Explaining a Machine-Learning Lane Change Model With Maximum Entropy Shapley Values

Meng Li, Yulei Wang, Hengyang Sun, Zhihao Cui, Yanjun Huang, and Hong Chen, *Fellow, IEEE*

***Abstract*—Artificial intelligence (AI) techniques have been widely implemented in the domain of autonomous vehicles (AVs). However, existing AI techniques, such as deep learning and ensemble learning, have been criticized for their black-box nature. Explainable AI is an effective methodology to understand the black box and build public trust in AVs. In this article, a maximum entropy-based Shapley Additive exPlanation (SHAP) is proposed for explaining lane change (LC) decision. Specifically, we first build an LC decision model with high accuracy using eXtreme Gradient Boosting. Then, to explain the model, a modified SHAP method is proposed by introducing a maximum entropy base value. The core of this method is to determine the base value of the LC decision model using the maximum entropy principle, which provides an explanation more consistent with the human intuition. This is because it brings two properties: 1) maximum entropy has a clear physical meaning that quantifies a decision from chaos to certainty, and 2) the sum of the explanations is always isotropic and positive. Furthermore, we develop exhaustive statistical analysis and visualization to present intuitive explanations of the LC decision model. Based on the explanation results, we attribute the causes of predictions with wrong results to model defects or sample sparsity, which provides guidance to users for model optimization.**

***Index Terms*—Autonomous driving, Lane change model, Machine learning, eXtreme Gradient Boosting, Explainable AI, Shapley Additive exPlanations, Maximum entropy.**

I. Introduction

AUTONOMOUS vehicles (AVs) have developed rapidly in recent years with far-reaching impacts on improving driving safety, enhancing travel efficiency, and reducing pollution [1], [2], [3], [4]. In particular, lane change (LC) decision systems of AVs are safety-critical for minimizing the occurrence of traffic accidents [5]. Currently, most LC systems rely on rule-based approaches, which are easy to implement but often not robust and adaptive to complex traffic environments [6], [7]. In contrast, machine learning (ML)-based approaches utilize real driving data to design LC decision models and some related research has been conducted with better performance. Based on instantaneous data, the literature [8], [9] conducted research on LC decision modeling. Gao et al. [8] proposed a personalized LC model using a Resnet, which can selectively emphasize important features and suppress unuseful features with regularization methods, thus achieving significantly better performance than other popular networks. Mousa et al. [9] implemented an LC model using the eXtreme Gradient Boosting (XGBoost) algorithm [10], which shows better performance compared to Decision Trees (DT) [11], Gradient Boosting Decision Trees (GBDT) [12] and Random Forest (RF) [13]. Based on time-series data, Wang et al. [14] used fuzzy inference system and long short-term memory (LSTM) neural network to predict LC behaviors. The method simulates drivers' cognitive processes and transforms driving environments into LC feasibility. LSTM neural network is used to predict LC behavior based on feasibility and vehicle trajectory.

However, due to the inherent black-box nature of complex ML models, the deployment of an LC decision model in the real world has been largely criticized for its deficiencies in transparency, reliability, and trustworthiness [15].

Providing explanations for AVs is an effective way to understand black-box decisions and build public trust [16], [17]. Several studies have made efforts in this domain. Kim et al. [18] established an end-to-end explainable decision-making system for AVs by introducing an intrinsic explanation model. The explanations were presented in the form of saliency and short texts. Furthermore, Xu et al. [19] proposed a decision-making system for AVs based on "object-induced behavior". The core idea was to simplify complex environments and focus only on the induced objects that are liable to cause vehicle hazards and influence vehicle decisions. The explanatory information was also conveyed by texts. However, these approaches require extensive textual annotation of scenario-specific decisions with the aid of experts' prior knowledge, which can not cover endless realistic scenarios. Bojarski et al. [20] proposed an approach called VisualBackProp to explain the impact of detected scene images on steering operations. The core idea of the VisualBackProp is to capture critical elements of the scene images by backpropagation. However, the VisualBackProp does not have ideal properties of efficiency, symmetry, dummy player, and linearity. To solve this problem, Lundberg et al. [21] proposed an explanation framework called Shapley Additive exPlanation (SHAP) that built on the Shapley value [22] in cooperative game theory. It was demonstrated that the SHAP is the only explainable method satisfying the ideal property. Liessner et al. [23] first used the SHAP method to calculate

Manuscript received 8 March 2023; revised 26 March 2023; accepted 29 March 2023. Date of publication 11 April 2023; date of current version 20 July 2023. This work was supported by the National Key Research and Development Program of China under Grant 2020AAA0108101, in part by the National Natural Science Foundation of China under Grant U1964201, in part by Shanghai Municipal Science and Technology Commission under Grant 23ZR1467700, and in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100. *(Corresponding author: Yulei Wang.)*

Meng Li, Yulei Wang, and Hong Chen are with the Department of Control Science and Engineering, Tongji University, Shanghai 200092, China (e-mail: m15764337083@163.com; wangyulei@tongji.edu.cn; chenhong2019@tongji.edu.cn).

Hengyang Sun, Zhihao Cui, and Yanjun Huang are with the Clean Energy Automotive Engineering Center, Tongji University, Shanghai 200092, China (e-mail: 1852014@tongji.edu.cn; c_z_hao@tongji.edu.cn; yanjun_huang@tongji.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2023.3266196.

Digital Object Identifier 10.1109/TIV.2023.3266196

the feature contributions of the ML-based longitudinal control model of AVs, and simulation results demonstrated that the SHAP is beneficial for understanding longitudinal control. Cui et al. [24] proposed an explanation method combining SHAP and RF, which is applied to a DQN-based vehicle-following model. Specifically, SHAP is applied to analyze the feature importance of the model, and then the explainable RF algorithm performs imitative learning on the DQN model to achieve a more transparent decision. Nahata et al. [25] built a collision risk assessment model using RF, and further obtained the importance sequence of input features based on Tree SHAP, on which the predicted risk results were described in natural language using counterfactual inference. Wang et al. [26] proposed a framework to reveal the general properties of human drivers' perceptual uncertainty reduction when performing interaction tasks, and validated it in a merging lane scenario. Specifically, a deep learning algorithm was used to construct an intention model of the merging lane. Then the importance of input features that drivers attend to in different merging situations is obtained by SHAP, and finally, it was verified that human drivers usually intentionally seek information to reduce their perceptual uncertainty based on the sequential changes in the importance of input features. However, the explanations of LC models have not been investigated with the SHAP yet. In addition, how determining the base value of the SHAP to obtain explainable results that are more consistent with human intuition also needs to be addressed.

