On safety assurance case for deep learning based image classification in highly automated driving

Himanshu Agarwal*†, Rafal Dorociak* and Achim Rettberg†‡
*HELLA GmbH & Co. KGaA, Lippstadt, Germany
Email: {himanshu.agarwal, rafal.dorociak}@hella.com
†Department of Computing Science, Carl von Ossietzky University Oldenburg, Germany
Email: achim.rettberg@iess.org
‡University of Applied Sciences Hamm-Lippstadt, Lippstadt, Germany

***Abstract*—Assessing the overall accuracy of deep learning classifier is not a sufficient criterion to argue for safety of classification based functions in highly automated driving. The causes of deviation from the intended functionality must also be rigorously assessed. In context of functions related to image classification, one of the causes can be the failure to take into account during implementation the classifier's vulnerability to misclassification due to high similarity between the target classes. In this paper, we emphasize that while developing the safety assurance case for such functions, the argumentation over the appropriate implementation of the functionality must also address the vulnerability to misclassification due to class similarities. Using the traffic sign classification function as our case study, we propose to aid the development of its argumentation by: (a) conducting a systematic investigation of the similarity between the target classes, (b) assigning a corresponding classifier vulnerability rating to every possible misclassification, and (c) ensuring that the claims against the misclassifications that induce higher risk (scored on the basis of vulnerability and severity) are supported with more compelling sub-goals and evidences as compared to the claims against misclassifications that induce lower risk.**

***Index Terms*—deep learning, neural networks, vulnerability, intended functionality, safety case, functional safety**

I. Introduction

Having witnessed an increased focus on embedding artificial intelligence techniques in the driving systems, the scope of functional safety also needs certain interpretations in context of highly automated driving. ISO26262 as a norm for functional safety in electrical and/or electronic (E/E) systems has garnered a worldwide acceptance by the automotive companies, ever since its release in 2011 [1]. It deals with the measures to mitigate the hazards in E/E systems that can lead to safety goal violations due to random hardware faults and systematic faults in hardware or software. The possibility of an unintended behaviour even if there are no faults in the hardware or software must not be overlooked. The deviation from the intended functionality can occur due to performance limitations, as addressed in ISO/PAS 21448 [2]. For machine learning (ML) based functions, the sources of performance limitations can be identified and a relevant safety argumentation can be established with a focus on mitigating the associated risk [3]. In [4], the approach to demonstrate the intended functionality of a ML function is based on reducing the risks associated with three sources of performance limitations, namely, underspecification, semantic gap and deductive gap. In order to support the corresponding arguments, the authors in [4] have proposed some validation targets as sub-goals in the safety assurance case. Our work in this paper is motivated with their work.

The deductive gap refers to the poor or incorrect learning of the features by the trained model due to which the intended functionality is not implemented properly [4]. For applications such as traffic sign classification, the features among the target classes have striking similarities. In this paper, we emphasize that the argumentation corresponding to the reduction of risks associated with deductive gap must also address the classifier's vulnerability to misclassification due to similarity between the target classes. This vulnerability can be exploited by the intentional or unintentional threats to cause a hazard. An example of intentional threat is the adversarial attack wherein the aim of the attacker is to manipulate the input such that the classifier model is compelled to misclassify it into another class [5]. The unintentional threats can be natural adversaries, poor lightning or adverse weather conditions. It can become rather easier for these threats to exploit the class similarities in order to produce an erroneous output. The intentional threats are security-related, while the unintentional threats are safety-related [6]. In this paper, we focus on the latter (i.e., safety).

The rest of the paper is organized as follows. In Section II, we conduct the vulnerability assessment for a traffic sign classification functionality to estimate the classifier's vulnerability to a particular misclassification. In Section III, we discuss how the risk associated with a particular misclassification can be scored, and also provide some theoretical insights into the argumentation to demonstrate the misclassification's low on-operation probability. In Section IV, we propose some validation targets (sub-goals) to strengthen the corresponding claim against the misclassifications that were assessed to be inducing the highest level of risk. A sample safety assurance case is then structured in Section V. Finally, some key concluding remarks are presented in Section VI.

