Assuring the Safety of Machine Learning for Pedestrian Detection at Crossings

Lydia Gauerhof[1,2], Richard Hawkins[2], Chiara Picardi[2], Colin Paterson[2], Yuki Hagiwara[1], and Ibrahim Habli[2]

[1] Corporate Research Robert Bosch GmbH, Renningen, Germany
[2] University of York, York, United Kingdom

Abstract. Machine Learnt Models (MLMs) are now commonly used in self-driving cars, particularly for tasks such as object detection and classification within the perception pipeline. The failure of such models to perform as intended could lead to hazardous events such as failing to stop for a pedestrian at a crossing. It is therefore crucial that the safety of the MLM can be proactively assured and should be driven by explicit and concrete safety requirements. In our previous work, we defined a process that integrates the development and assurance activities for MLMs within safety-related systems. This is used to incrementally generate the safety argument and evidence. In this paper, we apply the approach to pedestrian detection at crossings and provide an evaluation using the publicly available JAAD data set. In particular, we focus on the elicitation and analysis of ML safety requirements and how such requirements should drive the assurance activities within the data management and model learning phases. We explain the benefits of the approach and identify outstanding challenges in the context of self-driving cars.

Keywords: Machine Learning · Safety Argument · Self-Driving Car · Safety Assurance Process

1 Introduction

The assurance of safety-related systems which utilise Machine Learnt Models (MLMs) can only be achieved when arguments concerning the safety of the MLM are provided in the context of the overall system into which the model is deployed. For safety-related applications, the performance of the model is just one aspect that may be of interest; we must also take a much broader view of which aspects are important to assure the safety of the MLM. These aspects should be defined in the form of explicit Machine Learning (ML) safety requirements and should drive the way in which the MLM is trained and verified, with a particular focus on the quality and suitability of the traning and verification data sources.

In [15] we introduced a process for generating assurance arguments for MLMs. This process integrates development and assurance activities and can be used to incrementally generate the safety assurance argument and evidence that can be used to form a safety case for the MLM within the safety related context. We also described the structure of such arguments in the form of safety argument patterns [15]. Although some simple illustrative examples were provided, the details of how to implement the process activities, and the nature of the evidence that is generated were not provided. This paper seeks to address this by considering the safety-related automated driving scenario of a self-driving car approaching a pedestrian crossing. For this scenario we use a MLM for detection of pedestrians at the crossing that is trained on a publicly available dataset [17]. In particular, by considering this credible scenario and its associated safety implications, the primary contribution of the paper is that it shows how safety requirements can be systematically and traceably generated and refined across the different lifecycle phases of the MLM, particularly focussing on the data management and model learning requirements.

The rest of this paper is structured as follows. Section 2 provides an overview of our MLM safety assurance process. In Section 3 we describe the autonomous driving scenario that we used for our experiment and introduce the safety requirements for the system. Section 4 details the ML requirements that we derived for the scenario. Section 5 assesses the degree to which these requirements are satisfied for the data management and model learning stages of the lifecycle respectively. Section 6 discusses related work, draws conclusions from the paper and discusses our future work.

2 Model Learning Safety Assurance Process

The process we developed for assuring the safety of MLMs was presented in [15]. We split the ML lifecycle into five stages: requirements elicitation, data management, model learning, model verification and model deployment. Traditionally ML development has focused on data collection and model performance. For safety-related systems, a much broader view of ML development is required. In particular, the requirements elicitation stage must ensure that the ML requirements reflect the intent of the broader system-level safety requirements [9]. The model verification stage must provide an independent check that the requirements are satisfied and this must be particularly focused on the verification of explicit safety requirements. The model deployment stage must ensure that the learnt model will be acceptably safe when integrated into the larger system. To ensure that each lifecycle stage provides what is required to support a safety case, we can define a set of desired properties (desiderata) for each stage. It is important to have a clear and sufficient set of desiderata. For the work reported in this paper, we have used the assurance desiderata proposed by Ashmore et. al. in [1].

To ensure the desiderata are satisfied, specific ML safety requirements must be specified for each lifecycle stage. This is the focus of this paper. These ML requirements must relate to the specific safety requirements determined for the system into which the MLM will be deployed. The relationship between safety requirements at a system level and detailed ML requirements is not always obvious. For example, a safety requirement may define the need to identify all stop signs in an urban environment in sufficient time for the vehicle to stop comfortably. Turning this into specific and meaningful ML requirements relating to desiderata such as data coverage, model robustness or model accuracy is challenging and rarely discussed in a way that is justifiably traceable to system safety requirements. This paper describes how this may be done for a credible automotive scenario, focussing on ML safety requirements for data management and model learning. As part of a safety case, it must be demonstrated that the defined ML safety requirements are met. We discuss the activities that may be performed during the ML lifecycle to generate evidence to support this.