Motivated by the aforementioned studies, an explainable method called maximum entropy SHAP is developed and applied to the XGBoost-based LC decision model. First, we build an LC decision model with high accuracy using XGBoost. Then, to explain this model, we further propose the maximum entropy SHAP, the core of which is to determine the base value of SHAP by the maximum entropy method. Distinguishing from the literature [21], [23], [27], the maximum entropy SHAP has two properties: 1) maximum entropy has a clear physical meaning that quantifies a decision from chaos to certainty, and 2) the sum of the explanations is always isotropic and positive. Finally, this work provides a comprehensive explainable analysis for an LC decision model by local explanations, global summaries and feature dependencies, which are helpful for users to gain insight into the motivation of the model decisions.

The contribution is summarized in four points:

- A novel maximum entropy-based framework is proposed for selecting the base value of SHAP, which improves the performance of explanations.
- The proposed framework is consistent with human intuition by a clear physical meaning of maximum entropy, which describes the decision from chaos to certainty with an isotropic way.
- The proposed framework allows users to identify model defect or sample sparsity by analysing the feature contributions of incorrectly predicted samples.
- Our study provides a comprehensive explainable analysis for an LC decision model by local explanations, global summaries and feature dependencies.

The remaining sections of this article are organized as follows. Section II outlines the naturalistic driving data processing and XGBoost-based LC decision modeling. The primary emphasis of Section III is on the proposed maximum entropy SHAP approach. Section IV implements the XGBoost-based LC decision model and the maximum entropy SHAP, providing a comprehensive analysis of the findings. Finally, Section V presents the conclusions drawn from this study and outlines future research directions.

II. Naturalistic Driving Data Processing and XGBoost-Based LC Decision Modeling

As depicted in the upper half of Fig. 1, this section provides an overview of the naturalistic driving data processing method and the core principles of the XGBoost algorithm.

A. Naturalistic Driving Data Processing

The HighD dataset [28] was utilized due to its extensive use in previous research, (including but not limited to [29], [30], [31]), and rich data with correlations to LC decisions.

In order to analyze the LC and lane holding (LH) behavior of vehicles, the lateral positions of the vehicles with id (58) and (17) within the valid recording frame are analyzed as an example. For the vehicle with id (58), as shown in the curve between the red dot and the green dot of Fig. 2(a), the lateral position of the vehicle increases significantly when performing an LC, and it continuously increments throughout the LC process until it reaches the LC frame as shown by the green dot. The time (in this example, it is the input frame represented by the red dot) when the lateral position of the vehicle starts to change continuously in one direction is taken as the beginning time point of the vehicle's LC execution process. In this article, the instantaneous data at this frame is extracted as the LC decision data. Fig. 2(b) shows the trajectory of the LH vehicle with id (17). It can be seen that the vehicle starts to perform a continuous lateral movement from the red dot, however, no LC is executed. If this continuous lateral motion lasts for more than 20 frames (a frame lasts 40 ms), the instantaneous data at this frame will be extracted as LH decision data.

Feature selection is important for modeling LC decision. In fact, there is more than one way to determine the feature vector, and here we mainly refer to the literature [32] and [33]. Their works have pointed out that the main influence on the LC behavior of the host vehicle (HV) includes the front vehicle ($\mathrm{FV_C}$) in the current lane, the front vehicle ($\mathrm{FV_T}$) in the target lane, and the rear vehicle ($\mathrm{RV_T}$) in the target lane. Normally, the lateral velocity is not considered as the feature because it is small before the LC decision and can be ignored. In fact, the literature [32], [33] have selected the same features to LC decisions. Therefore, in this article, the feature vector is defined as $s = [v_0, v_1, v_2, v_3, d_1, d_2, d_3] \in \mathbb{S}$. As is shown in Fig. 3, v_0, v_1, v_2, and v_3 denote the velocities (unit: m/s) of HV, $\mathrm{FV_C}$, $\mathrm{RV_T}$, and $\mathrm{FV_T}$, respectively. d_1, d_2, and d_3 denote the longitudinal distances (unit: m) of HV relative to $\mathrm{FV_C}$, $\mathrm{RV_T}$, and $\mathrm{FV_T}$, respectively. The decision corresponding to a given feature vector is represented as $a \in \{a_{\mathrm{LC}}, a_{\mathrm{LH}}\} = \mathbb{A}$, where a_{LC} (a_{LH}) denotes the LC (LH) decision. Since the LC decision modeling is a classification problem, the decision needs to be coded with the category label $y \in \{\text{“0”}, \text{“1”}\} = \mathbb{Y}$, which is “0” (“1”) for a_{LH} (a_{LC}). Finally, 3,000 groups of data are screened, including 1,306 groups of LC data and 1, 694 groups of LH data.

B. XGBoost-Based LC Modeling

Suppose there are n samples, the features of the i-th sample are represented by s_i, and the label is y_i. $F(s_i)$ represents the current model's predicted value. The XGBoost loss function can

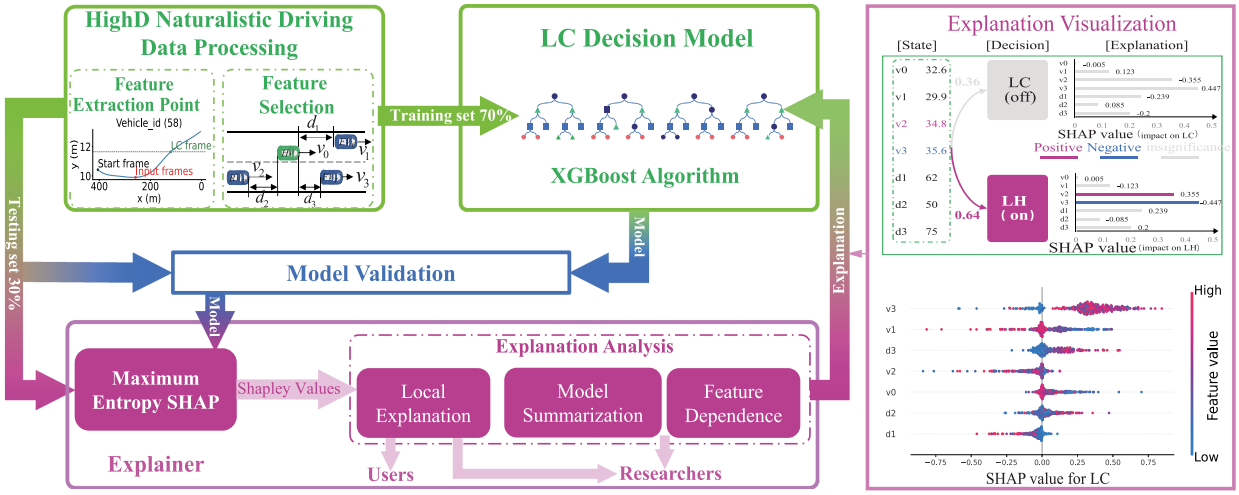


Fig. 1. The framework of explaining XGBoost-based LC decision model with maximum entropy SHAP. (1) Processing natural driving data for HighD: determination of feature extraction point and feature selection. (2) Decision modeling for LC using the XGBoost algorithm. (3) Maximum entropy SHAP design for explaining LC decision models in terms of local explanation, model summarization and feature dependence. (4) Visualisation of explanation results.

