
# AI and Other Decision Support Systems for Crash Preventability PAR Processing

## BACKGROUND

The Federal Motor Carrier Safety Administration (FMCSA) offers the Crash Preventability Determination Program (CPDP) to provide registered motor carriers an opportunity to have their Crash Indictor Behavior Analysis and Safety Improvement Category (BASIC) percentile modified to exclude crashes with not preventable determinations from the prioritization algorithm if they can provide compelling evidence that a crash is eligible for the program and was not preventable.

The process begins when the motor carrier submits a Request for Data Review (RDR) and involves labor-intensive manual review by DataQs analysts. This report describes techniques to automate or partially automate the CPDP. The report covers the developmental requirements needed for automation and discusses underlying machine learning techniques, such as document analysis based on computer vision (CV), optical character recognition (OCR), and natural language processing (NLP).

## OBJECTIVES

The goal of this project was to evaluate the feasibility of applying artificial intelligence (AI) and other decision support systems (DSSs) to automate portions of the CPDP.

Specifically, AI/DSSs would be used to identify relevant crash parameters by comparing Police Accident Reports (PARs) to State violation codes and apply the extracted information to make eligibility and preventability recommendations to DataQs analysts. Although the AI/DSS model can make determinations based on the machine-read PARs, individual processors must be developed for each State PAR, as the format, content, and codes differ between them.

## METHODS

A demonstration component of the project involved executing an AI/DSS model on all Texas PARs, acquiring and compiling relevant data to create machine-readable documents to use in a customized processor to provide eligibility and preventability determinations.

The specific workflow of the AI/DSS model was established through a literature review and trade-off analyses. The steps to conduct the process were as follows:

1. Environment Setup
2. PAR Retrieval
3. Document Parsing Prerequisites
4. Document Parsing Process
5. Narrative Analysis
6. Crash Diagram Analysis
7. Eligibility and Preventability Analysis
8. Summary Report Generation

## FINDINGS

- The customized solution recommended from the demonstration is estimated to result in an approximately 50 percent reduction in total time spent by DataQs analysts in their determination efforts.
- Although a 50 percent reduction would significantly ease the burden on the enforcement team, further refinement of the AI/DSS model that improves consistency may increase the reduction in time spent to an estimated 65 to 70 percent.
- The estimated costs are included in Table 1.

U.S. Department of Transportation
Federal Motor Carrier Safety Administration

FMCSA-RRT-22-009b
December 2024

**Table 1. Development estimations by organization for top 15 State PAR versions to achieve 75% accuracy across determinations.**

| Entity | Interpreter Logic (hours) | Computer Vision on Extracted Data (hours) | Narrative Analysis (hours) | Quality Assurance (hours) | Total (hours) | Estimated Cost (fully encumbered) | Years to Positive Return (75% accuracy, fully automated) |
|---|---|---|---|---|---|---|---|
| VTTI | 900 | 840 | 440 | 120 | 2,300 | $230,000 (at $100/hr) | 2.25 |
| USDOT-IT | 960 | 840 | 440 | 120 | 2,360 | $188,800 (at $80/hr) | 1.85 |
| Private Sector | 2,200 | 1,600 | 1,080 | 240 | 5,140 | $819,200 (at $160/hr) | 8.03 |

## CONCLUSIONS

### *Demonstration Recommendations*

While a learning-based system to read and assess different versions of PARs would be both unwieldy and unreliable, a more nuanced system with individual parsers for each version of the PAR in circulation could be developed. With the development of the automation process for Texas, several additional options are available for future efforts, such as machine learning, narrative analysis, and CV techniques.

The research team recommends the following efforts for future iterations of the automation processing:

1. Expand the current automation system from Texas (approximately 11 percent of total RDRs) to the top 15 States (approximately 70 percent of total RDRs).
2. Expand the narrative using additional NLP techniques to utilize the unit assignment automation and crash type database, assist in determinations of eligibility and preventability, and highlight pertinent information for the DataQs analyst.
3. Build a repository of document excerpts that can be used for training purposes to identify key eligibility or preventability information.
4. Build a web application that can be utilized by DataQs analysts or other users without the need to interact with a terminal window or coding environment.

The proposed corrections to prevent data loss could be implemented across the system process or as denoted for each specific version. This includes:

- Increasing accuracy by performing iterative investigation of incorrect values.
- Correcting web scraping issues to accommodate multiple files.
- Performing fuzzy matching or logic to ensure correct information is recorded.
- Performing CV OCR on selected excerpts from individual PAR versions that are not able to be read from Google's Document AI OCR processor.

### *Cost-Benefit Analysis for Continuation of Work*

This project evaluated only Texas PAR data, the top producing state, to quantify how well the developed technology assisted PAR reviewers. This section outlines the costs of a full implementation of the automated PAR processing for the remaining top 15 states that produce 70 percent of the RDRs.

A major goal of the project was to perform a cost-benefit analysis of the implementation of the AI/DSS model for all or most RDRs. Following the above recommendations, in order to attain 75 percent accuracy, customized solutions would have to be created for the top 15 PAR versions by submission magnitude (i.e., 70 percent of RDRs are accounted for by TX, FL, CA, IN, PA, OH, GA, NC, IL, TN, AL, MI, MO, KY, and NY). If further developed, a production-level system is expected to result in an approximately 40 to 50 percent reduction in total time spent by DataQs analysts in their determination efforts.

To read the complete report, please visit: https://rosap.ntl.bts.gov/view/dot/78877