A Training Framework for Enhancing the Robustness of Occupant Injury Prediction Algorithms

Tianle Lu, Dongyang Xu, Qingfan Wang, Shi Shang, Qing Zhou, Bingbing Nie*

I. INTRODUCTION

The application of data-driven algorithms in occupant injury prediction allows autonomous vehicles (AVs) to implement proactive protection strategies when confronting imminent collision scenarios, thereby mitigating injury severity [1]. Considering the significant impact of such protection strategies on occupant safety, the accuracy and robustness of occupant injury predictions are critical [2]. Classical machine learning models may misestimate occupant injuries, potentially resulting in critical decision-making errors [3]. To address these challenges, physics-informed neural network (PINN) has emerged as an effective training paradigm by integrating physical constraints into the model training process [4]. Motivated by this, we propose a PINN-based training framework that incorporates a vehicle dynamics model into the training process of machine learning methods to enhance the robustness of occupant injury prediction.

II. METHODS

The proposed PINN-based training framework contained four key modules (Fig. 1): (a) map information M and vehicles' historical trajectories A were collected and served as the inputs; (b) a pure neural network (PNN) was developed to further process the input data and predict injury severity $\hat{y}_m$; (c) a physical prediction model (PPM) was designed to output a weighting coefficient α, which guided the PNN's outputs to follow the vehicle dynamics; (d) overall, the model was trained using not only the prediction loss $\mathcal{L}$ but also by incorporating α to enforce physical constraints. This ensured that the model learned from both the ground truth y_{gt} and the physical prior knowledge. This method maps the collision velocity v_c to discrete injury severity categories defined by injury risk curves: *Slight* ($v_c \leq 20$ km/h), *Serious* (20 km/h $< v_c \leq 60$ km/h), and *Fatal* ($v_c > 60$ km/h) [5].

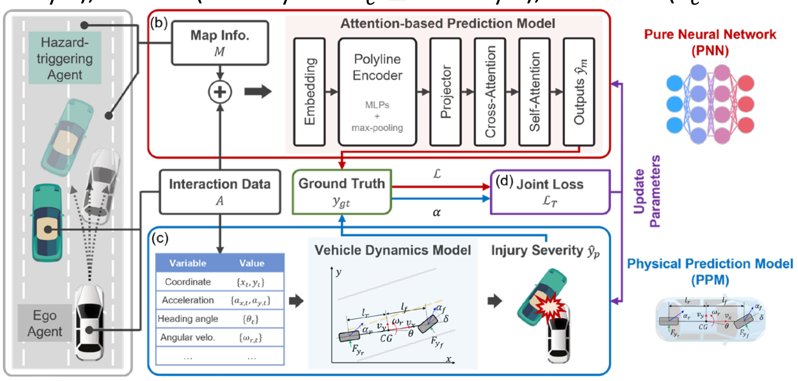

Fig. 1. The proposed PINN-based training framework of the occupant injury prediction model.

PNN. A Polyline Encoder, consisting of alternately stacked multi-layer perceptron (MLP) and max-pooling layers, was employed to extract key features A from the interaction process between the ego vehicle and the hazard-triggering vehicle. Concurrently, map information M was integrated to constrain the model to recognise interaction types. To process and integrate diverse features within a unified feature space, the projector mapped M into the embedding space of A. The fused features were subsequently input into the attention module, which comprised cross-attention and self-attention layers, to enhance the correlation of vehicle kinematic and road environment across different time intervals. The model's final output is the predicted injury severity $\hat{y}_m$.

PPM. A reachable set of trajectories for the target vehicle was predicted using vehicle dynamics theory, given its current driving state and control inputs. The control input $u_t^{(i)}$ was discretised into N predefined decisions, each of which involved two parameters: acceleration a, and steering angle δ. Specifically, the historical driving states of the target vehicle were fitted via cubic spline interpolation to predict the a and δ, which were then used to derive driving decisions. PPM assessed potential collisions by iteratively evaluating the overlap between the two vehicles at consecutive time steps along each trajectory generated by $u_t^{(i)}$. It then recorded the injury severity

B. Nie (e-mail: nbb@tsinghua.edu.cn) is an Associate Professor, T. Lu and Q. Wang are PhD students, D. Xu is a MSc student and Q. Zhou is a Professor, all in the School of Vehicle and Mobility at Tsinghua University and State Key Laboratory of Intelligent Green Vehicle and Mobility, China. S Shang is an Assistant Professor in the School of Technology at Beijing Forestry University, China.

