Can you trust your ML metrics? Using Subjective Logic to determine the true contribution of ML metrics for safety

Benjamin Herd
Fraunhofer Institute for Cognitive Systems (IKS)
Munich, Germany
benjamin.herd@iks.fraunhofer.de

Simon Burton
Fraunhofer Institute for Cognitive Systems (IKS)
Munich, Germany
simon.burton@iks.fraunhofer.de

ABSTRACT

Metrics such as accuracy, precision, recall, F1 score, etc. are generally used to assess the performance of machine learning (ML) models. From a safety perspective, relying on such single point estimates to evaluate safety requirements is problematic since they only provide a partial and indirect evaluation of the true safety risk associated with the model and its potential errors. In order to obtain a better understanding of the performance insufficiencies in the model, factors that could influence the quantitative evaluation of safety requirements such as test sample size, dataset size and model calibration need to be taken into account. In safety assurance, arguments typically combine complementary and diverse evidence to strengthen confidence in the safety claims. In this paper, we make a first step towards a more formal treatment of uncertainty in ML metrics by proposing a framework based on Subjective Logic that allows for modelling the relationship between primary and secondary pieces of evidence and the quantification of resulting uncertainty. Based on experiments, we show that single point estimates for common ML metrics tend to overestimate model performance and that a probabilistic treatment using the proposed framework can help to evaluate the probable bounds of the actual performance.

CCS CONCEPTS

• **Theory of computation** → **Machine learning theory**; *Automated reasoning*; • **Software and its engineering** → **Software safety**; • **Mathematics of computing** → *Probability and statistics*.

KEYWORDS

Machine learning, safety assurance, uncertainty, subjective logic

ACM Reference Format:
Benjamin Herd and Simon Burton. 2024. Can you trust your ML metrics? Using Subjective Logic to determine the true contribution of ML metrics for safety. In *The 39th ACM/SIGAPP Symposium on Applied Computing (SAC '24), April 8–12, 2024, Avila, Spain.* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3605098.3635966

1 INTRODUCTION

Machine learning (ML) is increasingly being employed in safety-critical applications, e.g. for deep learning-based perception in the context of autonomous driving. As a consequence, being able to argue the safety of ML-based functions is crucially important. This is also reflected in the growing body of literature concerned with safe and trustworthy AI [20]. Research can be essentially split into two camps: (1) the ML community which aims to propose concrete metrics and techniques to assess the quality of an ML model, e.g. performance metrics, robustness scores, verification approaches, or explainability techniques, and (2) the safety community which traditionally aims to understand causal relationships between faults, errors and failures [1] of different system components in order to provide an argument that the risk of safety-related failures of the system is acceptably low. From a conceptual perspective, the work done in the ML community can be seen as providing *evidence* in the form of either clearly measurable quantitative metrics or more qualitative insights into the underlying ML model which can then be used to support the overall *safety assurance argument.*

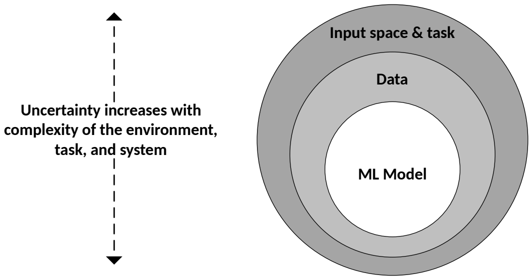

Uncertainty increases with complexity of the environment, task, and system
Input space & task
Data
ML Model

Figure 1: Dimensions of uncertainty impacting the safety assurance of ML

In a recent paper, the problem of ML safety assurance has been addressed from the perspective of *uncertainty* [5] and it has been argued that safety assurance requires the identification, and (if possible) reduction, of various manifestations of uncertainty. To this end, uncertainty is classified into three layers (see Figure 1): *model*, *data*, and *input space & task*. According to this conceptual model, any statement about the performance of a *model* is critically dependent upon the quality of the *data* used to train and test the model; any statement about the validity of the data is, in turn, conditional upon the representativeness of the *specification*, i.e. the formalised description of the input space and task. For example, assume a neural network-based perception model achieves 99% classification accuracy for the detection of pedestrians against a set of benchmark test data. Even if system-level measures are in place to mitigate the residual 1% of errors, the measurement of classification accuracy may be an insufficient basis for safety argumentation since it depends on (1) further aspects of the model such as its calibration

SAC '24, April 8–12, 2024, Avila, Spain
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0243-3/24/04.
https://doi.org/10.1145/3605098.3635966

quality and (2) the integrity and validity of training and testing data which, in turn, depends on (3) a sufficiently good understanding of the operational context and task to be solved. If any one or all of (1)–(3) are lacking, the *actual accuracy* of the model with respect to its intended functionality may vary from the *measured accuracy*.

In order to be convincing, an overall safety argument must therefore provide a holistic perspective, include sufficiently many pieces of evidence for all three layers. In general, a safety argument will be based on a range of both quantitative and qualitative pieces of evidence and associated argument structures. The challenge is to combine those pieces of evidence into an overall assessment of risk that reflects the combined uncertainty associated with the evidences. Our work aims to address this problem and presents an idea for a formal approach to evidence combination and uncertainty quantification within and across the layers mentioned above in a safety argument using *Subjective Logic (SL)* [14], a formalism combining probabilistic logic with the concepts of uncertainty and subjectivity. In particular, we make the following contributions:

(1) We describe how commonly used quantitative metrics on different layers of the uncertainty model shown in Figure 1 can be formulated as *opinions* in the framework of Subjective Logic (Section 3) from which Beta probability distribution functions can be derived that capture the base uncertainty associated with the measurement.
(2) We describe how uncertainty propagation within and across the layers of the uncertainty model can be modelled and quantified using the notion of transitive trust chains (Section 4). By incorporating additional knowledge on higher layers, overconfidence in a metric on a lower layer can be detected and previously "hidden" uncertainty can be identified.
(3) We illustrate the approach using a realistic and safety-critical example in the area of ML-based classification for traffic sign recognition. We show how uncertainty in the measured true positive rate (= recall) can be revealed by combining it with other pieces of evidence (Section 5).