II. Vulnerability Assessment

A. Purpose

If the image classification functionality is to be implemented for n number of classes, the total number of possible misclassifications is $n(n-1)$. Some of the pairs of classes can

978-3-9819263-5-4/DATE21/©2021 EDAA

have strong overlap between their dominant visual features. Irrespective of the classifier to be used, these pairs of classes can induce high vulnerability to misclassification into each other. The purpose of vulnerability assessment here is to serve as a systematic way of categorising a possible misclassification based on the level of vulnerability it can potentially induce in the classifier due to similarity between the corresponding pair of target classes.

One can conduct this vulnerability assessment for a specific trained classifier model by estimating the similarity based on the features that influence the decision of the model the most. But, on the other hand, an a priori assessment based on comparing the target classes with respect to their attributes of the dominant visual features can also help in identifying the critical cases (not specific to any model) and recognizing the corresponding steps or measures (e.g., related to designing or training the deep neural networks) that can contribute in making the function robust. In this paper, the scope of such an a priori assessment is limited to use cases like traffic sign classification, where the identification of the common dominant visual features in the target classes and the comparison between their corresponding attributes are plausible.

B. Target Classes

We have taken into consideration the traffic sign classes from the German Traffic Sign Recognition Benchmark (GTSRB) dataset [7]. There are 43 different classes: 8 speed limit, 4 prohibitory, 4 derestriction, 15 danger, 8 mandatory and 4 unique (priority road, yield to cross, stop, do not enter) signs.

C. Procedure

1) Identify the dominant visual features: There are two dominant visual features in all the traffic signs (target classes), namely, shape and color [8]. The 43 traffic sign classes in GTSRB can be grouped on the basis of these dominant visual features, as shown in Table I. As mentioned in Section II-A, the similarity between the classes will be assessed a priori, i.e., without using the actual data from GTSRB. It will be based only on the specification of the target classes (Table I).

2) Estimate the similarity between the classes based on all the dominant visual features individually: In context of traffic signs, this implies the similarity between their shapes and color combinations. The similarity between the two shapes S_1 and S_2 can be estimated by determining the area that is left uncovered when S_1 is superimposed on S_2 (or vice versa), as shown in Fig. 1(a) - 1(j). Higher the area of shaded region, lesser is the similarity between the two given shapes. Here, we assume the height (and width) of all the shapes to be the same. To analyse the possibility of a detected traffic sign getting misclassified into a traffic sign of an another shape, this assumption is sufficient. We calculated the area of shaded region (δ_s) for $l = 10\,\mathrm{cm}$ in Fig. 1(a) - 1(j). The obtained values are recorded in Table II. The observed maximum area is $\delta_s = 50\,\mathrm{cm}^2$, i.e., when a regular triangle is superimposed on an inverted triangle or vice versa (Fig. 1(e)). We divide the range of 0 (minimum area) to 50 (maximum area) into three equal divisions, such that a value of δ_s lying within:

- $0 \leq \delta_s < 16.67$ represents high,
- $16.67 \leq \delta_s \leq 33.33$ represents moderate, and
- $33.33 < \delta_s \leq 50$ represents low vulnerability

to misclassification for the corresponding pair of traffic signs.

The similarity between the two color combinations can be determined by taking the average of the background color difference and the border color difference. Mathematically:

$$\delta_c = \frac{1}{2}\Big[\underbrace{\lambda(r_1 g_1 b_1, r_2 g_2 b_2)}_{\text{color difference background}} + \underbrace{\lambda(r_1 g_1 b_1, r_2 g_2 b_2)}_{\text{color difference border}}\Big], \qquad (1)$$

TABLE I
DOMINANT VISUAL FEATURES IN TRAFFIC SIGN CLASSES

	No.	Geometric Shape	Traffic Signs
SHAPE	1.	Circle	Speed Limit (all 8), Prohibitory (all 4), Derestriction (all 4), Mandatory (all 8), Do not enter
	2.	Triangle	Danger (all 15)
	3.	Inverted Triangle	Yield to cross
	4.	Rhombus	Priority Road
	5.	Octagon	Stop

