the two crash types was validated by the ML model with qualitative data. Additionally, the comparative data attributes for VRUs prior actions and location were used to attest to the proposed collision types by using the ML model. In summary, the proposed collision types can be listed: Front-to-Rear, Front-to-Front, Angle, Sideswipe, Rear-to-Side, Rear-to-Rear, VRU-pedestrians, and VRU-bicyclists.

More niche types for VRU were tested with a classification model (random forest) such as the VRU-pedestrians-InRoadway (VRU-[user type]-[action prior to the accident]). Due to the small amount of niche collision types, other ML models might not fit well into these underrepresented classes. In the dataset acquired, out of 35,157 cases, only 43 cases are labeled as "VRU-bicyclists-AlongRoadway", while 7,959 cases belong to Front-to-Rear type. Within the training dataset, the number is even smaller. Thus, for this dataset, it is logical to classify using more general VRU types. As shown in Figure 3, when adopting the niche VRU collision types, the numbers for each VRU type range from 0.16% to 2.42%. Meanwhile, other non-VRU types can be as high as 30.47% of the entire dataset. In the training subset, the number of each VRU type is even smaller. Thus, for this dataset, it is better to classify using more general VRU types.

Out of 26,122 Crash Cases

Collision Type	Number of Occurrence
VRU-Pedestrian	1478
VRU-Bicyclist	660

Collision Type	Number of Occurrence
VRU-Pedestrians-InRoadway	203
VRU-Pedestrians-AlongRoadway	117
VRU-Pedestrians-CrossingRoadway	632 --> 2.42%
VRU-Pedestrians-Other	526
VRU-Bicyclist-InRoadway	85
VRU-Bicyclist-AlongRoadway	43 --> 0.16%
VRU-Bicyclist-CrossingRoadway	296
VRU-Bicyclist-Other	236

Figure 3 More General VRU Types (Left) & Niche VRU Types (Right)

3.2. Machine Learning (ML) Model

3.2.1. Model Selection

There are many ML algorithms that can perform text classification tasks, such as Support Vector Machine (SVM), Convolutional Neural Network (CNN), random forest, K Nearest Neighbors (KNN), etc. According to Shah et al. (2020), random forest is not always the best performing model, but it has the most balanced performance across different classes. A comparative analysis was performed with candidate ML models and random forest. Random forest was selected for collision type classifications because of its better accuracy score.

Random forest is an ensemble model and it is comprised of multiple decision trees. A simple example of the decision tree is shown in Figure 4. After preprocessing the data, such as vectorization, the input text is fed into the model. At the first split node, the model checks if the word “pedestrian” is included in the text, if yes, the model moves to the left node and checks if the word “crossing” is included. If both words are found, then the model predicts the collision types associated with this textual narrative would be VRU-Pedestrian. This is a simplified process of how a decision tree works. The real decision trees in random forest model can be much deeper and complex than this. The random forest model contains more than one decision trees. Each decision tree in the random forest model below has its own parameters and possibly a different structure. The text is fed into each decision tree in the same model and the final prediction, or classification result, is the mean value of all trees included in the random forest model. Figure 5 shows the process of random forest algorithm.