2 BACKGROUND AND RELATED WORK

2.1 Assurance uncertainty

Burton *et al.* [4] express the task of assuring the safety of ML (according to SOTIF[1]) in terms of demonstrating the fulfillment of a safety contract based on the following definition.

$$\forall i \in I. A(i) \Rightarrow G(i, M(i)) \tag{1}$$

Where, for all inputs i that fulfil the set of assumptions A on the operating domain and system context, the output of model M must fulfill a set of conditions defined by guarantees G. Absolute perfection is neither achievable nor required to achieve a tolerable level of residual risk according to an acceptance criterion (*AC*), therefore safety is defined in terms of a *probability of success* for a given *distribution of inputs* in the operational design domain (ODD), as reflected by the following equation:

$$\frac{\sum_{i \in I, A(i) \wedge G(i, M(i))} \mathbb{P}_{ODD}(i)}{\sum_{i \in I, A(i)} \mathbb{P}_{ODD}(i)} \geq AC \tag{2}$$

where $\mathbb{P}_{ODD} : I \rightarrow [0, 1]$ is the *input probability distribution function* of the ODD that assigns every input $i \in I$ with a probability value, with the condition that $\sum_{i \in I} \mathbb{P}_{ODD}(i) = 1$. For realistic systems, the guarantees G (e.g. avoidance of hazardous system actions) cannot be directly evaluated through observations of the model outputs during development. Instead, *measurable properties* P of M (e.g. accuracy, precision, recall, robustness, calibrated error rate) are evaluated for a finite number of samples j of I (e.g. our test dataset). The safety assurance problem can thus be refined from Equation 2 as follows:

$$\frac{\#\{j \in I : A(j) \wedge P(j, M(j))\}}{\#\{j \in I : A(j)\}} \approx \frac{\sum_{i \in I, A(i) \wedge G(i, M(i))} \mathbb{P}_{ODD}(i)}{\sum_{i \in I, A(i)} \mathbb{P}_{ODD}(i)} \tag{3}$$

The left-hand side of the equation represents the *estimated failure rate* calculated using directly measurable properties, and the right-hand side represents the *actual failure rate* that occurs during operation. Assurance uncertainty manifests itself as a lack of knowledge of the difference between the two failure rates, as expressed by the '≈' approximately equal relation. The quantification of the difference between the estimated and actual safety (or inversely, risk) is referred to as *assurance confidence estimation*.

2.2 Assurance confidence estimation

Assurance confidence estimation aims to reduce the uncertainties associated with the safety argument and work in this area can be classified into *qualitative* and *quantitative* approaches [8]. Qualitative approaches aim to decrease uncertainty by strengthening the argument itself, e.g. through additional confidence-specific claims, sub-claims, and evidences; quantitative approaches use uncertainty quantification techniques such as frequentist or Bayesian probabilities or Dempster-Shafer belief functions to quantify and aggregate uncertainty in order to come up with an overall confidence score. We focus here on quantitative approaches that operate on hierarchical safety arguments, i.e. tree-like structures where a top-level safety claim is recursively subdivided into sub-claims, often represented using the Goal Structuring Notation (GSN) [21].

Goodenough *et al.* [10] present an approach to safety confidence quantification based on the idea of *eliminative induction* and provide an overall confidence score through the use of an adapted form of *Baconian probability*. In this approach, *defeaters* represent statements that cast doubt on the validity of claims. Eliminating those defeaters one by one then contributes to successively strengthening the overall claim. The confidence in a claim C is then defined as the ratio of the number of eliminated defeaters to the overall number of defeaters specific to C. E.g., given five defeaters, three of which have been eliminated successfully, the resulting confidence in C is 3/5. It is important to note that this does not mean that the confidence is 3/5 = 60%. As a consequence, the resulting value does not represent an absolute mass of confidence and cannot be combined with other directly measured metrics, but can be used as a relative value of the strength of argument.

An approach based on *Dempster-Shafer theory (DST)* [24] was presented by Ayoub *et al.* [2]. The authors emphasise the avoidance of confirmation bias which is particularly problematic when only *supporting* evidence is taken into account. Therefore, they do not

[1] ISO 21448 "Road vehicles - Safety of the intended functionality (SOTIF)"

just focus on the assessment of overall sufficiency of a given safety argument, but also on the assessment of insufficiency. The approach consists of two steps: (1) assessment of sufficiency and insufficiency for each part of an assurance argument and assignment of degrees of belief; and (2) aggregation of degrees of belief of subordinate claims into an overall confidence measure using Dempster's rule of combination. Wang *et al.* [25] propose an approach that subdivides confidence into *trustworthiness of individual claims* and *appropriateness of inference rules* and uses DST to calculate an overall confidence score. Trustworthiness in a claim C is represented as a triple $(m(C), m(\bar{C}), m(C, \bar{C}))$ comprising belief, disbelief, and uncertainty in C. Appropriateness influences the propagation of trustworthiness from sub-claims to the top claim and is defined as a tuple comprising *contribution weights* for each sub-claim, the *cooperative contribution* of sub-claims and the *overall reliability* of sources of information or the completeness of premises. Technically, the overall reliability is represented as a discount factor in DST.

