Symbolic Knowledge Quality Evaluation with WInd

Federico Sabbatini[1,*], Roberta Calegari[2]

[1]National Institute for Nuclear Physics – Section in Florence, Sesto Fiorentino, Italy
[2]Alma Mater Studiorum—University of Bologna

Abstract

In multi-agent systems and intelligent environments, agents often rely on symbolic knowledge to reason, interact, and make decisions in a transparent and trustworthy manner. Ensuring the quality of such symbolic knowledge is crucial, especially when it is automatically extracted from opaque models through explainable AI techniques. However, the literature still lacks comprehensive and unbiased evaluation metrics that jointly account for predictive accuracy, human interpretability, and semantic completeness — three pillars of effective knowledge for agents. In this work, we introduce WInd, a novel and flexible scoring metric designed to assess the overall quality of symbolic knowledge in agent-based systems. WInd combines performance, readability, and completeness into a unified score, and further enables task-oriented customisation through the integration of user feedback. Its formulation supports automated knowledge tuning and facilitates knowledge sharing and comparison among agents with diverse goals and perspectives. We present the formal definition of WInd and provide a thorough comparative analysis against existing, yet limited, metrics. Our findings show that WInd offers a principled and adaptable framework for evaluating symbolic knowledge quality, paving the way for more autonomous, collaborative, and cognitively grounded intelligent agents.

Keywords

Symbolic knowledge, Explainable AI, Quality metrics, AutoML, Knowledge extraction

1. Introduction

Nowadays, symbolic knowledge-extraction (SKE) techniques are widely exploited to tackle interpretability issues of sub-symbolic artificial intelligence (AI), which despite being prediction-effective, typically relies on complex models (e.g., deep neural networks) difficult to interpret and explain [1, 2, 3]. SKE involves the extraction of knowledge out of a *black box* [4, 5] to create a surrogate symbolic representation. These techniques are used to improve the interpretability and explainability of machine learning (ML) models, allowing humans to better understand and trust their decisions, as well as to facilitate knowledge transfer between domains. In the context of intelligent agents and multi-agent systems, symbolic knowledge also serves as a fundamental enabler of autonomous reasoning, verifiability, and communication between agents operating in open and dynamic environments.

The literature on SKE techniques is quite extensive [6, 7, 8, 9, 10, 11, 12, 13, for instance], and there is no one-size-fits-all solution for every applicative scenario. Each technique has its strengths and limitations, and the selection of the best one depends on the specific requirements of the application and the data peculiarities. Indeed, the extracted knowledge quality depends on a variety of factors, e.g., the input data distribution, the applied pre-processing strategy, and the adopted feature selection technique. As a result, it is often necessary to experiment with multiple SKE techniques and compare their performance on the specific combination of (processed) data set and black-box model to identify the best approach. Selecting the best technique for the case at hand is thus a complex task requiring careful consideration of the specific application requirements and a deep understanding of the strengths and limitations of each technique. This challenge becomes even more critical when symbolic knowledge is intended to support deliberation, norm reasoning, or shared mental models amongst agents, which demand high-quality and consistent symbolic representations.

26th Workshop From Objects to Agents (WOA 2025)
*Corresponding author.
f.sabbatini1@campus.uniurb.it (F. Sabbatini); roberta.calegari@unibo.it (R. Calegari)
0000-0002-0532-6777 (F. Sabbatini); 0000-0003-3794-2942 (R. Calegari)

According to the literature, the quality of knowledge obtained through SKE can be assessed upon several indices [14, 15]: *(i)* accuracy [16] – evaluable through a comparison between the extracted knowledge and reference data to highlight how well the extracted knowledge matches the actual one in the data –, *(ii)* completeness [17] – extent to which all relevant information is captured –, *(iii)* clarity or readability [18]—ease of understanding and interpretation, generally assessed via comprehensibility of the extracted knowledge. These dimensions are essential not only for human interpretability, but also to ensure that agent-oriented reasoning over symbolic knowledge remains transparent, tractable, and semantically rich.

Knowledge extracted via SKE is usually compared manually and by considering these indices separately. Manual evaluation can be time-consuming and prone to subjective biases, as different human evaluators may have different opinions about relevance, completeness, clarity, and other knowledge aspects. Such limitations hinder the deployment of autonomous agents capable of self-assessing and improving their own symbolic knowledge bases.

