Interpretable confidence measures for decision support systems

Waa, Jasper van der; Schoonderwoerd, Tjeerd; Diggelen, Jurriaan van; Neerincx, Mark

DOI
10.1016/j.ijhcs.2020.102493

Publication date
2020

Document Version
Final published version

Published in
International Journal of Human Computer Studies

International Journal of Human-Computer Studies 144 (2020) 102493

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Interpretable confidence measures for decision support systems

Jasper van der Waa[*,a,b], Tjeerd Schoonderwoerd[a], Jurriaan van Diggelen[a], Mark Neerincx[a,b]

[a] TNO, Soesterberg, Kampweg 55, the Netherlands
[b] Technical University of Delft, Delft, Mekelweg 5, the Netherlands

ARTICLE INFO

Keywords:
Machine learning
Decision support systems
Confidence
Explainable AI
Artificial intelligence
Transparency
Interpretable
User study
Interpretable machine learning
Trust calibration

2018 MSC:
00-01
99-00

ABSTRACT

Decision support systems (DSS) have improved significantly but are more complex due to recent advances in Artificial Intelligence. Current XAI methods generate explanations on model behaviour to facilitate a user's understanding, which incites trust in the DSS. However, little focus has been on the development of methods that establish and convey a system's confidence in the advice that it provides. This paper presents a framework for Interpretable Confidence Measures (ICMs). We investigate what properties of a confidence measure are desirable and why, and how an ICM is interpreted by users. In several data sets and user experiments, we evaluate these ideas. The presented framework defines four properties: 1) accuracy or soundness, 2) transparency, 3) explainability and 4) predictability. These characteristics are realized by a case-based reasoning approach to confidence estimation. Example ICMs are proposed for -and evaluated on- multiple data sets. In addition, ICM was evaluated by performing two user experiments. The results show that ICM can be as accurate as other confidence measures, while behaving in a more predictable manner. Also, ICM's underlying idea of case-based reasoning enables generating explanations about the computation of the confidence value, and facilitates user's understandability of the algorithm.

1. Introduction

The successes in Artificial Intelligence (AI), Machine Learning (ML) in particular, caused a boost in the accuracy and application of intelligent decision support systems (DSS). They are used in lifestyle management (Wu et al., 2017), management decisions (Bose and Mahapatra, 2001), genetics (Libbrecht and Noble, 2015), national security (Pita et al., 2011), and in prevention of environmental disasters in the maritime domain (van Diggelen et al., 2017). In these high-risk domains, a DSS could be beneficial as it can reduce workload of a user, and increase task performance. However, the complexity of current DSS (e.g. based on Deep Learning) impedes users' understanding of a given advice, often resulting in too much or too little trust in the system, which can have catastrophic consequences (Burrell, 2016; Cabitza et al., 2017).

The field of Explainable AI (XAI) researches how a DSS can improve a user's understanding of the system by generating explanations about its behaviour (Guidotti et al., 2018; Kim et al., 2015; Miller, 2018b; Miller et al., 2017; Ridgeway et al., 1998). More specifically, the goal of these explanations is to increase understanding of the system's rationale and certainty of an advice that it provides (Holzinger et al., 2019a; 2019b; Miller, 2018a). It is hypothesized that the understanding that a user gains from these explanations facilitates adequate use of the DSS (Hoffman et al., 2018), and calibrates the user's trust in the system (Cohen et al., 1998; Fitzhugh et al., 2011; Hoffman et al., 2013).

Although understanding of the system can help a user to decide when to follow the advice of a DSS, it is often overlooked that a confidence measure can achieve the same effect (Papadopoulos et al., 2001). In this paper, we define a *confidence measure* as a measure that provides an expectation that an advice will prove to be correct (or incorrect). To help develop such measures, we introduce the Interpretable Confidence Measure (ICM) framework. The ICM framework assumes that a confidence measure should be 1) *accurate*, 2) able to *explain* a single confidence value, 3) use a *transparent* algorithm and 4) providing confidence values that are *predictable* for humans (see Fig. 1).

To illustrate the ICM framework, we will define an example ICM. We evaluated its accuracy, robustness and genericity on several classification tasks with different machine learning models. In addition, we applied the concept of an ICM on the use case of Dynamic Positioning (DP) within the maritime domain (van Diggelen et al., 2017). Here, a human operator supervises a ship's auto-pilot while receiving assistance from a DSS that provides a warning when human intervention is deemed necessary (e.g. based on weather conditions). It can be catastrophic if the operator fails to intervene in time. For example, an oil

* Corresponding author at: TNO, Soesterberg, Kampweg 55, the Netherlands.
E-mail address: jasper.vanderwaa@tno.nl (J.v.d. Waa).

https://doi.org/10.1016/j.ijhcs.2020.102493
Received 20 September 2019; Received in revised form 26 May 2020; Accepted 6 June 2020
Available online 09 June 2020

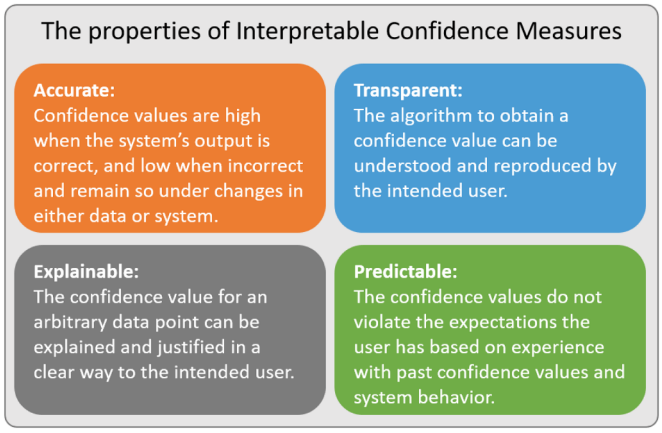


Fig. 1. The four properties of an Interpretable Confidence Measure to perform effective trust calibration.

tanker might spill large amounts of oil in the ocean, because the operator failed to intervene to prevent the ship from rupturing its connection to an oil rig. This use case provided a realistic dataset to evaluate our example ICM, as well as a context for a qualitative usability experiment with these operators. In this experiment, we evaluated the transparency and explainability properties of the ICM framework. To further substantiate these results, we performed a quantitative online user experiment in the context of self-driving cars.

We provide the ICM framework in Section 3, describe our example ICM in Section 4, our evaluations on the data sets in Section 4.1, and the two user experiments in Section 5 and 6. The next Section presents related work in the field of XAI and confidence measures in Machine Learning, which defines many current DSS.

2. Related work

Explainable AI (XAI) researches how we can improve the user's understanding in a DSS to reach an appropriate level of trust in its advice (Herman, 2017; Kim et al., 2015; Miller, 2018b; Miller et al., 2017; Ridgeway et al., 1998). For example by allowing users to detect biases (Doshi-Velez and Kim, 2017; Gilpin et al., 2018; Goodman and Flaxman, 2016; Zhou and Chen, 2018). Some XAI research focuses on these aspects from a societal perspective, trying to identify how intelligent systems should be implemented, when they should be used, and who should regulate them (Doshi-Velez and Kim, 2017; Lipton, 2016; Zhou and Chen, 2018; Zliobaite, 2015). Other researchers approach the field from a methodological perspective, and aim to develop methods that solve the potential issues of applying intelligent systems in society. See for example the overview of methods from Guidotti et al. (2018).

To generate explanations, many XAI methods use a meta-model that describes the actual system's behaviour in a limited input space surrounding the to be explained data point (Ribeiro et al., 2016). It only has to be accurate in this local space and can thus be less complex and more explainable than the actual system. A disadvantage of these approaches is that the meaningfulness of the explanation is dependent on the size of the local space and the brittleness of the used meta-model. When it is too small, the explanation cannot be generalized, and when it is too large, the explanation may lack fidelity. The advantage is that these methods can be applied to most systems (i.e. they are system- or model-agnostic). A second advantage is that the fidelity of explanations can be measured, since the meta-model's ground truth is the output of the system, which is readily available. This can be exploited to measure a meta-model's accuracy through data perturbation. In our proposed ICM framework, we apply the idea of system-agnostic local meta-models to obtain an interpretable confidence measure, not a post-hoc explanation of an output.

Confidence measures allow DSS to convey when an advice is trustworthy (Papadopoulos et al., 2001). However, a user's commitment to follow a DSS' advice is linked to his or her own confidence and that conveyed by the DSS (Landsbergen et al., 1997). A confident user confronted with a low system confidence reduces the user's confidence in his or herself, and vice versa. The work from Ye and Johnson (1995) and Waterman (1986) shows that this can be mitigated by explaining the DSS' confidence value by using a transparent algorithm. The work from Walley (1996) shows users tend to change their confidence when evidence for a correct or incorrect decision is gained or lost. Users expect the same predictable behaviour from a DSS' confidence measure. Hence, it should not only be transparent with explainable values but also behave predictable for humans.

