Proceeding of the 18th INFATS & 27th SAE-C/STC, 2024, pp1-9
2024.103

Research on Driver Injury Prediction for Passenger Car Rear-End Collisions

Xin RAN[1], Zhenfei ZHAN[2]
[1] *CountryChongqing Jiaotong University, Chongqing, China, 400074*
Email: *zhenfei_zhan@163.com*

Abstract: The current focus of intelligent automotive safety development is on actively preventing collisions. When an accident is unavoidable, intelligent vehicles require a safety prediction system to provide decision-making references for the adaptive restraint system within the vehicle. To this end, this paper establishes a driver injury prediction model for rear-end collisions in passenger cars based on the real accident database CISS. Utilizing a Stacking ensemble learning algorithm, the model employs five types of classifiers—Random Forest, LightGBM, XGBoost, GBDT, and CatBoost—as base learners, with a Logistic Regression classifier serving as the meta-learner, thereby constructing a two-layer ensemble learning prediction model. The model takes vehicle motion characteristics, restraint system features, and driver characteristics as inputs for the driver injury prediction model, with the maximum injury level of the driver as the output. Experimental results indicate that the constructed driver injury prediction model achieves an accuracy of 82.9%, surpassing that of other individual machine learning classifiers. This model can provide decision-making recommendations for the adaptive restraint system of intelligent vehicles during collision events.

Keywords: Intelligent safety; Accident data; Ensemble learning;Injury prediction

面向乘用车追尾事故的驾驶人损伤预测研究

冉　鑫[1]，詹振飞[1]
[1]重庆交通大学，重庆，中国，400074
Email: zhenfei_zhan@163.com

摘　要: 目前汽车智能安全的发展重点关注主动避免发生碰撞，当事故无法避免时，智能汽车需要安全预测系统为车辆内部自适应约束系统提供决策参考。为此，本文基于真实事故数据库 CISS 来搭建乘用车追尾事故驾驶员损伤预测模型。本文基于一种 Stacking 集成学习算法，以随机森林、Lightgbm、Xgboost、GBDT、Catboost 五类分类器为基学习器，逻辑回归分类器为元学习器，构建了双层集成学习预测模型。该模型以车辆运动特征、约束系统特征以及驾驶人特征等作为驾驶人损伤预测模型的输入，以驾驶人的最大损伤等级作为输出。试验结果表明，本文所构建的驾驶员损伤预测模型在准确率方面达到了 82.9%，高于其他单一机器学习分类器。能够为智能车辆在碰撞发生时的自适应约束系统提供决策建议。

关键词: 智能安全；事故数据；集成学习；损伤预测

1 引言

近年来，随着自动驾驶技术的不断发展，汽车安全也在不断向智能化方向发展[1]。智能汽车感知系统、车载计算系统等技术的完善，更是极大的促进了汽车智能安全的发展。但是由于交通环境复杂、司机违规操作等因素，事故的发生依旧无法完全避免[2]。汽车智能安全系统一方面在事故发生前通过感知决策规划等技术来控制车辆进行主动避撞，另一方面在无法避撞时通过调整乘员的约束系统，车辆碰撞角度等，来减少碰撞过程中人员受到的损伤。

因此当碰撞发生无可避免时，如何预测乘员或驾驶员的损伤情况从而对车辆约束系统做出正确调整对汽车智能安全的发展至关重要。

乘员损伤预测近年来已经成为车辆安全领域的研究热点之一[3]。关于乘员损伤预测模型构建的研究主要是基于数据驱动的统计模型与机器学习模型。统计模型由于其可解释强，广泛应用于事故严重程度预测与乘员损伤风险预测，例如有序 Probit 模型[4-5]、Logit 统计学模型[6]等等。与统计模型需要预先确定自变量与因变量的分布关系相比，机器学习模型更加灵活，适用范围更广，对待数据噪声、缺失值处理具有更好的鲁棒性。Ejima[7]基于 NASS-CDS 公布的事故数据集建立了 ISP-R 乘员损伤预测模型。Lu[8]基于 2019 年美国国家公路交通安全管理局（NHTSA）死亡分析报告系统 （FARS） 的年度数据构建了遗传算法和 BP 神经网络的驾驶员损伤预测模型，能够改进车辆碰撞自动呼叫系统中的算法和性能。Zhang[9]提出了一种基于混合特征选择的机器学习分类方法，构建了基于 XGBoost 的车车事故和单车事故损伤预测模型，发现该模型在预测性能方面优于随机森林、朴素贝叶斯等分类器。综上所述，随着人工智能的兴起，机器学习模型已经成为车辆碰撞中乘员或驾驶员损伤预测模型构建的一种趋势。