The importance of automating such a process of evaluation should also be considered in the light of an automated ML (AutoML) perspective [19]. In the context of SKE, AutoML techniques can automate the quality-evaluation process of extracted knowledge and thus the selection of the most suitable SKE technique for the task at hand, saving time and reducing the potential introduction of subjective biases. By automating this process, SKE systems can become more efficient and effective at extracting relevant, complete, consistent, clear, and readable knowledge from unstructured data sources[1]. Such automation also lays the foundation for adaptive agents capable of selecting, refining and integrating symbolic knowledge dynamically as new data or contexts emerge.

Some recent works have started to highlight these issues and have introduced metrics for automated evaluation [20, 21, 22]. Nonetheless, the metrics proposed thus far are still limited in scope and fail to encompass all the necessary evaluation criteria and simultaneously integrate user feedback and customisation. Therefore, there is still the need of a comprehensive and flexible score to automatically evaluate the quality of SKE output knowledge, by considering multiple evaluation criteria simultaneously as well as user customisation. A reliable and customisable metric would not only support more effective knowledge extraction pipelines, but also promote the emergence of autonomous agents capable of verifying and refining their own symbolic models in a principled way. Accordingly, in this paper we propose WInd as a comprehensive scoring metric designed to assess knowledge quality, to the benefit of automated evaluation and comparison of symbolic knowledge, including the outputs provided by SKE procedures.

2. Background and Motivations

SKE techniques are currently adopted to face several different real-world problems, especially in critical areas [23, 24, 25, 26, 27]. They generally provide output knowledge according to a symbolic representation that can be exploited to obtain interpretable predictions. It is widely acknowledged in the literature to assess knowledge quality based on its predictive performance, human-readability extent and completeness [28, 9, 10, 29, 30, 31, for instance]. The observed measurements for these indices vary depending on the chosen SKE algorithm as well as on the user-defined parameters of the algorithm itself, but also on the predictive capabilities of the underlying opaque predictor that needs to be explained. Comparisons may thus be carried out between different extraction procedures, but also between instances of the same extractor differently parametrised or applied to different black boxes.

In order to identify the best knowledge, the three aforementioned quality indices have to be compared over a possibly large set of candidates, during a time-consuming task susceptible to being affected by human biases. Therefore, this task may surely benefit from an automated selection technique, based on a formal scoring metric.

[1] A potential approach is to use reinforcement learning techniques to iteratively improve the quality of the extracted knowledge over time, based on feedback from users or domain experts. This could enable SKE systems to adapt to changing data sources

To achieve high quality, the evaluated knowledge should exhibit at the same time high predictive performance, high human readability, and high completeness. Predictive performance is related to the knowledge capability of providing accurate outputs when queried with instances to be predicted. Readability expresses the human effort required to understand the rationale behind the predictions. Completeness refers to the rate of predictions that the knowledge can offer in relation to the user queries (it is not relevant to consider the prediction goodness, but only the presence/absence of output responses).

The comparison of a knowledge set is trivially performed when it is possible to find a candidate knowledge maximising all three indices. Such a candidate results being the best knowledge in the set. Unfortunately, in real-world applications, it is very common to face a fidelity/readability trade-off, intended as the comparison between knowledge having high predictive performance but small readability and knowledge with higher human readability but smaller predictive performance [20]. The selection of the best knowledge in this scenario should carefully consider both parameters and be subject to a rigorous comparison that is not biased by humans. Nonetheless, it is important to let human users choose an adequate weight for the different quality indices, in order to adapt the comparison with respect to the sensitivity and the goal of the task at hand. In other words, in the same set of knowledge, it is possible to have more than a unique best candidate, given that depending on the given scenario users may want or need to privilege, for instance, the knowledge with the highest predictive performance despite a suboptimum readability extent, or, vice versa, the knowledge with highest human readability despite its predictions are not the most effective. A comprehensive and flexible scoring metric should thus accept some kind of user feedback to be applicable in real-world scenarios without limitations.

This issue has been debated in [20], where a knowledge quality scoring metric named FiRe is presented. FiRe is a flexible metric, since it accepts a user-defined parameter to tune the relevance of the knowledge readability with respect to its predictive performance. However, it is not comprehensive, given that it neglects the completeness index when calculating the knowledge quality score. FiRe is a multiplicative scoring function considering predictive performance and human readability expressed as *losses*—i.e., predictive loss and readability loss, calculated as predictive error and knowledge size, respectively. Examples of knowledge sizes may be the number of rules in a list, the number of leaves in a decision tree, or the number of rows in a decision table, depending on the knowledge representation. This implies that good knowledge quality is associated with small losses and thus with small FiRe scores, given that losses are multiplied.