	No.	Color (RGB Code)[a]		Traffic Signs
		Background	*Border*[b]	
COLOR	1.	White (255,255,255)	Signal Red (155,36,35)	Speed Limit (all 8), Prohibitory (all 4), Danger (all 15), Yield to cross
	2.	White (255,255,255)	Signal Black (43,43,44)	Derestriction (all 4)
	3.	Signal Blue (0,83,135)	Signal Blue (0,83,135)	Mandatory[c] (all 8)
	4.	Signal Yellow (249,168,0)	White (255,255,255)	Priority Road
	5.	Signal Red (155,36,35)	Signal Red (155,36,35)	Stop[c], Do not enter[c]

[a] the actual RGB color codes may vary slightly.
[b] ignoring the outer thin white border on the traffic sign boards.
[c] border color considered same as the corresponding background color.
Note: In practice, these specifications will have to be formally acquired in coordination with the sign board manufacturers and/or regulatory bodies.

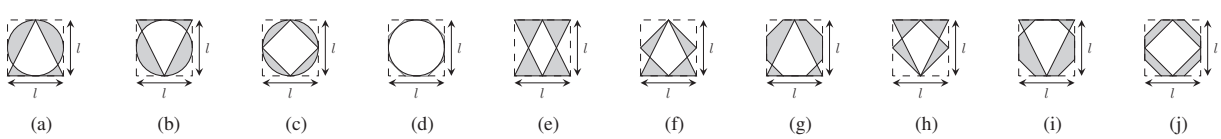

Fig. 1. Superimposition of a shape S_1 on another shape S_2 or vice versa: (a) Circle ↔ Triangle, (b) Circle ↔ Inverted Triangle, (c) Circle ↔ Rhombus, (d) Circle ↔ Octagon, (e) Triangle ↔ Inverted Triangle, (f) Triangle ↔ Rhombus, (g) Triangle ↔ Octagon, (h) Inverted Triangle ↔ Rhombus, (i) Inverted Triangle ↔ Octagon, and (j) Rhombus ↔ Octagon. Higher the area of the shaded region, lesser is the similarity between the two given shapes.

TABLE II
AREA OF SHADED REGION (δ_s)

Shape	Circle	Triangle	Inverted Triangle	Rhombus	Octagon
Circle	0	42.18	42.18	28.54	4.26
Triangle	42.18	0	50	33.3	44.23
Inverted Triangle	42.18	50	0	33.3	44.23
Rhombus	28.54	33.3	33.3	0	32.84
Octagon	4.26	44.23	44.23	32.84	0

Note: The values are in cm^2; obtained for $l = 10\,\mathrm{cm}$ in Fig. 1(a) - 1(j).

where δ_c is the required average color difference, (r_1, g_1, b_1) and (r_2, g_2, b_2) are the RGB codes, and λ is the color difference. The value of λ is estimated as the euclidean distance between the color coordinates in RGB space. To compensate for human eye sensitivity to certain colors, suitable sensitivity coefficients can be used in the calculation of euclidean distance [9]. Here, we do not see an unavoidable need to consider the human eye sensitivity for assessing the classifier's vulnerability to a misclassification; therefore, we calculate λ simply as:

$$\lambda(r_1 g_1 b_1, r_2 g_2 b_2) = \left[(r_2 - r_1)^2 + (g_2 - g_1)^2 + (b_2 - b_1)^2\right]^{\frac{1}{2}}, \tag{2}$$

Lower the value of δ_c in (1), higher is the similarity between the two given color combinations. Table III presents the values of δ_c between the five color combinations. Again, we divide the range of 0 (minimum average color difference) to 318.06 (maximum average color difference) into three equal divisions, such that a value of δ_c lying within:

- $0 \leq \delta_c < 106.02$ represents high,
- $106.02 \leq \delta_c \leq 212.04$ represents moderate, and
- $212.04 < \delta_c \leq 318.06$ represents low vulnerability

to misclassification for the corresponding pair of traffic signs.

In this way, we can first estimate the vulnerability to misclassification in terms of all the dominant features individually.

