

EMS Data Analysis and Forecasting

For CSIS 4495 Applied Research, Section 002

In Partnership with Riipen
Tony Tsui, Solaris Canada

A Project Proposal Submitted to
Douglas College
New Westminster, BC, Canada

by
Dundee Adriatico
300393449
January 26, 2026

I. Introduction

The Emergency Medical Services (EMS) refer to the first responders on the public health frontlines. They handle everything from routine incidents to life-threatening emergencies like heart attacks and major trauma. In North America, millions of EMS dispatches are happening annually, which generates a large volume of operational data stored within the National Emergency Medical Services Information System (NEMSIS) (National Association of State EMS Officials, 2020; NEMSIS TAC, 2025a). The NEMSIS Public-Use Research Dataset provides de-identified and large-scale records, which makes evidence-based analyses of EMS performance and workforce health possible.

A. Problem Framing

The Canadian EMS paramedics deal with significantly high injury rates from incidents like lifting of heavy patients, patient assaults, mental health stress from difficult scenes amongst others. Given these, several key questions can be asked:

- How can we leverage on NEMSIS data to forecast EMS demand and peak loads?
- What models and dashboards can unravel and spot injury-linked patterns, for instance overloads causing accidents, for better QA protocols?
- What kind of QA improvements can reduce field injuries? How can AI or ML forecasting and automated reporting enhance safety monitoring?

These questions are valuable because a forecast of EMS demand peaks brings about a couple of things ranging from preventing fatigue-driven injuries, optimizing rostering of responders, and mitigating risks during high-volume shifts where risky solo lifts are likely to happen. This then boost patient care with more reliable responses. This project also aims to addresses disabling injuries, which is defined as any physical or psychological injury causing a worker to miss at least one day of work. Generally, the EMS services show higher rates compared to other industries and ambulance services, reaching 21% annually in some cases (Edmonton Metro EMS blog, 2025).

B. Literature Summary & Gaps

The following literature shows the occupational hazards encountered by Canadian EMS paramedics:

- WorkSafeBC reports detailed the prevalence of musculoskeletal disorders from repetitive patient handling tasks, resulting in over 60,000 lost workdays in British Columbia alone (WorkSafeBC, n.d.).
- The analysis of Canadian Union of Public Employees (CUPE) of Ontario paramedic services revealed approximately 2,700 annual WSIB claims, worsened by chronic staffing shortages

that force extended shifts and increase exposure to violence and psychological trauma (CUPE, 2020).

- The Journal of Emergency Medical Services (JEMS) published studies to further quantify violence risks, finding that 48% of providers in regions like Peel near Toronto have experienced assaults, accentuating the multi-layered injury burden (JEMS, 2024).

While the National Emergency Medical Services Information System (NEMSIS) provides one of the largest standardized EMS datasets available for research, existing literature using NEMSIS data has largely focused on retrospective descriptions and statistical analyses of past events rather than leveraging modern machine learning and AI techniques for operational forecasting or predictive modeling (NEMSIS TAC, 2025a). Although machine learning methods have been applied in some contexts, such as binary outcome imputation within the NEMSIS dataset, these efforts do not extend to forecasting future demand peaks, injury risk prediction, or real-time operational insights based on the national data (National Association of State EMS Officials, 2020). Broader EMS research employing ML / AI exists, but predominantly uses regional or clinical data sources outside the NEMSIS Public-Use Research Dataset, highlighting a clear gap in predictive analytics research directly built on national EMS data. This gap underscores the opportunity in my RIPEN project (Riipen, 2026) to build scalable ML models that forecast demand surges, identify patterns linked to paramedic injuries, and support real-time quality improvement systems.

