# Final Report

**STIC Incentive Project Name: Use of AI to Improve Vulnerable Road User Safety**

**Submitted by:**

Daniel Carter
NCDOT Traffic Safety Unit
North Carolina Department of Transportation
750 N. Greenfield Parkway
Garner, NC 27529

**Date Submitted: November 6, 2025**

**Description of Project:**

This project was focused on data extraction using artificial intelligence. The origin of this project idea was based in the fact that NCDOT has extensive amounts of data that can be used to improve pedestrian safety. The department collects turning movement counts at over a thousand locations per year around the state. Many of these counts contain the count of pedestrians and bicyclists crossing the road. However, this data is contained in static PDF documents meant only for visual reading by NCDOT staff. As such, many years' worth of data on non-motorist travel exposure are not accessible for analysis on a statewide basis. NCDOT also has crash report data for each non-motorist crash, but the coded values on the crash report do not capture sufficient detail to provide insights into the factors and causes behind the crashes. The reports do contain written text narratives that capture additional data that could be used for further insights. The purpose of this project was to develop a methodology to use artificial intelligence (AI) to assemble data for vulnerable road user safety from these existing DOT data sources. The development of the AI methodology was intended to be a proof-of-concept. Further continuation of a successful proof-of-concept in a future project could implement the methodology within NCDOT to process and extract the data statewide.

**Overall Budget**:

The overall budget for this project was $145,000 and the project stayed within budget. The Project received authorization for $125,000.00 in National STIC Incentives Program funding, with the remainder funded by NCDOT. The budget covered the work of the consultants (Ernst and Young, LLP) working within the proposed timeframe.

**How the Work Specifically Meets the Program Criteria:**

The use of advanced machine learning and artificial intelligence techniques in this project fits very well with the goal of innovation that is the core of the STIC program and other FHWA initiatives, such as Every Day Counts (EDC). It also integrates with one of the goals of NCDOT's traffic safety programs to address pedestrian safety by improving the data underpinning the program and the subsequent project selection.

**Results of the Project:**

The project produced a successful proof-of-concept for the use of AI in data extraction for pedestrian safety, accompanied by usable datasets developed during the course of the project. The project addressed two areas of pedestrian safety related data – pedestrian counts extracted from turning movement count reports and pedestrian crash attributes extracted from crash report narratives.

<u>Pedestrian Counts</u>

NCDOT provided approximately 2,000 turning movement count reports (PDF format) to the project team. These were reports produced by NCDOT vendors based on typically 13-hour counts of vehicles, pedestrians, and bicyclists at intersection locations. The project team developed an AI model that used optical character recognition to extract pedestrian counts and accompanying details from nine different report formats. The project team worked with NCDOT to identify the priority data elements to extract from the reports. The model successfully extracted 18 elements from the turning movement count reports, such as location description, weather condition, and duration of count, in addition to the necessary data regarding the number of pedestrians observed in the count.

Using the developed model, the project team created a summary output dataset that consolidates the data from over 1,400 turn count reports, encompassing data from multiple count data vendors over the past five years. The project team also developed a

web-based tool to allow NCDOT to upload turn count reports and have the data extracted, thereby equipping NCDOT with the ability to run this tool in the future for extraction of pedestrian count data from additional reports.

Pedestrian Crash Attributes

NCDOT provided the project team with data on approximately 28,000 crashes that involved pedestrians. The dataset included crash narrative (the officer's written description of the crash scene and sequence of events) as well as many other relevant data fields at the crash level as well as unit and person level. In order to teach the model how to extract crash elements, NCDOT also provided a dataset of pedestrian crash specific coding that had been conducted for those crashes through a separate prior effort. The contractor worked with NCDOT to refine the data to be extracted, such as the coding of individual data elements to be consistent with current NCDOT data.

The project team developed an AI-driven model to extract four data elements from crash narratives:

- Crash type – this field indicated the specific type of crash, such as dart-out, vehicle failed to yield, and walking along roadway. The coding used in this field was configured to match the coding used in NCDOT's current non-motorist crash data, which uses crash types as laid out in the Pedestrian and Bicycle Crash Analysis Tool (PBCAT).
- Non motorist type – this field contained the type of non-motorist involved in the crash, either pedestrian, scooter, or skateboard user.
- Pedestrian activity – this field was coded in an open ended manner and served to extract a short phrase to describe the pedestrian's activity prior to the crash, such as "ran out into the street", "pushing vehicle", or "standing in the middle of the road".
- Crosswalk presence – this field indicated the presence of a crosswalk at the crash location, if such information could be derived from the narrative.

The contractor developed a web interface to allow a user to submit either individual narratives or multiple narratives. If submitted multiple narratives, the interface features bulk extraction of Crash Narratives, processing an input spreadsheet from SharePoint and generating an output file containing the four coded data fields for each crash narrative. This tool allows NCDOT to extract the crash data elements for future crashes.

**Challenges:**

One of the major challenges was the variety of ways in which the input data could be provided. For the turning movement count data, there were nine different vendors represented, all of whom produced reports in different formats and with differing levels of detail. Significant work went into training the model to consistently read the required fields from this variety of formats. It also represents a limitation of the model, since it can only extract pedestrian count data from the nine formats that were used in the training set. For the crash data extraction, the crash report narrative field is an open text descriptive field. This means that the amount of detail about the crash scene and sequence of events varies quite a bit since it depends on how the reporting officer wrote the report. The result of this was that some data elements could not be extracted for certain crashes, since the required indications in the narrative were either not present or unable to be understood in context.

**Lessons Learned:**

The process of developing an AI model is relatively new to NCDOT, and one of the lessons learned was the amount of time and interaction that was necessary to assist the contracted project team. The NCDOT staff spent significant time with the project team to clearly lay out the data needs, review the reports, and help the project team interpret the data contained in them.

**Next Steps:**

NCDOT plans to continue periodic use of the tool provided for data extraction from turning movement counts. This will be useful to the pedestrian data program, since the current count program still largely receives data via static PDF reports. In the area of crash narrative extraction, the findings will guide the development of the subsequent research on this topic.