Surrogate Deep Learning to Estimate Uncertainties for Driver Intention Recognition

Koen Vellenga
University of Skövde
& Volvo Car Corporation
Sweden

Alexander Karlsson
University of Skövde
Sweden

H. Joe Steinhauer
University of Skövde
Sweden

Göran Falkman
University of Skövde
Sweden

Anders Sjögren
Volvo Car Corporation
Sweden

ABSTRACT

Real-world applications of artificial intelligence that can potentially harm human beings should be able to express uncertainty about the made predictions. Probabilistic deep learning (DL) methods (e.g., variational inference [VI], VI last layer [VI-LL], Monte-Carlo [MC] dropout, stochastic weight averaging - Gaussian [SWA-G], and deep ensembles) can produce a predictive uncertainty but require expensive MC sampling techniques. Therefore, we evaluated if the probabilistic DL methods are uncertain when making incorrect predictions for an open-source driver intention recognition dataset and if a surrogate DL model can reproduce the uncertainty estimates. We found that all probabilistic DL methods are significantly more uncertain when making incorrect predictions at test time, but there are still instances where the models are very certain but completely incorrect. The surrogate DL models trained on the MC dropout and VI uncertainty estimates were capable of reproducing a significantly higher uncertainty estimate when making incorrect predictions.

CCS CONCEPTS

• **Computing methodologies → Machine learning approaches**; *Artificial intelligence*; Supervised learning by classification.

KEYWORDS

Driver intention recognition, probabilistic deep learning, surrogate modeling, uncertainty quantification

ACM Reference Format:
Koen Vellenga, Alexander Karlsson, H. Joe Steinhauer, Göran Falkman, and Anders Sjögren. 2023. Surrogate Deep Learning to Estimate Uncertainties for Driver Intention Recognition . In *2023 15th International Conference on Machine Learning and Computing (ICMLC 2023), February 17–20, 2023, Zhuhai, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3587716.3587758

ICMLC 2023, February 17–20, 2023, Zhuhai, China

ACM ISBN 978-1-4503-9841-1/23/02.
https://doi.org/10.1145/3587716.3587758

1 INTRODUCTION

Artificial intelligence (AI) based systems have demonstrated to be helpful in various fields. For example, in a medical context, AI can support disease detection or drug development [46], or in an automotive context, it can support a driver or drive by itself [4, 38]. However, an incorrect recommendation in these contexts can potentially cause harm to a human being. Hendrycks et al. (2022) [17] state that one of the unsolved problems in applying AI systems safely is the robustness of the underlying machine learning (ML) models and the capacity to handle rare events [51]. For example, advanced driving safety applications are struggling to predict unlikely and unseen complex road scenarios [31, 59], even with petabytes of naturalistic driving data available. Thus, expressing how certain an underlying ML model is about the made predictions is vital for AI systems that are likely to encounter situations that differ from the ones it has been trained on [5, 43, 55, 58]. Thus, EU regulations [6] have been proposed to enforce risk-mitigation measures, transparent decision-making, performance monitoring, and documenting the design choices and limitations of the system to evaluate and manage the risk to ensure safe deployment of AI systems that potentially can cause harm to humans.

Prior to all cars on the roads being fully autonomous, we will face a hybrid era, where both human and assisted driving vehicles will share the roads [11]. For an advanced driver assistance system (ADAS) to safely and reliably support its driver and to reduce potential traffic conflicts between road users, it is important to understand what other road users aspire to do. An example of such an ADAS is driver intention recognition (DIR). The goal of DIR is to timely recognize what a driver of the own car aims to do and to evaluate if it is safe for the driver to pursue that intention. DIR relies on observations from within the car from a driver monitoring system, the driving scene (through radar, lidar, or camera observations), and the vehicle dynamics sensors. Recent DIR studies (e.g., [2, 16, 42, 45, 57]) use deep learning (DL) to recognize the intentions, but do not communicate any uncertainty of the recognized intention, nor do they consider the computational limitations of the hardware in a car.