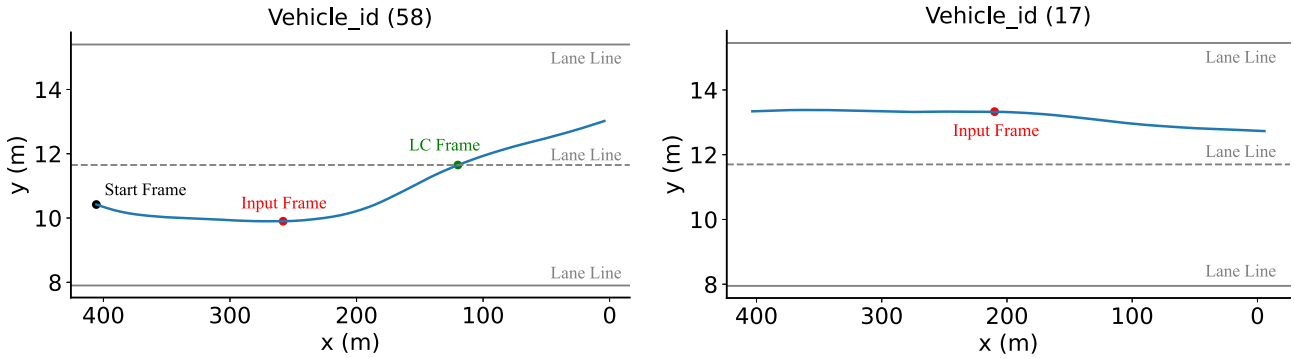


Fig. 2. Trajectories of vehicles. (a) Trajectory of vehicle No. 58 (LC) in valid recording frame. (b) Trajectory of vehicle No. 17 (LH) in valid recording frame.

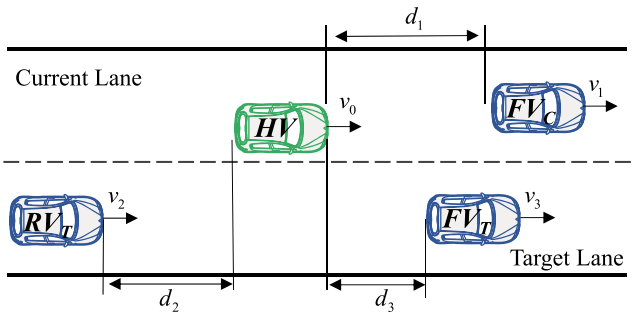


Fig. 3. Schematic of feature space.

be defined as:

$$L = \sum_{i=1}^{n} l(y_i, F(s_i)) + \sum_{k=1}^{K} (\Omega(f_k)), \tag{1}$$

where l is the loss function between the predicted value and the true value, K is the number of decision trees, and $\Omega(f_k)$ is the regularization term. To facilitate solution, we can expand the loss function using Taylor expansion:

$$L = \sum_{i=1}^{n} \left[l(y_i, F^{t-1}(s_i)) + g_i f_t(s_i) + \frac{1}{2} h_i f_t^2(s_i) \right] + \Omega(f_t), \tag{2}$$

where $F^{(t-1)}(s_i)$ represents the predicted value of the i-th sample in the $t-1$-th iteration, g_i and h_i represent the first and second derivatives of $l(y_i, F^{(t-1)}(s_i))$, respectively, $f_t(s_i)$ represents the predicted value of the t-th decision tree for the i-th sample, and $\Omega(f_t)$ represents the regularization term of the t-th decision tree. To construct the decision tree, XGBoost calculates the gain of the split to determine the splitting direction and position of each node using the following formula:

$$Gain = \frac{1}{2} \left[\frac{(G_{\mathrm{L}})^2}{H_{\mathrm{L}} + \lambda} + \frac{(G_{\mathrm{R}})^2}{H_{\mathrm{R}} + \lambda} - \frac{(G_{\mathrm{L}} + G_{\mathrm{R}})^2}{H_{\mathrm{L}} + H_{\mathrm{R}} + \lambda} \right] - \gamma \tag{3}$$

where G_{L} and H_{L} represent the sum of the first and second derivatives of the left subtree, G_{R} and H_{R} represent the sum of the first and second derivatives of the right subtree, λ is the regularization coefficient, and γ is a pre-set value used to control the number of leaf nodes for splitting. Finally, the greedy algorithm is used to find the optimal splitting point.

Although the LC model can be constructed using XGBoost, its decision lacks transparency, impeding its implementation in safety-critical AVs. Following, we will introduce the maximum entropy SHAP method for insight into the LC decision model as demonstrated in the lower half of Fig. 1.

Remark 1: Compared with instantaneous data, time-series data can further improve the accuracy of LC decision models with the price of increasing features. For example, if an instantaneous model has 7 features, then its time-series version with a time window of 10 has to consider 70 features, which are difficult for SHAP explanations. To be honest, high-dimensional explanation considering time-series data is still in its infancy and need future research.

III. Proposed Maximum Entropy SHAP Method

This section begins by reviewing the Shapley value and then proposes the maximum entropy SHAP method for explaining the XGBoost-based LC decision model introduced in Section II.

A. Shapley Value

Shapley value is a classical concept in cooperative game theory, which is used to distribute the total gains to each player in a coalition. It is a fair distribution because it uniquely has ideal properties of efficiency, symmetry, dummy player, and linearity [34].

To obtain the Shapley value of i-th feature, it is necessary to calculate the marginal contribution. It can be denoted as $v(S\cup\{i\})-v(S)$, which is the difference between the mapping values of value function v with and without i-th feature on all feature subsets $S\subseteq M$, where M is the set of all features. Then, the Shapley value of i-th feature is computed by weighting all marginal contributions:

$$\phi_i=\sum_{S\subseteq M\backslash\{i\}}\underbrace{\rho(m,|S|)}_{\text{weight factor}}\underbrace{(v(S\cup\{i\})-v(S))}_{\text{marginal contribution}}, \tag{4}$$

where $\rho(m,|S|)=\frac{|\mathcal{S}|!(m-|\mathcal{S}|-1)!}{m!}$, m is the total number of all features, $|S|$ is the number of features in S, $i=1,\ldots,m$.