统计模型难以满足智能车辆对乘员损伤预测的要求，而使用单一机器学习构建驾驶员损伤预测模型，由于各种机器学习算法性能不同，往往会导致在不同分类任务中表现出差异性。因此本文采用已有的真实事故数据进行模型构建，真实事故数据库 CISS 的数据样本充足，记录详细，能够满足模型构建的要求。考虑到单一机器学习模型在事故数据集上泛化性能不稳定的问题，本文采用基于 Stacking 的集成方法，通过集成不同基学习器的优势性能，构建高精度高可靠性的乘用车驾驶人损伤预测模型。选择能够通过车载传感器、雷达等手段获取的车辆运动特征、约束系统特征以及驾驶人特征等作为驾驶人损伤预测模型的输入，以驾驶人的最大损伤等级作为输出，构建驾驶人损伤模型。希望能够在碰撞无法避免时，为智能车辆作出减缓驾驶人损伤的决策提供一些建议。

2 数据预处理

2.1 数据来源

本文采用的数据来源于美国国家公路交通安全管理局（National Highway Traffic Safety Administration，NHTSA）的碰撞调查采样系统（Crash Investigation Sampling System）公开的 CISS 数据库[10]。该数据集是通过对美国选定地区 2016-2022 年发生的碰撞事故随机采样组成，采样事故必须涉及至少一辆被拖走的机动车。该采样系统每年公布数据约 5000 条，且对采样事故碰撞信息记录详细，事故记录信息如表 1 所示。

Table 1. CISS Accident Information Description

表 1. CISS 事故信息描述

数据集	描述
Crash Data	记录了车辆碰撞具体情况
General Vehicle Data	质量、航向角、降速等车辆参数
Exterior Vehicle Data	车辆外部变形情况
Interior Vehicle Data	车辆驾驶舱内损伤情况
Person Data	人员属性、安全措施及损伤等情况

本文研究对象为乘用车追尾事故中的驾驶员损伤预测问题，因此需要对 CISS 数据库公布的进行筛选，通过以下筛选规则对数据库的数据进行筛选：

（1）事故车辆类型为乘用车，仅考虑乘用车与乘用车碰撞

（2）不考虑多车追尾事故，仅选择包含两辆乘用车追尾的事故数据

（3）选择记录了驾驶人损伤最大部位的 AIS 等级的事故，去掉未记录驾驶人损伤的事故数据

从原始数据集中选择出了 1379 条符合上述筛选规则的数据作为模型构建的数据集。

2.2 数据预处理

2.2.1 特征选择

事故数据被广泛应用于交通安全领域，在事故损伤预测研究中，大多是对事故严重程度预测，故而特征选择往往是基于人-车-路-环境四个纬度来进行特征变量的初步选择。而本文研究重点在驾驶人损伤预测上，在进

行特征选择时，考虑到构建的损伤预测模型需要在车辆碰撞过程中能够实时进行损伤预测，因此仅仅选择了驾驶人特征，车辆特征选取能够从车辆传感器计算获取的航向角、降速、碰撞位置、以及驾驶舱约束系统特征等信息。初步选择特征变量如表 2 所示

Table 2. Preliminary Characteristics
表 2. 初步特征

特征变量	描述
DVLONG	碰撞降速
CURBWT	自车整备质量
DVANGTHIS	最大降速时自车航向角
DVANGOTH	最大降速时碰撞车航向角
SURFTYPE	道路类型
PART	车辆碰撞位置
BELTUSE	是否使用安全带
BAGDEPLOY	方向盘气囊是否展开
SEX	驾驶人性别
BMI	驾驶人身体质量指数
AGE	驾驶人年龄

在汽车追尾碰撞过程中，汽车质量会对人员安全性造成一定影响，较重的车辆在碰撞时可能会因为更大的动能而对乘员造成更大的冲击力，但同时也可能因为更重的车身结构而在碰撞中保持更好的稳定性，所以自车质量并不一定能够对模型精度起到正向作用，因此本文通过组合特征自车与碰撞车的整备质量比来替换自车整备质量，整备质量比能够体现出乘用车碰撞车辆的质量差异性。