3) Assign a classifier vulnerability rating to every possible misclassification: The vulnerability *high*, *moderate* and *low* estimated above are assigned the ratings of 3, 2 and 1, respectively. For instance, the vulnerability rating corresponding to the misclassification from *20 speed limit* sign into *go straight mandatory* sign (or vice versa) will be assigned as:

- $v_s = 3$ (high) in terms of shape similarity since $\delta_s = 0$ (refer to Table I for shapes and Table II for δ_s), and
- $v_c = 1$ (low) in terms of color similarity since $\delta_c = 260.26$ (refer to Table I for color and Table III for δ_c).

Assuming equal contribution of both shape and color similarity, the final classifier vulnerability rating (v) for a misclassification can be derived by taking the average of the corresponding v_s and v_c. Thus, $v = 2$ for the example above. Table IV categorizes the possible misclassifications among the GTSRB classes on the basis of derived vulnerability rating v.

TABLE III
AVERAGE COLOR DIFFERENCE (δ_c)

Background & Border	White & Red	White & Black	Blue & Blue	Yellow & White	Red & Red
White & Red	0	56.29	260.26	297.81	163.07
White & Black	56.29	0	219.24	318.06	219.35
Blue & Blue	260.26	219.24	0	312.94	190.35
Yellow & White	297.81	318.06	312.94	0	245.96
Red & Red	163.07	219.35	190.35	245.96	0

Note: All the above values are calculated using the formulae given in (1) and (2), and using the RGB codes given in Table I.

III. SCOPE OF THE SAFETY ARGUMENTATION

The vulnerability (v) of a misclassification derived above can be considered analogous to its likelihood in risk assessment matrix (Fig. 2). The other component, i.e., severity or impact (s) has to be assessed seperately. For instance, the severity of the misclassification of 20 speed limit sign into 120 speed limit sign can be considered as *high*, as it can trigger an undesirable driving action (i.e., the vehicle driving too fast). Whereas, the severity of the misclassification of *slippery road ahead* danger sign into *snow possible ahead* danger sign can be deemed as *low*, as it would still result in a desirable driving action (i.e., slow driving). The risk of misclassification can then be scored as *low*, *moderate* or *high* (Fig. 2). To demonstrate the low

TABLE IV
CATEGORISATION OF ALL THE POSSIBLE MISCLASSIFICATIONS AMONG THE GTSRB TRAFFIC SIGN CLASSES ON THE BASIS OF VULNERABILITY

Vulnerability		**Misclassification**
Level	***Rating*** (v)	**from ↔ into**
High	**3.0**	Speed Limit ↔ Speed Limit, Prohibitory ↔ Prohibitory, Derestriction ↔ Derestriction, Danger ↔ Danger, Mandatory ↔ Mandatory, } d Speed Limit (all) ↔ Prohibitory (all), Speed Limit (all) ↔ Derestriction (all), Prohibitory (all) ↔ Derestriction (all), Stop ↔ Do not enter.
Moderate	**2.5**	Speed Limit (all) ↔ Stop/ Do not enter, Prohibitory (all) ↔ Stop/ Do not enter, Mandatory (all) ↔ Stop/ Do not enter.
	2.0	Speed Limit (all) ↔ Danger (all), Speed Limit (all) ↔ Mandatory (all), Prohibitory (all) ↔ Danger (all), Prohibitory (all) ↔ Mandatory (all), Derestriction (all) ↔ Danger (all), Derestriction (all) ↔ Mandatory (all), Derestriction (all) ↔ Stop/ Do not enter, Yield to cross ↔ Speed Limit (all)/ Prohibitory (all)/ Derestriction (all)/ Danger (all).
	1.5	Priority Road ↔ Every other class, Danger (all) ↔ Stop/ Do not enter, Yield to cross ↔ Stop/ Do not enter.
Low	**1.0**	Danger (all) ↔ Mandatory (all), Mandatory (all) ↔ Yield to cross.

[d] misclassification into another sign of the same category.