3 Pedestrian Detection at Crossings Scenario: Vehicle-Level Safety Requirements

We consider a MLM that is being used to identify pedestrians at pedestrian crossings so that an autonomous vehicle is able to stop safely. We consider that a car (the Ego vehicle) is driving autonomously in an urban environment and is approaching a crossing. We can specify a safety requirement on the Ego vehicle as follows:

Ego shall stop at the crossing if a pedestrian is crossing.

At this level the safety requirement is defined for the vehicle as a whole. It is important to note that this safety requirement would apply to the vehicle irrespective of the use of ML as part of the implementation. Based on system level safety analysis, other safety requirements could be identified (such as that the Ego vehicle should not stop unnecessarily at a crossing) but we do not consider those within this paper.

In order to elicit safety requirements for the MLM it is first necessary to identify the safety requirements that apply to the relevant system component, in this case the object detection component. The safety process decomposes the system level safety requirement to the different components of the Ego vehicle. This takes account of the proposed system architecture for the vehicle as well as the relevant operating scenarios and operating environment as discussed below.

Ego is able to sense the environment using a Bosch stereo video camera [12] that is fitted above the rear view mirror. The camera has an image size of 1280 x 960 pixels and a frame rate of 30 images per second. The images are sent to the object detection component that identifies pedestrians in the images and creates bounding boxes around each pedestrian. Figure 1 shows an example of an image in which all pedestrians in the scene were successfully identified. These are indicated by the green bounding boxes in the image. In this case, the image has also been annotated with white bounding boxes which show the ground truth. This indicates that even though all objects were successfully detected errors in the bounding boxes still remain. By contrast, Figure 2 shows an image from the same crossing in which there are several identification errors in the object detection, with pedestrians who were not spotted by the object detection indicated by blue boxes in the scene.

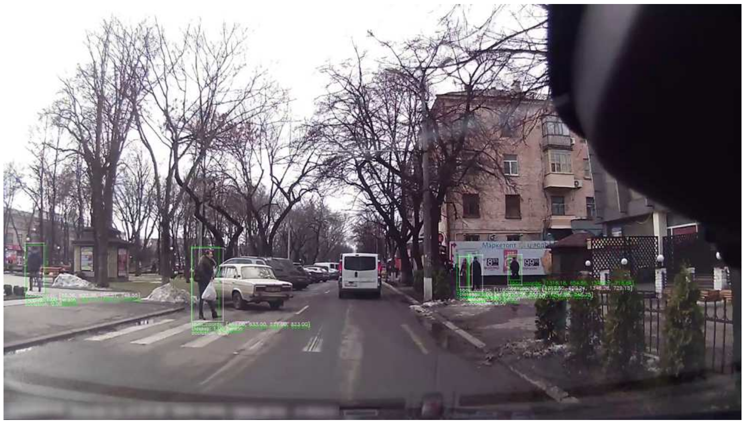


Fig. 1: An Ideal Pedestrian Detection at Crossings

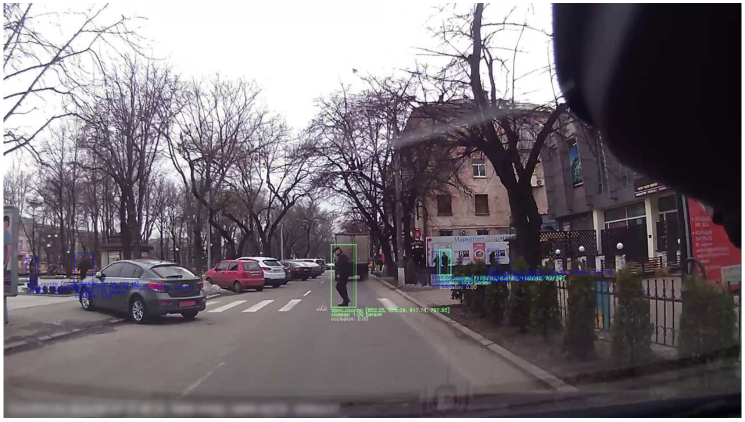


Fig. 2: An Example of Missed Detections at Crossings

It is crucial that the context within which the vehicle is expected to function is clearly and explicitly specified. For road vehicles this is normally done through the specification of the Operational Design Domain (ODD) [7]. J3016 defines an ODD as “operating conditions under which a given driving automation system or feature thereof is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristics” [19].