碰撞时的相对航向角能够体现出车辆碰撞的有效碰撞区域，通过与车辆碰撞位置，能够确定自车的有效碰撞区域，有研究通过分析车祸数据表明[11]1/3 的碰撞区域比 100%碰撞区域乘员损伤程度更加严重。因此本文通过相对航向角来替换自车与碰撞车辆的航向角作为模型的输入。组合特征如表 3 所示

Table 3. Combination Features
表 3. 组合特征

特征变量	描述
CURWTRATIO	整备质量比
DVHEADANG	碰撞时相对航向角

通过组合特征，确定了最终模型的输入特征变量，包括碰撞降速、整备质量比、碰撞时相对航向角、道路类型、车辆碰撞位置、是否使用安全带、方向盘气囊是否展开、驾驶人性别、驾驶人身体质量指数和驾驶人年龄十个特征变量作为损伤预测模型的输入。表 4 统计了各特征变量的分布，其中连续型变量统计了其均值与标准差，类别型变量统计了其类别的频数。

Table 4. Input feature variable
表 4. 输入特征变量

输入特征变量	描述	统计描述
DVLONG	碰撞降速	均值：21.25, 标准差：11.41
CURWTRATIO	整备质量比	均值：0.97, 标准差：0.27
DVHEADANG	碰撞时相对航向角	均值：8.87, 标准差：42.09
SURFTYPE	道路类型	混泥土(0):178; 沥青(1):1200
PART	车辆碰撞位置	前部(0):864; 尾部(1):515
BELTUSE	是否使用安全带	未使用(0):125; 使用(1):981
BAGDEPLOY	方向盘气囊是否展开	展开(0):712；未展开(1):453
SEX	驾驶人性别	男性(0):657; 女性(1):719
BMI	驾驶人身体质量指数	均值：28.03, 标准差：7.24
AGE	驾驶人年龄	均值：38.46, 标准差：17.47

2.2.2 缺失值处理

数据缺失值的处理在机器学习模型构建过程中是不可或缺的一步，特别是数据量不大时，缺失值的处理能够极大影响预测模型的精度。图 1 统计了数据缺失情况，可以看出驾驶人 BMI 与降速两个连续性特征变量缺失最多，缺失率分别为 36.77%、28.71%，均超过了数据样本的四分之一。其次分别是安全带是否使用与方向盘气囊是否展开两个离散型特征变量，缺失率分别为 19.8%、15.51%。其他特征变量的缺失率均未超过 1%。

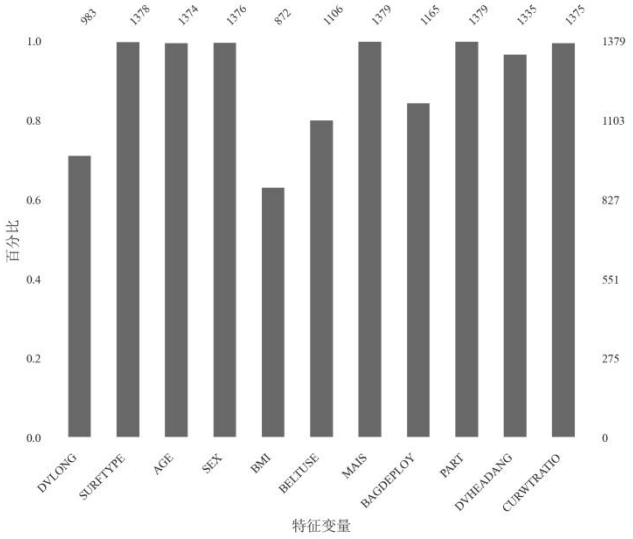

Figure 1. Distribution of missing values in the dataset

图 1. 数据集缺失值分布

缺失值处理方法一般分为两种，一种是直接删除含有缺失值的样本或缺失值较多的特征，另一种是通过各种插补法对缺失值进行填充[12]。由于本文数据量所限，不考虑删除样本或特征的，对缺失值采用插补法进行填充。常见的缺失值插补方法包括均值填充、中值填充、众数填充等统计方法或者是 KNN 填充、随机森林填充、多重插补等机器学习方法，考虑到追尾事故数据集的缺失值特征变量既有连续型又有离散型，且追尾事故数据样本之间具有一定的独立性，因此本文采用基于链式方程式（MICE）的多重插补来对混合型缺失数据进行插补。该方法可以与任意机器学习预测模型相结合进行插补数据，通过机器学习模型对缺失值进行预测填补能够在一定程度上克服样本间独立性，其更多的是基于特征间的相关性来进行插补，可以降低数据先验分布的影响。考虑到简便性与高效性，本文采用基于链式随机森林的多重插补算法（Miceforest）进行缺失值处理。图 2 所示为 Micefoest 插补原理示意图。