Uncertainties originate from multiple sources (e.g., due to measurement errors, missing or conflicting information, regularization effects, the statistical model induction, inference errors [7]). One can typically characterize the uncertainty either as aleatoric or as epistemic [8]. Aleatoric uncertainty refers to the inevitable natural randomness of the observations, and epistemic uncertainty

is the lack of knowledge or examples (that could potentially be reduced) [8, 35]. A normal deep neural network (DNN) produces a single-point estimate. However, there are multiple probabilistic DL approaches for DNNs that instead produce a distribution as a prediction (e.g., deep ensembles [29], variational inference [VI] [3, 15], Monte-Carlo [MC] dropout [12], and stochastic weight averaging - Gaussian [SWA-G] [21, 30]). These methods do require one to make a prediction for a single instance multiple times, which increases the computational costs [13].

A surrogate DNN can be used to quantify the uncertainty and avoids expensive MC sampling methods [61]. A surrogate model separates the uncertainty quantification task from the prediction task [13, 39, 40]. For example, Raghu et al. [39] implemented a separate network to directly predict the uncertainty based on the input to express the certainty of a medical diagnosis and Ramahlho and Miranda [40] used high-level feature representations to compare new images to the nearest neighbors from the training dataset to estimate an uncertainty score alongside the image classification task. To the best of our knowledge, a surrogate uncertainty model has not yet been evaluated in a DIR context and is hence the main contribution of this paper. We first use a probabilistic DL method to estimate the uncertainty for the test instances. We then evaluate if the probabilistic DL method produces higher uncertainty estimates for incorrectly predicted test instances, and conduct an exploratory analysis of the uncertainty decomposition. After that, we use the same probabilistic DL method to produce uncertainty values for the training dataset to train a surrogate DL model on. The goal of the surrogate DL model is to mimic the uncertainty estimation of the original probabilistic DL model. Lastly, we evaluate if the surrogate model produces higher uncertainty for incorrect test predictions. We use the openly available Brain4Cars [23] dataset that contains intention labels for lane-change and turn maneuvers.

2 PRELIMINARIES

2.1 Driver intention recognition

The term intention recognition is often used interchangeably with action, plan, or goal recognition. Nonetheless, intentions in a driving context refer to what a driver aspires to do (e.g., overtaking the car in front) [37, 44]. The goal (e.g., being in front of another car) refers to the desired end state and can be predicted from actions that are performed, (e.g., the steps of an overtaking maneuver), hence an intention is formed before the maneuver is pursued. To be able to act on an intention requires a plan that consists of multiple executable steps which then can be observed by other traffic participants (e.g., steering, accelerating, switching on the turn indicators).

2.2 Probabilistic deep learning methods

We select a subset of probabilistic DL methods that have been established in previous probabilistic DL studies (e.g., [22, 36, 50]) and that are practically applicable in a DIR context to generate uncertainty estimates. The methods are listed below with a basic intuition, and we refer to Zhou et al. (2021) [60] or Gawlikowski et al. (2021) [13] for a complete probabilistic DL overview.

- VI [3, 15, 26] – Replaces the weight variables of a DNN with variational distributions.
- Last Layer approximation [41, 48, 54] – Only applies the probabilistic DL method (e.g., MC dropout or VI) to the parameters of the last layer of the DNN.
- MC dropout [12] – Uses dropout during test time, which results in different network constellations for every forward pass.
- SWA-G [21, 56] – Performs Bayesian model averaging by marginalizing over the parameters of the network.
- Deep ensembles [29] – N networks are initialized and trained independently of each other.

2.3 Uncertainty quantification approach

The weights of a probabilistic DNN are sampled per forward pass (except for a deep ensemble where the different weights are the results of independently training the same DNN architecture N times). Multiple forward passes of the same input data result in a predictive uncertainty, which includes both the aleatoric and the epistemic uncertainty. Kwon et al. (2020) [28] proposed a direct mode-based uncertainty decomposition (Equation 1) based on Kendall and Gal's (2017) [24] approach. The aleatoric uncertainty is the average variance of a set of predictions, whereas the epistemic uncertainty can be interpreted as the average variance of a single prediction given the prediction mean of a set of multiple forward passes.

$$U_{pred} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}\left[diag(\hat{p}_i) - \hat{p}_i\hat{p}_i^{\top}\right]}_{\text{Aleatoric}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - \bar{p})(\hat{p}_i - \bar{p})^{\top}}_{\text{Epistemic}} \quad (1)$$

Where:
N = number of forward passes
$\hat{p}_i$ = the i^{th} instance from 1 to N
$\bar{p}$ = mean of the N stochastic predictions

3 EXPERIMENTAL SETUP

3.1 Problem formalization

On a highway or in an urban scenario a driver can form an intention to perform a maneuver (e.g., lane change or turn) $Y \in \{y_1, \ldots, y_n\}$ at a certain time step T. A set of data points $X_T = \{x_1, \ldots, x_m\}$ obtained from the vehicle sensors is collected at each time step T. Based on the input data X_T, a model aims to learn to recognize the driver's intentions.