To obtain ϕ_i, it is necessary to compute $v(\mathcal{S})$ without knowing all feature values. To do this, Janzing et al. [35] introduces the do-operator [36], which uses the model's intervention conditional expectation approximating the value function:

$$v(S)=E\left[F\left(X_S,X_{\bar{S}}\right)\mid \text{ do }\left(X_S=x_S\right)\right], \tag{5}$$

where X_S denotes the set of the observed features, $X_{\bar{S}}$ denotes the set of unknown features and x_S denotes the real values of observed features. In the case of feature independence, the intervening conditional expectation is equivalent to the marginal expectation, which is consistent with the SHAP [21]. Note that, in general, features from the background dataset are used to replace the corresponding unknown features in $X_{\bar{S}}$. However, arbitrary selection of replaced features (also called base value) leads to a situation that is inconsistent with human intuition, i.e. the sum of feature contributions $\sum_{i=1}^{M}\phi_{i,a}$ for probability function $F_a(s)$ corresponding to the classification result a in the current state s may have negative values:

$$\sum_{i=1}^{m}\phi_{i,a^*}=F_{a^*}(s)-\phi_{0,a^*}<0 \tag{6}$$

$$s.t.\ a^*=\arg\max_{a\in\mathbb{A}}F_a(s), \tag{6.a}$$

with $\phi_{0,a}=E[F_a(s)]$ the expectation of the original probability function. To address the above issues, a maximum entropy SHAP method is proposed in the next subsection to determine the appropriate base value.

B. Maximum Entropy SHAP

The SHAP is an explanation method based on feature attribution, which is to find a simple explanation model to approximate the original complex model. The original model can be expressed as the sum of the output expectation and the Shapley values of all the feature values. Based on the SHAP, we propose the maximum entropy SHAP method for the XGBoost-based LC model, that is

$$F_a\approx\phi_{0,a}+\sum_{i=1}^{m}\phi_{i,a} \tag{7}$$

$$\phi_{i,a}=\begin{cases}F_a\left(s^*\right) & i=0\\ \sum\limits_{S\subseteq M\backslash\{i\}}\rho(m,|S|)(v_a(S\cup\{i\})-v_a(S)) & i\neq 0\end{cases} \tag{8}$$

$$v_a(S)=E\left[F_a\left(\left(X_S,X_{\bar{S}}^*\right)\right)\mid do\left(X_S=x_S\right)\right], \tag{9}$$

where $X_{\bar{S}}^*$ denotes the subset of feature values of the maximum entropy state s^*, and s^* is derived by solving the maximum entropy problem with constraints:

$$s^*=\arg\max_{s\in\mathbb{S}}\left(\sum_{a\in\mathbb{A}}(-F_a(s)\log F_a(s))\right) \tag{10}$$

$$s.t.\ F_{a_{\text{LC}}}(s)\geq F_{a_{\text{LH}}}(s) \tag{10.a}$$

$$\sum_{a\in\mathbb{A}}F_a(s)=1, \tag{10.b}$$

where constraint (10.a) indicates that the maximum entropy state is selected from the LC samples, which is due to the LC behavior is of more interest than LH. Constraint (10.b) is the sum rule of probability. Since the maximum entropy function represented by (10) is convex [37] and the decision a is binary variable, a property of the maximum entropy SHAP is derived:

$$\begin{aligned}\sum_{j=1}^{M}\phi_{i,a^*}&=F_{a^*}(s)-\phi_{0,a^*}\\&=F_{a^*}(s)-F_{a^*}\left(s^*\right)\geq 0,\end{aligned} \tag{11}$$

which satisfies the isotropy because the probability of decision always has a non-negative contribution with respect to the base value and thus follows human intuition where scores is always positive relative to zero in a similar scoring system. Then, the probability of all classifications of the model at the maximum entropy state is

$$F_a\left(s^*\right)=\frac{1}{|\mathbb{A}|},\forall\, a\in\mathbb{A}. \tag{12}$$

However, in practice, (12) is usually not satisfied. Therefore, $\phi_{0,a}$ is a set of sequences corresponding to different predicted action a at the maximum entropy state s^*.

IV. Experiments

This section implements the XGBoost-based LC decision model and then uses the proposed maximum entropy SHAP

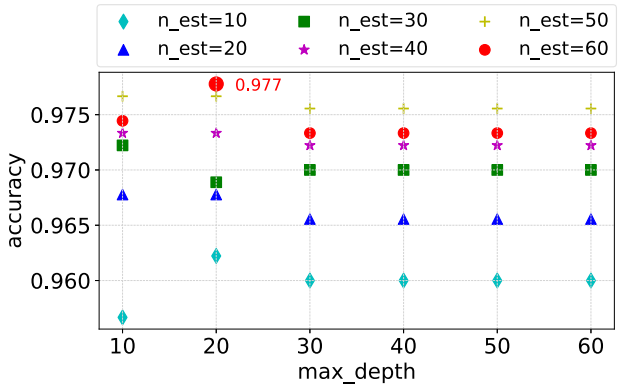


Fig. 4. Accuracy of XGBoost-based LC decision model with different parameters.

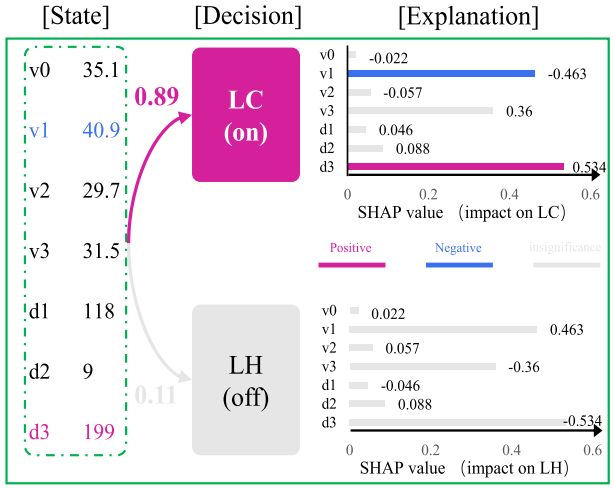


Fig. 5. Local explanation for an LC sample using a human-machine interface.

to explain the model. Finally, a comprehensive explanation of the LC decision model is performed in light of the simulation results.

A. The Results of Model Prediction

To obtain the reliable XGBoost-based LC decision model, the grid search method is used to optimize the key parameters including the maximum tree depth (max_depth) and the number of estimators (n_est) in a finite parameter space that relies on empirical design. Other parameters are set as follows: the weight of $L1$ regularization term $\gamma = 0$, the weight of $L2$ regularization term $\lambda = 0.5$. Fig. 4 shows the prediction accuracy of the XGBoost-based LC model during the testing with different parameters, where the red solid point achieves a local maximum value of 97.7% at $n\_est = 60$ and $max\_depth = 20$ in our experiments.