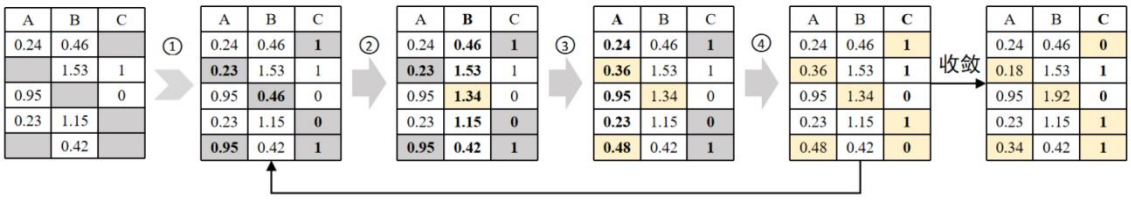

Figure 2. Schematic diagram of interpolation principle

图 2. 插补原理示意图

Miceforest 主要是通过随机森林分类模型与回归模型对缺失值进行链式插补，主要工作原理是首先通过随机采样对缺失值进行填充（步骤①所示），然后将缺失数据最少至最多的特征变量（图 2 所示依次为 B→A→C）依次作为标签，其他变量作为特征变量，通过训练随机森林分类和回归模型依次对缺失变量进行预测插（补步骤②、③、④所示），通过多次迭代步骤②~④，一般迭代 3~5 次即可使得特征间相关性收敛，此时完成一次插补。通过

n 次重复以上插补过程，可以获得 n 份插补数据集，一般采用均值的方法，对多个插补数据集进行合并得到最终插补数据，能够使插补数据与原始数据分布大致相同，图 3 展示了缺失率最大的两个特征变量的降速和 BMI 的插补效果，其中红色代表插补前数据分布，蓝色代表插补后数据分布，可以看出数据插补后数据分布与插补前数据分布大致相似，能够保证模型训练的数据可靠性。

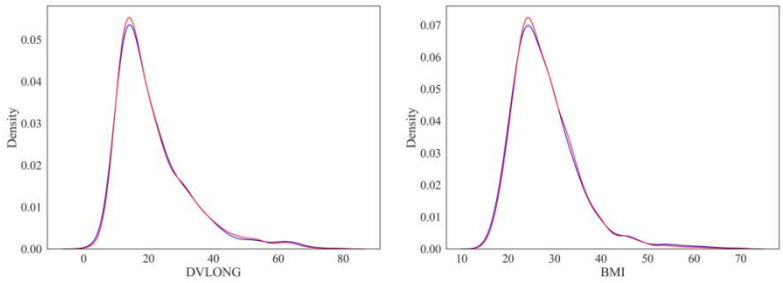

Figure 3. Distribution chart of data before and after interpolation

图 3. 插补前后数据分布图

2.2.3 预测标签

本文聚焦于追尾事故发生前驾驶人的损伤预测研究，为车辆在碰撞过程中调整车辆碰撞角度、驾驶舱约束系统等提供一定的参考建议，特别是当自车属于被追尾车辆时，驾驶员往往由于视线等原因并未注意到碰撞即将发生，来不及进行任何反应，此时安全预测系统可以对碰撞损伤进行预测从而调整自适应约束系统来减缓驾驶人损伤程度。本文考虑将驾驶人整体的损伤情况作为损伤模型预测标签，对驾驶人整体损伤进行预测，在车辆智能安全系统进行决策时能够较好的提供建议。

简明损伤分级标准（AIS）作为交通事故人员损伤评估标准，能够准确反映出在车祸事故中人员损伤程度，且 CISS 数据库记录了受伤人员最大损伤部位的 AIS 等级，因此本文提取出驾驶人的最大损伤部位的 AIS 等级作为损伤预测模型的标签，考虑到模型简化以及损伤等级类别不平衡性，将 AIS 为 0 的事故作为未受伤一类，AIS 为 1、2 等级的事故作为受轻伤的一类，AIS+3 等级的事故作为重伤及伤亡一类，将原来 0-6 七个类别划分为 3 个类别，能够简化驾驶人损伤预测模型以及提高预测精度。