C. Hypotheses, Assumptions & Benefits

Initial Hypothesis

By building data pipelines to clean and prepare the NEMSIS Public-Use Research Dataset and testing multiple forecasting models, starting with simple approaches like linear regression, then advancing to logistic regression and three or more complex algorithms (e.g., random forests, gradient boosting), I'll identify the best model for predicting EMS demand peaks and identifying peak-load periods with 75%+ accuracy. This will reveal high-risk patterns (e.g., call volume surges linked to lift injuries during overtime) and support QA recommendations that reduce disabling injury rates by 10–15% through smarter crew scheduling and resource allocation.

Key Assumptions

The NEMSIS dataset contains sufficient complete data (at least 70% after cleaning) in variables like call timestamps, incident types, and outcomes for reliable pattern detection; our forecasting models will run efficiently within the 2-month, 10-12.5 hours/week timeline; Toronto-area EMS services will find the demand insights actionable enough to pilot in practice; and demand peaks correlate closely enough with injuries that forecasting one helps prevent the other. I also assume our interactive dashboards will clearly communicate findings in ways paramedics and dispatchers can use day-to-day.

Potential Benefits

Within my tight timeline, this project delivers measurable value aligned with RIPEN goals: accurate demand forecasts and peak-load identification help EMS optimize scheduling to reduce fatigue-driven injuries and WSIB claims, protecting paramedic health and retention. For patients, reliable staffing ensures faster responses to emergencies like cardiac arrests. As a RIPEN student, I will gain hands-on experience across the full data science pipeline starting from data cleaning via pipelines, model comparison (simple to advanced algorithms), demand forecasting, until interactive dashboard design. These skills directly apply to healthcare analytics careers. The work is also scalable: other Canadian EMS regions can adopt our data pipeline approach, forecasting methods, and dashboard templates for their own QA improvements. Ultimately, I am transforming raw NEMSIS data into actionable operational insights and visual tools that make EMS safer for workers and patients while showcasing student proficiency in real-world data management and communication.

II. Proposed Research Project

A. Research Design & Objectives

My research is structured to follow a data science approach with five main phases: exploring the data to understand patterns, cleaning and preparing it for analysis, building ML models to forecast demand and predict injuries, creating automated reporting tools, and designing dashboards to show results. This design builds on what I learned in data processing and machine learning courses, plus research showing how analytics can improve EMS safety and operations.

My goals are straightforward:

- Clean and prepare the NEMSIS dataset into usable pipelines
- Build and test ML models that forecast EMS demand peaks and spot when lifting injuries are likely to happen
- Create automated reports that suggest QA improvements based on the data
- Design interactive dashboards that show what I found
- Give EMS services data-backed recommendations to cut on-duty injuries

B. Data Collection & Sample

I'm using the NEMSIS Public-Use Research Dataset covering 2019–2024, which has over 50 million EMS records. Since processing all of it would take too long with my 2-month, 10-12.5 hours/week timeline, I'll randomly sample about 200,000 incidents while keeping the mix of years, regions, and injury types realistic. I'll focus on records that mention workplace injuries—especially lifting or musculoskeletal strain—since those directly support my QA analysis.

C. Data Analysis Techniques & Methods

Data Cleaning: I'll use Python (Pandas) to ingest the data, spot outliers using the interquartile range method, and fill missing values with median or KNN techniques.

Exploratory Analysis: I'll look at time-series patterns between shift times and injuries, map high-risk areas geographically, and create new features like hour-of-day, day-of-week, call type, crew size, and weather variables.

Forecasting & Prediction: I'll test simple models like linear regression first, then move to more complex ones—logistic regression, random forests, and gradient boosting—to see which best predicts demand peaks and identifies high-risk shifts. Once I pick the best model, I'll use it to forecast busy periods and estimate injury risk windows.

Automated Reporting: I'll build a simple rule-based system using Python that reads the model outputs and automatically generates QA reports (e.g., "Recommend two-person lifts between 6–10 PM based on injury risk data").

D. Technologies & Tools

Platform & Environment: Jupyter Notebooks for my main analysis, working locally or on free cloud Python environments. I might use Databricks Community Edition for testing Spark-based data pipelines if I have time, though I'm aware of its limits.