Current DSS are often based on Machine Learning (ML). Different categories of confidence measures can be identified from this field, see Table 1 for an overview. The first, *confusion metrics* such as accuracy and the F1-score, are based on the confusion matrix. These tend to be transparent and predictable but lack accuracy and explainability for conveying the confidence of a single advice (Foody, 2005; Labatut and Cherifi, 2011). A ML model's *prediction score* such as the SoftMax output of a Neural Network, are also common as confidence measures. They represent the model's estimated likelihood for a certain prediction (Zaragoza and d'Alché Buc, 1998). They are highly accurate but their transparency and explainability is often low (Samek et al., 2017; Sturm et al., 2016). Furthermore, these measures tend to behave unpredictable as small changes in a data point can cause non-monotonic increases or decreases in the confidence value (Goodfellow et al., 2014; Nguyen et al., 2015). In *rescaling* such as with Platt Scaling (Platt and others, 1999) or Isotonic Regression (Zadrozny and Elkan, 2001; 2002), the prediction scores are translated into more predictable and accurate values (Hao et al., 2003; Liu et al., 2004). However, these are used to enable post-processing and not intended to be explainable or transparent (Niculescu-Mizil and Caruana, 2005). Some ML models are inherently *probabilistic* and output conditional probability distributions over its predictions. Examples are Naive Bayes (Rish et al., 2001), the Relevance Vector Machine (Tipping, 2000) and using neuron dropout (Gal and Ghahramani, 2016) or Bayesian inference (Fortunato et al., 2017; Graves, 2011; Paisley et al., 2012) on trained Neural Networks. Although they are accurate, they are also opaque and difficult to predict as conditional probabilities are difficult to comprehend by humans

Table 1
Categories and examples of commonly used confidence measures in Machine Learning and if their adherence to the four properties of an ICM.

	Category				
Property	Confusion metrics	Prediction scores	Rescaling	Probability	Voting
Accurate	-	+	+	+	+
Predictable	+	-	+	-	-
Transparent	+	-	-	-	-
Explainable	-	-	-	+	+
Example	F1-score Foody (2005)	SoftMax Papernot and McDaniel (2018)	Platt Scaling Platt and others (1999)	RVM Tipping (2000)	Random Forest Bhattacharyya (2013)

(Evans et al., 2003; Pollatsek et al., 1987). There are efforts to make such values more explainable for specific model types, see for example (Qin, 2006) and (Ridgeway et al., 1998). Finally, ML models are known to use voting to arrive at a confidence value (Polikar, 2006; Tóth and Pataki, 2008; Van Erp et al., 2002). Known examples are Random Forest, Decision Trees and ensembles of Decision Stumps (Stone and Veloso, 1997). These confidence values can be explained through examples (Florez-Lopez and Ramon-Jeronimo, 2015). However, their algorithmic transparency depends on the model and their values tend to change step-wise given continuous changes to the input, making them hard to predict by humans.

As can be seen in Table 1, neither category is accurate, predictable, explainable and transparent in a DSS context. A likely reason is that the purpose of these measures is to convey performance of an ML model to a developer, not the confidence of a DSS in an advice to a user. As a consequence, many of these measures are tailored to work for a specific or subset of model types. Only the confusion metrics of these categories are system-agnostic. In the next section we propose a system agnostic approach to confidence measures based on case-based reasoning that are not only as accurate as the above described measures, but also transparent, explainable and predictable.

3. A framework for interpretable confidence measures

In this section we propose a framework to create Interpretable Confidence Measures (ICM) that are not only accurate in their confidence assessment, but whose values are predictable as well as explainable based on a transparent algorithm. The ICM framework relies on a system-agnostic approach and performs a regression analysis with the correctness of an advice as the regressor. It does so based on case-based reasoning.

Case-based reasoning or learning provides a prediction by extrapolating labels of past cases to the current queried case (Atkeson et al., 1997). The basis of many case-based reasoning methods is the *k*-Nearest Neighbours (*k*NN) algorithm (Fix and Hodges Jr, 1951). This method follows a purely lazy approach (Wettschereck et al., 1997). When queried with a novel case, it selects the k most similar cases from a stored data set and assigns the case with a weighted aggregation of the neighbour's labels. The advantage of case-based learning methods is that its principle idea is closely related to that of human decision-making (Harteis and Billett, 2013; Hodgkinson et al., 2008; Schank et al., 2014). This makes such algorithms easier to understand and interpret (Freitas, 2014). In addition, they allow for example-based explanations of a single prediction (Doyle et al., 2003). These properties are exploited in the ICM framework to define a confidence measure as performing a regression analysis with case-based reasoning.

3.1. The ICM framework

In this section we formally describe the ICM framework. We assume the DSS as a function $f: \mathbb{R}^l \rightarrow \Upsilon$ that assigns an advice $y \in \Upsilon$ to data points $\vec{x}$ of l dimensions. It does this with a certain accuracy relative to the ground truth or label $y^* \in \Upsilon$. An ICM goes through four steps to define the confidence value $C(\vec{x})$ for $\vec{x}$: 1) an *update* step, 2) a *selection* step, 3) a *separation* step, and 4) a *computation* step. Below we discuss these steps, and an overview is shown in Fig. 2.

In the first step, the *update*, a memory $D = \{(\vec{x}_1, y_1^*), ..., (\vec{x}_n, y_n^*)\}$ is updated. This D forms the set of cases from which the confidence is computed. Given an update procedure u and new data-label pairs $(\vec{x}, y^*)$, an ICM continuously updates this memory $D' = u((\vec{x}, y^*), D)$ such that $|D| = n$. This ensures that D adapts to changes in the DSS over time. The initial D is initialized with a training set but is expanded and replaced with novel pairs during DSS usage. The size of D is fixed to n, and maintained by u. Examples of u can be as simple as a queue (newest in, oldest out) or based on more complex sampling methods (e.g. those that take the label and data distributions into account).

In the *selection step* a set S is sampled from D such that $S = s(\vec{x}, y|D)$, where s is some selection procedure. The purpose of s is to select all relevant data-label pairs to define the ICM's confidence value for the current $(\vec{x}, y)$. For example, following *k*NN, the k closest neighbours to $\vec{x}$ can be selected based on a similarity or distance function.

In the *separation step, S* is split into S^+ and S^- based on the current $(\vec{x}, y)$. The S^+ contains all $(\vec{x}, y^*)$ where $y = y^*$, with $S^- = S \setminus S^+$. In other words, S^+ contains all data points whose advice was similar to the current advice and correct. The S^- contains all data points with a different correct advice.

In the *computation step*, the S^+ and S^- are used to calculate the confidence value $C(\vec{x}|S^+, S^-)$ with a weighting scheme $w: \mathbb{R}^l \rightarrow \mathbb{R}$ (often abbreviated as $C(\vec{x})$):

$$C(\vec{x}|S^+, S^-) = Z(\vec{x}|S)^{-1} \sum_{\vec{x}_i \in S^+} w(\vec{x}, \vec{x}_i) - \sum_{\vec{x}_j \in S^-} w(\vec{x}, \vec{x}_j) \tag{1}$$

The weights w represent how much a data point in S^+ or S^- influences the confidence of the advice for $\vec{x}$. Again, taking *k*NN as an example, the w can simply contain a delta-function to 'count' the number of points in S^+ and S^-. Although, more complex weighting schemes are possible and advised. The Z^{-1} is a normalization factor:

$$Z(x)^{-1} = \frac{1}{\sum_{\vec{x}_i \in S} w(\vec{x}, \vec{x}_i)} \tag{2}$$

This ensures that the confidence value is bounded; $C \in [-1, 1]$, with -1 and $+1$ denoting the confidence that some y would prove to be incorrect or correct respectively. Intermediate values represent the surplus of available evidence for a correct or incorrect advice relative to all available evidence. For example, when $C(\vec{x}) = -0.5$ there is 50% surplus evidence that the advice y will be incorrect, relative to all available evidence. What constitutes as 'evidence' is determined by s to select relevant past data-label pairs and the weighting scheme w to assign their relevance. An ICM allows w and s to be any weighting scheme or selection procedure. Following other case-based reasoning methods, w and s often use a similarity or distance measure (e.g. Euclidean distance).

3.2. The four properties of ICM

In this section we explain why the above proposed ICM framework results in confidence measures that are not only accurate, but also predictable, transparent and explainable.