2.3 样本不平衡处理

经处理后的乘用车追尾事故数据集样本分布依旧不均，从表 4 可以看出重伤及死亡的类别占比不到 5%，属于严重样本不平衡，这会导致训练出的损伤预测模型预测结果更加倾向于未受伤或受轻伤。驾驶员损伤预测模型可能将重伤或死亡预测成未受伤或受轻伤，这在碰撞发生前将是灾难性结果，导致智能车辆避撞决策失误，从而引发严重交通事故发生，这是不可接受的。因此为解决类别不均衡问题，本文采用 SMOTE 过采样的手段对重伤及伤亡类别进行合成数据提升其样本量，使最终训练数据集达到平衡。SMOTE 算法原理如图 4 所示，它主要通过在少数类别样本特征区间进行取值来合成新的少数样本，能够极大程度的保证新合成的少数类别样本的特征分布与原少数类别的特征分布相似，保证整体数据类别特征分布最小误差变化的前提下使数据重新达到类别平衡，确保模型不会以牺牲少数类别为代价而偏向多数类别。

SMOTE[13]过采样采样过程如下所示：1）通过欧氏距离计算少数类别样本两两距离，确定 K 近邻；2）根据不平衡比例确定采样倍率；3）从少数类样本x_i以采样倍率与其 K 近邻内的样本$\tilde{x}$根据公式 1 进行线性插值，作为新的少数样本。

$$x_{new} = x_i + rand(0,1) \times |x_i - \tilde{x}| \tag{1}$$

数据集经过采样处理前后分布如表 5 所示。

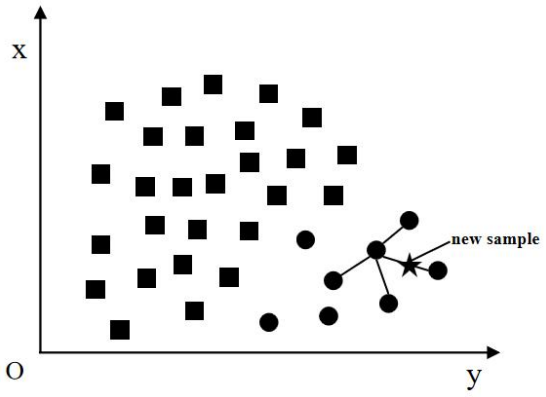

Figure 4. SMOTE algorithm diagram
图 4 SMOTE 算法原理图解

Table 5. Dataset label distribution
表 5. 数据集标签分布

标签	说明	原数据集样本数	采样后数据集样本数
0	未受伤	896	896
1	受轻伤	454	896
2	重伤或死亡	29	896

2.4 数据标准化处理

数据集中降速、年龄等特征变量量纲不一致，会导致模型在训练中会使某些量纲大的特征对预测结果的影响程度变大。因此在训练损伤预测模型前，需要对数据进行标准化处理来消除量纲的影响。本文采用去均值和方差归一化的方法来标准化数据，其公示如下所示：

$$x^* = \frac{x-\mu}{\sigma} \tag{2}$$

其中μ为某特征变量所有样本的均值，σ为某特征变量所有样本的标准差。

3 Stacking 集成学习训练预测模型

Stacking 集成学习模型作为使用在交通事故严重程度预测研究中使用广泛[14]，该模型主要由一层任意个数的基学习器、一层唯一元学习器堆叠组成，基学习器的输出结果作为元学习器的输入，因此基学习器的个数决定了元学习器的特征数量。这种堆叠式的集成方法通过结合多个基学习器模型的预测结果可以显著提高整体模型的准确性和泛化能力。可以集成多个不同的学习器模型，克服单个学习器在不同数据类型上表现的差异性，从而适应不同类型的数据和问题。

3.1 基学习器

Stacking 集成模型的基学习器的选择在原则上要求具有多样性[15]，尽可能多的选择表现好的不同类型的学习器，不同基学习器模型的优势和特点可以得到充分的发挥，从而提高整体模型的性能和泛化能力。本文初步选择多层感知机（MLP）、K 近邻（KNN）、决策树（DT）、随机森林（RF）、支持向量机（SVM）、轻量级梯度提升机（Lightgbm）、极度梯度提升树（XGBoost）、梯度提升决策树 （GBDT）、自适应提升树（AdaBoost）、类别型特征梯度提升算法（Catboost）共十种常见的分类学习器作为备选基学习器。