Various Bayesian approaches to quantify the confidence in safety arguments have been presented. Guo *et al.* [11] are among the ones to introduce the idea of using a *Bayesian Belief Network (BBN)* for safety assessment. They conclude that most of the problems related to applying safety standards are due to uncertainty in the assessment. Therefore they suggest the use of a BBN as a first attempt to combine expert knowledge with evidence. Denney *et al.* [7] combine safety arguments formulated using GSN with confidence assessments using BBNs. To this end, the authors first define sources of uncertainty (e.g. aleatory ones such as uncertainty in sensor values or epistemic ones such as uncertainty with respect to implementation correctness) in the GSN tree. Each source of uncertainty is then associated with a leaf node in the BBN and quantified as a discrete prior probability distribution over five confidence states (very low, low, medium, high, very high). Each confidence state maps to an interval in the range [0,1]. This mapping allows for the integration of both quantitative and qualitative confidence data. Higher-level probabilities (i.e. those of non-leaf nodes in the BBN) are obtained by exploiting the conditional independence of lower-level probabilities and computing the joint probability distribution. The authors of [7] emphasize that quantifying confidence and selecting an appropriate prior distribution is problematic in the presence of merely subjective judgment. Due to its ability to incorporate both quantitative knowledge (in the form of evidence) and qualitative knowledge (in the form of a fine-grained mapping of likelihood and confidence levels to opinion triangles [14]), we believe that Subjective Logic provides a more appropriate framework for the purpose of assurance confidence modelling.

2.3 Representing uncertainty in binary metrics

In this paper, we are concerned with uncertainty associated with typical binary ML-related quality metrics. This does not imply that the underlying model is itself performing a binary task (e.g. binary classification or regression). Even in the case of a multiclass problem, quality metrics used as evidence in a safety argument are often still binary. Performance metrics such as accuracy, precision, recall as well as other metrics about model calibration, data coverage, etc. are often calculated as ratios between the number of successes (evidence in favor of the claim) and failures (evidence against the claim) and thus provide a probabilistic result over a binary domain, expressed as a percentage. Given a success probability p, the number k of successes in n trials is therefore binomially distributed according to the following formula:

$$\mathrm{Bin}(n, k, p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4}$$

In an inductive setting, p is generally unknown, so we are interested in the opposite problem: given r successes and s failures, what is the actual probability p of success? To this end, we can employ the *Beta distribution.* Let $\alpha = r + 1$ and $\beta = s + 1$. The probability density function of the probability of success can then be defined as follows:

$$\mathrm{Beta}(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \tag{5}$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}$ and $\Gamma(n) = (n-1)!$ is the gamma function that helps to ensure that the probability density function integrates to 1 over its defined range $[0, 1]$. The Beta distribution can thus be used to represent uncertainty in the probability of success of a single, specific binary measurement. Things get more complicated when multiple measurements need to be combined. Subjective logic described in the next section offers a convenient solution to this problem.

2.4 Subjective Logic (SL)

Subjective Logic (SL) [14] is a framework for artificial reasoning with uncertain beliefs. It combines ideas from probabilistic logic and evidence theory, in particular Dempster-Shafer theory (DST), and aims to address some of the well-known issues with the latter approach [26]. DST – mostly its combination operator – has often been criticised as producing counterintuitive results in certain situations of belief fusion [23]. However, as argued by Jøsang [18], this originates in the failure of correctly interpreting the nature of situations to be modelled. He argues that the DST combination operator is actually a method for fusing *belief constraints* and also produces intuitively correct results in that case. In order to allow for a more nuanced treatment of different types of belief combination and fusion situations, SL provides not just one but a wide range of operators with different semantics as described further below.

The atomic building blocks of SL are *subjective opinions* (or *beliefs*) and SL offers a wide range of combination operators that allow for algebraic reasoning. Some of the combination operators can be mapped to their binary logical equivalents (e.g. AND, OR, XOR, Modus Ponens, or Modus Tollens), some of them go beyond the capabilities of binary logic (e.g. belief fusion and trust transitivity). Due to the variety of operators, SL is particularly well-suited for the representation of *trust networks* [6, 15].

Subjective opinions in SL express beliefs about the truth of propositions under degrees of uncertainty. Opinions can be *binomial* (i.e., referring to a binary target frame $X = \{x, \bar{x}\}$), *multinomial* (i.e., referring to a target frame X of cardinality > 2 with singleton elements only), or *hypernomial* (i.e., referring to a target frame X of cardinality > 2 with elements $x \in \mathcal{R}(X)$[2]). We restrict the focus to binomial opinions in this paper.

Definition 2.1 (Binomial opinion)**.** Let $X = \{x, \bar{x}\}$ be a binary domain. A binomial opinion about the truth of X is a tuple $\omega_X = (b, d, u, a)$ where

- b (belief): the belief mass in support of x being ***true***
- d (disbelief): the belief mass in support of x being ***false***
- u (uncertainty): the uncommitted belief mass
- a (base rate): the *a priori* probability in the absence of committed belief mass (often set to 0.5 for binary domains).

The components have to satisfy $b, d, u, a \in [0, 1]$ and $b + d + u = 1$.

From evidence to opinions: Given a binary domain $X = \{x, \bar{x}\}$, let r denote the number of observations supporting x and let s denote the number of observations supporting its negation, i.e., $\bar{x}$. Let further $a = 0.5$ be the default base rate and $W = 2$ a *non-informative prior weight*[3]. The belief, disbelief, and uncertainty values can then be calculated as follows:

$$b_X = \frac{r}{r + s + W} \tag{6}$$

$$d_X = \frac{s}{r + s + W} \tag{7}$$

$$u_X = \frac{W}{r + s + W} \tag{8}$$

From opinions to Beta distributions and vice versa: Each binomial opinion corresponds with a Beta probability density function (PDF). The α and β parameters of the Beta distribution (see Equation 5) can be derived from the base rate a, the observation evidence r and s, and the non-informative prior weight W as follows [16]:

$$\alpha = r + aW, \qquad \beta = s + (1 - a)W \tag{9}$$

The expectation value $E(X)$ of the Beta PDF can then be calculated as follows:

$$E(X) = \frac{\alpha}{\alpha + \beta} = \frac{r + aW}{r + s + W} \tag{10}$$

Combining opinions: SL provides a wide range of combination operators [14], including, e.g, addition, subtraction, conjunction, disjunction, negation, multi-source belief fusion, deduction, abduction, Bayesian reasoning, and trust discounting and is thus significantly more versatile than DST. Combining opinions provides an elegant and intuitive way to combine the underlying Beta distributions, a direct manipulation of which would be significantly more complex. In this paper, we are primarily interested in *trust discounting*, a type of belief combination that tends to increase uncertainty.

