MMDLNet: A Multimodal Deep Learning Network for Post-Collision Injury Prediction

Zhicheng Zou, Martin Elsegood, Giulio Ponte, Sam Doecke, Arnab Majumdar, Claire E. Baker*

I. INTRODUCTION

Road traffic collisions (RTCs) represent a critical global health burden, yet a persistent "care gap" remains between when a crash occurs and the first care for the casualty [1, 2]. Rapid and reliable post-crash triage is essential for improving patient outcomes. While current Automatic Collision Notification (ACN) systems offer basic location data that has been shown to improve response times and subsequently patient outcomes, ACN lacks the granularity to provide real-time remote triage decision-support. Advanced ACN additionally provide an estimate of overall severity, however they lack granularity to inform emergency operations centre staff dispatch to RTCs [3,4]. Contemporary, rich crash-related data and deep learning approaches have increasing potential to support early injury assessment. We assess how static crash descriptors, irregularly sampled event data recorder (EDR) data and post-crash images input to a compact multimodal deep learning framework predict occupant- and vehicle-level treatment type, and body-region-specific injury level to better align with emergency response needs.

II. METHODS

Data Source The model was developed and validated using the CASR-EDR dataset maintained by the Centre for Automotive Safety Research in South Australia [5]. This unique repository links high-frequency vehicle telemetry with police reports and verified hospital outcomes coded via the 2015 Abbreviated Injury Scale (AIS). The cohort consists of 1,063 crashes involving 2,260 vehicles and 3,134 individual occupants.

Model Architecture Transformer architecture with multimodal to process heterogeneous data streams (Fig 1) to predict occupant- and vehicle-level treatment type (none, doctor, hospital type, fatal), and body-region-specific injury (MAIS to AIS body regions including head and neck, face, chest, abdomen, extremities, external).

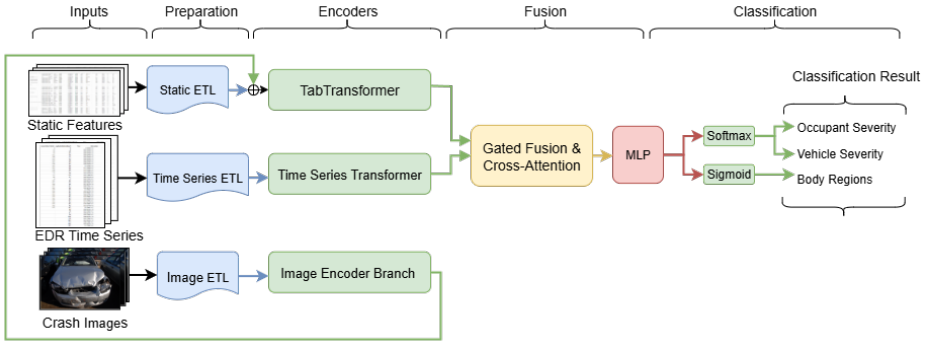

Figure 1 MMDLNet model transformer architecture visual summary.

Static features include environmental (e.g. lighting, road type), vehicle (e.g. safety systems, model age) and occupant variables (e.g. demographics) which are input to a TabTransformer encoder that embeds categorical variables and projects numerical descriptors into a contextualised sequence, capturing non-linear interactions between variables such as seating position and restraint use. Time series crash kinematics are input to a Time Series Transformer that uses continuous positional encodings (CPE) to handle the irregular sampling inherent in EDR data from a variety of vehicles. Further irregularity arises from the transition between low-frequency pre-crash buffering (e.g. 1–2 Hz) and the high-frequency crash-pulse triggers (e.g. >100 Hz) recorded during the impact phase. Where images are available, they are input to a frozen ResNet-50 backbone which extracts features from post-crash photographs. The resulting modality-specific representations are then fused within MMDLNet using a gated residual mechanism, in which a sigmoid gate controls the dimension-wise mixture of representations to blend these modalities. This allows the model to adaptively reduce reliance on noisy modalities, for example relying on static context when telemetry is absent, or emphasising dynamic EDR signals when the crash pulse is informative. A two-token cross-attention block facilitates bidirectional exchange between the static and dynamic streams, and the fused representation is processed by an MLP with hidden dimensions of 1024 and 512.

Optimisation, training and assessment The model is trained using a multi-task objective that combines losses for occupant-, vehicle- and body-region severity. Optimisation is performed using AdamW with a cosine-annealed learning-rate schedule. To improve stability and reduce overfitting, early stopping, dropout regularisation and gradient clipping are adopted. Class imbalance is addressed through class-weighted cross-entropy for severity tasks and per-region positive weights for body-region predictions. For the visual branch, images are processed with standard ImageNet preprocessing and augmentation. Model performance is assessed on the CASR-EDR

C Baker (c.baker17@imperial.ac.uk) is a Design Engineering Research Fellow, A Majumdar is a Civil Engineering Professor and Z Zou was a Security & Resilience masters student aat Imperial College London, UK. Z Zou is a PhD student at the University of Bristol, UK. M Elsegood and G Ponte are Research Engineers and S Doecke is a Senior Research Fellow at Adelaide University Centre for Automotive Safety Research.

dataset and compared with representative classical and learning-based baselines: XGBoost, LightGBM, CatBoost, random forest (RF), multinomial logistic regression (LR), and gated recurrent unit (GRU). We report Area Under the Receiver Operating Characteristic curve (AUROC), Area Under the Precision-Recall curve (AUPRC), F_1 score, recall-oriented F_3score, Expected Calibration Error (ECE) and Brier score.