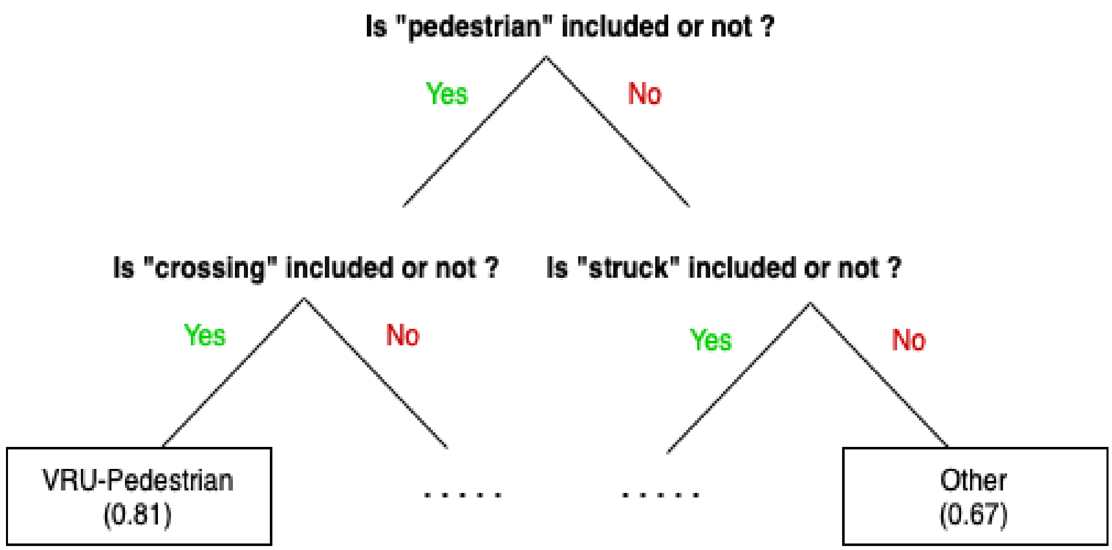

Figure 4 Decision Tree Example

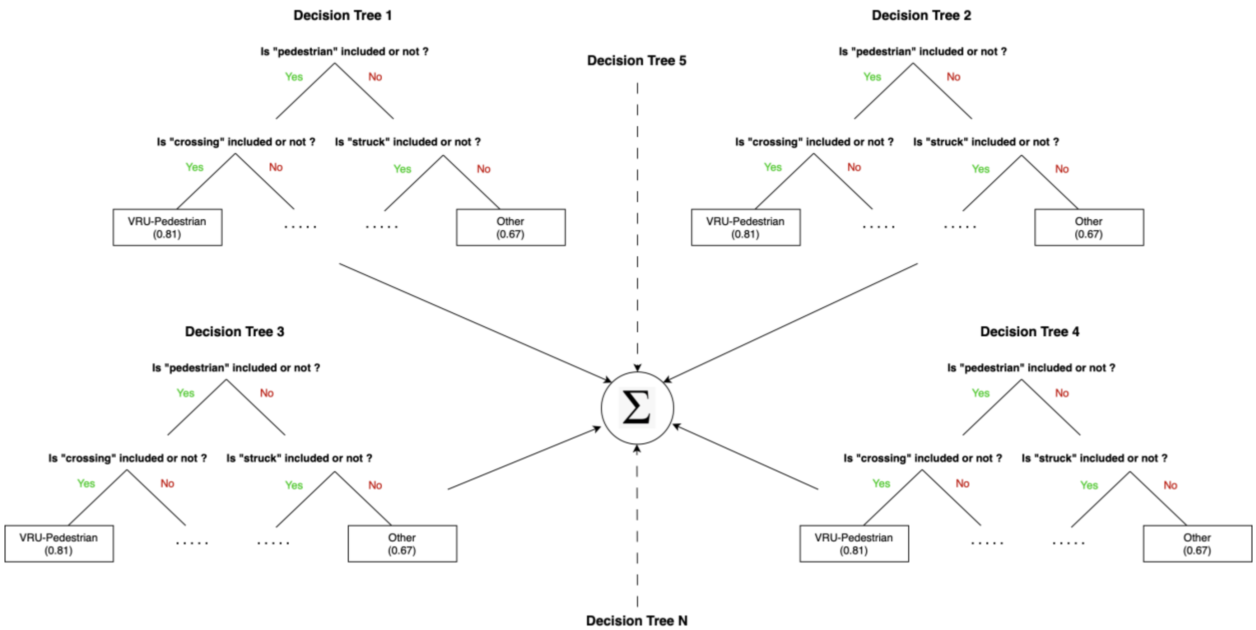

Figure 5 Random Forest Algorithm

3.2.2. Model Validation

A labeling process was required to prepare for the ML modeling. The labeling process was automated using variables such as "person involved" and "first harmful event". If there was pedestrian or bicyclist involved in the crash case and/or the first harmful event happened to a pedestrian/bicyclist, this case can be labeled as VRU types. The collision types are the dependent variable. There is only one independent variable which is the textual summary narrative.