In SL, various trust discounting operators have been proposed, among them uncertainty favouring trust transitivity (UFTT) [13], opposite belief favouring (OBF) [17], base rate sensitive discounting (BRSD) [17], and probability-sensitive discounting (PSD) [14]. OBF and BRSD take into account the beliefs that an agent has about the trustworthiness of other agents. As such, they are particularly suitable for modelling human trust and therefore not directly relevant for this work. Similar to BRSD, PSD takes into account the base rate for the calculation of belief and disbelief which is also not relevant here. As a consequence, we focus on UFTT in this paper.

Definition 2.2 (Trust discounting). Let $\omega^A_X = (b_X, d_X, u_X, a_X)$ be an opinion that agent A holds about a domain of interest $X = \{x, \bar{x}\}$. Further, let $\omega^B_A = (b_A, d_A, u_A, a_A)$ be an opinion that agent B holds about the trustworthiness of agent A. We can now calculate a combined opinion $\omega^{B;A}_X$ that discounts agent A's opinion about x by agent B's opinion about agent A's trustworthiness as follows:

$$b^{B;A}_X = b^B_A b^A_X \tag{11}$$

$$d^{B;A}_X = b^B_A d^A_X \tag{12}$$

$$u^{B;A}_X = d^B_A + u^B_A + b^B_A u^A_X \tag{13}$$

$$a^{B;A}_X = a^A_X \tag{14}$$

By using the symbol '⊗' to designate the trust discounting operator, we define $\omega^{B;A}_X \equiv w^B_A \otimes w^A_X$. The effect of trust discounting in a transitive chain is that uncertainty increases at the expense of belief and disbelief. The operator is associative but not commutative [13].

3 MODELLING ML METRICS IN SL

In order to assess the performance of an ML model, the quality of a data set, or the completeness and appropriateness of a specification, a wide range of both quantitative and qualitative metrics have been proposed. In order to incorporate the outcome of these metrics as *evidence* into a safety argument, it is important to represent them in a formal and uniform way. SL represents a framework that allows for the formulation and combination of both quantitative and qualitative insights. In this work, we limit our attention to quantitative metrics.

As described in Section 2.4, there are essentially two ways to construct an opinion in SL. We can either set the belief, disbelief, and uncertainty masses directly, if appropriate information is available. Or we can derive these masses from information about the underlying evidence (see Equations 6–8). Several of the basic ML performance metrics such as accuracy, precision, recall, F1 score, specificity, etc. are based on the confusion matrix by considering *true positive (TP)*, *true negative (TN)*, *false positive (FP)*, and *false negative (FN)* results and can thus be directly translated into opinions using the latter approach. Hacker and Seewig [12] have recently proposed an error detection approach for DNNs that allows for the calculation of scores in the [0,1] range for typical SOTIF[4]-related insufficiencies such as data completeness, data quality, model interpretability, model correctness, model robustness, and model uncertainty representation, which can also be turned into opinions in a similar way. However, also metrics aiming at other aspects of the model, e.g. the Brier score [3] for model calibration and data coverage metrics follow a similar scheme. Modelling the outcome of such metrics as opinions in SL is straightforward, as we illustrate in

[2] $\mathcal{R}(X)$ denotes the reduced powerset of X, i.e., the set of all subsets excluding the empty set $\emptyset$ and the full set X.

[3] The non-informative prior weight ensures that when evidence begins to accumulate, uncertainty decreases accordingly. W is typically set to the same value as the cardinality of the domain (2 in our binary case), thus artificially adding one "success" and one "failure". Higher values are equally possible but that would mean that more evidence is required for uncertainty to decrease. $W = 2$ is thus chosen to maintain a balance between belief, disbelief, and uncertainty in a way that is not overly sensitive to small amounts of evidence [13, 19]. Ultimately, the choice of W is application-specific

[4] ISO 21448: "Road vehicles - Safety of the intended functionality (SOTIF)"

the following paragraphs. It is important to note that our choice of metrics is by no means complete and a realistic network of metrics would have to be significantly more comprehensive.

Recall: Recall (also known as *true positive rate)* is a well-known metric that assesses the prediction performance of a classification model and is calculated as $TP/(TP+FN)$. Due to its specific focus on false negatives, recall is particularly relevant from a safety perspective. In order to formalise a recall measurement as a binomial opinion, we can consider $r = TP$ as positive evidence and $s = TP + FN$ as negative evidence and use Equations 6–8 to form opinion $\omega_{rec} = (b_{rec}, d_{rec}, u_{rec})$ as follows:

$$b_{rec} = \frac{TP}{TP+FN+W} \tag{15}$$

$$d_{rec} = \frac{FN}{TP+FN+W} \tag{16}$$

$$u_{rec} = 1 - b_{rec} - d_{rec} \tag{17}$$

Note that the recall metric does not take into account true negatives and false positives. This effectively reduces the amount of overall evidence and increases the influence of the non-informative prior weight W, resulting in higher uncertainty. Additional metrics derived from the confusion matrix such as accuracy, precision, or the F1 score can be formalised in a similar way.