3.2 Dataset

Brain4Cars is an open-source dataset [23] that is annotated with the driver's intention as labels for maneuvers such as left lane changes (124), right lane changes (123), left turns (58), right turns (55), and driving straight (234). The pre-processed dataset has five pre-defined folds and contains aggregations of facial landmark trajectories, the number of lanes, an intersection indicator, and the speed of the car.

3.3 Underlying model architecture

We will use the same underlying model architecture for each of the probabilistic DL and surrogate trained models. A recurrent

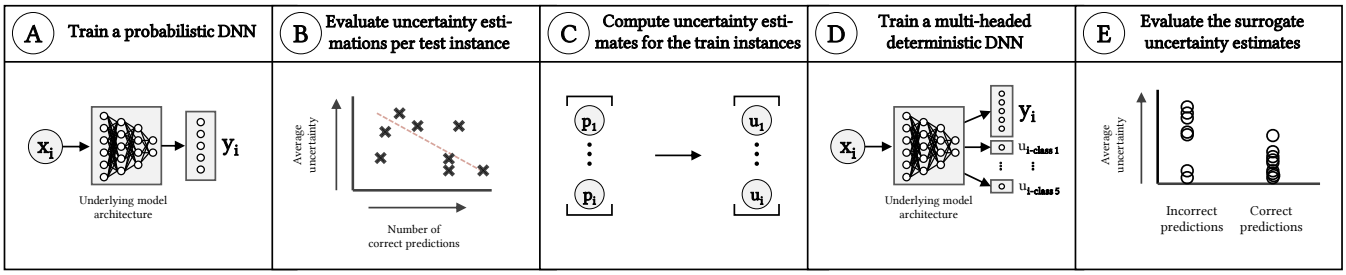

Figure 1: Schematic overview of the surrogate uncertainty estimation approach. First (step A) a probabilistic deep learning model is trained to predict the driver intentions (x_i represents an input sequence and y_i the intention labels). Second (step B), the uncertainty estimates are evaluated. Third (step C), the same probabilistic deep learning model is used to compute a set of predictions P for each train instance. The set of predictions p_i of a single train instance are used to produce a predictive uncertainty value u_i for each class in y_i. Fourth (step D), the surrogate model with the modified output layer is trained on y_i and u_i. Fifth (step E), we compare the average predicted uncertainty values between the incorrectly and correctly predicted test instances.

neural network (RNN) architecture is used to learn the DIR task. The long-short-term memory (LSTM) [18] cell reuses the output of a previous time step in a RNN architecture [47]. The LSTM controls which information to memorize and which elements are relevant for the prediction. A neural architecture search (NAS) [10] has been performed to motivate the complexity of the architecture. The search space was based on previous DIR studies and a Monte Carlo Tree Search [53] was deployed as a search strategy. A two-layer RNN (input-LSTM(60)-LSTM(60)-output) was established as the best-performing and least complex architecture.

3.4 Surrogate uncertainty model

Figure 1 shows a schematic overview of the approach we take to train a surrogate uncertainty estimation model. First, we train a probabilistic DL model and use MC sampling to compute a set P of 100 stochastic predictions for every test instance. Then, we evaluate if the uncertainty estimates U are higher for incorrectly predicted test instances compared to correct predictions. If the model produces significantly higher uncertainty estimates for the incorrect predictions, we use the model to compute the predictive uncertainty values for the training dataset instances. A surrogate model is then trained on both the DIR labels and the uncertainty estimates. The output layer of the surrogate model is modified to enable the prediction of an uncertainty estimate for every DIR label. Different from [39, 40], we learn the uncertainty based on the approximations of probabilistic DL methods and directly predict the uncertainty based on the same high-level feature representation that is used to predict the class distributions. Lastly, the uncertainty estimates of the incorrectly predicted test instances by the surrogate model are compared to the correctly predicted test instances.