Accurate. We define the accuracy of a confidence measure as its ability to convey a high confidence for either a correct or incorrect advice, when the advice is indeed correct or incorrect. For an ICM, this can be defined as:

$$a = \frac{1}{|D|} \sum_{i=0}^{|D|} \delta(\vec{x}_i, y_i^*, C(\vec{x}_i)) \tag{3}$$

Where δ is the Kronecker delta, with $\delta = 1$ when $f(\vec{x}) = y^*$ and $C(x) \geq 0$, or when $f(\vec{x}) \neq y^*$ and $C(x) < 0$. Overall, case-based reasoning methods are often accurate enough for realistic data sets (McLean, 2016). However, the accuracy depends on the choice for the selection procedure s and weighting scheme w. If one chooses a simple *k*NN paradigm, one may expect a lower accuracy then when using a more sophisticated s and w. More complex options could include learning a complex similarity measure (Papernot and McDaniel, 2018). This potentially increases the accuracy, but at the cost of ICM's transparency and predictability.

Predictable. A confidence measure should behave predictable; it should monotonically increase or decrease when more evidence or data becomes available for an advice being correct or incorrect respectively. For an ICM to be predictable, it must use a monotonic similarity function. Any step-wise or non-monotonic similarity function creates

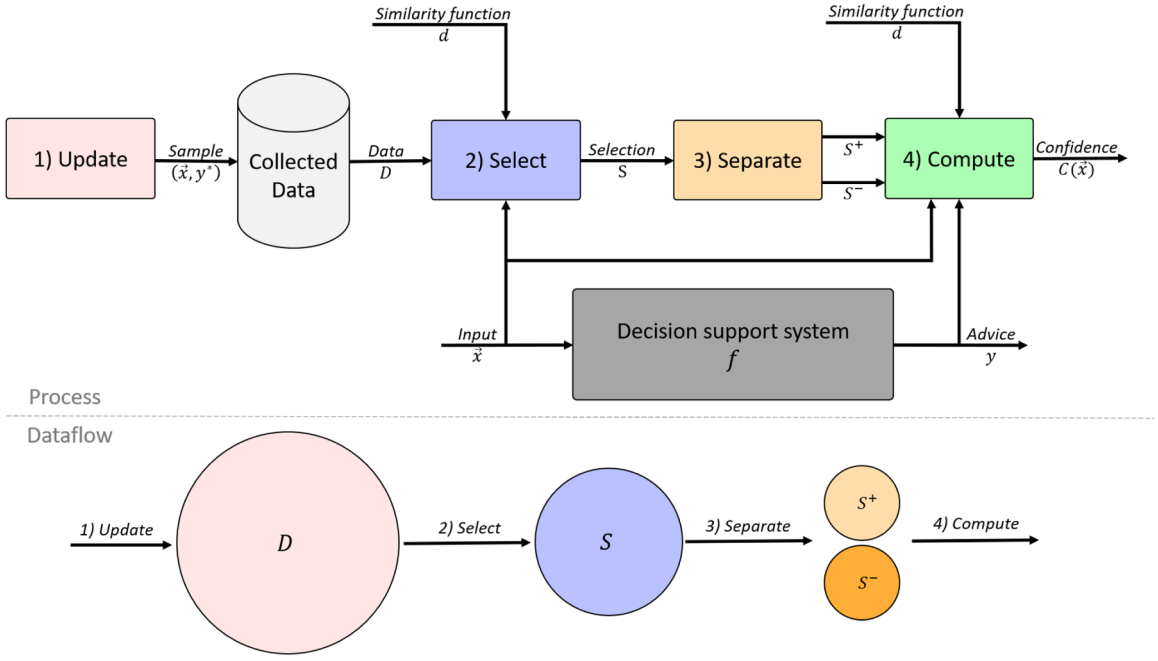


Fig. 2. A visual depiction of the ICM framework and its four steps for computing the confidence value C for a data point $\vec{x}$ and advice y. Given a continuously updated data set D, set S is selected containing relevant data-label pairs. This S is separated in S^+ and S^- with the points that show y is correct or incorrect respectively. The confidence value $C(\vec{x})$ is computed using S^+ and S^-.

confidence values that suffer from changes that are unexpected for humans. In addition, with the update procedure u an ICM adjusts its confidence according to any changes in the data distribution or DSS itself.

Transparent. An ICM is transparent; its algorithm can be understood relatively easily by its users. Case-based reasoning is often applied by humans themselves (Schank et al., 2014). This makes the idea of an ICM, recall past data-label pairs and extrapolate those to the current data point into a confidence value, relatively easy to comprehend. A deeper understanding of the algorithm may be possible, but depends on the complexity of the similarity measure, the selection procedure s and weighting scheme w.

Explainable. The confidence of an ICM can be easily explained using examples as selected from S^+ and S^-. It allows for a template-based explanation paradigm, for example:

> "I am $C(\vec{x})$ confident that y will be correct based on $|S|$ past cases deemed similar to $\vec{x}$. Of these cases, in $|S^+|$ cases the advice y was correct. In $|S^-|$ cases the advice y would be incorrect."

These cases can then be further visualized through a user-interface, for example with a parallel-coordinates plot (Artero et al., 2004). Such plots provide a means to visualize high-dimensional data and convey the ICM's weighting scheme. They allow users to identify if the selected past data points and their weights make sense and evaluate if what the ICM constitutes as evidence should indeed be treated as such. It may even enable the user to interact with the ICM by tweaking its potential hyper parameters (e.g. parameters for the selection procedure and weighting scheme).

Existing research such as that by Mandelbaum and Weinshall (2017), Subramanya et al. (2017) and Papernot and McDaniel (2018) can be framed as an ICM. All are based on case-based learning and can be described by the four steps of the framework. However, their transparency and predictability tends to be limited due to their choice to use a Neural Network to define their similarity measure. This hinders the ICM's transparency and predictability, but still allows the generation of explanations.

4. ICM Examples

In this section we propose three examples of implementing an ICM using relatively simple techniques from the field of case-based reasoning. To define our ICM, we need to define the update procedure u, the selection procedure s and the weighting scheme w. The u remains unchanged: A queue mechanism that stores the latest $(\vec{x}, y^*)$ pair and removes the oldest from D.

The first example, ICM-1, is based on kNN and use it to define both s and w. The selection procedure is $S = s(\vec{x}|D, k, d)$ which selects the k closest neighbours in D to $\vec{x}$ with d being a distance function. The weighting scheme becomes $w(\vec{x}, \vec{x}_i) = 1, \forall\, \vec{x}_i \in S$. When applied to Eq. (1), the resulting ICM counts and the relative number of points in S^+ and S^- to arrive at a confidence value:

$$C(x|S^+, S^-) = \frac{1}{k}(|S^+| - |S^-|) \tag{4}$$

This reflects the idea that confidence is ≥ 0 when the majority of k nearest neighbours are in favor of the given advice, and < 0 otherwise.

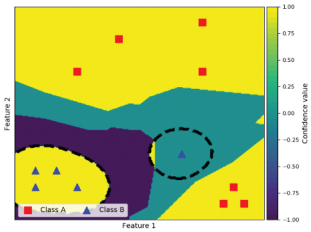
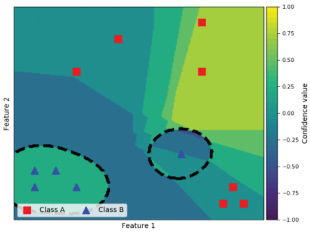
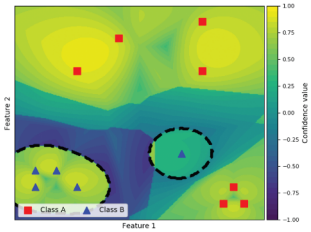
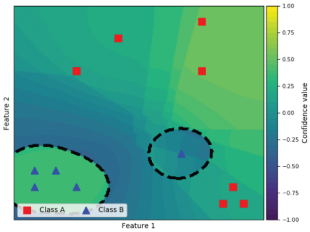
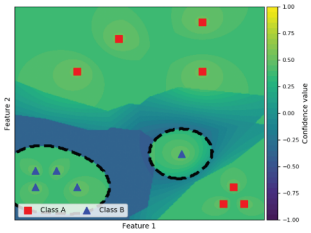
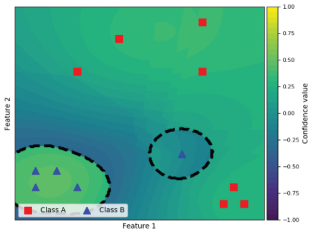
For our second example, ICM-2, we extend ICM-1 with the idea of Weighted kNN (Dudani, 1976; Hechenbichler and Schliep, 2004). It weights each neighbour with a kernel based on its similarity to $\vec{x}$ according to a distance function d. Given a Radial Basis Function (RBF) as kernel, the weighting scheme becomes $w(\vec{x}, \vec{x}_i) = \exp\left[-\left(\frac{1}{\sigma} d(\vec{x}, \vec{x}_i)\right)^2\right]$. The σ is the standard deviation of the RBF and we set it to $\sigma = d(\vec{x}, \vec{x}_{k+1})$ where $\vec{x}_{k+1}$ is the $k+1$ distant neighbour of $\vec{x}$. If we choose to use the Euclidean distance $d = \|\vec{x} - \vec{x}_i\|^2$, the confidence value becomes:

$$C(x|S^+, S^-, d) = \frac{1}{|S^+|}\sum_{\vec{x}_i \in S^+} \exp\left[-\left(\frac{1}{\sigma}\|\vec{x} - \vec{x}_i)\|^2\right)^2\right] - \frac{1}{|S^-|}\sum_{\vec{x}_j \in S^-} \exp\left[-\left(\frac{1}{\sigma}\|\vec{x} - \vec{x}_j)\|^2\right)^2\right] \tag{5}$$

These values depend not only on the number of points in S^+ and S^-, but also on their similarity to $\vec{x}$. With this RBF kernel neighbours are weighted exponentially less important as they become dissimilar to $\vec{x}$.