Brier score: The Brier Score [3] quantifies the calibration of a model by measuring the mean squared difference between the predicted probabilities and the actual outcomes. It ranges from 0 to 1, with lower scores indicating better calibration. For binary measurements, it is calculated as follows:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2 \tag{18}$$

where N is the number of samples, f_i is the predicted probability of the positive class for sample i, and o_i is the actual outcome (0 or 1) for sample i. Intuitively, the Brier score quantifies the calibration error of a model and therefore the *disbelief* in the model being correctly calibrated. One straightforward option to turn this calculation into an opinion would thus be to use the Brier score directly as the disbelief mass and the inverse of the Brier score as the belief mass. This would result in an opinion $\omega_{BS} = (1 - BS, BS, 0)$. However, this approach does not reflect the uncertainty associated with the amount of evidence used for the calculation. A better alternative would thus be to 'break up' the Brier score formula and define the disbelief in the model accuracy as the sum of squared errors, divided by the total amount of evidence plus the non-informative prior weight, resulting in the following opinion components:

$$b_{BS} = 1 - d_{BS} - u_{BS} \tag{19}$$

$$d_{BS} = \frac{\sum_{i=1}^{N}(f_i - o_i)^2}{N+W} \tag{20}$$

$$u_{BS} = \frac{W}{N+W} \tag{21}$$

<table>
<tr><td>Daytime</td><td colspan="2">morning</td><td colspan="2">day</td><td colspan="3">evening</td><td>night</td></tr>
<tr><td>Haze/fog</td><td colspan="4">no</td><td colspan="4">yes</td></tr>
<tr><td>Street condition</td><td>dry</td><td colspan="2">wet</td><td colspan="2">icy</td><td>snow</td><td colspan="2">broken</td></tr>
<tr><td>Sky</td><td colspan="3">cloudy</td><td colspan="3">clear</td><td colspan="2">no</td></tr>
<tr><td>Rain</td><td colspan="4">no</td><td colspan="4">yes</td></tr>
<tr><td>Reflection on road</td><td colspan="4">no</td><td colspan="4">yes</td></tr>
<tr><td>Shadow on road</td><td colspan="4">no</td><td colspan="4">yes</td></tr>
</table>

Table 1: Example ODD specification (adapted from [9])

Dataset coverage: Ensuring good dataset coverage with respect to the underlying specification of the input space (often referred to as *operational design domain (ODD)* in automotive applications) that the ML model is supposed to operate in is critically important to ensure the safety of the resulting system. ODD specifications are often given as a range of *dimensions*, each with its specific *aspects* as shown in Table 1. One way to ensure coverage is to employ *combinatorial testing* [9]. Here, t-wise combinations of aspects (= equivalence classes) across dimensions are constructed and a coverage metric can be obtained. As noted by [9], the approach can also be used to rate an existing dataset w.r.t. its coverage of an abstract specification. It is easy to see how such a metric can be turned into an opinion, similar to the other metrics above. Let N be the total number of combinations and let C the number of covered cases. The opinion components can then be defined as follows:

$$b_{DC} = \frac{C}{N+W} \tag{22}$$

$$d_{DC} = 1 - d_{DC} - u_{DC} \tag{23}$$

$$u_{DC} = \frac{W}{N+W} \tag{24}$$

It is debatable whether the non-informative prior weight W should be included in that case. If there is a fair amount of certainty regarding the choice of dimensions and aspects, W may be omitted. This has to be decided based on the nature of the specific use case[5].

Numeric examples of some of the calculations above will be given as part of the example analysis in Section 5. In the next section, we discuss how measurement results formulated as opinions can be combined in order to study uncertainty propagation.

4 COMBINING ML METRICS IN SL

ML performance metrics such as the ones described in the previous are generally used independently to assess different aspects of a trained model. Good results for these metrics are expected to increase trust in the capabilities of the model. However, as described in Section 1, the quality of a model is always dependent upon the quality of the data it has been trained and tested on which, in turn, is dependent upon the quality of the specification of the input space and task. The metrics described in the previous sections can thus be associated with different layers of the uncertainty hierarchy shown in Figure 1. As a consequence, the confidence in an asserted 'primary' evidence on the model layer is a function of the confidence in certain further, 'secondary' evidences on the model, data, and

[5] A subdivision of the overall ODD into μODDs as suggested by Koopman *et al.* [22] might be helpful here.

input space levels and cannot be obtained in isolation. The purpose of this section is to illustrate how Subjective Logic can be used to combine different opinions to model and quantify uncertainty propagation across different metrics.

Let ω^A_{rec} be a binomial opinion that a hypothetical agent A (for notation purposes) formed about the recall of binary classification model M. Furthermore, let ω^B_{bs} be a binomial opinion that a hypothetical agent B formed about the calibration of M. A badly calibrated classification model can impact all values derived from the confusion matrix. Bad calibration may result in misclassifications, leading to incorrect counts of true positives and false negatives in the confusion matrix, and may thus bias recall. B's knowledge about the calibration of M can be intuitively understood as the level of trust that B has in A's opinion. By combining A's and B's measurements appropriately using the trust discounting operator introduced in Section 2.4, we can thus formulate a refined opinion that represent A's recall measurement discounted by B's opinion about the model calibration according to Equations 12–14 as $\omega^{B;A}_{rec} = \omega^B_{bs} \otimes \omega^A_{rec}$. This new opinion provides a more refined view on the original recall measurement since it incorporates not just the uncertainty resulting from the recall measurement itself but also the uncertainty associated with the Brier score measurement as well as the influence of the latter on the former. This aspect is illustrated in more detail using a concrete example in Section 5.

The process does not have to stop here and opinion $\omega^{B;A}_{rec}$ can be further discounted by knowledge about factors on the same or higher layers, resulting in arbitrarily complex transitive trust chains across multiple layers in Figure 1. This allows for the modelling of uncertainty propagation and the calculation of overall residual uncertainty in an flexible manner, as exemplified in the next section.