通过将数据集以 7：3 的比例划分训练集与测试集。采用 5 折交叉验证的方法对备选基学习器进行训练，考虑到样本不均衡性，以 F1 分数作为指标，选出其中 5 个 F1 分数较高的学习器作为本文损伤预测模型的基学习器。图 5 展示了各基学习器的 F1 分数得分情况，可以直观的看出随机森林、轻量级梯度提升机、极度梯度提升树、梯度提升决策树、类别型特征梯度提升算法这 5 类集成决策树的算法模型得分情况明显高于其他几类单一学习器，这说明了集成学习算法往往更适合于基于事故数据的驾驶员损伤预测模型的构建。

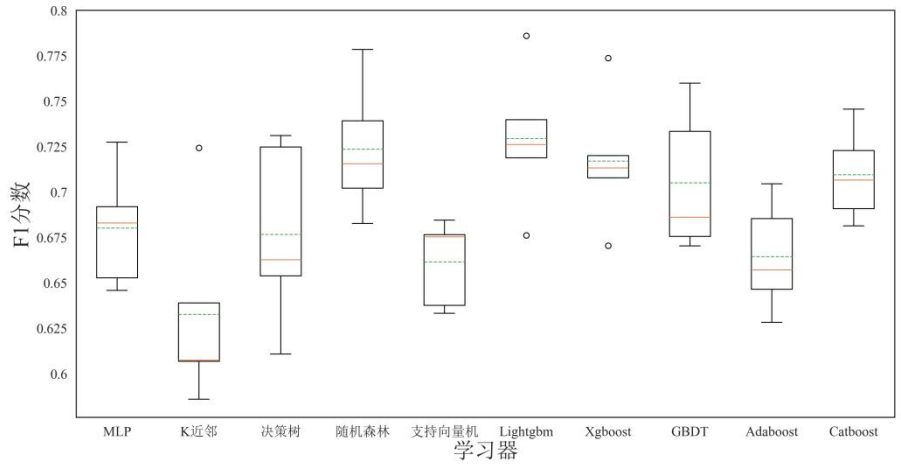

Figure 5. The score situation of each learner's F1 score
图 5. 各学习器 F1 分数的得分情况

3.2 元学习器

在 Stacking 集成学习中，元学习器的选择是至关重要的。元学习器的输入是各个基学习器的输出，因此选择合适元学习器能够消除基学习器的差异，使最终的预测模型效果到达更优[16]。在选择元学习器时应遵循以下几点原则：1）避免复杂模型和基学习器相同的模型，这可以减少过拟合的风险；2）计算高效，应尽可能选择能够快速训练与预测的模型；3）预测性能，具有较好的泛化性能与集成性能。

逻辑回归算法作为一种线性模型，其泛化性能好，模型简单，计算速度快且计算稳定，不易受到随机数据干扰。因此被广泛用于 Stacking 模型的元学习器。因此本文采用逻辑回归算法作为元学习器[17]。

3.3 模型评估

在机器学习中，混淆矩阵常常被用于检验机器学习分类模型在测试集上的预测性能。表 6 为二分类混淆矩阵示意图，TP 表示被正确分类的正样本数、FP 表示被错误分类的正样本数、FN 表示被错误分类的负样本数、TN 表示被正确分类的正样本数。

Table 6. confusion matrix
表 6. 混淆矩阵

	预测值	
真实值	正例	反例
正例	TP(True Positive)	FN(False Negative)
反例	FP(False Positive)	TN(True Negative)

机器学习常见的分类指标有准确率、精确率、召回率、F1 分数、受试者工作特征（Receiver Operating Characteristic, ROC）曲线等等。机器学习分类模型的评价指标有准确率,其计算公式如下所示，

机器学习常见的分类指标有准确率、精确率、召回率、F1 分数、受试者工作特征（Receiver Operating Characteristic, ROC）曲线等等。机器学习分类模型的评价指标有准确率,其计算公式如下所示，

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \tag{3}$$

准确率能够评价模型的整体性能、但当样本不均衡，少数类样本被分到多数类别中时，准确率的变化起伏不大，难以衡量模型对于少数类样本的预测性能，因此本文不以准确率作为模型的评价指标，而是选择精确率、召回率、F1 分数作为本文损伤预测模型的评价指标。精确率、召回率、F1 分数计算公式如下：