$\hat{y}_p$. If $\hat{y}_p$ matched y_{gt}, a high weight $\alpha = \alpha_{boost}$ was assigned to enhance the adherence of the PNN to the physical constraints imposed by the PPM during training. Otherwise, a low weight $\alpha = \alpha_{abate}$ was allocated to mitigate the influence of errors. Consequently, the joint loss $\mathcal{L}_T$ was calculated as the product of α and $\mathcal{L}$:

$$\mathcal{L}_T = \alpha \cdot \mathcal{L} = \begin{cases} \alpha_{boost} \cdot CE(y_{gt}, \hat{y}_m), & \hat{y}_p = y_{gt} \\ \alpha_{abate} \cdot CE(y_{gt}, \hat{y}_m), & \hat{y}_p \neq y_{gt} \end{cases} \quad (1)$$

where $CE(\cdot,\cdot)$ denoted the cross-entropy loss.

III. INITIAL FINDINGS

The datasets used for training and testing were obtained from a previous driving simulator experiment, which included 1,879 high-risk conflict cases [6]. To validate the proposed PINN-based framework, the performance of the following models was compared: (1) classical machine learning models, such as XGBoost and LSTM; (2) a risk prediction model, specifically, an Att-Bi-LSTM (ABL) model based on a self-attention mechanism [7]; and (3) the proposed training framework (PNN & PINN-based). In addition to classification metrics, the linear weighted Kappa coefficient κ was introduced to measure the consistency between predicted labels and ground truth labels, with a κ value approaching 1 indicating higher prediction robustness. The results are summarised in Table I.

TABLE I

PERFORMANCE COMPARISON OF VARIOUS OCCUPANT INJURY PREDICTION APPROACHES

Model	***Slight***		***Serious***		***Fatal***		**Accuracy**	**κ**
	F1-score	**AUC**	**F1-score**	**AUC**	**F1-score**	**AUC**		
XGBoost	0.947	0.930	0.905	0.926	**0.933**	0.955	0.933	0.882
LSTM	0.931	0.909	0.851	0.885	0.844	0.926	0.899	0.825
ABL	0.948	0.935	0.911	0.937	0.911	0.944	0.931	0.880
PNN	0.938	0.962	0.897	0.970	0.899	**0.978**	0.921	0.858
PINN-based	**0.963**	**0.981**	**0.933**	**0.982**	0.932	0.977	**0.951**	**0.912**

The PINN-based model obtained best performance in prediction accuracy and robustness metrics. From the total eight metrics, our proposed model showed suboptimal results on only two metrics, yet it remained very close to optimal performance. The average F1-score and AUC increased by 3.7% and 4.5%, respectively, compared with baseline models, indicating better overall prediction performance (F1-score, which balances correct predictions and missed cases) and enhanced ability to correctly differentiate between injury severity categories (AUC). Accuracy and κ improved by 3.3% and 6.0%, respectively. Notably, the integration with the PPM further refined the model's structural design, enhancing classification performance in the *Slight* and *Serious* categories and increasing κ from 0.858 to 0.912.

IV. DISCUSSION

The accuracy and robustness of occupant injury prediction are vital for AVs to implement protection strategies in unavoidable collisions. Previous studies focused on accuracy-related metrics to assess model performance, neglecting the robustness. To address this gap, this study proposes a PINN-based model training framework. This approach mitigates the risk of the model relying solely on data distribution patterns to make predictions and ensures that outputs align more closely with physical principles, thereby enhancing the prediction robustness. However, integrating physical constraints increases model training time. Additionally, this preliminary study relies on a simplified representation of injury severity. Future work should evaluate its effectiveness using more detailed and representative injury metrics while optimizing training efficiency.

V. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (52131201).

VI. REFERENCES

[1] Wang, Q., *et al., iScience*, 2022.

[2] Shen, J., *et al.*, *IRCOBI*, 2024.

[3] Geng, M., *et al.*, *Transp Res Part C: Emerg Technol*, 2023.

[4] Rajamani, R., *Springer Sci Bus Media*, 2011.

[5] Richards, D., *Road Saf Web Publ No. 16*, 2010.

[6] Kuang, G., *et al.*, *IRCOBI*, 2024.

[7] Xie, Z., *et al.*, *Accid Anal Prev*, 2024.