One of the reasons for specifying the ODD is to reduce the complexity of the input space. For instance, particular geographical areas and country-specific circumstances, such as traffic signs, can be excluded. Also weather conditions such as snow, and time of day such as night, may be excluded, meaning that Ego would not operate autonomously under such conditions. Measures are put in place to ensure operation does not occur under the excluded conditions [5]. There are a number of approaches to structuring the ODD, such as equivalence classes [18]. There are also a number of ODD ontologies that have been suggested [11, 7].

In the driving scenario in this paper, the ODD specifies that Ego operates on roads in the UK and in daylight, and that the weather conditions may be variable. In order to make our scenario concrete, we assume that pedestrians will only cross the road at the crossing, so we do not here consider pedestrians stepping off the pavement into the road.

Based on this we are able to specify safety requirements on the object detection component. This component is implemented in our example system using an MLM, in this case classification using a Convolutional Neural Network (CNN) based on SqueezeNet and localisation based on a Region Proposal Network (RPN). It is important to note however that at this point, the safety requirements could apply equally to the component whether it was a MLM or a more traditional software component.

To elicit the safety requirements we first consider the performance required of the object detection in order to satisfy the high-level safety requirement. Table 1 defines three performance related requirements. The justification for these requirements is provided below.

Table 1: Performance and Robustness requirements for object detection

	Performance
RQ1:	When Ego is 50 metres from the crossing, the object detection component shall identify pedestrians that are on or close to the crossing in their correct position.
RQ1.1:	In a sequence of images from a video feed any object to be detected should not be missed more then 1 in 5 frames.
RQ1.2:	Position of pedestrians shall be determined within 50cm of actual position
	Robustness
RQ2:	The object detection component shall perform as required in all situations Ego may encounter within the defined ODD.
RQ3:	The object detection component shall perform as required in the face of defined component failures arising within the system

For RQ1, 50m is specified as this is the minimum distance at which a decision to stop must be made if Ego is to stop comfortably at the maximum assumed speed. Stopping safely at a crossing requires consideration of this comfortable braking distance for the Ego vehicle; it would not be acceptable to brake excessively for pedestrians. The maximum comfortable braking distance will depend upon the speed of Ego and the road surface conditions. We assume for this scenario that comfortable braking loses roughly 20kph per second on a damp road, so if Ego is travelling at 60kph in the urban area it will take around 50 metres to stop comfortably. This requires that Ego has sufficient confidence in the identification of pedestrians at 50 metres, prior to this point Ego will be detecting the possible presence of pedestrians, however the uncertainty in those identifications may be relatively high.

RQ1 assumes that any pedestrian close to the crossing is intending to cross. This is certainly a conservative assumption that may result in some unrequired stopping, but is made here to simplify the scenario. In practice this could be mitigated through trajectory prediction for pedestrians (so for example pedestrians close to, but moving away from, the crossing would be rejected). It is taken that any pedestrian within one metre of the crossing is considered to be close to the crossing for the purposes of this scenario. Any pedestrians further away than this are assumed to be not intending to cross prior to Ego arriving at the crossing.

RQ1.1 and RQ1.2 further refine RQ1 by considering how good the performance of pedestrian identification and positioning needs to be in the context of the high-level system safety requirement and the system architecture. RQ1.1 is based upon the frame rate of the video feed as described above, and considers the fact that the ML model is deployed to a pipeline in which computational power is limited. As such the model may be unable to identify all objects in the scene for every frame at run-time. However the frame rate is such that the subsequent component into which the output of object detection is fed will ignore single frame changes in detections. RQ1.2 is based upon an assessment that 50cm discrepancy in position provides a sufficient safety margin for pedestrians.

In addition to requirements on performance, it is also necessary for the performance of the object detection to be robust to the different situations that Ego may encounter. Table 1 defines two requirements relating to robustness. RQ2 is justified on the basis that if a situation that Ego encounters is outside of the defined ODD then the system will revert to a fail-safe or a manual drive mode (it is not required for object detection to cope with such situations). The safety of such transitions would be handled at the vehicle level. RQ3 acknowledges that the system components cannot be assumed to always perform perfectly. Object detection must therefore be able to cope with some defined failures or degradation. It should be noted that any failures in other system components that are not specified or are unanticipated must still be dealt with, but this would be done as part of the vehicle level safety case.

As the object detection is implemented using a MLM, these safety requirements on object detection must be interpreted to be meaningful for ML to enable assurance of the MLM to be demonstrated. In the next section we describe how ML safety requirements are derived.