Another quality metric, Q_s, also based on index loss multiplication has been proposed in [21]. The main differences with respect to FiRe are the inclusion of the knowledge completeness loss in the metric and the inability to let users tune the relative loss weights. Q_s is thus comprehensive, but it offers no flexibility.

To our knowledge, no other metrics assessing symbolic knowledge quality have been proposed in the literature. A complete metric, encompassing predictive performance, human readability, and completeness indices with the possibility to tune their relative importance in the overall score calculation, is thus still missing. Such a metric is the basic brick for enabling an impartial, standardised, and concise evaluation of symbolic knowledge quality. It is worth noting that the capability of evaluating symbolic knowledge quality with these properties is essential for AutoML procedures, as it would enable the automatic selection of high-quality symbolic knowledge representations, which in turn would lead to more interpretable and trustable ML models. Without such evaluation metrics, AutoML algorithms may select suboptimal symbolic knowledge representations that could result in poor model performance and wasted resources.

This study introduces WInd as a comprehensive and flexible scoring function that addresses the gap in the literature concerning the assessment of symbolic knowledge. WInd merges the advantages of the two previously mentioned metrics, i.e., the flexibility of FiRe and the comprehensiveness of Q_s. Indeed, WInd incorporates all three indices found in the literature as common proxies to evaluate symbolic knowledge and it accepts a user-defined weighting parameter for each index. As a result, different symbolic knowledge can be easily compared in terms of predictive performance, human readability

Table 1
Functions involved in the definition of the WInd scoring metric: definition, meaning and parameters.

Function	Description	Parameters	
Accuracy $P(p, \varphi)$	Capability to give the correct predictions	p	Predictive performance loss of the knowledge e.g., misclassification rate, mean absolute error
		φ	Importance of the predictive performance loss
Readability $R(r, \rho)$	Human-readability extent	r	Readability loss expressed as knowledge size e.g., number of rules/leaves in a list/tree
		ρ	Importance of the readability loss
Completeness $C(c, \xi)$	Capability to give an output response	c	Completeness loss of the knowledge e.g., rate of unprovided predictions
		ξ	Importance of the completeness loss

and completeness by exploiting the WInd metric, providing a quantitative and formal score easily customised by users according to their needs.

Quality indices. Knowledge quality is generally evaluated through the aforementioned indices, i.e., predictive performance, human readability, and completeness [14, 15]. There is no unique method to compute them.

Predictive performance may be assessed through the same methods adopted for any predictors. It may be evaluated with respect to the ground truth of a data set or the outputs of an opaque model that the symbolic knowledge is mimicking. Assessments are task-dependent. For classification tasks, the accuracy and F_1 scores are generally adopted. For regression tasks, the most common choices are the mean absolute/squared error (MAE/MSE) and the R^2 score.

Readability is usually related to knowledge size, e.g., an SKE algorithm producing a list of n rules is more readable than another one providing a list(tree) having $2n$ rules(leaves) [32]. However, we acknowledge that this simplification does not fully capture the notion of readability, which the internal complexity of each individual rule or symbolic element can also influence. For instance, a larger set of individually simple rules may be easier to interpret than a smaller set of structurally complex ones, involving nested logical constructs, non-linear thresholds, or fuzzy predicates. In this work, we adopt knowledge size as a proxy for readability due to its objective measurability, ease of interpretation, and broad applicability across different symbolic representations. Nevertheless, we explicitly recognise that this is a coarse-grained approximation. A finer-grained assessment of readability — accounting for both the number and the syntactic/semantic complexity of individual knowledge items — remains an open challenge. We view the integration of such refined metrics as a natural extension of the WInd metric, and outline it as a direction for future work. Further readability information can be included, as the complexity of individual knowledge items. However, there are no available techniques to quantitatively and formally assess item readability, e.g., a tree leaf describing an M-of-N logic rule with respect to a decision table entry related to a fuzzy rule [20]. For this reason, the knowledge size is usually considered sufficient to express readability thanks to its straightforward interpretation, even though any other more refined readability assessment, also considering the readability of each individual knowledge item, can be exploited.