III. INITIAL FINDINGS

MMDLNet demonstrated superior performance across all metrics compared to the strongest classical baselines across vehicle, occupant and body-region task prediction. Vehicle-level prediction demonstrated the strongest overall results, achieving an AUROC of 0.9821 and a recall-oriented F_3 score of 0.8417. Due to space constraints, only occupant severity is shown (Table 1). For occupant severity, MMDLNet attains the highest F_3 score (0.6314), improving on the strongest baseline (CatBoost, 0.5162) by 0.1152, indicating superior sensitivity to severe outcomes. Body-region injury severity prediction achieved macro-averaged AUPRC (0.5410) and F_1 score (0.4401) were lower than the other two tasks, though it maintained a strong F_3 score of 0.6911.

Table 1 Occupant severity (six-class) representative baselines versus our multimodal model. Test set results averaged over three independent runs. Thresholds for and are tuned on validation to maximise F3 score. Arrows show whether high or low is preferable, while bold is best performance and italic-underlined is second-best performance.

Methods	Accuracy ↑	AUROC ↑	AUPRC ↑	F1 ↑	F3 ↑	ECE ↓	Brier ↓
LR	0.5243	0.8268	0.4452	0.4091	0.4522	0.0952	0.6259
RF	0.7187	0.8597	0.5450	0.4199	0.3956	*0.0261*	*0.3846*
LightGBM	0.7161	*0.8752*	0.5419	0.5190	0.5091	0.0959	0.3911
XGBoost	*0.7263*	0.8438	0.5467	*0.5408*	0.5083	0.0476	0.4157
CatBoost	0.7187	0.8580	*0.5659*	0.5366	*0.5162*	0.0415	0.4014
GRU	0.6240	0.8062	0.3626	0.3460	0.3567	0.3150	0.6719
MMDLNet	**0.7806**	**0.8863**	**0.6913**	**0.5950**	**0.6314**	**0.0124**	**0.2118**

Ablation studies demonstrate that integrating static descriptors, EDR timelines, and images significantly improves injury prediction accuracy and calibration compared to single-modality baselines. The fusion mechanism, which uses gated residuals and cross-attention, provides the most substantial boost in recall-oriented performance by adaptively balancing contextual information with realized crash dynamics. Telematics data from EDRs offer decisive value, especially when continuous positional encodings are used to preserve the irregular temporal structure of kinematic signals (Table 2). Finally, the image branch provides the largest gains in ranking-based metrics by capturing visual evidence of structural intrusion that sensor data alone might miss.

Table 2 Summary component contributions. Note: all values are macro-averaged across occupant-, vehicle-, and body-region severity.

Configuration	AUPRC	F3 (Recall-biased)	ECE (Calibration)
Baseline (naïve fusion)	0.4485	0.4626	0.0913
+ Gated fusion and cross-attention	0.4831	0.5600	0.0415
+ Continuous positional encodings	0.5550	0.6120	0.0319
Full model (inc. image branch)	0.6992	0.7214	0.0142

IV. DISCUSSION

The inclusion of time-series and image data alongside traditional static predictors into our multimodal deep learning network architecture demonstrated consistent improvements over representative classical and learning-based baselines in terms of both discrimination and calibration, while maintaining computational efficiency compatible with dispatch-time deployment. In emergency medical services, the target for missing severe cases (undertriage) is typically <5% [5]. MMDLNet's recall-weighted architecture is specifically tuned to meet this safety-critical threshold by interpreting crash pulses which are expected to relate to trauma presentation through injury biomechanics. Furthermore, low ECE is vital for dispatcher confidence. In a high-stress emergency call centre where false positives are a key current set back, well-calibrated risk scores provide better decision support tools. Model robustness is high from utilising learned "no-image" and "no-telemetry" embeddings, the framework remains functional even when data streams are incomplete: a common reality in the immediate aftermath of a collision. Future work will compare this to a limited feature set based on literature review [3]. Our results indicate that combining increasingly available vehicle sensor data with visual evidence can enhance early estimates of injury risk and support automated post-crash triage.

V. REFERENCES

[1] WHO, Global Road Safety Report, 2023 [2] Nutbeam & Stassen, Inj, 2025 [3] Baker et al., SSRN, 2026
[4] Baker et al., SJTREM, 2026. [5] ACS, Committee on Trauma, 2021 [6] Reed, e-Call report, 2025