4 ML Safety Requirements Elicitation

In order to create a safety argument for the MLM, it is necessary to specify concrete and meaningful ML safety requirements, i.e. traceable to the vehicle and component-level safety requirements as discussed in Section 3. That is, the ML requirements must be sufficient to ensure that the safety requirements identified in Section 3 are satisfied. The ML safety requirements are defined with a consideration of each phase of the ML lifecycle and the identified desiderata for each phase. In this paper we focus on the requirements for the data management and model learning phases.

Tables 2-4 show the ML requirements that we have derived for these phases of the lifecycle. The tables enumerate requirements for each of the identified desiderata. For the data management phase, the desiderata we use are that the data should be relevant (Table 2), complete (Table 3), accurate (Table 4), and balanced (Table 4). These desiderata are consistent with the work of Ashmore et. al. in [1] where the desiderata are discussed in more detail. The ML requirements reflect these ML desiderata within the context of the safety requirement we have identified for object detection in our scenario. A justification for these requirements is provided below.

Table 2: ML requirement elicitation for the **Relevant** desiderata of the **Data Management** lifecyle phase

RQ4:	All data samples shall represent images of a road from the perspective of a vehicle.
RQ5:	Crossings included in data samples shall be of a type found on UK roads.
RQ6:	Pedestrians included in data samples shall be of a type that may use crossings on UK roads.
RQ7:	The format of each data sample shall be representative of that which is captured using sensors deployed on the ego vehicle.
RQ8:	Each data sample shall assume sensor positioning which is representative of that be used on the ego vehicle.

Table 3: ML requirement elicitation for the **Complete** desiderata of the **Data Management** lifecyle phase

RQ9:	The data samples shall include sufficient range of environmental factors within the scope of the ODD.
RQ10:	The data samples shall include sufficient range of pedestrians within the scope of the ODD.
RQ11:	The data samples shall include images representing a sufficient range of distances from the crossing up to that required by the decision making aspect of the perception pipeline. .
RQ12:	The data samples shall include examples with a sufficient range of levels of occlusion giving partial view of pedestrians at crossings.
RQ13:	The data samples shall include a sufficient range of examples reflecting the effects of identified system failure modes.

If we first consider the requirements relating to the 'Relevant' desiderata, we must specify requirements that define which data is relevant to the safety requirements. Any data that is not relevant should be excluded from the data set. In order to have relevance in this context, the data sample must be an image

Table 4: ML requirement elicitation for the **Accurate** and **Balanced** desiderata of the **Data Management** lifecyle phase

	Accurate
RQ14:	All bounding boxes produced shall be sufficiently large to include the entirety of the pedestrian.
RQ15:	All bounding boxes produced shall be no more than 10% larger in any dimension than the minimum sized box capable of including the entirety of the pedestrian
RQ16:	All pedestrians present in the data samples must be correctly labelled.
	Balanced
RQ17:	The data shall have a comparable representation of samples for each relevant class and feature (any class must not be under-represented with respect to the other classes or features)

that features a road as it may appear to Ego vehicle, and where this includes features of interest these should be relevant to the operational domain. In this case the features of interest are crossings and pedestrians. Relevant images would be expected to include some or all of these features. RQ4 to RQ6 capture this requirement for relevant data samples.

In addition the format of each image must be relevant. Since we understand the way in which images will be captured on the Ego vehicle, we can identify factors that are important to ensure the images are of a relevant format. In this case the relevant factors are the type of image created by the sensors and the position of the sensors in the vehicle. Physical properties of sensors can have a profound impact on the data gathered and it is often easier to collect data from publicly available sets or test harnesses which differ from the final deployed system. For example, the lenses on two different cameras will have different levels of distortion, vignetting and chromatic aberration. In order to ensure that issues of distributional shift, due to sensor variation, are avoided we can specify a requirement to ensure that the sensors used in training and deployment are not materially different (RQ7). The images, even if not generated from the Ego vehicle itself, must reflect the position of Ego's sensors. RQ8 defines this requirement.