B. Model Explanation

This section implements the maximum entropy SHAP to explain the LC model. The maximum entropy base value is calculated:

$$\begin{cases} s^* = [32.6\,\text{m/s}, 33.3\,\text{m/s}, 26.6\,\text{m/s}, 23.9\,\text{m/s}, \\ \qquad 25.1\,\text{m}, 30.0\,\text{m}, 7.8\,\text{m}] \\ \Rightarrow \phi_{0,a_{\text{LH}}} = F_{a_{\text{LH}}}(s^*) = 0.49 \\ \Rightarrow \phi_{0,a_{\text{LC}}} = F_{a_{\text{LC}}}(s^*) = 0.51. \end{cases} \tag{13}$$

1) Local Explanation: Local explanation is reflected as the reason for the difference between the predicted value of a given sample and the base value ϕ_0, as expressed in (7). For the LC sample shown in Fig. 5, since its probability of LC is 0.89 > 0.5, the resulting decision is LC, in which $v_1 = 40.9$ m/s and $d_3 = 199$ m are the most critical features contributing -0.463 and 0.534, respectively. The intuition is that since $v_1 = 40.9$ m/s is large, there is usually no need for the host vehicle to perform LC to increase its velocity, which leads to a large negative contribution to LC. $d_3 = 199$ m implies that the host vehicle has sufficient space for a safe LC, which contributes to the LC by 0.534. These results are in close agreement with the human driver's intuition for LC decisions in highway scenarios. In addition, a human-machine interface design in Fig. 5 can facilitate the human user's timely understanding of how the AI model drives.

To analyse the explanation of the differences between mandatory and discretionary LC in reality, our work filters the "mandatory LC" data according to the Time to Collision (TTC) defined as TTC $= d_1/(v_0 - v_1)$ with a threshold of 5. A sample of LC with a TTC of 4.47 s is shown in Fig. 6(a), where $v_1 = 7.04$ m/s and $v_0 = 10.09$ m/s contribute the most to the probability of LC, reflecting that the low velocity of the front vehicle in current lane and the relatively fast velocity of the host vehicle are the key reasons for the mandatory LC. In contrast, an LC sample with TTC of 29.01 s is shown in Fig. 6(b), where $v_3 = 31.09$ m/s and $d_3 = 189.15$m contribute the most to the LC probability, reflecting that the high velocity of the front vehicle in the target lane and the large relative distance d_3 are the decisive factors leading to the discretionary LC in this sample. These findings suggest that mandatory LCs prioritize features that increase the likelihood of avoiding an imminent collision, whereas discretionary LCs prioritize features that enhance overall driving efficiency.

More interestingly, local explanations can help users analyze samples of misprediction and attribute incorrect predictions to model defects or sample sparsity. As shown in Table I, for sample 1 of LC, its probability is $0.35 < 0.5$, which leads to a wrong prediction. It can be seen that the distance to the front vehicle in the target lane is $d_3 = 120.6$ m, corresponding to $\phi_7 = 0.21$, which is consistent with the intuition that a large distance increases the probability of LC for the host vehicle. For $v_2 = 24.5$ m/s, the intuition is that a small velocity of the rear vehicle makes it safer to change lanes, thus increasing the LC probability. However, it is inconsistent with the result that its large negative contribution to LC. This inconsistency indicates the limitation of the XGBoost-based LC model to characterize this class of sample features. We attribute this type of misprediction to a defect in the model itself. For sample 2 of LC, $v_1 = 43.8$ m/s makes a large negative contribution $\phi_2 = -0.36$, which is consistent with the intuition that the host vehicle generally does not change lanes when the velocity of the front vehicle in the current lane is large. This result illustrates the sparsity of the LC behavior of this sample i.e., a minority would make similar LC choices. We attribute this type of misprediction to the sparsity of the sample. For the LH sample, the most important feature $d_3 = 152$ m contributes $\phi_7 = -0.49$ to LH. which is consistent with the intuition that the host vehicle is

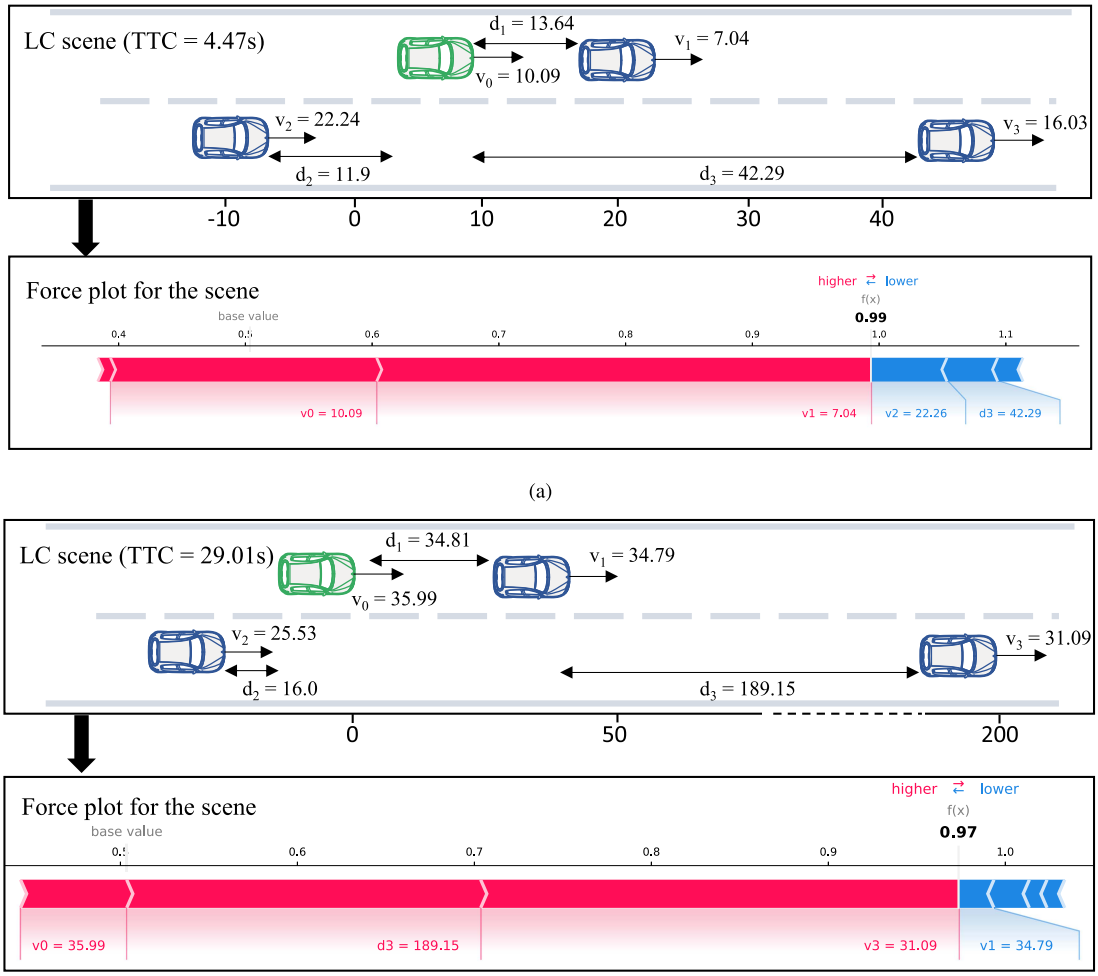
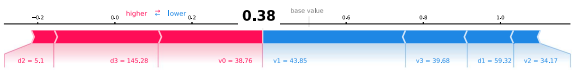
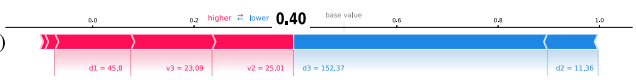