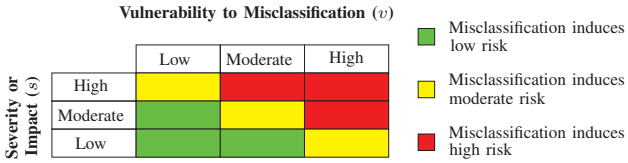

Fig. 2. Risk assessment matrix for a misclassification.

on-operation probability of a misclassification in the image classification function, following evidences can be used:

1) Against the misclassifications that induce low risk: The results of rigorous testing, e.g., from on-road driving in different weather conditions, geographic locations, etc. could be sufficient evidence to demonstrate the low on-operation probability of the misclassifications that pose low risk (green zone in Fig. 2). Nevertheless, an acceptable threshold for the metrics such as *probability of misclassification*[1] p_{k_1,k_2} must be defined explicitly. Also, it is important to cover all the elements of the operational design domain (ODD) during testing [10]. Here, we make an assumption on the test data representativeness such that the metrics acquired using the test data can be considered relevant to the function's on-operation performance. In real practice, it can be quite challenging; hence, the topic of coverage or completeness of the test data will have to be dealt comprehensively.

2) Against the misclassifications that induce moderate risk: One of the possible ways to demonstrate the low on-operation probability can be the monitoring of performance metrics (e.g., p_{k_1,k_2}) for the corresponding classes after implementation of additional measures such as redundancy of the deep learning models, use of independent and diverse machine learning techniques or image processing methods that work in parallel with the main deep learning classifier. Though the vulnerability due to class similarity can also affect these secondary classifiers or methods, but the fusion of the individual predictions made by the main classifier and the secondary classifiers (or methods) can greatly improve the overall decision-making in the classification functionality. The application of unanimity voting can provide the function an ability to refrain from yielding a decision, i.e., when the predictions made by the main classifier and the secondary classifiers (or methods) are not in agreement. In this case, an appropriate action such as transition into a safe state can be implemented at system-level.

Let us assume the number of test images belonging to class k_1 as N, and the main classifier misclassifies $M_1 \leq N$ images into class k_2. The corresponding probability of misclassification is $p_{k_1,k_2} = M_1/N$. When the secondary classifiers (or methods) are working in parallel with the main classifier, let us say, for R out of N images, the function yields *no decision* via unanimity voting. The *rejection rate*[2] becomes $\bar{r}_{k_1} = R/N$, where $0 \leq \bar{r}_{k_1} \leq 1$. Also, let us assume that $M_2 \leq M_1$ images still suffer misclassification into class k_2. Hence, the probability of misclassification, when additional measures work simultaneously, is $\bar{p}_{k_1,k_2} = M_2/(N-R)$, and it should expectedly be lower than p_{k_1,k_2}. An acceptable threshold must be set for the values of p_{k_1,k_2}, $\bar{r}_{k_1}$ and $\bar{p}_{k_1,k_2}$.

3) Against the misclassifications that induce high risk: Some additional compelling validation targets against the misclassifications that induce high risk (red zone in Fig. 2) can be beneficial. This is further discussed in Section IV.

The abovementioned points are summarized in Table V.

[1]Probability of misclassification p_{k_1,k_2} is defined as the number of test images belonging to class k_1 that were misclassified into class k_2, divided by the total number of test images belonging to class k_1.

[2]Rejection rate is defined as the number of test images for which the function yields *no decision*, divided by the total number of test images.

IV. Validation Targets

As mentioned in Table V, additional compelling validation targets can be established to strengthen the claim against the misclassifications that pose high risk. Such validation targets can be related to low-level intricate details or complex notions that are interpretable to humans. Consider the misclassification of a 20 speed limit sign into a 120 speed limit sign. It can be assigned *high* vulnerability ($v = 3.0$ as per Table IV) and *high* severity (s) ratings. Thus, it belongs to the high risk zone in Fig. 2. The validation target against the misclassification of a 20 speed limit sign into a 120 speed limit sign, for instance, can be formulated in a way such that it highlights the function's ability to identify if the posted speed limit has 2 or 3 digits. Among all the misclassifications that induce high risk, the intricacy of the notions formulated as validation targets would be more for the misclassifications with *high* vulnerability ratings, owing to higher similarity between the dominant visual features of their corresponding classes. Table VI presents some examples of such specific validation targets. Note that the list is not complete and can be further extended. The techniques of deep visualization (e.g., in [11]) and concept embedding analysis (e.g., in [12]) can be used to ensure that such notions interpretable to humans have been adequately grasped by the classifier models used in the traffic sign classification function. In the scope of this paper, the validation targets in Table VI serve only as examples; thus, the confidence measure will have to be quantified when specifying an actual validation target.