We next consider the desiderata 'Complete'. From the robustness requirement RQ2 we know the data must include sufficient examples to reflect different situations Ego may encounter. Through consideration of the defined ODD we know these must include, for example, variations in the environment (a defined range of lighting and weather conditions), and in pedestrians (a defined range of ages, sizes, numbers of people and variations in gait and pose). It should be noted that an explicit enumeration of the scope of such variables is particularly critical when using MLMs in order to ensure robustness can be achieved. Experience has shown us that complex ML models can become over reliant on features in the image (over-fitting) if insufficient variation in those features is present in the data. By ensuring that a range of pedestrian features are present in the data sets we are less likely to produce models which fail to perform appropriately in the real world. RQ9 and RQ10 capture these requirements with referenced to the ODD, it is crucial therefore that the ODD is clearly documented and validated as part of the vehicle safety process. As well as exploring the scope of the ODD to consider different situations, we must also consider the impact on the images of the distance of Ego from the crossing (affecting the size of image features), and the possibility of occlusions in the image (we have discussed these effects in more detail in [2]. RQ11 and RQ12 address this issue.

From the robustness requirement RQ3 it has already been identified that the object detector must perform acceptably in the face of system failures. We acknowledge that the performance of a sensor will degrade over time, for example a camera lens will become scratched. Since this is generally unavoidable we must be confident that the performance of the object detection is not impacted by normal wear and tear. This means that the data used in the ML lifecycle must include sufficient examples that reflect the effects of these system failures on the images that are obtained. The relevant failures must be identified through failure analysis of the system (for example this could be linked to the outputs of an FMEA). RQ13 is specified to address this issue.

Another desiderata that must be considered is 'Accurate'. The performance of MLMs is highly dependent on the quality of the data from which they learn and as such all labelling should be accurate. The performance requirement RQ1.2 specifies a performance requirement on the prediction of the pedestrian's position. In order to assess this performance it is necessary to compare model predictions with the ground truth labels encoded in the training and testing data sets. RQ14 is therefore specified to ensure that the bounding box added to the dataset contains the whole of the pedestrian. If any part of the pedestrian, for example an arm or a leg, were not included inside the bounding box then when the model performance were assessed with reference to the bounding box a model could be deemed to meet the performance requirements when it actually breached the 50cm required by RQ1.2. Whilst this requirement specifies a minimum size for the bounding box, it does not consider a maximum size. It would be possible to meet RQ14 by creating very large boxes around every pedestrian, however this is likely to make the system unusable as free space is essentially identified as containing a pedestrian. RQ15 addresses this issue by specifying a limit on the size of the bounding box.

The performance requirement RQ1.1 may be interpreted as an ML requirement to avoid false negatives. This leads to a requirement on the accuracy of the training data. The training data is labelled (by a human) to identify the pedestrians in each image. Manual labelling of data is error prone and drawing bounding boxes in particular is difficult. If the images are labelled incorrectly such that the pedestrians are not identified in the image then this can lead to false negatives in the output of the MLM as well. RQ16 is specified to address this.

Finally RQ17 addresses the desiderata 'Balanced'. The requirements have already specified the need for relevance and coverage in the data, it is also important that certain features are not over or under represented in the data set. Again the relevant classes and features can be identified through consideration of the ODD.

Having defined explicit ML safety requirements it is then necessary to demonstrate that the ML requirements are satisfied. In Section 5 we discuss whether the data we used in this experiment meet the ML safety requirements and whether additional activities are required to support a safety case. Section 5.2 then discusses this for the model that is learnt.

5 Satisfying ML Safety Requirements

In order to investigate the sufficiency of the requirements defined in this paper we considered an experimental object detection MLM consisting of a CNN trained using the JAAD dataset [17]. In this section the data management and model learning for this MLM is assessed against the defined requirements to determine whether the requirements are satisfied. This highlighted areas where the MLM is insufficient from a safety assurance perspective, and identified additional assurance activities that would be required. This highlights the key role of an explicit elicitation of ML safety requirements in assuring MLMs.

5.1 Assessing the Data Management Safety Requirements

In this section, we discuss each data requirement presented in Tables 2-4 with reference to the JAAD dataset used in our experiment.

RQ4–5: In the dataset there are 25 videos relative to pedestrian crossing at designated and signalised crossings. For each of these videos approximately 82,000 image samples can be extracted. The recordings were done during 240 hours of driving across several locations in North America and Eastern Europe. Even if some of the crossings could be considered similar between Eastern Europe and UK (e.g. zebra and pegasus crossings), the data does not meet this requirements because UK locations are not included in the recording and therefore not all UK pedestrians crossings types are considered. In particular it can be easily noted that Pelican crossings are not included in the data. Augmenting the data by synthesising missing images can partially solve the problem, but the data samples generated must be very close to real world images. A better solution could be to undertake additional data collection in different UK crossing locations.
RQ6: When considering the JAAD dataset, we see that many classes of pedestrian are included, e.g. examples of children are included as is a man pushing a buggy. There are however some relevant omissions from this including people with disabilities, and people with different colour skin or ethnicity. When considering if there are any particular characteristics of UK pedestrians it seems important given the UK climate to ensure there are images of people carrying umbrellas or wearing waterproof jackets. These are not found in the data. Therefore the dataset could not be said to meet this requirement as more data would need to be collected that included the missing categories.