Fig. 6. Local explanation of mandatory and discretionary LCs. (a) Local explanation of a mandatory LC. (b) Local explanation of a discretionary LC.

TABLE I
ANALYSIS OF INCORRECTLY PREDICTED SAMPLES (HERE DENOTED MODEL DEFECT AS ♢ AND DENOTED SAMPLE SPARSITY AS △)

Sample	Force plot of local explanation	Shapley values							Model check
		ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	
LC (0.35)		0.07	0.02	-0.20	-0.12	-0.10	-0.03	0.21	♢
LC (0.38)		0.26	-0.36	-0.07	-0.16	-0.12	0.06	0.27	△
LH (0.40)		0.01	0.02	0.16	0.15	0.15	-0.10	-0.49	♢

more likely to change lanes when the distance is large between the front vehicle in the target lane and the host vehicle. Therefore, we attribute it to a model defect. Overall, the analysis and attribution of incorrect predictions provide guidance for users and researchers to optimize the model.

2) Model Summarization: Model Summarization provides insight into the general impact of features on prediction by showing the distribution of SHAP values for all features and the corresponding trends. The x-axis shown in Fig. 7 represents specific SHAP values, and the y-axis represents the categories of features in order of importance. The dots indicate all samples, and the feature values are indicated from small to large by the corresponding blue to red color. It can be seen that for the LC case shown in Fig. 7, in general, the feature values of v_3, d_3, and d_2 and their corresponding SHAP values are approximately positively correlated. The intuition of these phenomena is reflected

TABLE II
COMPARISON BETWEEN OUR PROPOSED MAXIMUM ENTROPY SHAP (HERE DENOTED AS ME-SHAP) AND THE ORIGINAL SHAP (HERE DENOTED AS SHAP)

Sample	Feature values							Method	Base	Shapley values							
	v_0	v_1	v_2	v_3	d_1	d_2	d_3		ϕ_0	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	$\sum_{i=1}^{7}\phi_i$
LC (0.86)	32.9	28.0	31.2	32.1	31.7	19.0	15.6	SHAP	**0.35**	0.06	0.17	0.15	0.15	0.09	-0.07	-0.04	**0.51**
								ME-SHAP	**0.51**	-0.02	0.12	-0.07	**0.56**	-0.13	-0.06	-0.05	**0.35**
LH (0.54)	25.4	22.8	34.9	32.1	25.2	90.6	28.5	SHAP	**0.65**	0.00	0.10	0.02	-0.04	-0.03	-0.20	0.06	**-0.11**
								ME-SHAP	**0.49**	-0.02	0.06	**0.34**	-0.15	0.00	-0.16	-0.01	**0.05**
LH (0.51)	31.1	35.2	20.6	31.1	29.6	92.6	95.8	SHAP	**0.65**	0.01	0.02	0.23	-0.08	-0.05	-0.19	-0.08	**-0.14**
								ME-SHAP	**0.49**	-0.10	0.12	**0.84**	-0.56	0.07	-0.20	-0.15	**0.02**

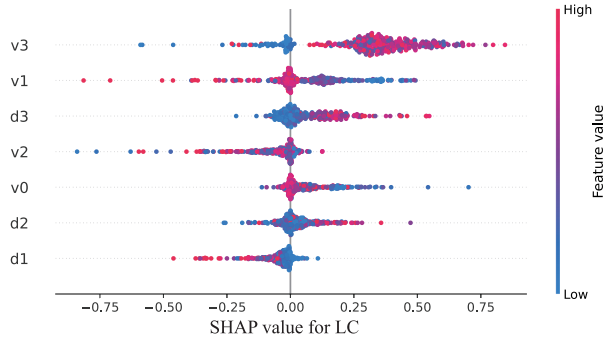


Fig. 7. Model Summarization for LC samples.

as follows respectively: 1) the greater the velocity of the front vehicle in the target lane, the higher driving efficiency the host vehicle may obtain after changing lanes; 2) the large distance between the host vehicle and the front vehicle in the target lane can ensure there is enough space for the host vehicle to perform a safe LC; 3) the large distance between the host vehicle from the rear vehicle in the target lane can support the host vehicle to complete the LC. In comparison, the feature values of v_1, v_0, and d_1, are approximately inversely proportional to their corresponding SHAP values. These phenomena reflect the following intuitions: 1) the small velocity of the front vehicle means that the driving efficiency of the current lane is low, so it is easier to push the host vehicle to change lanes for high driving efficiency; 2) the smaller the velocity of the host vehicle, the more inclined it is to change lanes to avoid the danger of high-velocity LC; 3) the small relative distance between the host vehicle and the front vehicle is likely to lead to collisions, thus the host vehicle is more inclined to change lanes to improve driving efficiency without collisions. In general, the model summarization provides developers with the overall law of the model.

3) Feature Dependence: Feature dependence, as the complement to the model summarization, can reveal how the SHAP values of the interested features vary with other variables. For all LC instances, as shown in Fig. 8(b), the x-axis in the dependency plot portrays the values of the interested features, and the y-axis on the left side represents its SHAP values under the effect of the dependent variables, whose value is represented by the color histogram on the right side, with small to large corresponding to a color change from blue to red. Taking the front vehicle in the current lane as an example, the effect of d_1 on the distribution of SHAP values for velocity v_1 is shown in Fig. 8. From three scenes shown in Fig. 8(a), we can observe a regularity: when the velocity of front vehicle is approximately less than 20 m/s and the relative distance between the host vehicle and the front vehicle is small, the velocity's SHAP value of the front vehicle is larger. This regularity of the samples is represented by blue dots in the blue box in feature dependency plot as shown in Fig. 8(b). This observation highlights that the model has learned the rule that a sufficiently low velocity of front vehicle and a small relative distance between the host vehicle and front vehicle are more likely to result in an LC, which aligns with human intuition.