In our third example, ICM-3, we build further on ICM-2. In it, we

Table 2
These figures show the confidence values for the three example ICM implementations on a 2D synthetic binary classification task for $k = 2$ and $k = 3$. The background of each figure represents the confidence value at that point, the classification model's decision boundaries are shown by the dashed lines and D is plotted as points coloured by their true class label.

	$k = 2$	$k = 8$
ICM-1		
ICM-2		
ICM-3		

estimate σ for each confidence value as the average similarity between the k neighbours. Hence, ICM-3 remains equal to Eq. (5), instead with $\sigma = \frac{1}{k}\sum_{\vec{x_i}\in S} ||\vec{x} - \vec{x_i}||^2$. With it, ICM-3 provides confidence values that take the number of data points in S^+ and S^- into account, but also weighs their similarity to $\vec{x}$ according to how similar the k neighbours are to each other. Meaning that the neighbour most similar to $\vec{x}$ contributes the most to the confidence estimation relative to the other $k - 1$ neighbours.

4.1. Comparison of exemplar ICM behaviour

In this section we evaluate ICM-1, ICM-2 and ICM-3 and assess their behaviour, accuracy and predictability over changes in the data.

See Table 2 for the confidence values of all three example ICM on a synthetic 2D binary classification solved by standard SVM. This data set was generated using Python's SciKit Learn package (Pedregosa et al., 2011). The table contains six plots of ICM-1, ICM-2 and ICM-3 with $k = 2$ and $k = 8$. ICM-1 shows a high confidence when we would expect it. As points with a certain prediction approaches memorized points (in Euclidean space) with that prediction as their label, the confidence for a correct predictions increases. As opposed to an increasing confidence for an incorrect prediction when such points approach memorized data points whose label is different than the prediction. ICM-1 does show abrupt confidence changes with $k = 2$, that decrease for $k = 8$. Similar behaviour can be seen for ICM-2 and ICM-3. The difference is that both show even smaller abrupt changes due to their RBF kernel, with ICM-3 being the smoothest as the kernel adapts to the local density. For $k = 8$ we see that ICM-2 and ICM-3 result in an overall lack of confidence. With higher k values, S starts to contain nearly all data points from D. The summed weights for S^+ and S^- begin to represent the label ratio and confidence goes to zero. This sensitivity is likely unique to our ICM examples, and state of the art case-based reasoning algorithms are less likely to be as sensitive to k or use a different mechanism than kNN.

Next, we evaluate the accuracy of ICM-3 on two benchmark classification tasks each solved by a Support Vector Machine (SVM), Random Forest and Multi-layer Perceptron (MLP). We chose for ICM-3 as the most sophisticated ICM example. The confidence accuracy of ICM-3 was computed using Eq. (3). The confidence values of the SVM were computed using Platt scaling (Platt and others, 1999), of the Random Forest using its voting mechanism, and of the MLP by setting SoftMax as its output layer's activation function. Since neither of these confidence values could express a high confidence for an incorrect classification, the accuracy from Eq. (3) was adjusted to measure zero confidence as correct for an incorrect classification. The two classification tasks were a handwritten digits recognition task (Alimoglu et al., 1996) and the diagnoses of heart failure in patients (Detrano et al., 1989). The data set properties, trained models and their hyper parameters are summarized in Tables 3 and 4 respectively.

Fig. 3 shows the results from ten different runs per test set and model combination. The ICM performs equally well in confidence estimation as the models on both data sets. It shows that an ICM can be applied to a variety of models and performs equally well in terms of estimating when a classification would be correct. In addition, an ICM conveys also its confidence in a classification being incorrect and tends to be more transparent, predictable and explainable.

Fig. 4 shows the accuracy of the example ICM over different values

Table 3
The properties of the two benchmark data sets used to evaluate the three ICM examples. Also shows the properties of the synthetic data used to evaluate the robustness to changes in data distributions.

Name	Classes	Type	Features	Train/test
Heart Detrano et al. (1989)	3	Tabular	4	227/76
Digits Alimoglu et al. (1996)	10	Images	64	1347/450
Synthetic	6	Tabular	2	100/300

Table 4
Shows the hyper parameters and accuracy on train- and test set for each model and data set combination used to compare our example ICM with. We used the SciKit Learn package from Python as the implementation of each model (Pedregosa et al., 2011).

Data	Model	Accuracy (train)	Parameters
Heart	SVM	77.63% (94.27%)	RBF kernel, $\gamma = 0.1$, $C = 1.0$
Heart	MLP	76.32% (91.85%)	Adam optimizer ($\alpha = 1e^{-3}$, decay= $1e^{6}$), 100 epochs, 16 batch size, Softmax output layer, 2 hidden Layers ([16, 3], ReLu, 20% dropout)
Heart	Random Forest	73.68% (100%)	Gini, Bootstrapping, 50 estimators
Digits	SVM	98.22% (100%)	RBF kernel, $\gamma = 0.01$, $C = 10.0$
Digits	MLP	99.33% (99.73%)	Adam optimizer ($\alpha = 1e^{-4}$, decay= $1e^{6}$), 250 epochs, 16 batch size, Softmax output layer, 3 hidden layers ([64, 32, 6], ReLu, 20% dropout)
Digits	Random Forest	97.33% (100%)	Gini, Bootstrapping, 100 estimators
Synthetic	Random Forest	97.33% (100%)	Gini, Bootstrapping, 100 estimators

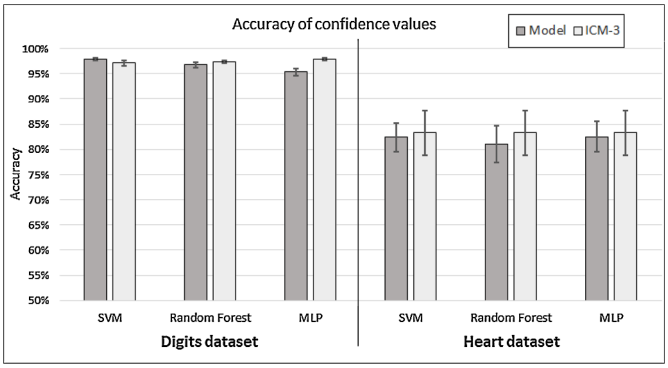


Fig. 3. The accuracy of ICM-3 on two data sets with the accuracy of the confidence estimates from various models. It shows that ICM implementations are applicable to different models and can be equally accurate as the model itself. The error bars represent the 95% confidence intervals.

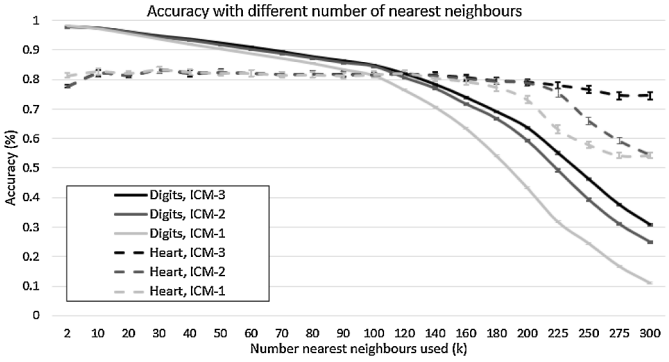


Fig. 4. The accuracy of ICM on two data sets for different numbers of nearest neighbours used. It shows our proposed Robust Weighted *k*NN algorithm (ICM-3) compared to ICMs with weighted *k*NN (ICM-2) and *k*NN (ICM-1). It illustrates the robustness of ICM-3 against different *k*.

for *k*. The *n* was set to encapsulate the entire training data set. This figure shows that ICM-3 is most robust against different values for *k*. More state of the art algorithms based on *k*NN can be applied to increase this robustness further, or algorithms based on an entirely different paradigm can be used to define the selection procedure *s*.

Fig. 5 shows how the example ICM behaves with different numbers of memorized data points. The *k* was fixed to its optimal value of 10 neighbours for both data sets. These results show that even these simple ICM are accurate at memory sizes around 10% of the data the models needed for training.