3.5 Evaluation metrics

The predictive uncertainty produced by the probabilistic DL models is quantified by using Equation 1. The performance of the models to predict the driver intention is also evaluated. The cross-entropy loss (CE$\downarrow$[1]) is a proper scoring rule [14] (meaning it only reaches its optimum when the predicted distribution is equal to the ground truth) and measures the difference between two discrete distributions. Additionally, we also report the precision ($\uparrow$) for the models.

3.6 Training configuration

As mentioned by Ovidia et al. (2019) [36], it is difficult to compare the performance of the probabilistic DL approaches given the differences in the training procedures. However, we aim to keep the hyperparameters fixed as much as possible. We do not change the pre-defined data splits as defined by Jain et al. (2016) [23]. The cross-entropy loss is used for the DIR classification task. The surrogate model regresses five uncertainty estimates (one for each class) and uses the mean-squared error loss. The ADAM optimizer [25], with a learning rate of $1e^{-3}$, and a 0.3 dropout rate [49] is used. The SWA-G models first learn for 100 epochs and start with the SWA after that, and we train 100 models for the deep ensemble approach (matching the number of forward passes used for the other probabilistic DL methods). For implementing the models we use TensorFlow (2.9) [1], Edward 2.0 [52] and TensorFlow Probability [20].

3.7 Experiments

The aim is to understand how well the probabilistic DL methods express the uncertainty and how well a surrogate model can replicate that. For the probabilistic DL models, we use MC sampling to produce a set of 100 stochastic predictions for each test instance. First, we analyze the correlation between the number of correct predictions and the average epistemic, aleatoric, and predictive uncertainty estimate per test instance. This gives us an indication of whether there is a linear relation between correct predictions and the uncertainty estimate. After that, we split the test instances into a group with more than 75 correct predictions, a group with less than 25 correct predictions, and a group with instances that are predicted correctly 25 times or more but less than 75 times. The rationale for making three groups instead of two is that we

[1]The arrow ($\uparrow$ /$\downarrow$) indicates per metric if higher or lower indicates a better result.