To verify the superiority of the proposed maximum entropy SHAP, this work compares it with the original SHAP. Table II shows the SHAP values calculated from origin SHAP and maximum entropy SHAP for each of the three samples (containing one LC sample and two LH samples), where the sum $\sum_{i=1}^{7}\phi_i$ of SHAP values of features always remains positive in both LC and LH decisions using the maximum entropy SHAP. However, the original SHAP has either positive or negative value on the sum $\sum_{i=1}^{7}\phi_i$ of feature contributions of a decision, which causes a lack of isotropy and confuses human intuition that scores is always positive relative to zero in a similar scoring system. Moreover, compared to the original SHAP method, the results of using maximum entropy SHAP show that the SHAP value of a feature is significantly greater than the SHAP values of other features as depicted in ϕ_i with red font. This is beneficial for identifying the key factors in model decisions. In addition, the point of maximum entropy corresponds to the smallest decision probability (0.51 for LC), which implies the highest uncertainty in the tested samples. In comparison, the decision probability of the other samples are larger than the base value, which reflects a decision process from chaos to certainty.

V. CONCLUSION

We have proposed a novel framework for selecting the base value of SHAP, which is based on the principle of maximum entropy. The explanation provided by this framework is more consistent with human intuition because maximum entropy describes the uncertainty of a decision and the sum of the contributions of the features is isotropic. By comparison with the original SHAP method, the advantages of the maximum entropy SHAP are validated in our experiments. Furthermore, we perform a thorough visual analysis of the LC decision model to help users gain insight into the motivation of the LC decisions. In addition, our work attributes the causes of predictions with wrong results to model defects or sample sparsity, which provides guidance for users to optimise their models.

Explainable AI research in AVs is still in its infancy. There are still several challenges that need to be addressed to truly help the users to understand the decision-making process, one of which is to provide an explanation of the impact of features on decision in the spatio-temporal dimension. Future research

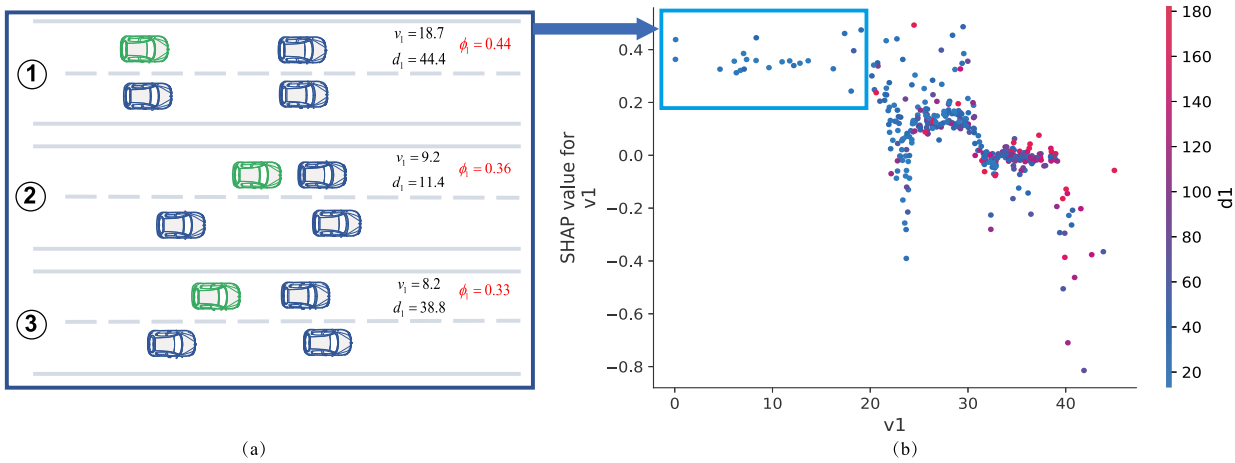


Fig. 8. Feature dependence. (a) Three scenes of LC samples. (b) The feature dependency of v_1 and d_1 (shows the influence of d_1 on the distribution of SHAP value of v_1).

will use the Shapley flow to reveal the spatio-temporal flow of feature importance. In addition, causal inference has a powerful ability to build causal maps between features. Therefore, for a time-series model, integrating causal inference with the Shapley flow can provide improved explanations to human operators.

REFERENCES

[1] Z. Wang et al., "Driver behavior modeling using game engine and real vehicle: A learning-based approach," *IEEE Trans. Intell. Veh.*, vol. 5, no. 4, pp. 738–749, Dec. 2020.

[2] G. Gunter, C. Janssen, W. Barbour, R. E. Stern, and D. B. Work, "Model-based string stability of adaptive cruise control systems using field data," *IEEE Trans. Intell. Veh.*, vol. 5, no. 1, pp. 90–99, Mar. 2020.

[3] C. E. Tuncali, G. Fainekos, D. Prokhorov, H. Ito, and J. Kapinski, "Requirements-driven test generation for autonomous vehicles with machine learning components," *IEEE Trans. Intell. Veh.*, vol. 5, no. 2, pp. 265–280, Jun. 2019.

[4] Y. Huang, J. Du, Z. Yang, Z. Zhou, L. Zhang, and H. Chen, "A survey on trajectory-prediction methods for autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 652–674, Sep. 2022, doi: 10.1109/TIV.2022.3167103.

[5] C. Vallon, Z. Ercan, A. Carvalho, and F. Borrelli, "A machine learning approach for personalized autonomous lane change initiation and control," in *Proc. IEEE Intell. Veh. Symp.*, 2017, pp. 1590–1595.

[6] X. Xu, L. Zuo, X. Li, L. Qian, J. Ren, and Z. Sun, "A reinforcement learning approach to autonomous decision making of intelligent vehicles on highways," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 50, no. 10, pp. 3884–3897, Oct. 2018.

[7] J. Chen, S. E. Li, and M. Tomizuka, "Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5068–5078, Jun. 2022.

[8] J. Gao, H. Zhu, and Y. L. Murphey, "A personalized model for driver lane-changing behavior prediction using deep neural network," in *Proc. IEEE 2nd Int. Conf. Artif. Intell. Big Data*, 2019, pp. 90–96.

[9] S. R. Mousa, P. R. Bakhit, O. A. Osman, and S. Ishak, "A comparative analysis of tree-based ensemble methods for detecting imminent lane change maneuvers in connected vehicle environments," *Transp. Res. Rec.*, vol. 2672, no. 42, pp. 268–279, 2018.

[10] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.