In a separate study (van Diggelen et al., 2017) we applied ICM-3 to a real-world DSS in the Dynamic Positioning case described in the introduction. This case was also used in one of our user experiments, described in detail in Section 5. Here, a Deep Neural Network predicted

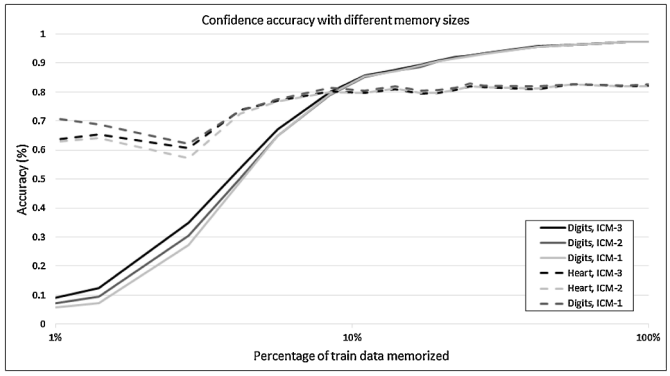


Fig. 5. The accuracy of the example ICM on the two benchmark data sets for different values of *n*, the number of memorized data points. It illustrates the robustness of each ICM with different *n*.

when an ocean ship was likely to drift of course and notified a human operator to intervene. The ICM-3 was used to express more information to the operator on whether a prediction could be trusted to prevent under- or over-trust. In this study, we showed that ICM was able to compute a confidence value of the Deep Neural Networks prediction with 87% accuracy (van Diggelen et al., 2017).

Finally, we evaluated how well ICM-3 was able to adjust its confidence values when faced with a shift in the data and label distribution. As stated in Section 4, the update procedure *u* used a simple queuing method to update *D*. To test the effects this *u* has on the confidence accuracy, we constructed a synthetic non-linear classification task and artificially shifted its data distribution after having computed the confidence of the first 100 data points. We compared this confidence accuracy over time with the performance of the Random Forest model with and without continuously updating that model.

The results of are shown in Fig. 6, repeated ten times with different random seeds to obtain the confidence bounds that are shown. The plot shows that ICM-3 is capable of adjusting its confidence estimation to abrupt changes in the data distribution. It performs nearly the same to continuously retraining the model when obtaining a new data point, however the ICM requires no explicit update.

The above results illustrate that even simple ICM can already perform surprisingly accurate on two benchmark data sets and on different models. In addition, even a simple update of the memorized data points result in an confidence estimation adaptive to changes in the data. It shows that ICM can provide a common framework to devise system-agnostic confidence measures.

5. A qualitative user experiment: Interviews with domain experts

This section summarizes the first of two user experiments. This experiment is explained in more detail in our previous work (van der Waa et al., 2018). In the experiment, several domain experts

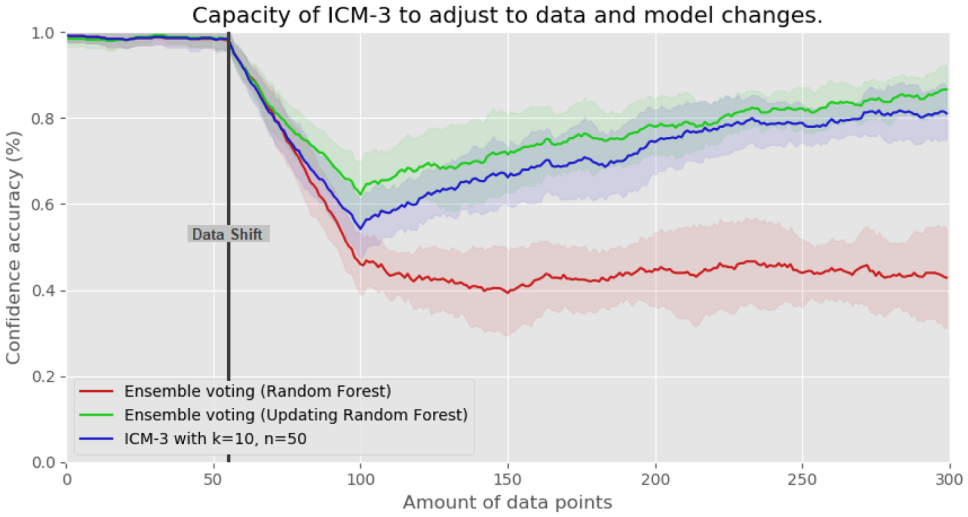


Fig. 6. The moving average accuracy of querying confidence values from ICM-3, a static Random Forest and a continuously updated Random Forest. It shows a shift in the label distribution after 100 data points in the synthetic data. The plot shows that ICM-3 is capable of adjusting its confidence values nearly as well as the confidence from the continuously updating model.

were interviewed to evaluate the transparency of the case-based reasoning approach underlying an ICM compared to other confidence measures.

Dynamic Positioning (DP) formed the use case of the experiment. Here, a ship's bridge operator is responsible for maintaining the ship's position aided by an auto-pilot and a DSS (van Diggelen et al., 2017). The DSS warns the user when the ship's position is expected to deviate from course and human intervention is required. Structured interviews with DP operators were conducted where we elicited their understanding and needs of a confidence value that accompany the DSS' prediction. Three confidence measure categories were evaluated; 1) ICM, 2) Platt Scaling and 3) SoftMax activation functions.

The interview was structured in three phases. In the first phase we provided a layman's - but complete - description of each confidence measure. Participants were asked to select their preferred method followed by explaining each measure in their own words. This enables us to discover which algorithm they preferred, but also which they could reproduce accurately (signifying a better understanding). We found that they understood ICM best, but preferred the SoftMax measure. When asked, participants mentioned that estimating confidence in their line of work is difficult and as such they expected a confidence measure to be very complex. This result points towards what users might prefer in a confidence measure (complexity), may not necessarily be what they need (transparency).

The second phase provided examples of realistic situations, the DSS' prediction and a confidence value. Each example was accompanied by three explanations, one from each measure. Participants were asked which explanation they preferred for each example. On average, they preferred the explanations from ICM as it specifically addressed past examples and explained their contribution to the confidence value. Afterwards, participants were asked to explain in their own words how each confidence measure would compute their values for unseen situations. The results showed that the operators could replicate ICM's explanations more easily than that of the other two.

The third and final phase allowed the participants to describe their ideal confidence measure for the DSS. Several participants described a case-based reasoning approach as used by ICM. Others preferred a combination of both an ICM and SoftMax. When asked why, they replied that they preferred the case-based reasoning approach but they believed it to be too simplistic on its own to be accurate in their line of work. They tried to add their interpretation of a SoftMax activation function to ICM to satisfy their need for added complexity.

These results may indicate that domain experts are able to understand a case-based reasoning approach for a confidence measure more easily than the DSS' prediction scores defined by a SoftMax output layer, or the scaled prediction scores with Platt Scaling.

6. A quantitative user experiment: an online survey on user preferences

The second experiment was performed using a quantitative online survey. We evaluated the users' interests and preferences concerning explanations about the confidence of an advice as provided by a DSS. Moreover, we investigated if the proposed ICM, based on case-based reasoning, is in line with what humans desire from a confidence measure and explanations.

Below we describe the use case, participant group, stimuli, design and analyses in more detail, followed by the results.

6.1. Use case: Autonomous driving

In the survey, participants were provided a written scenario describing an autonomous car. This scenario stated that the car could provide an advice to turn its self-driving mode on or off, given the current and predicted road, weather and traffic conditions. The advice would be accompanied by a confidence value as calculated by the car. Participants were instructed to assume several years of experience with the car and that the car showed to be capable of driving autonomously on frequently used roads. At some point on such a familiar road, the car would provide the advice to turn on automatic driving mode. The experiment followed with a questionnaire revolving around this advice and the given confidence value.

6.2. Participants

Recruitment was done via Amazon's Mechanical Turk, and each participant received $0.45 for participating in the survey based on the estimated time for completion and average wages. Only participants of 21 years or older were included. A total of 26 men and 14 women aged between 24 and 64 years ($M = 35.6$, $SD = 9.4$) were recruited, who were all (self-rated) fluent English speakers. On average, participants indicated on a 5-point Likert scale that they had some prior knowledge with self-driving cars ($M = 3.00$, $SD = 0.68$). Hence, participants could be biased towards answering questions based upon knowledge about self-driving cars, instead of using the description in the questions. However, the scores on the dependent variables of the participants indicated they were knowledgeable ($n = 6$) or very knowledgeable ($n = 1$) did not significantly differ from the scores of others and were included.