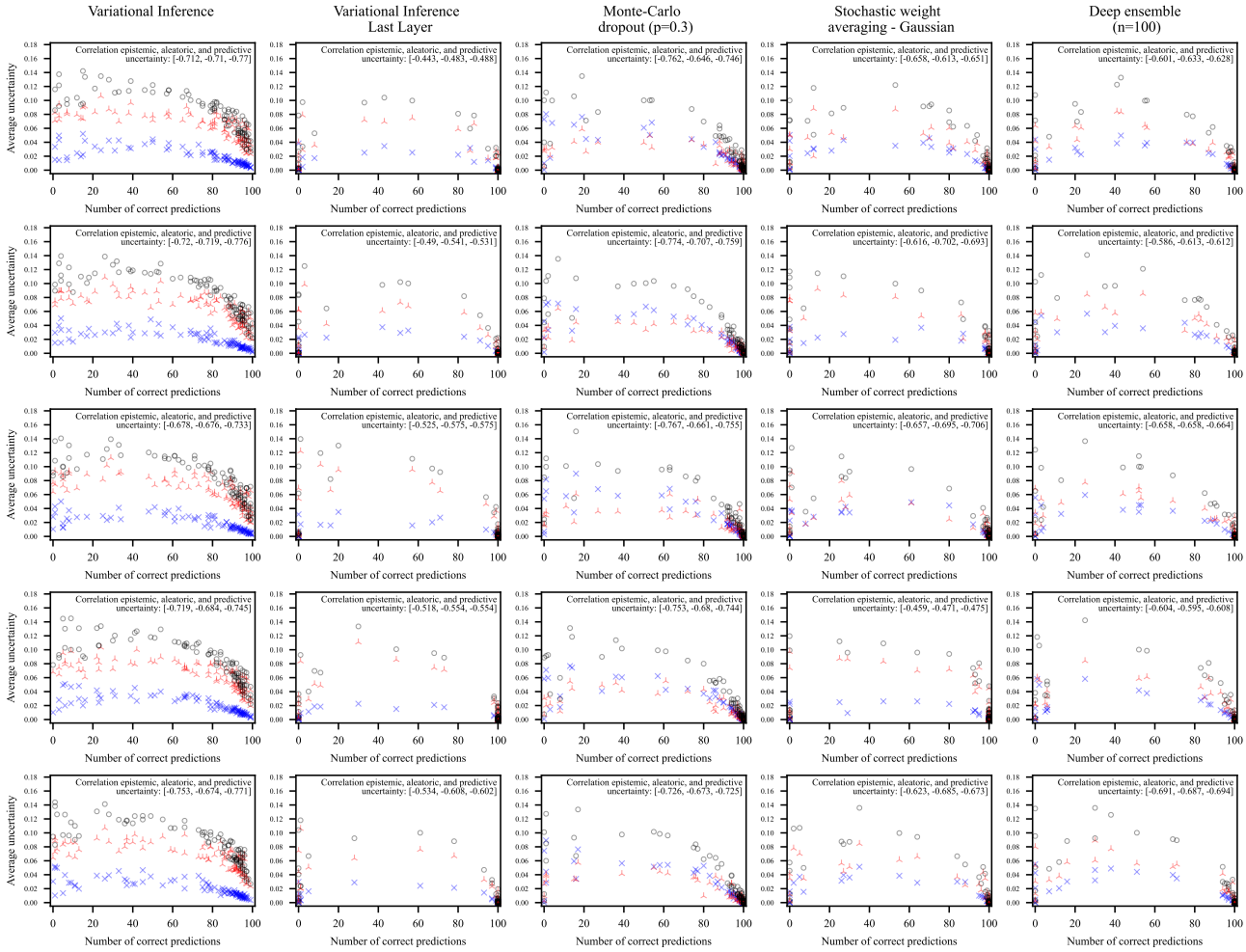

Figure 2: Visualization of the number of correct predictions and average uncertainty per test instance. The columns represent the probabilistic deep learning methods and the rows represent the five respective folds. For each test instance, the blue crosses represent the average epistemic uncertainty, the red three-pointed stars the average aleatoric uncertainty, and the black circles represent the average predictive uncertainty per class. The markers are made a bit transparent (alpha=0.4), which means that the color seems darker if there are multiple test instances with the same values.

want to explore if the group that is sometimes wrong or right produces higher uncertainty (due to inconsistent predictions). We use the non-parametric Kruskal–Wallis test [27] to observe if there are any differences between the groups, and the Dunn's test [9] for the posthoc pairwise comparisons with Holm's p-value adjustment method [19] to avoid falsely observing significant differences between the pairs.

Next, we evaluate how each respective probabilistic DL surrogate model can replicate the predictive uncertainty estimates. That means that, for each of the probabilistic DL methods, we compare if the uncertainty for incorrectly predicted test instances is higher compared to correctly predicted test instances. We use the non-parametric Mann-Whitney U test [32] to establish if there is a significant difference between these groups.

Table 1: Overview of the model performance for the five respective folds. The precision is the macro average of all classes.

Method	Model type	Precision ↑	CE ↓
VI	Probabilistic	83.94 ± 1.22	0.54 ± 0.02
	Surrogate	88.92 ± 1.21	0.51 ± 0.06
VI - LL	Probabilistic	88.96 ± 1.05	0.61 ± 0.06
	Surrogate	89.00 ± 1.71	0.85 ± 0.22
MC – dropout (p=0.3)	Probabilistic	**90.32 ± 0.30**	**0.41 ± 0.02**
	Surrogate	89.04 ± 0.81	0.61 ± 0.09
SWA-G	Probabilistic	89.42 ± 1.23	0.52 ± 0.04
	Surrogate	88.38 ± 1.28	0.89 ± 0.11
Deep ensemble (n=100)	Probabilistic	89.92 ± 0.66	0.48 ± 0.01
	Surrogate	88.94 ± 1.25	0.75 ± 0.10

4 RESULTS

Table 1 contains the average performance of all trained models for the five folds. While the MC dropout method yielded the highest precision and lowest CE, there is in general an overlap in the performance of the other models. The VI model has a lower precision compared to the other models. The surrogate models take roughly 75 times less time to compute a prediction and uncertainty estimate compared to the probabilistic DL models.