Completeness can be measured as the percentage of input feature space that is covered by the knowledge, equivalent to the input feature subspace where the knowledge is able to draw predictions. When this measurement requires too much effort, e.g., for data sets with a large number of input features, it is possible to estimate the completeness by querying the knowledge with a set of instances and calculating the percentage of provided responses.

3. The WInd Metric for Knowledge Quality

The WInd (**W**eighted quality **Ind**ex) score has been designed to provide a concise knowledge quality evaluation based on predictive performance, human readability and completeness, all expressed as *losses*. In the following we refer to these assessments as *raw quality indices*. Flexibility is ensured by three weighting parameters that can be tuned by users to influence the metric's behaviour according to their application-specific needs. The adoption of parametrised metrics that adapt to end-users' needs is an established practice in the ML literature. Examples are the F-measure and the pinball loss, inspiring this work [33, 34].

WInd is a multiplicative function of three terms, each one constituted by an exponential function aimed at weighting a raw quality index with the corresponding user-defined weight and then squashing the result within the $(0, 1]$ half-open interval. The reason behind the exploitation of a multiplicative function for WInd rather than other statistical aggregation functions (e.g., minimum or maximum) descends from the need to avoid the prevalence of a single term over the others, resulting in equivalent WInd scores even for knowledge pieces with non-equivalent quality. WInd is formally defined as the following continuous and differentiable function:

$$\text{WInd} : \left(\mathbb{R}^+ \times \mathbb{R}^+ \times [0,1] \times \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+\right) \to (0,1], \tag{1}$$

$$\text{WInd}(p, r, c, \varphi, \rho, \xi) = P(p, \varphi) \cdot R(r, \rho) \cdot C(c, \xi), \tag{2}$$

where p, r and c are the raw indices for the knowledge predictive performance, human-readability extent and completeness, respectively, expressed as losses. φ, ρ and ξ are the corresponding user-defined weights. $P(\cdot)$, $R(\cdot)$ and $C(\cdot)$ are the three exponential functions denoting the accuracy, readability and completeness scores, respectively, that are multiplied together to obtain the final WInd score. Table 1 resumes the meaning and parameters of the WInd underlying functions. Given the multiplicative nature of WInd, the scaling introduced by the parametrised exponentials is propagated to the overall metric score, which shares the same $(0, 1]$ range. The WInd functions are formally defined as follows:

$$P: \quad (\mathbb{R}^+ \times \mathbb{R}^+) \to (0,1], \qquad P(p, \varphi) = e^{-3\varphi p^2}, \tag{3}$$

$$R: \quad (\mathbb{R}^+ \times \mathbb{R}^+) \to (0,1], \qquad R(r, \rho) = e^{-0.01\rho r^2}, \tag{4}$$

$$C: \quad ([0,1] \times \mathbb{R}^+) \to (0,1], \qquad C(c, \xi) = e^{-8\xi c^2}. \tag{5}$$

The fixed values appearing in Equations (3)–(5) (i.e., 3, 0.01, 8) have been fine-tuned after a thorough study involving the function properties and the range of admissible values for the raw quality indices. They were selected following a systematic analysis combining mathematical behaviour, value ranges of expected inputs, and sensitivity requirements for each index. The value 3 for predictive loss was chosen to ensure sufficient steepness in penalising moderate to high prediction errors (e.g., misclassification rates above 0.2), especially under high importance settings ($\varphi \geq 1$). This allows WInd to strongly differentiate symbolic knowledge with low accuracy. The constant 0.01 in the readability term accounts for the much larger numerical range of the readability loss (e.g., knowledge size ranging from 1 to 30 or more rules/leaves). A smaller multiplier ensures that the readability contribution is not overly dominant or suppressed in typical symbolic structures. The value 8 for completeness loss was selected to reflect its bounded domain in [0, 1], where even small losses (e.g., 0.1 coverage loss) may have strong semantic significance in certain agent-based applications. This setting emphasises responsiveness to even partial coverage gaps when completeness is weighted heavily ($\xi \geq 1$). These values were empirically validated through grid-based simulations over realistic ranges of losses and weights (see Figure 1 and 2), to ensure desirable monotonicity, boundedness, and discriminative behaviour. Although these constants offer reasonable default behaviour, the design remains modular: users may replace or adjust these constants if domain-specific tuning is required — a feature aligning with wind's principle of flexibility. We emphasise here that completeness, readability and predictive performance loss have very different domains, possibly bounded to the user preferences. Furthermore, the same value for distinct losses may