5 APPLICATION TO TRAFFIC SIGN CLASSIFICATION

We will now illustrate the concepts introduced in the previous two sections using a realistic safety-critical example and demonstrate the modelling of uncertainty propagation of a performance measurement across the layers in Figure 1. To this end, we consider the problem of ML-based Traffic Sign Recognition (TSR) as an example function whose failure may have safety-relevant consequences and therefore requires safety assurance. We assume that the TSR is used as part of a highway pilot function that is enabled to automatically control the vehicle under certain conditions on a highway and should adjust its driving strategy based on prompts it receives from roadside traffic signs. We are particularly concerned with the recognition of *construction site signs* so that the Highway Pilot can be deactivated and control passed back to the driver. We therefore aim to ensure that as few construction signs as possible are missed.

We trained a deep neural network on the German Traffic Sign Recognition Benchmark (GTSRB) dataset[6]. The training set contains 51,840 images, subdivided into 43 classes. Class 25 represents construction signs and contains 1,500 samples. The test dataset contains 12,629 images, 480 of which belong to the class of construction signs. We trained a ResNet model of depth 18 on the dataset with 38 epochs, a drop rate of 0.3, Adam optimiser, an initial learning rate of 0.001, scheduler step size of 20, and scheduler decay rate of 0.5. The model achieves an overall test accuracy of 98% and, specifically for the construction sign class, 97% precision and 98% recall.

From a safety perspective, recall is a particularly crucial metric as its inverse expresses the fraction of misclassifications. Since missing a construction sign may lead to a hazardous situation, both a high recall rate and a high level of confidence in the recall rate are essential for safety. As for the first point, 98% is a good value but certainly not sufficient. However, the focus of this paper is on the second point, i.e. on the question how much this value can be trusted. To this end, we need to factor in the hidden influencing factors. In reality, there are many of such influencing factors associated with the three layers mentioned above. For simplicity, we restrict the focus in the following analysis to the following aspects:

1. The calibration of the model (→ *Model* layer in Figure 1).
2. The amount of evidence used for classification (→ relationship between *Model* and *Data* layer).
3. The coverage of the dataset w.r.t. the specification (→ relationship between *Data* and *Input space* layer).

We will now address these different points and examine how they affect the overall uncertainty in the accuracy metric.

Step 1: Modelling recall as an opinion: We start by forming an opinion about the recall measurement by incorporating the evidence as described in Equations 15–17. An analysis of the confusion matrix of our model yields the following results:

- Num. of true positives (TP): 470
- Num. of false positives (FP): 10
- Num. of true negatives (TN): 12,149
- Num. of false negatives (FN): 10

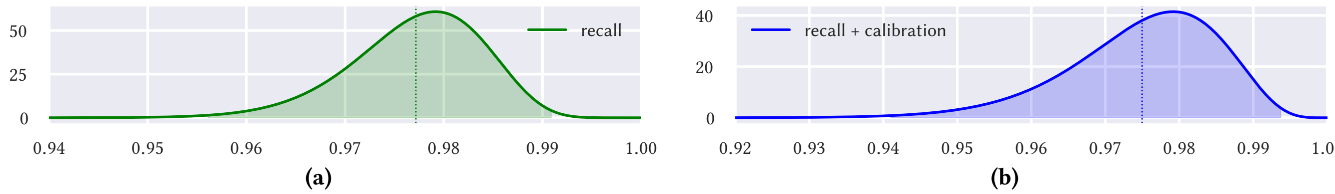
Given that information, we apply Equations 15–17 and calculate the following opinion (we omit the base rate of 0.5 from the opinion since it will not be relevant for the analysis). Note that we again introduce a hypothetical observer agent A for notation purposes.

$$\omega^A_{rec} = (0.975, 0.021, 0.004) \tag{25}$$

We see that the belief mass $b^A_{rec} = 0.975$ is slightly lower than the original recall value of 0.98 and the initial uncertainty is set to 0.004. This is due to the influence of the non-informative prior weight $W = 2$ in Equations 6 and 7 and reflects uncertainty resulting from the comparatively small amount of evidence of 480 data points for the construction sign class. We have thus already implicitly addressed point 2 (amount of evidence) in the list above: had the amount of evidence been larger, the resulting uncertainty would have been smaller. A visualisation of the Beta distribution associated with ω^A_{rec} is shown in Figure 2 (a). The 95% confidence interval ranges from 0.956 to 0.991. From a safety perspective, it might thus be advisable to assume a more conservative value of 95.6% for our recall metric instead of 98%.

Step 2: Including model calibration: As the next step, we address the quality of the confidence that the model has in its own predictions using the Brier Score described in Section 3 and discount the opinion formed above appropriately. Since we have a multi-class classification problem but only focus on the class of construction signs, we need to make sure we only sum up the squared errors in classifications of images that belong to our focus class (TP and

[6] https://benchmark.ini.rub.de/index.html



(a)

(b)

Figure 2: Beta PDF of the recall opinion for the traffic sign classification example (a) and Beta PDF of recall discounted by model calibration (b) including confidence interval (shaded area) and expectation value (vertical dotted line).

.

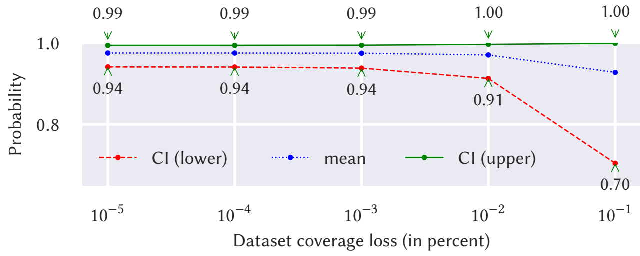


Figure 3: Sensitivity analysis for dataset coverage: upper bound, mean, and lower bound of the 95% confidence interval of the resulting Beta distributions.