6.3. Stimuli

We composed a survey in which we asked participants about their interests and preferences concerning explanations about the confidence

of an advice as provided by a self-driving car. The system was presented as being able to drive perfectly without assistance from a user within most situations, but unable to drive fully autonomously in some other undefined situations. We asked participants to indicate how much importance they would attach to: 1) understanding the confidence measure's underlying algorithm, 2) their past experience with other advice from the car, and 3) predictions about future conditions (e.g. weather). The importance was indicated on a 7-point Likert scale with 1 meaning 'not at all important' and 7 meaning 'very much important'.

Moreover, we asked participants to rank five methods of presenting the advice that the car could provide (with 1 being most preferred, and 5 being least preferred):

a) No additional information;
b) A general summary of prior experiences;
c) General prior experience accompanied by an illustrative specific past experience;
d) Current situational aspects that played a role;
e) Predicted future situational characteristics that could affect the decision's outcome.

Fig. 7 shows a screenshot that contains the question in that asked users to rank different types of explanations according to their preference. Advice 2 and 3 provide illustrative examples of the type of information that an ICM can provide to a user (corresponding to b) and c) in the above enumeration). That is, the confidence of the DSS is explained in terms of similar stored past experiences with its own performance. The difference between advice 2 and 3 is that the latter includes a specific example of a situation in which the advice appeared not to be correct, while the former does not.

6.4. Experimental design

We investigated two variables. 1) The importance of different information in determining when to follow an advice: information about the confidence measure's algorithm, information about prior experience, or information about the predicted future situation. 2) The information preference in an accompanying explanation: no additional explanation, general prior experience, specific prior experience, current situation, or predicted future situation. Both dependent variables were investigated within-subjects, meaning that all participants indicated their importance rating and preference rankings for all types of information and explanations respectively.

6.5. Analyses

We performed two non-parametric Friedman tests with post-hoc Wilcoxon signed rank tests on the ordinal Likert scale data to investigate two topics: 1) The relative importance of information that taken into account when deciding whether or not to follow the advice, and 2) the difference between preference ratings of the types of explanation.

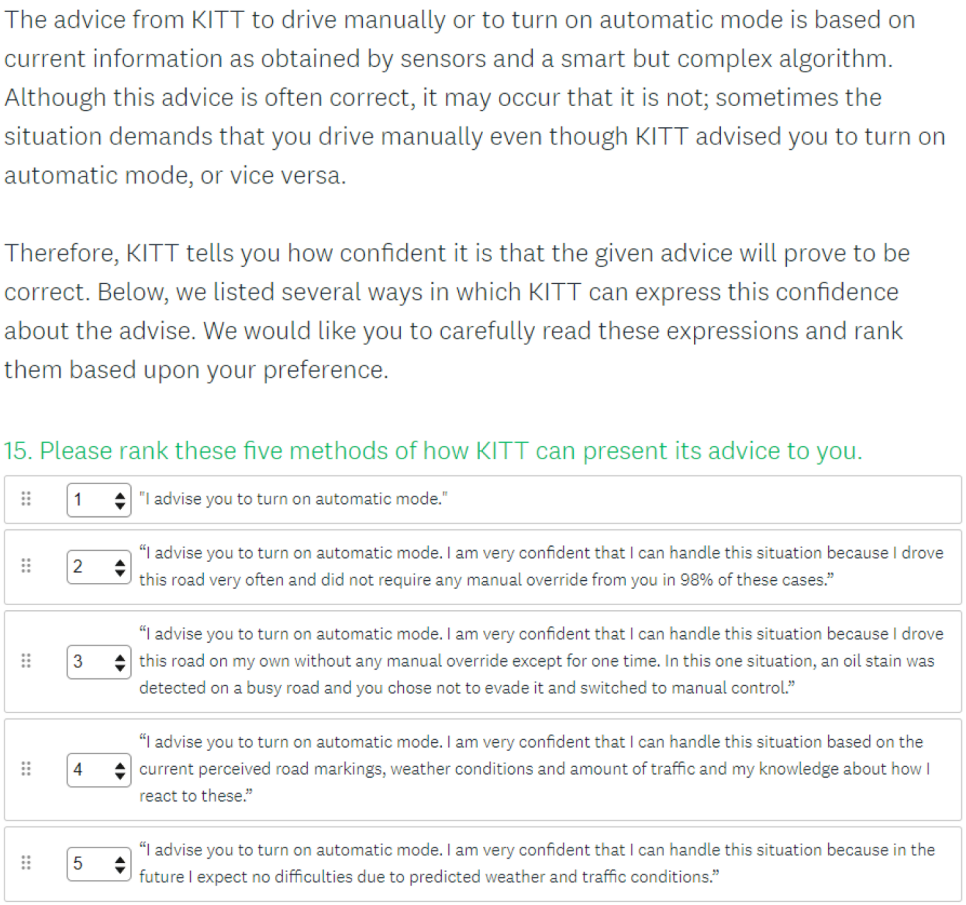


Fig. 7. Screenshot of the section of the survey in which participants were asked to rank different kinds of explanations based on their preference. Advice 2 and 3 provide illustrative examples of the type of information that an ICM can provide to a user.

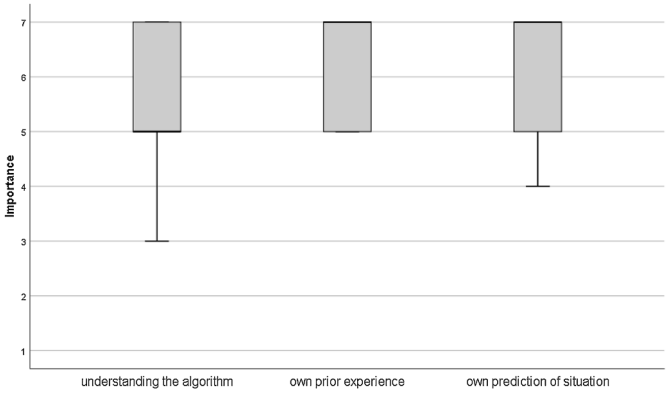


Fig. 8. Boxplot of the Likert scale ratings indicating the importance of different types of information used to determine to follow the advice to turn on automatic driving mode. The higher the ratings, the more the information was preferred.

6.6. Results

Fig. 8 shows the distribution of Likert scale ratings concerning the importance of information in the advice. Ratings are high in general, as indicated by the high medians and the minor deviations from these median scores.

There is a statistically significant difference in importance ratings of the considered information when evaluating an advice, $\chi^2(2) = 16.77$, $p < 0.001$. Wilcoxon signed-rank tests showed that participants rated prior experience with the system as more important for deciding about following an advice than understanding the advice system ($Z = -3.71$, $p < 0.001$), but not more important than predictions about future situational circumstances ($Z = -1.58$, $p = 0.115$). However, predictions about future circumstances were rated as being more important than understanding the advice system ($Z = -2.89$, $p = 0.004$).

Fig. 9 shows the means and 95% confidence intervals of the rankings concerning the preferences of participants for different types of additional information given in an advice.

There is a statistically significant difference in rankings of the five types of advice, $\chi^2(4) = 39.38$, $p < 0.001$. Table 5 shows the results of the post-hoc tests. Importantly, participants preferred the explanation that contained general prior experiences over the one that presented a specific experience of a case in which the advice was not followed. They

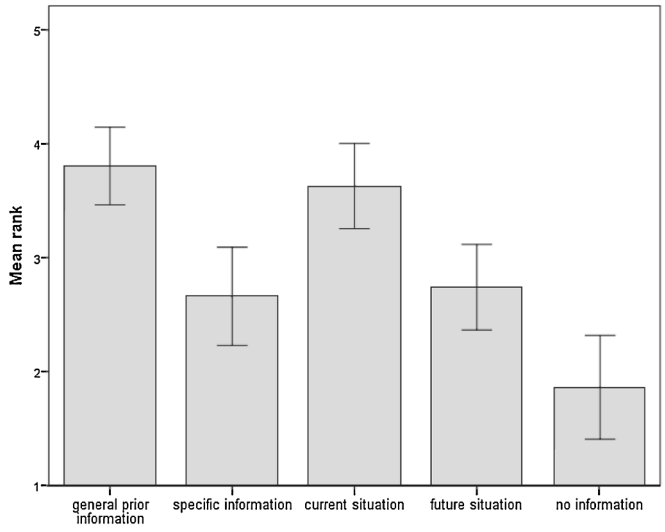


Fig. 9. Means and 95% confidence intervals of the preference rankings concerning the different types of advice that are provided by the autonomous car. The rankings are inverted, the higher the rank the more preferred.

also preferred general prior information over information concerning the future situation, and over no additional information. However, preference ratings for using general prior experience as explanation about an advice were, on average, not higher than using information about present situational circumstances.

In this user experiment, we investigated how participants judged different types of information a confidence measure may use and include in an explanation. Overall, the use of relevant prior experiences was judged as important in both defining confidence values and explaining them. Equally important was the information contained in the current situation. This indicates that ICM and its explanations match peoples expectations and preferences of a confidence estimation. It also underlines the importance of confidence measures providing an explanation about its values, something ICM readily supports. However, confidence measures may also require to explain how the current situation relates to those past experiences. For ICM that entails explaining the similarity function and why it selected those past experiences given the current situation. To do so, the similarity function needs to be easily understood or explained.