assume different or even opposite meanings. For instance, a readability loss of 1 is always an optimum achievement (corresponding to very concise knowledge with a single human-interpretable item) and a completeness loss of 1 is always the worst case (knowledge incapable of providing predictions for any input query). Conversely, depending on the specific scenario, a predictive performance loss of 1 may be catastrophic (e.g., when expressed as a misclassification rate it represents 100%) or acceptable (e.g., when expressed as a mean absolute error with respect to a variable ranging between 100 and 200). Consequently, we aimed to parametrise Equations (3)–(5) by considering these issues, with the ultimate goal of designing a versatile and flexible scoring metric imposing no particular constraints on the loss definitions and enabling users to tune the loss relevance coherently. In other words, users can adopt the same values to set the importance of all losses, e.g., importance equal to 3, 1 and 0 to represent losses with high, medium and no relevance in assessing knowledge quality. Although the values of the three WIND score underlying losses vary in magnitude and meaning, this setting is possible thanks to the fixed values of the parameters appearing in Equations (3)–(5).

When considered from an analytical standpoint, the rationale behind the optimisation of these values is to obtain three "well-behaved" exponential functions having some clear characteristics: *(i)* tend to 1 when associated with desirable knowledge properties (e.g., high predictive performance or human readability), *(ii)* tend to 0 for indices denoting poor quality, and *(iii)* have a steepness tunable through an individual user-defined weight parameter. Therefore, each exponential term of WIND has a high value (close to 1) only when related to a raw index expressing good quality and/or to a low user-defined importance for that raw index. Otherwise, terms are dragged towards 0 by quality depletions of the raw indices and/or increases in their weights. Accordingly, the WIND metric assumes values close to 1 for high-quality knowledge and values towards 0 when evaluating poor-quality knowledge.

From the properties of multiplication, it can be also noticed that knowledge is deemed with high quality via WIND only if *all three* underlying exponentials have high scores. Conversely, low quality is associated with a small value of *at least one* exponential. The WIND score trend is shown in Figure 1 for varying values of its parameters.

3.1. Underlying Functions of WIND

As mentioned above, the WIND metric is based on three functions expressing as many scores for the knowledge predictive performance, human-readability extent and completeness. Without loss of generality, only the properties for the accuracy function are discussed here since they also hold for the readability and completeness functions.

The accuracy function $P(p, \varphi)$ requires as parameters a raw predictive loss (p) and its importance in the knowledge quality estimation (φ). WIND imposes no constraints on how the predictive loss should be expressed. The only requirement is that it should be encoded as a value directly proportional to the knowledge predictive error. When performing regression it is possible to adopt the mean absolute or squared errors, whereas for classification tasks the rate of wrong predictions may be a suitable choice. Losses inversely proportional to the F_1 and R^2 scores are also acceptable. It is evident that the domain of $P(\cdot)$ strictly depends on the adopted loss metric. For instance, the misclassification rate ranges in $[0, 1]$, whereas there is no upper bound for the mean errors, whose range is $[0, +\infty]$.

The φ importance has been designed to be a user-defined non-negative real value. When $\varphi = 1$, the predictive loss has a medium impact on the overall WIND score. $\varphi < 1$ assigns small relevance to the predictive loss, implying that a good accuracy score is still possible even if the loss is not very close to 0. Conversely, if $\varphi > 1$ the predictive loss is crucial in the knowledge quality evaluation and thus it must be as close as possible to 0 to enable a good accuracy score.

As a result of the aforementioned considerations, the domain of $P(\cdot)$ is $(\mathbb{R}^+ \times \mathbb{R}^+)$ and from Equation (3) its range can be trivially calculated as $(0, 1]$. The function is thus always positive and bounded from below by 0 and from above by 1. $P(\cdot)$ is continuous and differentiable and from the same equation, it is possible to trivially obtain its first partial derivatives with respect to p and φ, for which the following