FN) and ignore FP and TN resulting from the classification of non-construction sign images. This is important in our case since only the errors in TP and FN will impact the recall score. The resulting measured Brier score for class 25 is 0.004, indicating fairly good calibration. Following Equations 19–21, we model the disbelief mass d_{bs} as the sum of squared errors of TP and FN classifications which amounts to 2.148 for our focus class of construction signs and the belief mass as $1-d_{bs}$. The uncertainty mass is calculated by dividing the prior weight $W = 2$ by the sum of all construction sign input images (= 480) plus W. The resulting opinion ω_{bs}^{B} is shown below:

$$\omega_{bs}^{B} = (0.995, 0.004, 0.0002) \tag{26}$$

Using the trust discounting operator, we can now calculate a discounted opinion $w_{rec}^{B;A}$ that expresses agent B's view on A's recall measurement, taking into account its own opinion about model calibration:

$$\omega_{rec}^{B;A} = \omega_{bs}^{B} \otimes \omega_{rec}^{A} = (0.971, 0.021, 0.01) \tag{27}$$

The resulting Beta PDF is shown in Figure 2 (b). We see that calibration errors have introduced an additional amount of uncertainty into the discounted recall measurement and the confidence interval now ranges from 0.94 to 0.994. Our safety-aware conservative estimate of the actual recall should thus be further reduced to 94%.

Step 3: Including data coverage: Incorporating data coverage works in the same way as the integration of calibration just described. We assume a hypothetical agent C that performs a data coverage measurement as described in the end of Section 3, resulting in an opinion ω_{DC}^{C}. This opinion can then be used to further discount the previously computed opinion $\omega_{rec}^{B;A}$ as follows:

$$\omega_{rec}^{C;B;A} = \omega_{DC}^{C} \otimes \omega_{rec}^{B;A} \tag{28}$$

In order to illustrate how data coverage further impacts the previously computed opinion $\omega_{rec}^{B;A}$, we do not assume a single measurement for data coverage like we did in the previous two steps but instead perform a sensitivity analysis and investigate the relationship between different values for data coverage and the size of the confidence interval of the resulting final opinion $\omega_{rec}^{C;B;A}$. We start with an (unrealistically) high assumed data coverage value of 0.99999 and reduce it down by several orders of magnitude to 0.9 and inspect how the upper bound, mean, and lower bound values of the 95% confidence intervals vary. The results are shown in Figure 3. The x axis shows the *data coverage loss*, i.e. the inverse of data coverage, the y axis shows the respective probability value. We can see that the narrowest confidence interval ranges from 0.94 to 1.0 for (unrealistically) high levels of data coverage but widens quickly, even for large data coverage values of ≥ 99%.

Summary: The resulting opinion $\omega_{rec}^{C;B;A}$ represents a combined measurement that quantifies the uncertainty of our primary recall metric by taking into account aspects of model calibration and data coverage. For each of the measurements, the influence of sample size is also considered. The example illustrates that the initially measured recall value of 98% is subject to a significant uncertainty margin that needs to be taken into account when viewing the perception capabilities of the underlying ML model from a holistic safety perspective. Incorporating knowledge about secondary aspects such as sample size, model calibration, and data coverage into the original recall measurement reveals that, from a safety perspective, a much more conservative value of at most 94% should be used to estimate the *actual* performance of the model.

6 DISCUSSION AND CONCLUSIONS

In the ML community, performance metrics such as accuracy, precision, or recall are generally used as *primary evidence* to assess the quality of a model. However, from a safety perspective, such point estimates are only partially meaningful since their trustworthiness depends on other secondary factors such as sample size, the calibration of the model, the quality of the dataset, or the appropriateness of the specification of the operational domain. In order to obtain confidence in any piece of primary evidence, further arguments about secondary aspects of model, dataset, or specification quality and completeness are thus required.

In this paper, we proposed a formal framework based on Subjective Logic (SL) that aims to quantify the uncertainty associated with primary evidence by combining them with other, secondary, pieces of evidence into an overall assessment of risk represented as a probability distribution with clearly measurable properties such as variance, standard distribution, or confidence interval. We illustrate the approach by modelling the relationship between a commonly used base metric (recall) and additional properties about sample size, model calibration, and dataset coverage and show how they impact the uncertainty in our primary evidence. Our experiments show that, from a safety perspective, single point estimates for base metrics such as recall draw an overly optimistic picture of model performance that changes to the worse if other aspects of the model such as sample size, calibration, or dataset coverage are taken into account. As a consequence, from a safety perspective, significantly more conservative estimates should be assumed.

Our work aims to be a starting point towards a more formal treatment of uncertainty in the safety assurance of ML and, as such, is necessarily preliminary and incomplete. While we focussed on less than a handful of commonly used types of quantitative evidence, a more thorough investigation of relevant metrics and their relationship is required. Second, we aim to study how other types of evidence that cannot be easily described as ratios can be effectively turned into opinions in SL. Even more challenging, many important types of evidence (e.g. in the context of explainability) are purely qualitative in nature; whilst SL generally supports the integration of qualitative insights [14], this problem requires further investigation. Third, we just focussed on modelling relationships between metrics by means of trust discounting, i.e. relationships that tend to *increase* uncertainty. However, it is also possible to combine a base metric with a secondary piece of evidence that helps to *strengthen* the former, i.e. to *reduce* uncertainty in it. An example could be to complement an accuracy measurement a with further insights about the robustness of the model which may be used to "argue away" some of the residual uncertainty in a. SL provides useful fusion operators for such situations and we aim to investigate this direction further as part of our future work. And finally, while the resulting probability distributions represent a quantification of 'subjective trust' according to the semantics of the trust discounting operator, further (empirical) analyses are required to assess how well they capture the actual statistical uncertainty of the underlying primary evidence.