Programming Languages & Libraries: Python 3.11, Pandas (data handling), scikit-learn (ML models), XGBoost or similar (advanced forecasting), and MLflow (to track which models work best).

Database: I'll start with Excel or CSV stored locally; if using Databricks, I'll explore Delta Lake for organizing data versions.

Front-End & Visualization: Streamlit (my main choice because it works smoothly with Python) for interactive dashboards. Worst case scenario, if Streamlit doesn't fit, I can switch to Tableau Public or Power BI depending on what works best.

E. Expected Results & Practical Impact

I expect to deliver cleaned datasets, forecasting models that predict demand peaks with 75%+ accuracy, and interactive dashboards showing busy times and injury hot spots by region or shift. The automated reporting tool will generate QA summaries and safety recommendations automatically from the data.

Real-world benefits: Better demand forecasts help EMS schedule crews smarter, reducing fatigue and lifting injuries, which cuts WSIB claims and keeps experienced paramedics on the job. Patients benefit from reliable staffing, especially for time-critical emergencies like cardiac arrests. As a student, I'm building a complete data science portfolio starting from pipelines to ML model testing until dashboard design. This is directly relevant to healthcare analytics jobs. Because I'm using a

public dataset and open-source tools, my approach and dashboards can be shared with other Canadian EMS services looking to improve their own safety systems and operational efficiency.

III. Riipen External Partners

This Riipen-approved project partners with Tony Tsui of Solaris Canada or Solaris Innovations Canada Inc. in Markham, ON. The company is a minority-owned tech startup in smart accessories. It also partners with NEMSIS (National EMS Information System) for data standards and research applications.

IV. Project Planning and Timeline

Overview

With 2 months (8 weeks) and approximately 10–12.5 hours per week (80–100 total hours), I've structured the project into five overlapping phases with clear milestones and deliverables. This timeline includes time to explore Databricks while maintaining Jupyter Notebooks as my primary fallback, plus a 5–10 hour contingency buffer built into Week 8. This prioritizes core deliverables (cleaned data, forecasting models, dashboards) while remaining realistic about solo scope and learning curve.

Phase Breakdown & Milestones

Phases	Phase Points	Description
Phase 1: Setup, Learning & Exploratory Data Analysis (Weeks 1–2, 20 hours)	<i>Deliverables</i>	Project environment setup, NEMSIS dataset loaded, EDA report with visualizations, Databricks exploration notes
	<i>Milestones</i>	<ul style="list-style-type: none"> • Week 1a (3 hrs): Download NEMSIS public dataset; set up Jupyter Notebooks and Python libraries • Week 1b (4 hrs): Study Databricks Community Edition—features, limits, cluster setup • Week 2a (5 hrs): Complete exploratory analysis in Jupyter—understand data structure, identify missing values, visualize injury/demand patterns • Week 2b (5 hrs): Test NEMSIS sample on Databricks; decide if it's worth continuing or pivot to Jupyter-only
	<i>Decision Point</i>	If Databricks setup takes >5 hours or cluster limits frustrate progress, commit fully to Jupyter Notebooks and reallocate those 2–3 hours to modeling