V. Safety Assurance Case

The Goal Structuring Notation (GSN) in Fig. 3 illustrates a sample safety assurance case for traffic sign classification

TABLE V
Evidence(s) for demonstrating the low on-operation probability of a misclassification

Evidence	Risk of misclassification		
	Low	*Moderate*	*High*
Results of rigorous testing (covering all the elements of the specified ODD)	✓	✓	✓
Results/evidence associated with the use of additional measures	e	✓	✓
Results/evidence associated with additional compelling validation targets			✓

[e] might be required if the corresponding value of the metric (e.g., p_{k_1,k_2}), observed via rigorous testing, is not within an acceptable threshold.

TABLE VI
SOME EXAMPLES OF SPECIFIC VALIDATION TARGETS AGAINST THE MISCLASSIFICATIONS THAT WERE ASSESSED TO BE INDUCING HIGH RISK

Misclassification from → into	**v: Vulnerability** **s: Severity[f]**	**Validation targets for the traffic sign classification functionality in an autonomous vehicle[g]**
	v: High s: High	It is able to correctly visualize with high confidence the number of digits depicted on a circular sign board having white background and red border (i.e., two digits in *20* speed limit and three digits in *120* speed limit sign).
	v: High s: High	It is able to correctly visualize with high confidence the difference between the digits '2' and '8' depicted on a circular sign board having white background and red border.
	v: High s: High	It is able to correctly visualize with high confidence whether a circular sign board having white background and red border depicts any symbol(s) or is blank (i.e., *no vehicles of any kind permitted* prohibitory sign is blank, whereas *no passing for any vehicle* prohibitory sign depict symbol of two cars).
	v: High s: High	It is able to correctly visualize with high confidence the difference between the symbols: car (light vehicle) and truck/trailer (heavy vehicle over 3.5 t), depicted on a circular sign board having white background and red border. It is able to correctly differentiate between the color coding of the symbols (light and heavy vehicle) depicted on a circular sign board having white background and red border (i.e., red color of symbol implies *passing prohibited*).
	v: High s: High	It is able to correctly visualize with high confidence the direction of arrow head depicted on a circular sign board having blue background, even in case of minor perturbations due to geometric transformations (e.g., rotation).
	v: High s: High	It is able to correctly visualize with high confidence whether the circular sign board has a red or a black border (i.e., red for *80* speed limit sign and black for *end of 80 speed limit* derestriction sign). It is able to correctly visualize with high confidence whether a circular sign board depicts a band of diagonal lines from north-east to south-west (i.e., *no* for *80* speed limit and *yes* for *end of 80 speed limit* derestriction sign).
	v: Moderate s: High	It is able to correctly visualize with high confidence the color of the background on a circular sign board (i.e., red for *do not enter* sign and blue for *go straight* mandatory sign). It is able to correctly visualize with high confidence the orientation of the bar/arrow depicted on a circular sign board (i.e., horizontal bar on *do not enter* sign and vertical arrow on *go straight* mandatory sign).
	v: Moderate s: High	It is able to correctly visualize with high confidence the difference between the triangular and circular shapes of a sign board having white background and red border. It is able to correctly visualize with high confidence whether the sign board depicts pictoral symbol or numeric digits (i.e., pictoral symbol on *snow possible ahead* danger sign and numeric digits on *120* speed limit sign).
	v: Moderate s: High	It is able to correctly visualize with high confidence the difference between the (a) shapes: octagon and rhombus, and (b) color combinations: {red background, red border} and {yellow background, white border}. It is able to correctly visualize with high confidence whether the detected sign board depicts alphabetical characters or not (i.e., *yes* for *stop* sign and *no* for *priority road* sign).