ACKNOWLEDGMENTS

We thank Balahari Balu and João Vitor Zacchi of Fraunhofer IKS for helpful comments and discussions. We also thank the anonymous referees for their valuable comments and helpful suggestions. This work was performed as part of the ML4Safety project supported by the Fraunhofer Internal Programs under Grant No. PREPARE 40-02702.

REFERENCES

[1] Algirdas Avizienis, J-C Laprie, Brian Randell, and Carl Landwehr. 2004. Basic concepts and taxonomy of dependable and secure computing. *IEEE transactions on dependable and secure computing* 1, 1 (2004), 11–33.

[2] Anaheed Ayoub, Jian Chang, Oleg Sokolsky, and Insup Lee. February 2013. Assessing the overall sufficiency of safety arguments. *21st Safety-Critical Systems Symposium (SSS'13)* (February 2013), 127–144.

[3] Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review* 78, 1 (1950), 1–3.

[4] Simon Burton, Lydia Gauerhof, Bibhuti Bhusan Sethy, Ibrahim Habli, and Richard Hawkins. 2019. Confidence Arguments for Evidence of Performance in Machine Learning for Highly Automated Driving Functions. In *Computer Safety, Reliability, and Security*, Alexander Romanovsky, Elena Troubitsyna, Ilir Gashi, Erwin Schoitsch, and Friedemann Bitsch (Eds.). Springer International Publishing, Cham, 365–377.

[5] Simon Burton and Benjamin Herd. 2023. Addressing uncertainty in the safety assurance of machine-learning. *Frontiers in Computer Science* 5 (2023). https://doi.org/10.3389/fcomp.2023.1132580

[6] Tong Cheng, Guangchi Liu, Qing Yang, and Jianguo Sun. 2019. Trust assessment in vehicular social network based on three-valued subjective logic. *IEEE Transactions on Multimedia* 21, 3 (2019), 652–663.

[7] Ewen Denney, Ganesh Pai, and Ibrahim Habli. 2011. Towards measurement of confidence in safety cases. In *2011 International Symposium on Empirical Software Engineering and Measurement*. IEEE, 380–383.

[8] Lian Duan, Sanjai Rayadurgam, Mats PE Heimdahl, Anaheed Ayoub, Oleg Sokolsky, and Insup Lee. 2017. Reasoning about confidence and uncertainty in assurance cases: A survey. In *Software Engineering in Health Care: 4th International Symposium, FHIES 2014, and 6th International Workshop, SEHC 2014, Washington, DC, USA, July 17-18, 2014, Revised Selected Papers 4*. Springer, 64–80.

[9] Christoph Gladisch, Christian Heinzemann, Martin Herrmann, and Matthias Woehrle. 2020. Leveraging combinatorial testing for safety-critical computer vision datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 324–325.

[10] John B Goodenough, Charles B Weinstock, and Ari Z Klein. 2013. Eliminative induction: A basis for arguing system confidence. In *2013 35th International Conference on Software Engineering (ICSE)*. IEEE, 1161–1164.

[11] Baofeng Guo. 2003. Knowledge representation and uncertainty management: applying Bayesian belief networks to a safety assessment expert system. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*. IEEE, 114–119.

[12] Lukas Hacker and Jörg Seewig. 2023. Insufficiency-Driven DNN Error Detection in the Context of SOTIF on Traffic Sign Recognition Use Case. *IEEE Open Journal of Intelligent Transportation Systems* 4 (2023), 58–70.

[13] Audun Jøsang. 2001. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 9, 03 (2001), 279–311.

[14] Audun Jøsang. 2016. *Subjective logic*. Vol. 3. Springer.

[15] Audun Jøsang and Touhid Bhuiyan. 2008. Optimal trust network analysis with Subjective Logic. In *2008 Second International Conference on Emerging Security Information, Systems and Technologies*. IEEE, 179–184.

[16] Audun Jøsang, Jin-Hee Cho, and Feng Chen. 2018. Uncertainty characteristics of subjective opinions. In *2018 21st International Conference on Information Fusion (FUSION)*. IEEE, 1998–2005.

[17] Audun Jøsang, Stephen Marsh, and Simon Pope. 2006. Exploring different types of trust propagation. In *International Conference on Trust Management*. Springer, 179–192.

[18] Audun Jøsang and Simon Pope. 2012. Dempster's rule as seen by little colored balls. *Computational Intelligence* 28, 4 (2012), 453–474.

[19] Audun Jøsang, Dongxia Wang, and Jie Zhang. 2017. Multi-source fusion in Subjective Logic. In *2017 20th International Conference on Information Fusion (Fusion)*. 1–8. https://doi.org/10.23919/ICIF.2017.8009820

[20] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. 2022. Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)* 55, 2 (2022), 1–38.

[21] Tim Kelly and Rob Weaver. 2004. The Goal Structuring Notation – a safety argument notation. In *Proceedings of the dependable systems and networks 2004 workshop on assurance cases*, Vol. 6. Citeseer.

[22] Philip Koopman, Beth Osyk, and Jack Weast. 2019. Autonomous vehicles meet the physical world: RSS, variability, uncertainty, and proving safety. In *Computer Safety, Reliability, and Security: 38th International Conference, SAFECOMP 2019, Turku, Finland, September 11–13, 2019, Proceedings 38*. Springer, 245–253.

[23] Judea Pearl. 1990. Reasoning with belief functions: An analysis of compatibility. *International Journal of Approximate Reasoning* 4, 5-6 (1990), 363–389.

[24] Glenn Shafer. 1976. *A mathematical theory of evidence*. Vol. 42. Princeton University Press.

[25] Rui Wang, Jérémie Guiochet, Gilles Motet, and Walter Schön. 2019. Safety case confidence propagation based on Dempster–Shafer theory. *International Journal of Approximate Reasoning* 107 (2019), 46–64.

[26] Lotfi A Zadeh. 1984. Review of a mathematical theory of evidence. *AI magazine* 5, 3 (1984), 81–81.