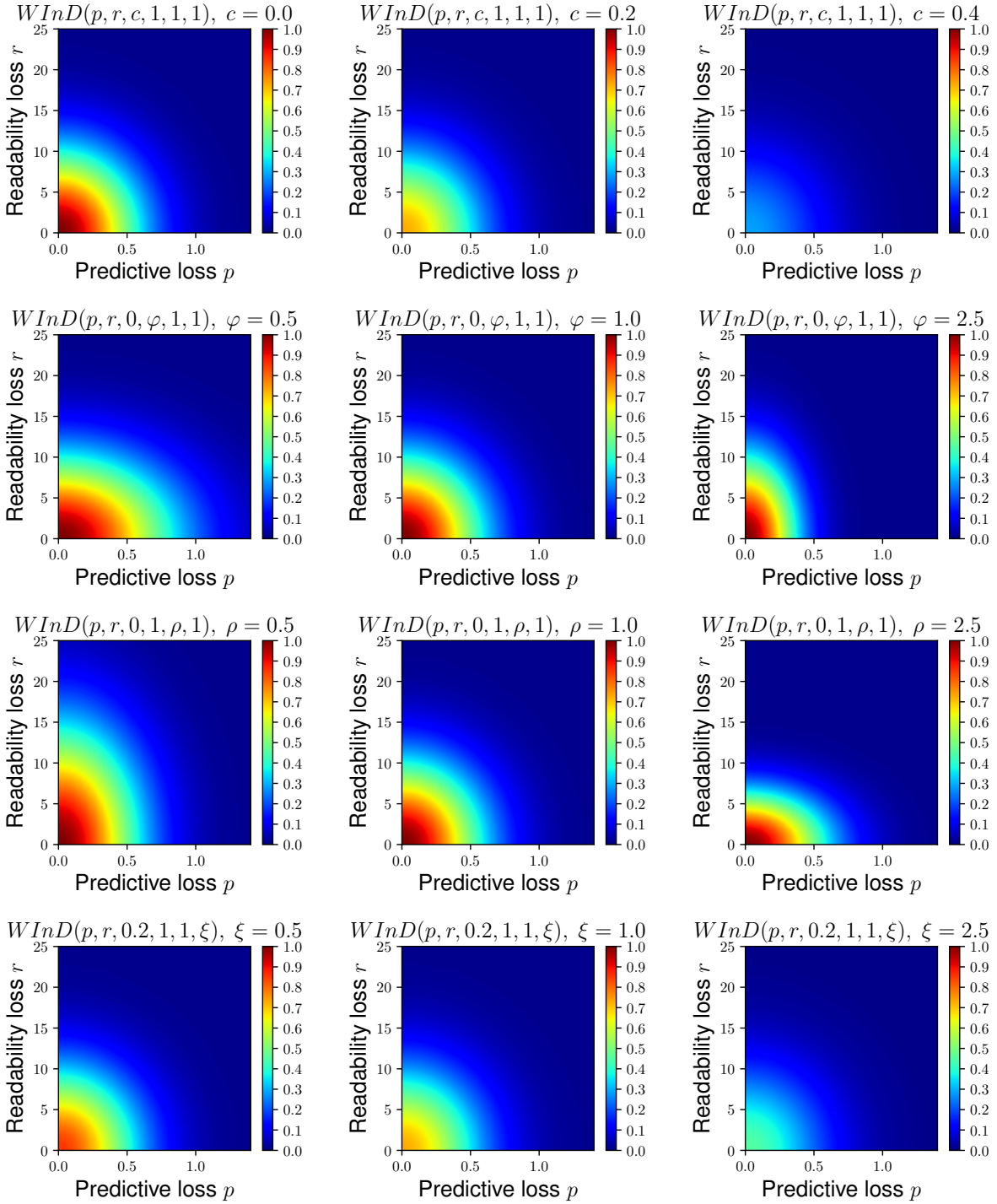

Figure 1: WInd scores (colour) for different values of knowledge predictive loss p (x-axis), readability loss r (y-axis), completeness loss c (top row), predictive loss importance φ (second row), readability loss ρ (third row), and completeness loss importance ξ (bottom row). Parameter values are increased from left to right. Values for the raw indices and losses are selected to show the behaviour of WInd with common real use-case configurations.

properties hold:

$$\frac{\partial P}{\partial p} P(p, \varphi) < 0 \quad \forall p, \varphi \in \mathbb{R}^+, \tag{6}$$

$$\frac{\partial P}{\partial \varphi} P(p, \varphi) < 0 \quad \forall p, \varphi \in \mathbb{R}^+. \tag{7}$$