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{5}$$

$$\text{F1} = \frac{2\text{TP}}{2\text{TP+FP+FN}} \tag{6}$$

通过前文选择出随机森林、Lightgbm、Xgboost、GBDT、Catboost 五个模型作为集成学习损伤预测模型的基学习器。为进一步提升集成学习预测性能与泛化性能，需要对基学习器进行超参数寻优，一方面可以提高各基学习器的性能，另一方面通过参数调节可以防止基学习器过拟合。本文采用网格搜索的方法对各个基学习器的主要超参数进行寻优，表 7 展示了各基学习器的最优参数。通过各基学习器最优超参数重新训练各基学习器模型，使用双层堆叠的方法进行集成学习预测模型的构建。通过统计基学习器与集成预测模型在测试集上的准确率、精确率、召回率和 F1 分数来对模型进行综合性能的评价，表 8 展示了统计结果。结果显示，基于 Stacking 的集成学习损伤预测模型在各个指标上均优于其他单体基学习器，这表明基于 Stacking 的集成学习算法在乘用车追尾事故种构建驾驶员损伤预测模型比单一的机器学习模型更加可靠，证明了集成学习算法能够集成各学习器的在数据集上的性能，提高损伤预测的性能与泛化能力。基于事故数据的乘员损伤预测模型在选择特征输入时往往是基于事故宏观层面上进行选择，利用解释性强的机器学习模型构建乘员损伤预测模型，从而在事故级别的层面上来挖掘人员损害的影响因素，如 Chen 等[18]通过墨西哥州追尾事故数据集构建了基于决策树和朴素贝叶斯的混合驾驶员损伤预测模型，其在测试集上精度为 62.73%，难以满足汽车智能安全系统安全预测系统的精度要求，且其在特征上也未选择反映碰撞降速的变量，这可能也是其精度较低的原因之一。本文基于车辆碰撞层面的特征构建的 Stacking 集成损伤预测模型在各项指标上均达到了上达到了 82.8%以上，能够满足汽车智能安全对于延缓驾驶员损伤的决策的要求。stacking 模型的追尾事故驾驶员损伤预测能够在一定程度上为汽车智能安全系统的决策提供参考，但由于其数据收集的局限性，并未记录气囊起爆时间、主动预紧安全带的预紧时刻、预紧力等，难以对自适应约束系统的调整具体方案作出直接指导。对于此，已经有学者通过构建有限元仿真数据库，Bance 等[19]通过 Madymo 与 Lsdyna 等软件进行有限元仿真与多刚体仿真，通过设计车辆碰撞脉冲、乘员体征、车辆约束系统参数矩阵，构建了单一车型的常见车辆碰撞场景的乘员损伤数据库，基于该数据库提出了一种 AIS 等级精确损伤风险估计框架，能够快速的进行乘员损伤预测，但通过精细化的数字仿真构建碰撞乘员损伤数据库构建损伤预测模型，其预测结果极大的受到数字仿真的精度影响。

Table 7. Optimal hyperparameters for individual learners

表 7. 个体学习器最优超参数以来

基学习器	最优参数
RF	决策树数量 301，树最大深度为 10，叶子节点最少样本数 2
Lightgbm	树最大深度 3，叶子节点数 4
Xgboost	决策树数量 101，树最大深度为 4，学习率 0.7，
GBDT	决策树数量 343，树最大深度 9，学习率 0.15
Catboost	迭代次数 39，学习率 0.18，树最大深度 12

Table 8. Prediction results of individual learners and ensemble learners

表 8. 个体学习器及集成学习器预测结果

	Accuracy	Precision	Recall	F1
RF	80.67	80.37	80.66	80.27
Lightgbm	75.71	74.99	75.71	75.21
Xgboost	79.18	78.93	79.18	79.03
GBDT	82.78	82.57	82.78	82.62
Catboost	78.93	78.63	78.93	78.61
Stacking	**82.90**	**82.80**	**82.90**	**82.80**

4 总结

本文基于乘用车追尾事故数据提出了一种基于 Stacking 集成学习乘用车追尾事故中的驾驶员损伤预测模型。首先通过对 CISS 数据库进行数据清洗、预处理、特征工程等，得到可以用于直接训练损伤预测模型的完整数据