[f] probable severity based on general rules, however, in real practice, the estimation of severity will involve a detailed analysis at the vehicle-level.
[g] assume the vehicle to be a car with gross weight less than 3.5 tonnes (t).

function. The claim that the function complies with its intended functionality is specified as a top-level goal G1. The contexts C1 and C2 constrain the scope to the specified ODD and the target classes for which the function is designed, respectively. The context C3 indicates that deep neural network is used as the main classifier, and the assumption A1 is stated to suggest that it is the most appropriate technique. One of the various sub-goals to support G1 is to highlight that the risk associated with the deductive gap (i.e., inadequate implementation of the intended functionality) has been reduced, as specified in G2. The corresponding acceptable residual risk will have to be explicitly defined (C4) in order to interpret the accomplishment of the sub-goal G2. Further sub-goals supporting G2 are linked via the argument over appropriate implementation (strategy S1). While the authors in [4] have derived some validation targets as sub-goals to support G2 via S1, here we illustrate further only our argument over the classifier's vulnerability to misclassification due to similarity between the target classes. The aspects covered in Section II - IV are structured within the strategy S2. The context C5 specifies the approach adapted for the systematic investigation of the similarity between the target classes and the estimation of the classifier vulnerability rating corresponding to every possible misclassification (Section II). The assumption A2 states that every possible misclassification among the target classes has been assigned a vulnerability (v) and a severity (s) rating, and the corresponding risk is scored. The assumption A3 states that the testing data include samples corresponding to every ODD element. It can be further elaborated with a justification, e.g., the elements for which the real-world data could not be collected were compensated by the data produced via synthetic data generation techniques (like Generative Adversarial Networks [13]). The context C6 specifies the use of additional measures, e.g., the secondary classifiers (or methods) that work in parallel with the main classifier. Let us assume that unanimity voting is used so as to enable the function to reject issuing a decision if the individual predictions made by the main classifier and the secondary classifiers (or methods) are not in agreement. An overview of the sub-goals G3 to G5 and their required evidence(s) is already presented in Table V. The sub-goal G5 corresponding to the misclassifications that induce high risk is supported by additional validation targets (examples in Table VI) and the corresponding evidences (results from feature visualization or concept analysis). In C7 and C8, the acceptable thresholds on the required metrics (p_{k_1,k_2}, $\bar{r}_{k_1}$ and $\bar{p}_{k_1,k_2}$) can be explicitly defined. However, how to deduce these thresholds has not been addressed in this paper.

VI. CONCLUSION

In this paper, we emphasize that the argumentation over appropriate implementation of the intended functionality for an image classification function must take into account the vulner-