RQ7: The cameras used for the recording of the dataset are describe in [16]. The resolution of the three cameras are compatible with the one deployed in the ego vehicle (see Section 3). Consequently, this requirement can be considered satisfied.

RQ8: The camera recording the data used is positioned inside the car below the rear view mirror as described in [16]. In the ego vehicle the camera is mounted inside the car above the rear view mirror. Although the position is not exactly the same of the ones used for the data recording the distance is not significant and as such the requirement is satisfied.

RQ9–10: The data represents some different weather conditions (e.g. snow and rain). They do not consider different positions of the sun or different daytime lighting (e.g. sunset). Limited visibility weather conditions like fog are also not included, even though this is part of the ODD. The data includes pedestrians of different ages and height, as well as different walking speeds. No running pedestrians are included however. Although there are a sufficient range of examples for some features, for others the data is found to be lacking. Augmentation techniques can be applied to address this, for example by varying the colour of pixel or the orientation of pedestrians to the camera as done in previous work (e.g. [6]). In particular, Zhang and colleagues [20] described a method, through the use of a Generative Adversial Network (GAN), to synthesize scenes for autonomous driving simulating different weather conditions and then different lighting conditions. Further, possible evidence for supporting the argument in order to satisfy the requirements can be represented by performance graphs showing the difference between original and augmented data and how the different features included influence the performance. The data include some busy crossings that have groups of up to a maximum of 11 people. The performance of the MLM in identifying pedestrians when in groups compared to individuals could be used as evidence of this requirement. If performance is seen to be worse for groups, then more data samples for groups of people should be included.

RQ11: In most of the images included in the dataset, the pedestrians are very close to the car so do not respect the distance necessary for the pipeline decision making (in excess of 50m). The dataset would therefore be inappropriate against this requirement.

RQ12: There is partial occlusion of pedestrians in some of the data samples. For example, some pedestrians are occluded by gates or by the car in front of the Ego vehicle. Again, the number of occluded data samples could be increased through synthesis. For example artificial masking of pedestrians could be used to help meet the requirement.

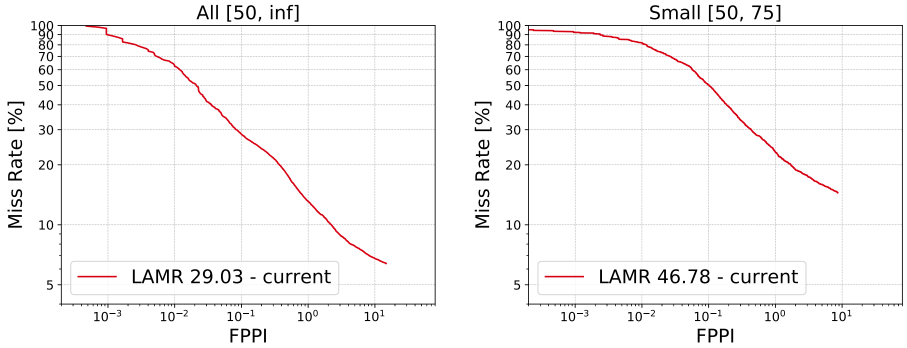
RQ13: Data samples derived from identified failures in the system are not present in the dataset. Also the classifier is not tested with adversarial attacks. For these reasons the requirements are not satisfied. In order to satisfy these requirements failures need to be identified and recorded in a report that can be used as evidence to support the argument. After failures are identified, corresponding data samples need to be added to the data set.

Req.14–16: While the process used to generate the dataset is described in [16], there is limited information regarding the generation of bounding boxes. Piotr's annotation toolbox [8] is used to define the bounding boxes and annotate the images. However, there is no information regarding the process to ensure that these are correct with respect to ground truth. The accuracy of the labelling is a function of the skills of the individuals undertaking the task and the validation processes used during labelling.
RQ17: Using a public dataset results in a lack of information and control in the number of data samples recorded for each feature of interest. Features that are under-represented have to be identified and possibly over-sampled in order to improve the performance of the classifier in presence of these features. Augmentation approaches can be used here, as well as other techniques for detecting and mitigating rare classes, such as [14].