Phases	Phase Points	Description
	<i>Tasks</i>	Data ingestion, summary statistics, initial time-series plots, injury type frequency analysis, Databricks trial, environment comparison
Phase 2: Data Cleaning & Pipeline Development (Weeks 2–4, 25 hours)	<i>Deliverables</i>	Cleaned dataset, reusable ETL pipeline code, preprocessing documentation
	<i>Milestones</i>	<ul style="list-style-type: none"> Week 2–3: Build Pandas pipelines for outlier detection, missing value imputation (median/KNN)—in Jupyter (primary) or Databricks (if smooth) Week 3–4: Test and validate pipelines on full 200K sample; document data quality metrics
	<i>Tasks</i>	Outlier removal, feature engineering (hour-of-day, weekday, call type), variable derivation, data versioning
	<i>Buffer</i>	If Databricks is slow, this phase stays in Jupyter; no time loss
Phase 3: ML Model Development & Evaluation (Weeks 4–6, 30 hours)	<i>Deliverables</i>	Trained forecasting models, model comparison report, best-model selection
	<i>Milestones</i>	<ul style="list-style-type: none"> Week 4: Build baseline—linear regression for demand forecasting Week 5: Develop advanced models—logistic regression, random forests, gradient boosting; train on 150K samples, validate on 50K Week 6: Compare accuracies, select best model (target 75%+); document hyperparameters and reasoning
	<i>Tasks</i>	Model training, cross-validation, accuracy/precision/recall metrics, MLflow logging, hyperparameter tuning
Phase 4: Automated Reporting & QA Integration (Weeks 6–7, 15 hours)	<i>Deliverables</i>	Rule-based reporting script, sample QA recommendations output
	<i>Milestones</i>	<ul style="list-style-type: none"> Week 6–7: Build simple rule engine using Python; generate automated QA alerts (e.g., "High-risk lift windows: 6–10 PM")

Phases	Phase Points	Description
	<i>Tasks</i>	Logic scripting, linking model predictions to QA rules, testing report generation on sample data
Phase 5: Dashboard Development & Final Deliverables (Weeks 7–8, 20 hours + 5–10 hour buffer)	<i>Deliverables</i>	Interactive Streamlit dashboard, final project report, code repository with documentation
	<i>Milestones</i>	<ul style="list-style-type: none"> Week 7: Build Streamlit app with demand forecasting viz, injury hotspot maps, peak-load charts Week 8a (12 hrs): Polish dashboard, add filters/interactivity, test end-to-end workflow Week 8b (5–10 hrs buffer): Debug issues, final documentation, repo cleanup, presentation prep, extended features if time permits
	<i>Tasks</i>	Streamlit coding, potentially geospatial visualization if possible, dashboard testing, final documentation, GitHub/repo upload, contingency work

Gantt Chart

Activities	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9
Setup and EDA									
Databricks Trial									
AI Integration Plan									
Data Cleaning									
Feature Engineering									
Model Development (Simple)									
Model Development (Advanced)									
Model Selection									
AI-Driven Reporting									
Dashboard Development & AI									
Final Polish / Buffer									

V. Project Contract

Project Overview

- **Student:** Dundee Adriatico
- **Project:** EMS Data Analysis and Forecasting (RIPEN)
- **Duration:** 8 weeks - 10 weeks (January 26 – April 10, 2026)
- **Time Commitment:** 10–12.5 hours per week

Core Deliverables

1. **Data Pipelines:** Clean and prepare NEMSIS dataset (Pandas ETL)
2. **Predictive Models:** Forecast EMS demand peaks (linear regression, random forests, gradient boosting)
3. **Interactive Dashboard:** Streamlit app showing forecasts, peaks, injury hotspots, AI-generated QA alerts
4. **AI Integration:** Rule-based reporting + TF-IDF text processing
5. **Code Repository:** GitHub with full documentation and data dictionary

Key Terms

- **Databricks trial:** 4–5 hours Week 1; pivot to Jupyter if time-consuming
- **AI scope:** Rule-based + lightweight NLP only; no LLM APIs
- **Model target:** 75%+ accuracy; best of 2–3 algorithms
- **Buffer:** 5–10 hours Week 8 for debugging and documentation
- **Final deadline:** April 10, 2026

Communication Schedule

- **Weekly:** Stakeholder meeting every Friday 7:30 AM PST
- **Monthly:** End-of-month check-in with course instructor

I Commit To

- Delivering all core deliverables on time
- Attending weekly stakeholder + monthly professor meetings
- Communicating blockers early
- Following code quality & documentation standards

Student Name: Dundee Adriatico

Student Signature: 