Considering an initial point of contact is essential to clarify collision types. However, it is not easy to conduct with narrative data because the narrative data might not have technical information. For example, several considerations are required to define collision types; sideswipe and angle. Thus, a label "Sideswipe+Angle" was created by combining sideswipe and angle. This preprocessing was expected to eliminate some impact of missing information for enhancing the accuracy of the model.

Based on 26,122 crash cases that were obtained from the data pre-processing, several datasets were created for the ML model. A training dataset prepares the model for our use case; a validation dataset is for fine-tunes the model; and a testing dataset reviews for model accuracy. The split ratio is 2 (training dataset): 1 (validation dataset): 7 (testing dataset).

The accuracy of the collision type classifications has reached 83.75% in the last run. Two additional tests are performed: (1) the classification of "first harmful event", and (2) the classification of "person type". The accuracy for "first harmful event" classification is lower than that of collision type classification, 70.64%. And the accuracy for "person type" is as high as 91.00%. Figure 6 indicates a summary of the accuracy of the collision type classifications.

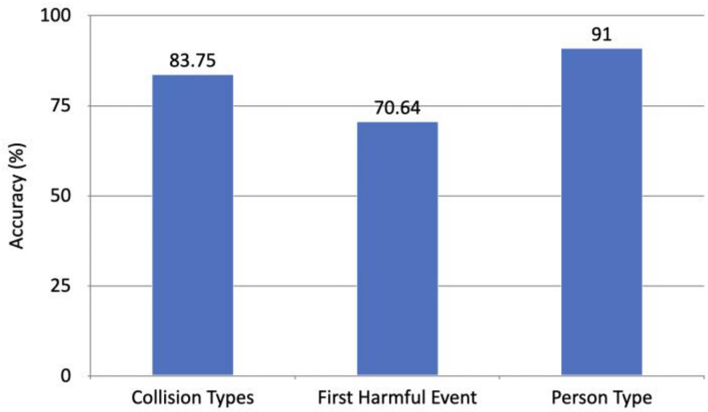

Figure 6 Accuracy of Classifications

4. Concept of Operations

An automatic process proposed to implement a new data processing of the crash data reporting focused on assigning the proposed collision types without a historical data processing. An expectation for the automatic process was reflected to be faster even without some complicated data pre-processing and to be accurate with narrative descriptions to identify collision types. Thus, it requires the use of a new dataset having only unstructured and qualitative data from other data sources as well as the MPD database. After the ML process is applied, identified collision types are sent to DDOT GIS for visualizing the collision types at crash locations on the D.C. map. A separated section for information on the collision types "Crash Types" is created with other information such as crash details, crash injuries, and general crashes. The new section "Crash Types" provides to the Open Data for user's information. This is a concept of operation for the automatic process of the crash data reporting described in Figures 7 and 8 demonstrating high-level architectures.