[11] S. B. Kotsiantis, "Decision trees: A recent overview," *Artif. Intell. Rev.*, vol. 39, no. 4, pp. 261–283, 2013.

[12] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front. Neurorobot.*, vol. 7, 2013, Art. no. 21.

[13] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogrammetry Remote Sens.*, vol. 114, pp. 24–31, 2016.

[14] W. Wang, T. Qie, C. Yang, W. Liu, C. Xiang, and K. Huang, "An intelligent lane-changing behavior prediction and decision-making strategy for an autonomous vehicle," *IEEE Trans. Ind. Electron.*, vol. 69, no. 3, pp. 2927–2937, Mar. 2022.

[15] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," *Int. J. Comput. Vis.*, vol. 130, pp. 2425–2452, 2022.

[16] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10142–10162, Aug. 2022.

[17] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions," 2021. [Online]. Available: https://arxiv.org/abs/2112.11561

[18] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 563–578.

[19] Y. Xu et al., "Explainable object-induced action decision for autonomous vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9523–9532.

[20] M. Bojarski et al., "Visualbackprop: Efficient visualization of CNNs for autonomous driving," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 4701–4708.

[21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 4765–4774.

[22] L. S. Shapley, "A value for n-person games," *Classics Game Theory*, vol. 69, pp. 69–77, 1997.

[23] R. Liessner, J. Dohmen, and M. A. Wiering, "Explainable reinforcement learning for longitudinal control," in *Proc. 13th Int. Conf. Agents Artif. Intell.*, 2021, pp. 874–881.

[24] Z. Cui, M. Li, Y. Huang, Y. Wang, and H. Chen, "An interpretation framework for autonomous vehicles decision-making via SHAP and RF," in *Proc. IEEE 6th CAA Int. Conf. Veh. Control Intell.*, 2022, pp. 1–7.

[25] R. Nahata, D. Omeiza, R. Howard, and L. Kunze, "Assessing and explaining collision risk in dynamic environments for autonomous driving safety," in *Proc. IEEE Int. Intell. Transp. Syst. Conf.*, 2021, pp. 223–230.

[26] H. Wang, H. Liu, W. Wang, and L. Sun, "On trustworthy decision-making process of human drivers from the view of perceptual uncertainty reduction," 2022. [Online]. Available: https://arxiv.org/abs/2210.08256

[27] S. M. Lundberg et al., "From local explanations to global understanding with explainable ai for trees," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020.

[28] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highd dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems," in *Proc. IEEE 21st Int. Conf. Intell. Transp. Syst.*, 2018, pp. 2118–2125.

[29] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet, and F. Nashashibi, "Attention based vehicle trajectory prediction," *IEEE Trans. Intell. Veh.*, vol. 6, no. 1, pp. 175–185, Mar. 2021.

[30] S. Mozaffari, E. Arnold, M. Dianati, and S. Fallah, "Early lane change prediction for automated driving systems using multi-task attention-based convolutional neural networks," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 758–770, Sep. 2022.
[31] X. Tang et al., "Prediction-uncertainty-aware decision-making for autonomous vehicles," *IEEE Trans. Intell. Veh.*, vol. 7, no. 4, pp. 849–862, Dec. 2022.
[32] D.-F. Xie, Z.-Z. Fang, B. Jia, and Z. He, "A data-driven lane-changing model based on deep learning," *Transp. Res. Part C: Emerg. Technol.*, vol. 106, pp. 41–60, 2019.
[33] X. Gu, J. Yu, Y. Han, M. Han, and L. Wei, "Vehicle lane change decision model based on random forest," in *Proc. IEEE Int. Conf. Power Intell. Comput. Syst.*, 2019, pp. 115–120.
[34] A. E. Roth, *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge, U.K.: Cambridge Univ. Press, 1988.
[35] D. Janzing, L. Minorics, and P. Blöbaum, "Feature relevance quantification in explainable AI: A causal problem," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 2907–2916.
[36] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[37] A. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

Meng Li received the B.S. degree from the College of Electronic and Information Engineering, Changchun University of Science and Technology, Changchun, China, in 2017, and the M.S. degree in Control Science and Engineering from Jilin University, Jilin, China, in 2020. He is currently working toward the Ph.D. degree with the Department of Control Science and Engineering, Tongji University, Shanghai, China. His research interests include intelligent driving, artificial intelligence, explainable AI and autonomous vehicle.

Yulei Wang received the Ph.D. degree in control science and engineering from the Harbin Institute of Technology, Harbin, China, in 2013. From 2013 to 2020, he was an Associate Professor with the College of Communication Engineering, Jilin University, Jilin, China. He is currently an Research Fellow with the Department of Control Science and Engineering, Tongji University, Shanghai. His main research interests include autonomous driving, artificial intelligence, vehicle control engineering and intelligent transportation systems.

Hengyang Sun is currently working toward the B.S. degree with the School of Automotive Studies, Tongji University, Shanghai, China. His research interests include machine learning, deep learning, explainable AI and autonomous vehicle.

Zhihao Cui received the B.S. degree from the School of Vehicle Engineering, Chongqing University of Technology, China, Chongqing, in 2021. He is currently working toward the M.S. degree with the School of Automotive Studies, Tongji University, Shanghai, China. His research interests include machine learning, deep learning, explainable AI, and autonomous vehicle.

Yanjun Huang is currently a Professor at School of Automotive studies, Tongji University, Shanghai, China. He received the Ph.D. degree in 2016 from the Department of MME, University of Waterloo, Waterloo, On, Canada. His research interest include the improving vehicle performance in terms of safety, energy-saving, and intelligence by using advanced control and learning methods. He has authored or coauthoreds several books, more than 60 papers in journals and conference. He was the recipient of IEEE VTS 2019 Best Land Transportation Paper Award and 2018 Best paper of Automotive Innovation. He is AE or EBM of IET Intelligent Transport System, SAE International Journal of Commercial vehicles, International Journal of Autonomous Vehicle System.

Hong Chen (Fellow, IEEE) received the B.S. and M.S. degrees in process control from Zhejiang University, China, in 1983 and 1986, respectively, and the Ph.D. degree in system dynamics and control engineering from the University of Stuttgart, Sttugart, Germany, in 1997. In 1986, she joined the Jilin University of Technology, Jilin, China. From 1993 to 1997, she was a Wissenschaftlicher Mitarbeiter with the Institut fuer Systemdynamik und Regelungstechnik, University of Stuttgart. Since 1999, she has been a Professor a with Jilin University and hereafter a Tang Aoqing Professor. Recently, she joined Tongji University as a distinguished Professor. Her research interests include model predictive control, nonlinear control, artificial intelligence and applications in mechatronic systems such as automotive systems.