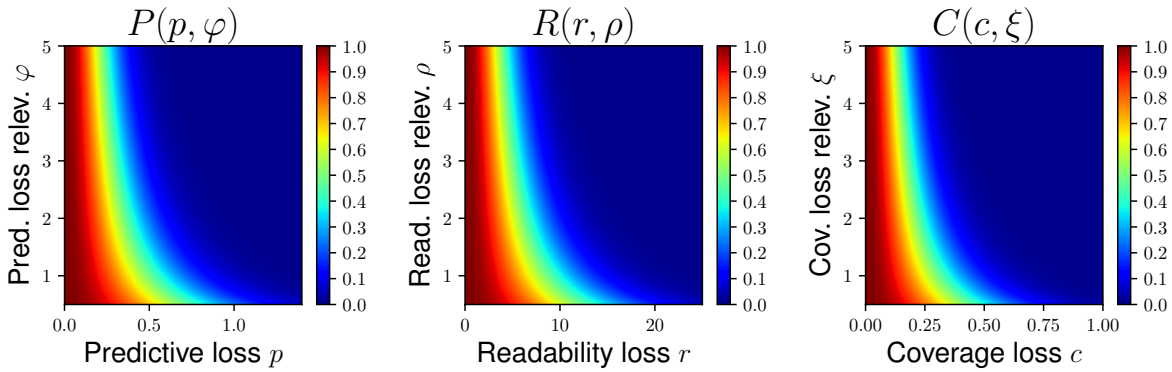

Figure 2: WIND accuracy, readability and completeness function trends in their domains. Functions' input parameters are reported in the axes, the corresponding value is represented by the colour, as depicted in the colourbars. The plots highlight that the underlying functions of WIND have been designed to support the different domains of the raw losses.