4.1 Probabilistic DL methods

Figure 2 shows the average aleatoric and epistemic uncertainty and the number of correct predictions per test instance. The probabilistic DL models make 100 predictions for each test instance. The columns represent the probabilistic DL models, and the rows represent the five different folds of the Brain4Cars dataset. The probabilistic DL models tend to produce less uncertainty when the number of correct predictions for a test instance is higher. The MC dropout and VI have the highest correlations between the correct predictions and average uncertainties. The VI model never produces very certain incorrect predictions, compared to the other models that have instances in the lower left corner of the graph (which means that the instance is predicted incorrectly and the average predictive uncertainty is close to 0). In general for the tested probabilistic DL models, the aleatoric and epistemic uncertainty estimates follow a similar pattern. However, the epistemic uncertainty produced by the MC dropout and SWA-G models is occasionally higher than the aleatoric uncertainty, which can hint at unstable predictions.

For all models, the Kruskal-Wallice test is significant, which means that there is a significant difference between some of the three groups (group 1: less than 25 correct predictions, group 2: more than 25 but less than 75 correct, group 3: more than 75 correct)[2]. For all models, we observe that group 3 is significantly different from both groups 1 and 2 for all folds, but that groups 1 and 2 are not significantly different from each other.

4.2 Surrogate model uncertainty estimation performance

The predictive uncertainty estimates are significantly higher for the incorrectly predicted test instances compared to the correctly predicted test instances for the MC dropout based surrogate model (fold 1: $\alpha = 0.010$, fold 2: $\alpha = 0.000$, fold 3: $\alpha = 0.004$, fold 4: $\alpha = 0.000$, fold 5: $\alpha = 0.010$) and the VI based surrogate model (fold 1: $\alpha = 0.013$, fold 2: $\alpha = 0.000$, fold 3: $\alpha = 0.009$, fold 4: $\alpha = 0.002$, fold 5: $\alpha = 0.001$). The deep ensemble based surrogate models did not produce significantly higher uncertainty estimates for incorrectly classified test instances compared to correctly classified test instances. The surrogate model trained on the VI-LL uncertainty estimates produced significant differences for one fold and the SWA-G based surrogate model for three folds[3].

[2]Splitting the data into thirds did not yield a different result.

[3]Tested the effect of clipping the uncertainty estimates from 0 to ∞ and a Softplus activation for the uncertainty regression heads, but it did not result in any differences.

5 DISCUSSION AND CONCLUSION

Using AI in situations where an incorrect prediction can lead to harming human beings requires a careful approach. We evaluated multiple probabilistic deep learning methods and a surrogate deep learning uncertainty estimation approach on the Brain4Cars [23] dataset. We found that all probabilistic deep learning methods are capable of capturing the aleatoric and epistemic uncertainty. The Monte-Carlo dropout and variational inference based surrogate models produced significantly higher uncertainty estimates for incorrectly predicted test instances compared to correct predictions.

The Hamiltonian Monte-Carlo (HMC) [33, 34] is considered state-of-art for Bayesian approximation and used to compare probabilistic deep learning methods to [22]. It could be interesting to investigate if a surrogate model would be able to replicate the HMC uncertainty estimates, given that the HMC is in general computationally too expensive to use for inferencing [28].

To understand the potential bias introduced by learning uncertainty estimations from a probabilistic deep learning method, one can explore if a performance decrease of the uncertainty estimations is observed when the surrogate model is forced to predict instances that originate from another dataset with different classes (similar to the experiment of Ovadia et al. [36]).

From a driver intention recognition perspective, the precision of the probabilistic models (except for the variational inference trained model) and surrogate models is not significantly different. Besides producing accurate predictions, it is essential to predict the driver's intentions far enough ahead in time to enable a DIR system to provide the driver enough time to process the information and anticipate their actions. To eventually improve road safety, we also need to investigate how to use the predictions and uncertainty estimates within an advanced driving assistance system. This means that future DIR methods should also be able to reject a prediction based on the uncertainty estimate to avoid the risk of being incorrect. One should be able to motivate why a certain threshold is used and, for example, why in certain scenarios only a limited set of potential driving intentions are considered for making a recommendation.

ACKNOWLEDGMENTS

This work was supported by the Intention Recognition in Real Time for Automotive 3D Situation Awareness (IRRA) Project (https://www.vinnova.se/p/intention-recognition-i-realtid-for-automotive-3d-situation-awareness-irra/).