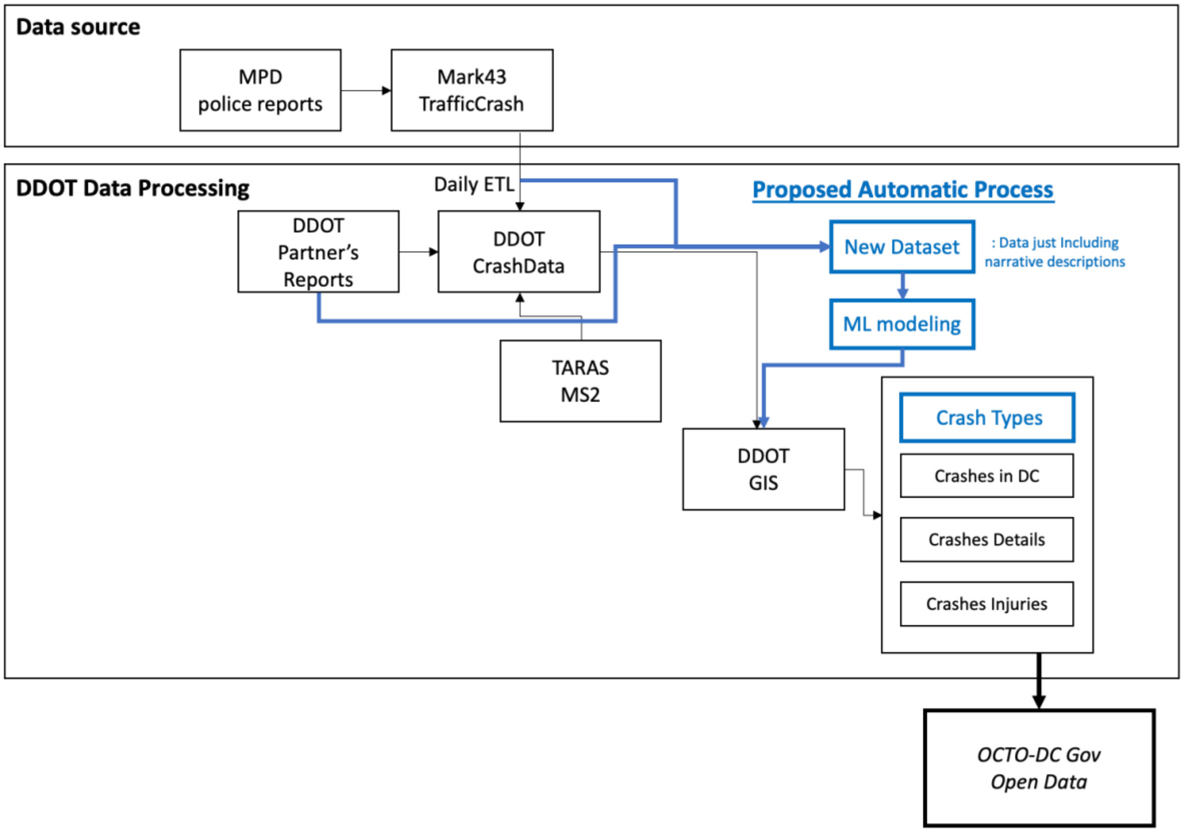

Figure 7 High-level Functional Architecture: Proposed Automatic Process

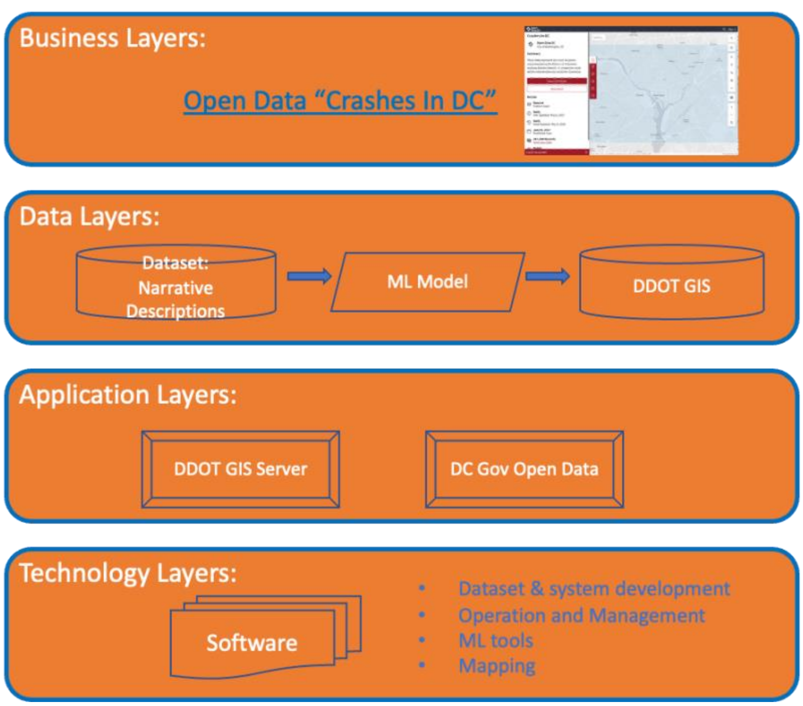

Figure 8 High-level Enterprise Architecture

5. System Construction

Preliminary cost and time estimates were required to specify with tasks for evaluating potential risk for the proposed system implementation and operation. Table 1 shows the details:

Table 1 Anticipated cost and time estimate

Tasks	Cost Estimate	Time Estimate
Task 1. Generating a new dataset	• Personnel cost; responsibility for developing a new dataset by integrating data sources. • ETL cost; an additional cost for the ETL process with other data sources. • Data Storage cost; a cost for adding a new dataset.	10-20 days
Task 2. Developing a ML process system	• Personnel cost; responsibility for developing a ML process system and validating the system by integrating with the current systems. • System modification and integration cost; a cost for modifying the current system to integrate with the new ML process system. • Data server upgrade cost; a cost for upgrading the data server adding other data sources and a new data processing with the ML model. • ML Software cost; a cost for the ML modeling.	20-30 days
Task 3. Generating a new section; Crash Types on the DC map	• Personnel cost; responsibility for generating an additional section "Crash Types" at DDOT GIS server and integrating the ML process system. • System modification cost; a cost for modifying the current GIS system to add the Crash Types section. • Data server upgrade cost; a cost for upgrading the data server by adding a data processing for Crash Types. • GIS server upgrade cost; a cost for upgrading the GIS server to add location information for Crash Types.	5-10 days
Task 4. Updating a function for Crash Types at the Open Data	• Personnel cost; responsibility for updating a function for Crash Types at the Open Data website. • Webserver upgrade cost; a cost for upgrading the webserver size with an additional processing.	5-10 days
Task 5. Evaluating system implementation/operation	• Personnel cost; responsibility for continuously evaluating each step of system implementation and validating system outputs by a quality standard. • System implementation and operation cost; a cost for reviewing several phases: planning, analysis, design, development, testing, implementation, and maintenance.	30-60 days

6. System Maintenance

A protocol for monitoring the proposed data processing system is highly recommended to prevent potential issues such as poor data quality, the ML process system failure, or system integration failure. The new dataset having unstructured and qualitative data and using integrated data sources should be supervised to guarantee the accuracy of the ML process. The

Open Data application should be periodically checked to verify information on crash types and locations.

7. Anticipated Benefits and Limitations

The proposed solutions were expected not only to satisfy the user's needs in terms of the crash data reporting process focusing on information of collision types, but also to benefit the stakeholders. The expected benefits of the proposed solutions are summarized in Table 2:

Table 2 Anticipated benefits

Benefits	**Highlights**
Operational benefits	• Reduce a time of the data processing for crash data reporting. • Simplify complicated steps for data pre-processing. • Improve issues of data quality and completeness by increasing data reliability.
Safety benefits	• Reduce crashes involving VRUs with safety awareness. • Reduce injury levels with safety awareness. • Improve overall road traffic safety.
Mobility benefits	• Reduce conflicts between vehicles and VRUs. • Improve and increase VRU mobility.
Environmental benefits	• Reduce emissions with increasing VRUs.

Some limitations remain on reliability of additional data sources and validation of the ML model to provide fast, accurate information of collision types for users.

8. Conclusion

This report described the proposed solutions for a high-level concept-of-operations: the automatic process of crash data reporting to automatically assign proper collision types based on unstructured and qualitative data. The proposed solutions can be summarized with two key parts: 1) to propose new collision types considering crashes involving VRUs, and 2) to develop a ML model for an automatic process of the crash data reporting to assign the collision types. The proposed collision types were present adding two new categories (VRU-pedestrians and VRU-bicyclists) with a consideration of prior actions of vulnerable road users and their location. The types were validated by the proposed ML model using crash data provided by DDOT. Consequently, the concept of operations: the automatic process was established by implementing the ML model process to assign the proposed collision types for user's information. The proposed solutions are expected to improve the existing issues of the crash data reporting system and to draw numerous benefits for road traffic safety in the Washington D.C area.