7. Discussion

Although the proposed ICM framework relies on a case-based reasoning approach, it is also closely related to the field of conformal prediction (Shafer and Vovk, 2008). Methods from this field define a set of predictions that is guaranteed to contain the true prediction with a certain probability (e.g. 95%). Conformal prediction methods share many similarities to ICM, such as their model-agnostic approach and use of (dis)similarity functions. Current research focuses on making these methods more explainable and transparent (Johansson et al., 2018). Our experimental work on these topics may provide valuable insights for future conformal prediction methods. In addition, future work may aim to explore how conformal prediction methods can be used in the ICM framework.

An important trade-off in an ICM is between its accuracy and transparency, as an increase in accuracy implies an increase in complexity. A concrete example is the similarity measure, it can be as straightforward as Euclidean distance or as complex as a trained Deep Neural Network (as in Mandelbaum and Weinshall, 2017). For some domains, a relative simple similarity measure may not suffice due to its high dimensional nature or less-than apparent relations between features (e.g. the many pixels in an image recognition task). A more complex or even learned similarity measure may solve such issues. However, it may may prevent users from adopting the system in their work due to a lack of understanding (Ye and Johnson, 1995). This is sometimes referred to as the accuracy and transparency trade-off in current AI. To solve this, simplified model-agnostic methods generating explanations may be a solution. However, it also requires exploring where users allow for system complexity and where transparency is required.

Besides such technical issues, an interesting finding from the online survey was that participants did not found it important for an explanation to refer to past situations in which the provided advice proved to be incorrect. This could indicate the tendency of people to favor information that confirms their preexisting beliefs and to be ignorant towards falsification, a phenomenon known as the confirmation bias (Gilovich et al., 2002). Importantly, such a preference does not necessarily mean that it is best to omit this kind of information. That is, the main goal of the transparency and explainability properties of an ICM is to enable users to better understand where the confidence value originates from in order to more accurately predict the extent to which an advice of the system can be trusted. In order to enable people to make an accurate assessment, it is essential to provide both confirming and contradictory information, precisely because we know that people are prone to ignore information that does not confirm their beliefs. Future work on confidence measures should not only conduct user

Table 5
Results of the Wilcoxon signed rank post-hoc tests on the preference rankings of information that is included in an explanation about the advice.

	Present Situation	General past experience	Future situation	Specific past experience
Present situation				
General past experience	*n.s.*			
Future situation	$Z = -2.00,\ p = .045$	$Z = -2.35,\ p = .019$		
Specific past experience	$Z = -2.77,\ p = .006$	$Z = -3.34,\ p = .001$	*n.s.*	
No information	$Z = -3.86,\ p < .001$	$Z = -4.39,\ p < .001$	$Z = -3.40,\ p = .001$	$Z = -2.28,\ p = .023$

experiments revolving around preferences, but also on how they affect system adoption, usage and task performance.

Moreover, findings from our user experiments implied that people prefer to know about the current situational circumstances. This preference holds even when a given confidence value was high and they said they trusted this estimation. This could indicate that people still want to be able to form their own judgement about the DSS' advice based on their own observations, in order to maintain a sense of control and autonomy (Legault, 2016). Hence, a confidence measure is not a substitute for a user's own judgement process and should be designed to facilitate this process. ICM's property of explainability may offer a vital contribution to this process. Further investigation is required to identify what should be explained in addition to an ICM and how this should be presented.

8. Conclusion

In this work we proposed the concept of Interpretable Confidence Measures (ICM). We used the idea of case-based reasoning to formalise such measures. In addition, we motivated the need for confidence measures to be not only accurate, but also explainable, transparent and predictable. An ICM aims to provide a user of a decision support system (DSS) information whether a DSS's advice should be trusted or not. It does so by conveying how likely it is that the given advice turns out to be correct based on past experiences.

Three straightforward ICM implementations were proposed and evaluated, to serve as concrete examples of the proposed ICM framework. Two user experiments were conducted that showed that participants were able to understand the idea of case-based reasoning and that this was in line with their own reasoning about confidence. In addition, participants especially preferred their confidence values to be explained by referring to past experiences and by highlighting specific experiences in the process.

Future work may focus on further expanding the ICM framework by incorporating more state of the art methods for confidence estimation. Especially methods from the field of conformal prediction may prove valuable. Additional user experiments could provide more insight in user requirements for confidence measures. Other user experiments could investigate the effects of confidence measures on actual task performance.

CRediT authorship contribution statement

Jasper van der Waa: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Tjeerd Schoonderwoerd:** Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Jurriaan van Diggelen:** Writing - review & editing, Supervision, Funding acquisition. **Mark Neerincx:** Writing - review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgements

This research was funded by TNO's early research programs (ERP) and risk-bearing exploratory research programs (RVO). We would like to thank our colleagues for insightful discussions. Special gratitude go to Alexander Boudewijn, Stephan Raaijmakers and Catholijn Jonker.

References

Alimoglu, F., Alpaydin, E., Denizhan, Y., 1996. Combining Multiple Classifiers for Pen-Based Handwritten Digit Recognition. Institute of Graduate Studies in Science and Engineering, Bogazici University Master's thesis.
Artero, A.O., de Oliveira, M.C.F., Levkowitz, H., 2004. Uncovering clusters in crowded parallel coordinates visualizations. IEEE Symposium on Information Visualization. IEEE, pp. 81–88.
Atkeson, C.G., Moore, A.W., Schaal, S., 1997. Locally weighted learning. Artif. Intell. Rev. 11 (June), 11–73.
Bhattacharyya, S., 2013. Confidence in predictions from random tree ensembles. Knowl. Inf. Syst. 35 (2), 391–410.
Bose, I., Mahapatra, R.K., 2001. Business data mining; a machine learning perspective. Inf. Manage. 39 (3), 211–225.
Burrell, J., 2016. How the machine thinks: understanding opacity in machine learning algorithms. Big Data Soc. 3 (1).
Cabitza, F., Rasoini, R., Gensini, G.F., 2017. Unintended consequences of machine learning in medicine. JAMA 318 (6), 517–518.
Cohen, M.S., Parasuraman, R., Freeman, J.T., 1998. Trust in decision aids: a model and its training implications. in Proc. Command and Control Research and Technology Symp. Citeseer.
Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K.H., Lee, S., Froelicher, V., 1989. International application of a new probability algorithm for the diagnosis of coronary artery disease. Am. J. Cardiol. 64 (5), 304–310.
van Diggelen, J., van den Broek, H., Schraagen, J.M., van der Waa, J., 2017. An intelligent operator support system for dynamic positioning. International Conference on Applied Human Factors and Ergonomics. Springer, pp. 48–59.
Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.
Doyle, D., Tsymbal, A., Cunningham, P., 2003. A Review of Explanation and Explanation in Case-Based Reasoning. Technical Report. Trinity College Dublin, Department of Computer Science.
Dudani, S.A., 1976. The distance-weighted k-nearest-neighbor rule. IEEE Trans. Syst. Man Cybern. 325–327.
Evans, J., Handley, S., Over, D., 2003. Conditionals and conditional probability. Exp. Psychol. 29 (2), 321.
Fitzhugh, E.W., Hoffman, R.R., Miller, J.E., 2011. Active Trust Management. Ashgate.
Fix, E., Hodges Jr, J.L., 1951. Discriminatory Analysis-Nonparametric Discrimination: Consistency Properties. Technical Report. California Univ Berkeley.
Florez-Lopez, R., Ramon-Jeronimo, J.M., 2015. Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal. Expert Syst. Appl. 42 (13), 5737–5753.
Foody, G.M., 2005. Local characterization of thematic classification accuracy through spatially constrained confusion matrices. Int. J. Remote Sens. 26 (6), 1217–1228.
Fortunato, M., Blundell, C., Vinyals, O., 2017. Bayesian recurrent neural networks. arXiv:1704.02798.
Freitas, A.A., 2014. Comprehensible classification models: a position paper. ACM SIGKDD Explor. Newsl. 15 (1), 1–10.
Gal, Y., Ghahramani, Z., 2016. A theoretically grounded application of dropout in re-current neural networks. Advances in Neural Information Processing Systems. pp. 1019–1027.
Gilovich, T., Griffin, D., Kahneman, D., 2002. Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge university press.
Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., Kagal, L., 2018. Explaining explanations: an approach to evaluating interpretability of machine learning. arXiv:1806.00069.
Goodfellow, I. J., Shlens, J., Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv:1412.6572.
Goodman, B., Flaxman, S., 2016. European union regulations on algorithmic decision-making and a "right to explanation". arXiv:1606.08813.
Graves, A., 2011. Practical variational inference for neural networks. Advances in Neural Information Processing Systems. pp. 2348–2356.
Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2018. A