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. (2015).

[2] Abdelmoudjib Benterki, Moussa Boukhnifer, Vincent Judalet, and Choubeila Maaoui. 2020. Artificial Intelligence for Vehicle Behavior Anticipation: Hybrid Approach Based on Maneuver Classification and Trajectory Prediction. *IEEE Access* 8 (2020), 56992–57002.

[3] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *International Conference on Machine Learning*. PMLR, 1613–1622.

[4] Markus Borg, Cristofer Englund, Krzysztof Wnuk, Boris Durann, Christoffer Lewandowski, Shenjian Gao, Yanwen Tan, Henrik Kaijser, Henrik Lönn, and Jonnas Törnqvist. 2019. Safely Entering the Deep: A Review of Verification and

Validation for Machine Learning and a Challenge Elicitation in the Automotive Industry. *Journal of Automotive Software Engineering* 1, 1 (2019), 1–19.
[5] J Quiñonero Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2009. Dataset shift in machine learning. *The MIT Press* 1 (2009), 5.
[6] E Commission et al. 2021. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. *COM (2021)* 206 (2021).
[7] Michael C Darling. 2019. Using Uncertainty To Interpret Supervised Machine Learning Predictions. (2019).
[8] Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? Does it matter? *Structural safety* 31, 2 (2009), 105–112.
[9] Olive Jean Dunn. 1964. Multiple comparisons using rank sums. *Technometrics* 6, 3 (1964), 241–252.
[10] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural architecture search: A survey. *The Journal of Machine Learning Research* 20, 1 (2019), 1997–2017.
[11] Lex Fridman. 2018. Human-centered autonomous vehicle systems: Principles of effective shared autonomy. *arXiv preprint arXiv:1810.01835* (2018).
[12] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. PMLR, 1050–1059.
[13] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2021. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342* (2021).
[14] Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association* 102, 477 (2007), 359–378.
[15] Alex Graves. 2011. Practical variational inference for neural networks. *Advances in neural information processing systems* 24 (2011).
[16] Zixu Hao, Xing Huang, Kaige Wang, Maoyuan Cui, and Yantao Tian. 2020. Attention-Based GRU for Driver Intention Recognition and Vehicle Trajectory Prediction. In *2020 4th CAA International Conference on Vehicular Control and Intelligence (CVCI)*. IEEE, 86–91.
[17] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021. Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916* (2021).
[18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[19] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.
[20] Google Inc. 2022. Tensorflow probability.
[21] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*. Association For Uncertainty in Artificial Intelligence (AUAI), 876–885.
[22] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. 2021. What are Bayesian neural network posteriors really like?. In *International conference on machine learning*. PMLR, 4629–4640.
[23] Ashesh Jain, Hema S Koppula, Shane Soh, Bharad Raghavan, Avi Singh, and Ashutosh Saxena. 2016. Brain4Cars: Car That Knows Before You Do via Sensory-Fusion Deep Learning Architecture. (2016). arXiv:1601.00740v1 [cs.RO]
[24] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems* 30 (2017).
[25] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[26] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
[27] William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* 47, 260 (1952), 583–621.
[28] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. 2020. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis* 142 (2020), 106816.
[29] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* 30 (2017).
[30] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. 2019. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems* 32 (2019), 13153–13164.
[31] Osama Makansi, Özgün Çiçek, Yassine Marrakchi, and Thomas Brox. 2021. On exposing the challenging long tail in future prediction of traffic actors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13147–13157.
[32] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.
[33] Radford Neal. 1992. Bayesian learning via stochastic dynamics. *Advances in neural information processing systems* 5 (1992).
[34] Radford M Neal et al. 2011. MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo* 2, 11 (2011), 2.
[35] William L Oberkampf, Jon C Helton, Cliff A Joslyn, Steven F Wojtkiewicz, and Scott Ferson. 2004. Challenge problems: uncertainty in system response given uncertain parameters. *Reliability Engineering & System Safety* 85, 1-3 (2004), 11–19.