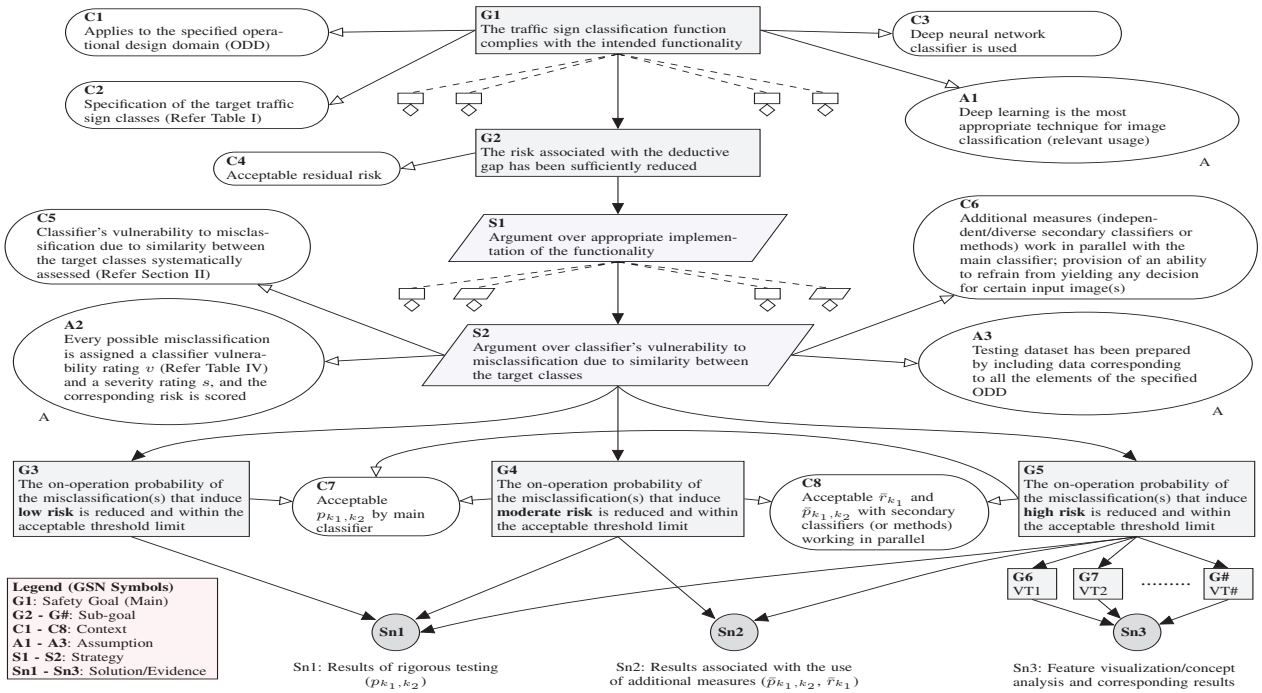

Fig. 3. GSN illustrating a part of the safety assurance case for deep learning based traffic sign classification function. The notations VT denote the validation targets (examples given in Table VI), and dashed arrows represent the connectivity to the other possible (undeveloped) sub-goals or strategies not specified here.

ability to misclassification due to similarity between the target classes. In context of traffic sign classification, we assessed this vulnerability through a systematic investigation based on the specification of the target traffic sign classes. Firstly, the similarity between the classes is assessed with respect to their dominant visual features, followed by which a vulnerability rating is assigned to every possible misclassification based on the similarity between the corresponding pair of classes. The obtained vulnerability ratings can be considered analogous to likelihood in risk assessment. To demonstrate the low on-operation probability of misclassifications that have high risk scores, the corresponding argumentation can be strengthened with the help of additional compelling sub-goals (validation targets related to intricate details) and evidences to make it more convincing. As an extension to this work, the arguments discussed theoretically in this paper need to be experimentally validated using trained deep neural networks. An empirical assessment of the relation between vulnerability to misclassification and it's corresponding probability on the real data is to be performed. Some elements (e.g., contexts, assumptions) and terms (e.g., acceptable threshold, residual risk) used in the safety assurance case presented here are addressed in a general manner, and will have to further elaborated and/or quantified in a formal way.

ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their valuable feedback and comments.

REFERENCES

[1] *ISO 26262:2011 Road Vehicles – Functional Safety*, ISO, 2011.
[2] *ISO/PAS 21448:2019 Road vehicles – Safety of the intended functionality*, ISO, 2019.
[3] S. Burton, L. Gauerhof, and C. Heinzemann, "Making the case for safety of machine learning in highly automated driving," in *International Conference on Computer Safety, Reliability, and Security*. Springer, 2017, pp. 5-16.
[4] L. Gauerhof, P. Munk, and S. Burton, "Structuring validation targets of a machine learning function applied to automated driving," in *International Conference on Computer Safety, Reliability, and Security*. Springer, 2018, pp. 45-58.
[5] C. Szegedy *et al.*, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.
[6] M. B. Line, O. Nordland, L. Røstad, and I. A. Tøndel, "Safety vs security?" in *PSAM conference, New Orleans, USA*, 2006.
[7] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition." *Neural Networks*, vol. 32, pp. 323-332, 2012.
[8] X. W. Gao, L. Podladchikova, D. Shaposhnikov, K. Hong, and N. Shevtsova, "Recognition of traffic signs based on their colour and shape features extracted using human vision models," *Journal of Visual Communication and Image Representation*, vol. 17, no. 4, pp. 675–685, 2006.
[9] W. Mokrzycki and M. Tatol, "Colour difference ΔE-A survey," *Machine graphics and vision*, vol. 20, no. 4, pp. 383–411, 2011.
[10] P. Koopman and F. Fratrik, "How many operational design domains, objects, and events?" in *SafeAI@ AAAI*, 2019.
[11] A. Nguyen, J. Yosinski, and J. Clune, "Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks," *arXiv preprint arXiv:1602.03616*, 2016.
[12] G. Schwalbe and M. Schels, "Concept enforcement and modularization as methods for the ISO 26262 safety argumentation of neural networks," in *10th European Congress on Embedded Real Time Software and Systems (ERTS 2020)*, 2020.
[13] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.