Reference

1. 2020 BICYCLISTS AND OTHER CYCLISTS Traffic Safety Fact Sheet. (2022). Dot.gov. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813322
2. 2020 FARS/CRSS Pedestrian Bicyclist Crash Typing Manual: A Guide for Coders Using the FARS/CRSS Ped/Bike Typing Tool. (2022). Dot.gov. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813250
3. 2020 PEDESTRIANS Traffic Safety Fact Sheet. (2022). Dot.gov. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813310
4. 2021 FARS/CRSS Coding and Validation Manual. (2022). Dot.gov. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813426
5. Comparing Demographic Trends in Vulnerable Road User Fatalities and the U.S. Population, 1980–2019. (2021). Dot.gov. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813178
6. EUROPEAN NEW CAR ASSESSMENT PROGRAMME (Euro NCAP) TEST PROTOCOL -AEB/LSS VRU systems Implementation 2023. (2022). Euroncap.com. https://cdn.euroncap.com/media/75436/euro-ncap-aeb-lss-vru-test-protocol-v43.pdf
7. GIS-Crash Data. (2023). District Department of Transportation.
8. James, R., Hammit, B., Pandey, V., Vincent, J., & Yahia, C. (2018). Non-Recurring Congestion Mitigation Solutions for the Washington District Department of Transportation. Transportationops.org. https://transportationops.org/sites/transops/files/Texas%20TTT%20Submission.pdf
9. PBCAT pedestrian and bicycle crash analysis tool version 3.0 user guide. (2022). U.S. DOT Federal Highway Administration. https://doi.org/10.21949/1521874
10. Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research, 5(1). https://doi.org/10.1007/s41133-020-00032-0

Appendix A: Datasets and Data attributes

DDOT provided 23 revised crash datasets listed in Table 3 and data attributes were selected for the proposed solutions.

Table 3 Used datasets and data attributes

Total Datasets (23)	Used Datasets (4)	Common Data Attributes	Comparative Data Attributes
Data	**Yes**	-crashed	-narrative -summaryNarrative
externalTrafficCrash	**Yes**	-crashed -locationId	-firstHarmfulEventValue -mannerOfCrashValue -initialPointOfContactAttrDisplayValue
Trafficcrashlocation	**Yes**	-locationId	
TrafficCrashPeople	**Yes**	-crashed -personId	-personTypeAttrDisplayValue -nonMotoristActionPriorToCrashAttrDisplayValue -nonMotoristLocationAtTimeOfCrashAttrDisplayValue
Entityorderedattribute	No	N/A	N/A
Entityorderedattribute	No	N/A	N/A
HomeAddress	No	N/A	N/A
InvolvedPerson	No	N/A	N/A
InvolvedVehicles	No	N/A	N/A
LinkedPerson	No	N/A	N/A
Person	No	N/A	N/A
Personattribute	No	N/A	N/A
Personinjury	No	N/A	N/A
Reportingperson	No	N/A	N/A
respondingOfficer	No	N/A	N/A
Subjectsintrafficcrash	No	N/A	N/A
submittedBy	No	N/A	N/A
Trafficcrashattribute	No	N/A	N/A
Trafficcrashentitydetail	No	N/A	N/A
Trafficcrashroadway	No	N/A	N/A
TrafficCrashVehicle	No	N/A	N/A
TriggerUpdate	No	N/A	N/A
VehicleItemAttributes	No	N/A	N/A
Vehicles	No	N/A	N/A

Appendix B: Demonstration on A Map

A visual model for mapping was developed with classifications of the collision types shown in Figure 9. The classifications were visualized on the DC map with diverse colors described in Table 4.