Date: 26 Jan 2026

VI. AI Use Section

AI Tool Name	Version Account Type	Specific Feature for which the AI Tool was Used	Value Addition <i>(What value did you add over and above what AI did for you?)</i>
Perplexity	Education Pro	Asked AI to surface literature that would support framing the problem that demonstrates occupational hazard encountered EMS personnel	Having done a coop role in VCH, I learned that work-related injuries are handled by WorkSafeBC so I also asked AI to check for any statistics coming from the organization, this is on top of the 2 it produced on CUPE and JEMS
Perplexity	Education Pro	Asked AI to confirm if there's any AI or ML related studies or projects published using the NEMSIS data in order to check uniqueness of my study	Gave input to AI about the focus of my studies which is really about forecasting work-related injuries for emergency personnels so that it serves as comparison for other studies that may have been done
Perplexity	Education Pro	Asked AI to evaluate my timeline if reasonable and for any crucial activities I may have missed	Provide the milestones I crafted as well as the activities envisioned so that AI provide further input on anything I missed

VII. Work Log

Date	Number of Hours	Description of Work Done
18-Jan-26	0.5	Meeting with Priya, initial discussion of topics
21-Jan-26	2	NEMSIS Data initial proposal
23-Jan-26	0.5	Meeting with Solaris, Tony Tsui
24-Jan-26	1	Initial works on the final proposal, intro and proposed research project
25-Jan-26	2	Align Intro and Proposed Research Project on the final proposal – Riipen link finally given
26-Jan-26	4	Final works on the final proposal – Estimate and update timelines, build contract, update references, update AI usage table and prompts and perform final review and revisions

VIII. Closing and References

Thank you to Tony Tsui for giving me this opportunity to work on this project. I appreciate Priya, my professor, for allotting effort, time, and guidance to initiate this project. Special thanks to VCH, because without the work I did for VCH, I wouldn't have the appreciation or drive to take on these kinds of projects. Finally, to my wife and twins for being my inspiration every day.

Web Reports

CUPE. (2020). *Under pressure: A statistical report on paramedic services in Ontario*. <https://cupe.on.ca/under-pressure-a-statistical-report-on-paramedic-services-in-ontario/>

Edmonton Metro EMS. (2025). *Disabling injury rate*. <https://www.edmontonmetrolocal.ca/blog/edmonton-metro-ems-disabling-injury-rate/>

JEMS. (2024, October 8). *EMS Week: Canadian paramedic service discovers 48% of providers reported experiencing violence*. <https://www.jems.com/ems-operations/ems-week-canadian-paramedic-service-discovers-48-of-providers-reported-experiencing-violence>

NEMSIS TAC. (2025a). *NEMSIS technical assistance center report: Current research applications*. <https://nemsis.org>

Riipen. (2026). *EMS data analysis and forecasting*. <https://douglascollege.riipen.com/projects/ALe7YnO6>

WorkSafeBC. (n.d.). *Evaluation of paramedics tasks and equipment to control musculoskeletal injury risk*. <https://www.worksafebc.com/en/resources/about-us/research/finding-solutions-archive/evaluation-of-paramedics-tasks-and-equipment>

Published Materials

National Association of State EMS Officials. (2020). *NEMSIS overview and standardization efforts*.

IX. Appendix

List of AI Prompts

Prompts
Given my project and objective on analyzing EMS data and forecasting information to improve QA protocols to lessen work-related injuries of emergency personnel, find relevant literature that would support demonstrating the occupational hazards encountered by these workers.
When the literature was given, I further prompted if AI can give me stats by WorkSafeBC, having learned about the organization during my coop role time in VCH.
Given that I am doing a project on NEMSIS data, provide confirmation on any studies done in the past that is related to AI or ML so that I understand the significance and uniqueness of my study. My study focuses on forecasting work-related injuries for emergency personnels.
Review and evaluate my timeline of the project from EDA, cleaning, model evaluation and selection until dashboard reporting and check if reasonable given the 10-12.5 hrs I can allot weekly for a 2-month period. Mention any activities I may have missed to include.