In short, a public dataset such as JAAD is not sufficient to satisfy the ML safety requirements for our scenario. This result is not unexpected, but it highlights the role of explicit ML safety requirements in both highlighting deficiencies, and identifying necessary actions. Public datasets can however be useful for an exploratory analysis in order to refine the requirements as suggested by Gelman and colleagues [10].

5.2 Assessing the Model Learning Safety Requirements

In this section we discuss the ML saftey requirements that relate to the learned model itself as presented in Table 1, with reference to the model used in our experiment.
RQ1: In order to evaluate classifiers in the automotive domain it is common practice to use the log average miss rate (LAMR) [13]. Having constructed a convolution neural network as an MLM, we calculated the LAMR for images in the dataset. When considering all pedestrians larger than 50 pixels in the image, we obtain an LAMR of 29.03%. We note that for those pedestrians between 50 and 75 pixels this increases to 46.78%. These results are shown in Figure 3a and Figure 3b with more detail provided in Table 5.
RQ1.1: The JAAD dataset does provide a labelling which allows for each object to be tracked through frames. However, at present we do not have access to a pipeline which allows us to generate evidence to evaluate whether the MLM meets this requirement. This remains as future work.
RQ2: The images in the data set only cover 5 locations with the vast majority of videos captured in one location. Some of the images included weather features, for example LAMR results are shown in Table 6 for LAMR under snow conditions. Even these cases are restricted since snow is lying on the ground, so variations such as falling snow are missing. The generation of the JAAD database used for training required considerable effort, especially in the labelling of objects within the scene. In order to assess the ability of the MLM to operate at locations other than those in the JAAD dataset would require additional data collection and significant labelling effort. Without this, it is impossible to assess if the requirement could be met using this MLM.



(a) MR vs. FPPI for pedestrians larger than $50px$ in height

(b) MR vs. FPPI for pedestrians detection of the size $50px$ to $75px$

Fig. 3: Miss rate (MR) vs. false positive per image (FPPI) for pedestrian detection with different hights

RQ3: The JAAD videos were not captured using sensors traditionally used for autonomous vehicles. Instead, consumer video cameras were employed. In order to evaluate the effects of sensor wear, we would need to either simulate wear on the images, which would require a wear model to be validated, or we would need to collect data using sensors which had been subjected to appropriate wear, e.g. lens scratches etc. This new set could then be used as a test set on the candidate MLM. At present no such wear model or testing set exists and we can not therefore assess if the requirement is met.

Table 5: Log average miss rate (LAMR) of pedestrian detection with different heights of bounding boxes and occlusion severity (the smaller, the better)

Heights	**in Pixels**	**LAMR in %**	**LAMR in %**	**LAMR in %**
		no Occlusion	**Occlusion 25% -75%**	**Occlusion > 75%**
Small	50 - 75	46.78	54.12	62.18
Medium	75 - 100	20.22	28.91	36.49
Large	100 - 200	7.96	16.14	25.72
Huge	200 - 400	7.47	13.18	19.05
Giant	400 - 600	10.76	21.18	31.03

6 Discussion and Conclusions

There is no established approach to the assurance of MLMs for use in safety-related applications. Within the automotive domain, established safety standards such as ISO26262 do not consider MLMs. Traditional testing methods and test coverage metrics used for safety-related software, such as Modified Condition Decision Coverage, are not applicable to Neural Networks [3]. To try to close this gap, Cheng et. al. introduced metrics for measuring NN dependability

Table 6: Log average miss rate (LAMR) of pedestrian detection with different heights and occlusion severity under the **snow conditions**

Heights	in Pixels	LAMR in % no Occlusion	LAMR in % Occlusion 25% -75%	LAMR in % Occlusion > 75%
Small	50 - 75	39.30	47.01	55.89
Medium	75 - 100	13.78	19.08	27.65
Large	100 - 200	9.08	19.92	32.21
Huge	200 - 400	4.44	7.92	12.39
Giant	400 - 600	-	15.77	34.00

attributes including robustness, interpretability, completeness and correctness. Building upon this and other works, in [4] they introduce an "NN-dependability-kit" that could be used to support the development of a safety argument. Their work is not however driven by specific requirements that are explicitly and traceably linked to system-level safety analysis. Being able to demonstrate and justify this link is crucial to creating a compelling safety case.

This traceable link between system safety requirements and ML safety requirements is the focus of our work reported in this paper. This is important for two reasons: to maintain the link with vehicle-level hazardous events (and their mitigation) and to ensure that safety considerations are addressed in the detailed ML lifecycle phases. In particular, as we have shown in this paper, the ML safety requirements can be used to drive and scope the safety assurance activities. In this paper we have focused on the ML safety requirements for the data management and model learning phases. In our ongoing work, we intend to extend this to consider ML verification and deployment, which are two crucial aspects for a compelling safety case. Furthermore, formalizing these requirements in contract-based design allows machine support for refinement checks within a component-based system [2]. We hope that this work is of benefit to both researchers and engineers and helps inform the current debate concerning the safety assurance and regulation of autonomous driving.