Table 4 Map Color Labels

Collision Types	Colors
Front to Rear	*Purple*
Front to Front	*Orange*
Sideswipe, Same Direction	*Green*
Sideswipe, Opposite Direction	*Dark Blue*
Angle	*Blue*
Rear to Rear	*Dark Purple*
VRU-Pedestrians	*Light Red*
VRU-Cyclists	*Cadet Blue*
VRU- Collision with Parked Vehicle	*Dark Green*
Others	*Red*
Unknown	*Dark Red*

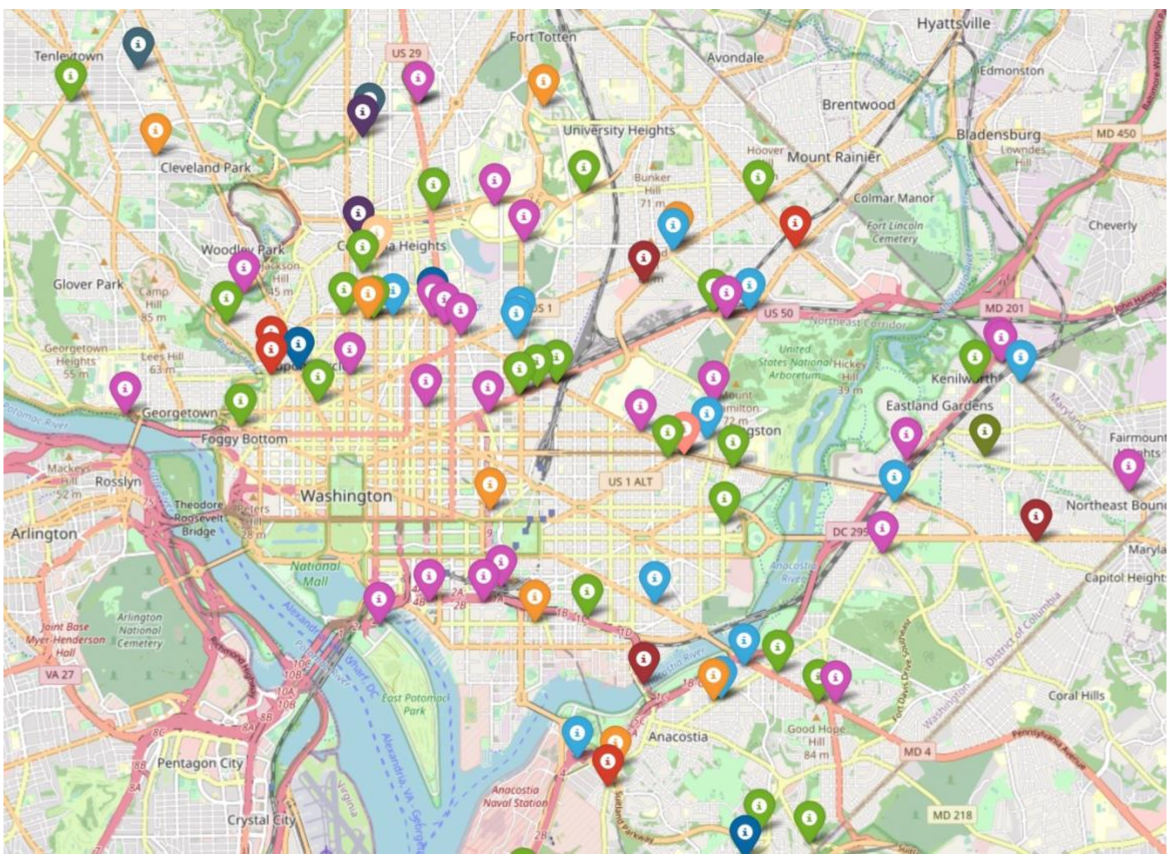

Figure 9 Map Visualization by collision types

The map visualization can provide a comprehensive understanding of road safety issues. The high-risk areas for non-VRUs and VRUs can be easily identified on the map. Such information can help the urban planners to improve the safety measures. Also, the distribution of specific collision types can provide valuable insights into the underlying causes for each type of collisions.

The future plan is to enable further statistical analysis using the timestamp information too. Traffic accidents have different patterns at different time ranges. It would be helpful if the map visualization can be done on a one-month time span, or select the accidents happened on a specific time range, every day, throughout the year. This can further enhance the effectiveness and efficiency of safety interventions.