集。通过 F1 指标从十种各类机器学习模型中选择出随机森林、Lightgbm、Xgboost、GBDT、Catboost 五种机器学习分类器作为基学习器，以逻辑回归模型作为元学习器。通过网格搜索的方法对所有基学习器进行超参数寻优，在进行最终集成学习预测模型的构建，结果显示该预测模型精度可达到 82.9%，均高于其他单一学习器且精确率、召回率等其他性能指标也高于单一学习器。因此在追尾事故无法避免时，有希望通过该追尾事故驾驶人损伤预测模型为自适应约束系统的决策提供一定的建议，减缓驾驶人的损伤情况。但本文的研究还有以下不足：1）通过机器学习构建损伤预测模型受数据集的大小以及完整性影响严重，本文构建的损伤模型数据量不大，在未来可以使用多种来源的事故数据集，增加数据集的量级，提高模型的的泛化性、鲁棒性。2）本文仅仅构建了追尾事故中的损伤预测模型，在未来更应该关注在其他事故场景中的损伤预测，如侧碰、角碰等，构建适用于任意事故场景中的的损伤预测模型。3）本文仅针对事故中驾驶人的研究，未对事故中其他人员损伤预测模型的研究，在智能车辆安全中，完整的安全预测模块应包含驾驶人在内的所有乘员的损伤预测以及行人损伤预测，因此在后续研究中应该关注如何整合不同人员的损伤预测模型，以提升智能车辆安全预测模块的功能的完整性。

参考文献 (References)

[1] Kapileswar Rana., Narendra Khatri. Automotive intelligence: Unleashing the potential of AI beyond advance driver assisting system, a comprehensive review. Computers and Electrical Engineering. 2024. 117: 109237.

[2] Amoadu M., Ansah E W., Sarfo J O. Psychosocial work factors, road traffic accidents and risky driving behaviours in low-and middle-income countries: A scoping review. IATSS Research. 2023. 47(2): 240-250.

[3] Santos K., Dias J P., Amado C. A literature review of machine learning algorithms for crash injury severity prediction. Journal of Safety Research. 2022. 80: 254-269.

[4] YU R J., ABDEL-ATY M. Using hierarchical Bayesian binary probit models to analyze crash injury severity on high speed facilities with real-time traffic data. Accident Analysis & Prevention. 2014. 62: 161-167.

[5] ZHANG Y., LI Z B., LIU P., et al. Exploring contributing factors to crash injury severity at freeway diverge areas using ordered probit model. Procedia Engineering. 2011. 21: 178-185.

[6] Ghasedi M., Sarfjoo M., Bargegol I. Prediction and Analysis of the Severity and Number of Suburban Accidents Using Logit Model, Factor Analysis and Machine Learning: A case study in a developing country. SN Applied Sciences. 2021. 3: 13.

[7] Ejima S., Goto T., Zhang P., et al. Comparison of Injury Severity Prediction Using Selected Vehicles From Real-World Crash Data//27th International Technical Conference on the Enhanced Safety of Vehicles (ESV). National Highway Traffic Safety Administration. 2023. 23-0034.

[8] Lu Y., Kuang R. A Driver Injury Prediction Model based on Genetic Algorithm and BP Neural Network//2023 7th International Conference on Transportation Information and Safety (ICTIS). Xi'an, China: IEEE, 2023: 1984-1989.

[9] Zhang S., Khattak A., Matara C M., et al. Hybrid feature selection-based machine learning Classification system for the prediction of injury severity in single and multiple-vehicle accidents. PLoS ONE. 2022. 17(2): e0262941.

[10] NHTSA Research & Data. National Automotive Sampling Syatem(NASS) [EB/OL] . [2023-02-10]. https://www. nhtsa. gov/research-data/national-automotive-sampling-system-nass.

[11] RAGLAND C.，DALRYMPLE G. Overlap car-to-car tests compared to car-to-half barrier and car-to-full barrier tests. Auto & Traffic Safety. 1991. 1(2):213-224.

[12] 柏伟.交通事故数据缺失机理和插补策略研究.西南交通大学,2019.

[13] 王梦娇.道路交通事故严重性成因分析及预测研究.长安大学,2023.

[14] WOLPERT D H. Stacked generalization. Neural Networks. 1992. 5(2): 241-259.

[15] 单永航,张希,胡川,等.基于集成学习的交通事故严重程度预测研究与应用.计算机工程. 2024. 50(02):3 3-42.

[16] CUI S Z，YIN Y Q，WANG D J，et al. A stacking-based ensemble learning method for earthquake casualty prediction. Applied Soft Computing. 2021. 101:107038.

[17] SRIMANEEKARN N., HAYTER A., LIU W., et al. Binary response analysis using logistic regression in dentistry. International Journal of Dentistr. 2022. 22: 5358602.

[18] Chen C, Zhang G, Yang J, et al. An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier. Accident Analysis & Prevention, 2016, 90: 95-107.

[19] Bance I, Yang S, Zhou Q, et al. A framework for rapid on-board deterministic estimation of occupant injury risk in motor vehicle crashes with quantitative uncertainty evaluation. Sci China Technol Sci, 2021, 64: 521-534.