survey of methods for explaining black box models. ACM Comput. Surv. (CSUR) 51 (5), 93.
Hao, H., Liu, C.-L., Sako, H., et al., 2003. Confidence evaluation for combining diverse classifiers. ICDAR. vol. 3. pp. 760–765.
Harteis, C., Billett, S., 2013. Intuitive expertise: theories and empirical evidence. Educ. Res. Rev. 9, 145–157.
Hechenbichler, K., Schliep, K., 2004. Weighted k-nearest-neighbor techniques and ordinal classification. Discussion Paper 399, SFB386
Herman, B., 2017. The promise and peril of human evaluation for model interpretability. arXiv:1711.07414.
Hodgkinson, G.P., Langan-Fox, J., Sadler-Smith, E., 2008. Intuition: a fundamental bridging construct in the behavioural sciences. Br. J. Psychol. 99 (1), 1–27.
Hoffman, R.R., Johnson, M., Bradshaw, J.M., Underbrink, A., 2013. Trust in automation. IEEE Intell. Syst. 28 (1), 84–88.
Hoffman, R. R., Mueller, S. T., Klein, G., Litman, J., 2018. Metrics for explainable ai: challenges and prospects. arXiv:1812.04608.
Holzinger, A., Carrington, A., Müller, H., 2019a. Measuring the quality of explanations: the system causability scale (SCS). Comparing human and machine explanations. arXiv:1912.09024.
Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H., 2019. Causability and explainability of artificial intelligence in medicine. WIREs Data Min. Knowl. Discov. 9 (4), e1312.
Johansson, U., Linusson, H., Löfström, T., Boström, H., 2018. Interpretable regression trees using conformal prediction. Expert Syst. Appl. 97, 394–404.
Kim, B., Glassman, E., Johnson, B., Shah, J., 2015. iBCM: Interactive Bayesian Case Model Empowering Humans via Intuitive Interaction. Technical Report. MIT-CSAIL-TR-2015-010.
Labatut, V., Cherifi, H., 2011. Evaluation of performance measures for classifiers comparison. arXiv:1112.4133.
Landsbergen, D., Coursey, D.H., Loveless, S., Shangraw Jr, R., 1997. Decision quality, confidence, and commitment with expert systems: an experimental study. J. Public Adm. Res.Theory 7 (1), 131–158.
Legault, L., 2016. The need for autonomy. Encyclopedia of Personality and Individual Differences. Springer, New York, NY, pp. 1120–1122.
Libbrecht, M.W., Noble, W.S., 2015. Machine learning applications in genetics and genomics. Nat. Rev. Genet. 16 (6), 321.
Lipton, Z. C., 2016. The mythos of model interpretability. arXiv:1606.03490.
Liu, C.-L., Hao, H., Sako, H., 2004. Confidence transformation for combining classifiers. Pattern Anal. Appl. 7 (1), 2–17.
Mandelbaum, A., Weinshall, D., 2017. Distance-based confidence score for neural network classifiers. arXiv:1709.09844.
McLean, S.F., 2016. Case-based learning and its application in medical and health-care fields: a review of worldwide literature. J. Med. Educ. Curric.Dev. 3, S20377.
Miller, T., 2018a. Contrastive explanation: a structural-model approach. arXiv:1811.03163.
Miller, T., 2018. Explanation in artificial intelligence: insights from the social sciences. Artif. Intell.
Miller, T., Howe, P., Sonenberg, L., 2017. Explainable AI: beware of inmates running the asylum or: how i learnt to stop worrying and love the social and behavioural sciences. arXiv:1712.00547.
Nguyen, A., Yosinski, J., Clune, J., 2015. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 427–436.
Niculescu-Mizil, A., Caruana, R., 2005. Predicting good probabilities with supervised learning. Proceedings of the 22nd International Conference on Machine Learning. ACM, pp. 625–632.
Paisley, J., Blei, D., Jordan, M., 2012. Variational bayesian inference with stochastic search. arXiv:1206.6430.
Papernot, N., McDaniel, P., 2018. Deep k-nearest neighbors: towards confident, interpretable and robust deep learning. arXiv:1803.04765.
Papadopoulos, G., Edwards, P.J., Murray, A.F., 2001. Confidence estimation methods for neural networks: a practical comparison. IEEE Trans. Neural Netw. 12 (6), 1278–1287.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.
Pita, J., Tambe, M., Kiekintveld, C., Cullen, S., Steigerwald, E., 2011. Guards: game theoretic security allocation on a national scale. The 10th International Conference on Autonomous Agents and Multiagent Systems. International Foundation for Autonomous Agents and Multiagent Systems, pp. 37–44.
Platt, J., others, 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv. Large Margin Classif. 10 (3), 61–74.
Polikar, R., 2006. Ensemble based systems in decision making. IEEE Circuits Syst. Mag. 6 (3), 21–45.
Pollatsek, A., Well, A.D., Konold, C., Hardiman, P., Cobb, G., 1987. Understanding conditional probabilities. Organ. Behav. Hum. Decis. Process. 40 (2), 255–269.
Qin, Z., 2006. Naive bayes classification given probability estimation trees. 2006 5th International Conference on Machine Learning and Applications (ICMLA'06). IEEE, pp. 34–42.
Ribeiro, M. T., Singh, S., Guestrin, C., 2016. Model-agnostic interpretability of machine learning. arXiv:1606.05386.
Ridgeway, G., Madigan, D., Richardson, T., O'Kane, J., 1998. Interpretable boosted Naïve bayes classification. KDD. pp. 101–104.
Rish, I., et al., 2001. An empirical study of the naive bayes classifier. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. pp. 41–46.
Samek, W., Wiegand, T., Müller, K.-R., 2017. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv:1708.08296.
Schank, R.C., Kass, A., Riesbeck, C.K., 2014. Inside Case-Based Explanation. Psychology Press.
Shafer, G., Vovk, V., 2008. A tutorial on conformal prediction. J. Mach. Learn. Res. 9 (Mar), 371–421.
Stone, P., Veloso, M., 1997. Using decision tree confidence factors for multiagent control. Robot Soccer World Cup. Springer, pp. 99–111.
Sturm, I., Lapuschkin, S., Samek, W., Müller, K.-R., 2016. Interpretable deep neural networks for single-trial eeg classification. J. Neurosci. Methods 274, 141–145.
Subramanya, A., Srinivas, S., Babu, R. V., 2017. Confidence estimation in deep neural networks via density modelling. arXiv:1707.07013.
Tipping, M.E., 2000. The relevance vector machine. Advances in Neural Information Processing Systems (NIPS' 2000). pp. 652–658.
Tóth, N., Pataki, B., 2008. Classification confidence weighted majority voting using decision tree classifiers. Int. J. Intell. Comput.Cybern. 1 (2), 169–192.
Van Erp, M., Vuurpijl, L., Schomaker, L., 2002. An overview and comparison of voting methods for pattern recognition. Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition. IEEE, pp. 195–200.
van der Waa, J., van Diggelen, J., Neerincx, M.A., Raaijmakers, S., 2018. ICM: An intuitive model independent and accurate certainty measure for machine learning. ICAART. 2. pp. 314–321.
Walley, P., 1996. Measures of uncertainty in expert systems. Artif. Intell. 83 (1), 1–58.
Waterman, D., 1986. A Guide to Expert Systems. Addison-Wesley Pub. Co., Reading, MA.
Wettschereck, D., Aha, D.W., Mohri, T., 1997. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artif. Intell. Rev. 11 (1–5), 273–314.
Wu, Y., Yao, X., Vespasiani, G., Nicolucci, A., Dong, Y., Kwong, J., Li, L., Sun, X., Tian, H., Li, S., 2017. Mobile app-based interventions to support diabetes self-management: a systematic review of randomized controlled trials to identify functions associated with glycemic efficacy. JMIR mHealth and uHealth 5 (3), e35.
Ye, L.R., Johnson, P.E., 1995. The impact of explanation facilities on user acceptance of expert systems advice. Mis Q. 157–172.
Zadrozny, B., Elkan, C., 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. 1. pp. 609–616.
Zadrozny, B., Elkan, C., 2002. Transforming classifier scores into accurate multiclass probability estimates. Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 694–699.
Zaragoza, H., d'Alché Buc, F., 1998. Confidence measures for neural network classifiers. Proceedings of the Seventh Int. Conf. Information Processing and Management of Uncertainty in Knowledge Based Systems.
Zhou, J., Chen, F., 2018. 2D transparency space; bring domain users and machine learning experts together. Human and Machine Learning. Springer, pp. 3–19.
Zliobaite, I., 2015. A survey on measuring indirect discrimination in machine learning. arXiv:1511.00148.