Acknowledgements. This work is funded by the Assuring Autonomy International Programme https://www.york.ac.uk/assuring-autonomy.

References

1. Ashmore, R., Calinescu, R., Paterson, C.: Assuring the machine learning lifecycle: Desiderata, methods, and challenges (2019), http://arxiv.org/abs/1905.04223
2. Burton, S., Gauerhof, L., Sethy, B.B., Habli, I., Hawkins, R.: Confidence arguments for evidence of performance in machine learning for highly automated driving functions. In: International Conference on Computer Safety, Reliability, and Security. pp. 365–377. Springer (2019)
3. Cheng, C., Nührenberg, G., Huang, C., Ruess, H., Yasuoka, H.: Towards dependability metrics for neural networks. In: 2018 16th ACM/IEEE International Conference on Formal Methods and Models for System Design (2018)

4. Cheng, C.H., Huang, C.H., Nührenberg, G.: nn-dependability-kit: Engineering neural networks for safety-critical autonomous driving systems. arXiv:1811.06746
5. Colwell, I., Phan, B., Saleem, S., Salay, R., Czarnecki, K.: An automated vehicle safety concept based on runtime restriction of the operational design domain. In: 2018 IEEE Intelligent Vehicles Symposium (IV) (2018)
6. Crispell, D., Biris, O., Crosswhite, N., Byrne, J., Mundy, J.L.: Dataset augmentation for pose and lighting invariant face recognition. arXiv:1704.04326 (2017)
7. Czarnecki, K.: Operational world model ontology for automated driving systems—part 1: Road structure. Waterloo Intelligent Systems Engineering Lab (WISE) Report (2018)
8. Dollár, P.: Piotr's computer vision matlab toolbox (PMT), https://github.com/pdollar/toolbox
9. Gauerhof, L., Munk, P., Burton, S.: Structuring validation targets of a machine learning function applied to automated driving. In: Computer Safety, Reliability, and Security. Springer International Publishing (2018)
10. Gelman, A., Loken, E.: The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University (2013)
11. Geyer, S., Baltzer, M., Franz, B., Hakuli, S., Kauer, M., Kienle, M., Meier, S., Weißgerber, T., Bengler, K., Bruder, R., et al.: Concept and development of a unified ontology for generating test and use-case catalogues for assisted and automated vehicle guidance. IET Intelligent Transport Systems (2013)
12. GmbH, R.B.: Stereo video camera product data sheet. Available at https://www.bosch-mobility-solutions.com/media/global/products-and-services/passenger-cars-and-light-commercial-vehicles/driver-assistance-systems/lane-departure-warning/stereo-video-camera/product-data-sheet-stereo-video-camera.pdf (2020/02/28)
13. Liu, S., Huang, D., Wang, Y.: Adaptive NMS: refining pedestrian detection in a crowd. CoRR **abs/1904.03629** (2019), http://arxiv.org/abs/1904.03629
14. Paterson, C., Calinescu, R.: Detection and mitigation of rare subclasses in neural network classifiers. arXiv preprint arXiv:1911.12780 (2019)
15. Picardi, C., Paterson, C., Hawkins, R., Calinescu, R., Habli, I.: Assurance argument patterns and processes for machine learning in safety-related systems. In: Proceedings of the Workshop on Artificial Intelligence Safety (SafeAI 2020). pp. 23–30. CEUR Workshop Proceedings (2020)
16. Rasouli, A., Kotseruba, I., Tsotsos, J.K.: Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
17. Rasouli, A., Kotseruba, I., Tsotsos, J.K.: It's not all about size: On the role of data properties in pedestrian detection. In: ECCVW (2018)
18. Rosenwald, G.W., Chen-Ching Liu: Rule-based system validation through automatic identification of equivalence classes. IEEE Transactions on Knowledge and Data Engineering (1997)
19. SAE: J3016, taxonomy and definitions for terms related to on-road motor vehicle automated driving systems (2013), https://saemobilus.sae.org/content/j3016_201401
20. Zhang, M., Zhang, Y., Zhang, L., Liu, C., Khurshid, S.: Deeproad: Gan-based metamorphic testing and input validation framework for autonomous driving systems. In: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering (2018)