[36] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems* 32 (2019), 13991–14002.
[37] Luís Moniz Pereira et al. 2013. State-of-the-art of intention recognition and its use in decision making. *AI Communications* 26, 2 (2013), 237–246.
[38] Martin Rabe, Stefan Milz, and Patrick Mader. 2021. Development methodologies for safety critical machine learning applications in the automotive domain: A survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 129–141.
[39] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. 2019. Direct Uncertainty Prediction for Medical Second Opinions. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 5281–5290. https://proceedings.mlr.press/v97/raghu19a.html
[40] Tiago Ramalho and Miguel Miranda. 2020. Density estimation in representation space to predict model uncertainty. In *International Workshop on Engineering Dependable and Secure Machine Learning Systems*. Springer, 84–96.
[41] Carlos Riquelme, George Tucker, and Jasper Snoek. 2018. Deep bayesian bandits showdown. In *International conference on learning representations*.
[42] Yao Rong, Zeynep Akata, and Enkelejda Kasneci. 2020. Driver Intention Anticipation Based on In-Cabin and Driving Scene Monitoring. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 1–8.
[43] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.
[44] Fariba Sadri. 2011. Logic-based approaches to intention recognition. In *Handbook of research on ambient intelligence and smart environments: Trends and perspectives*. IGI Global, 346–375.
[45] Omveer Sharma, NC Sahoo, and Niladri B Puhan. 2022. Kernelized convolutional transformer network based driver behavior estimation for conflict resolution at unsignalized roundabout. *ISA transactions* (2022).
[46] Dinggang Shen, Guorong Wu, and Heung-Il Suk. 2017. Deep learning in medical image analysis. *Annual review of biomedical engineering* 19 (2017), 221.
[47] Alex Sherstinsky. 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena* 404 (2020), 132306.
[48] Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. 2015. Scalable bayesian optimization using deep neural networks. In *International conference on machine learning*. PMLR, 2171–2180.
[49] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.
[50] Niclas Ståhl, Göran Falkman, Alexander Karlsson, and Gunnar Mathiason. 2020. Evaluation of uncertainty quantification in deep learning. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, 556–568.
[51] Nassim Nicholas Taleb. 2020. Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications. *arXiv preprint arXiv:2001.10488* (2020).
[52] Dustin Tran, Alp Kucukelbir, Adji B. Dieng, Maja Rudolph, Dawen Liang, and David M. Blei. 2016. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787* (2016).
[53] Linnan Wang, Yiyang Zhao, Yuu Jinnai, Yuandong Tian, and Rodrigo Fonseca. 2019. Alphax: exploring neural architectures with deep neural networks and monte carlo tree search. *arXiv preprint arXiv:1903.11059* (2019).
[54] Joe Watson, Jihao Andreas Lin, Pascal Klink, Joni Pajarinen, and Jan Peters. 2021. Latent Derivative Bayesian Last Layer Networks. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1198–1206.
[55] Oliver Willers, Sebastian Sudholt, Shervin Raafatnia, and Stephanie Abrecht. 2020. Safety concerns and mitigation approaches regarding the use of deep learning in safety-critical perception tasks. In *International Conference on Computer Safety, Reliability, and Security*. Springer, 336–350.
[56] Andrew G Wilson and Pavel Izmailov. 2020. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems* 33 (2020), 4697–4708.

[57] Yang Xing, Chen Lv, Huaji Wang, Dongpu Cao, and Efstathios Velenis. 2020. An ensemble deep learning approach for driver lane change intention inference. *Transportation Research Part C: Emerging Technologies* 115 (2020), 102615.

[58] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* 64, 3 (2021), 107–115.

[59] Weitao Zhou, Zhong Cao, Yunkang Xu, Nanshan Deng, Xiaoyu Liu, Kun Jiang, and Diange Yang. 2022. Long-Tail Prediction Uncertainty Aware Trajectory Planning for Self-driving Vehicles. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 1275–1282.

[60] Xinlei Zhou, Han Liu, Farhad Pourpanah, Tieyong Zeng, and Xizhao Wang. 2021. A Survey on Epistemic (Model) Uncertainty in Supervised Learning: Recent Advances and Applications. *Neurocomputing* (2021).

[61] Yinhao Zhu, Nicholas Zabaras, Phaedon-Stelios Koutsourelakis, and Paris Perdikaris. 2019. Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *J. Comput. Phys.* 394 (2019), 56–81.

Received 20 December 2022; accepted 5 January 2023