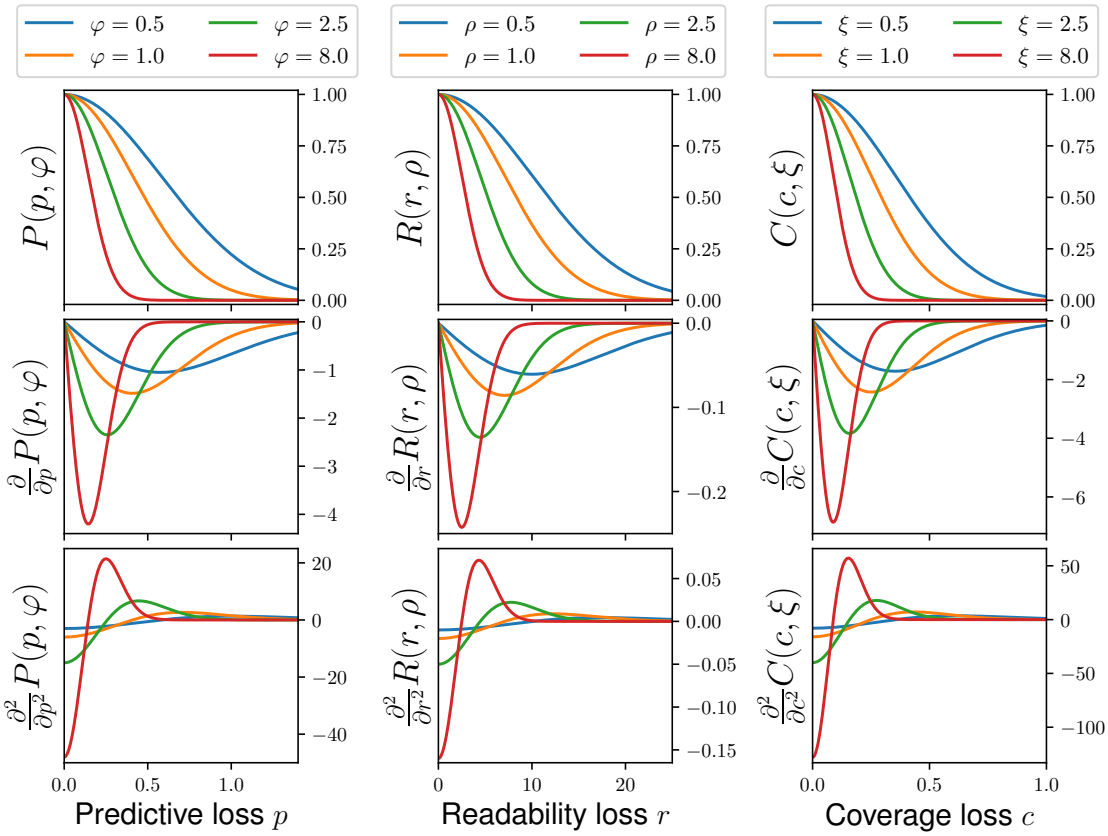

Figure 3: WIND accuracy, readability and completeness functions with respect to predictive performance p, knowledge size r and knowledge completeness c losses and user-defined weights φ, ρ, ξ (top panels). First and second partial derivatives (middle and bottom panels, respectively) with respect to p, r and c. The plots highlight that the underlying functions of WIND have been designed to support the different domains of the raw losses.

Consequently, $P(\cdot)$ is a decreasing monotonic function with respect to both p and φ, i.e., increasing losses lead to decreasing accuracy scores, and the same happens with increasing values for its user-defined weights, as expected.

The same properties of continuity, differentiability, boundedness and monotonicity hold for the readability and completeness functions. The readability function $R(r, \rho)$ requires as parameters a raw readability loss r and its relevance ρ. The knowledge size is a suitable proxy for the readability loss, but

any other measurement proportional to the human effort required to understand the knowledge and/or its predictions is acceptable. The completeness function $C(c, \xi)$ requires a raw completeness loss c and its relevance ξ. Percentages are particularly suited to express completeness losses, for instance via the number of data points not covered by the knowledge over their total amount, or via the uncovered input space volume over the whole volume. As for the ρ and ξ weight, they are subject to the same considerations presented above for φ.

The exponential functions of WInd are shown in Figure 2. Figure 3 depicts their first and second partial derivatives with respect to the knowledge raw losses.

As a final remark on the underlying exponential functions of WInd, it is important to point out that users may inject surrogates for them, or only for a subset of them, as long as substitute functions with the properties described here for our proposals are provided. This constitutes an essential source of flexibility, letting users customise any aspect of the WInd metric, if needed.

3.2. Properties of WInd

From Equation (2) and by considering the properties of the aforementioned exponentials, it is possible to derive the properties of WInd. The WInd metric is a continuous and differentiable function bounded from below by 0 and from above by 1. More in detail,

$$\text{WInd}(\cdot) \simeq 1 \iff P(\cdot) \simeq 1 \wedge R(\cdot) \simeq 1 \wedge C(\cdot) = 1, \tag{8}$$

meaning that the WInd score is equal (close) to 1 if all three exponential scores are equal (close) to 1. The score is exactly 1 only if all knowledge losses are equal to 0 or have an importance equal to 0. On the other hand,

$$\text{WInd}(\cdot) \to 0 \iff P(\cdot) \to 0 \vee R(\cdot) \to 0 \vee C(\cdot) \to 0, \tag{9}$$

meaning that the WInd score is dragged towards 0 by at least an exponential score close to 0 (asymptotic behaviour).

Monotonicity of the WInd metric function descends from the partial derivative analysis reflecting the scoring behaviour by varying individual parameters. WInd is thus a monotonically decreasing function with respect to p, r, c, φ, ρ and ξ for any possible values of these parameters within their domain.

4. Experiments

The effectiveness of the WInd scoring function to assess symbolic knowledge is demonstrated here via the comparison of several outputs provided by the GridEx [35] and CART [7] SKE algorithms applied to naive Bayes (NB) and k-nearest neighbours (k-NN) classifiers. We relied on the ML models and SKE techniques implemented within the scikit-learn[2] and PSyKE[3] Python libraries [36, 37, 38, 39, 40, 41]. WInd is thus exploited to perform quantitative quality evaluations and its results are compared to those of the FiRe and Q_s scores [20, 21]. Experiments are carried out on the Wisconsin breast cancer (WBC) data set [42], a binary classification data set having 30 continuous input features and 569 instances. Both the NB and the k-NN classifiers have been trained with half of the whole data set. A k equal to 15 has been adopted. The same training instances have been fed to SKE models to extract human-interpretable knowledge. The remaining data samples have been used to assess the quality of the extracted knowledge. Quality raw indices observed to evaluate the knowledge are the accuracy score for the predictive performance, the knowledge size as a proxy of the human readability and the coverage in terms of the amount of provided predictions with respect to the test set cardinality. We point out here that CART is based on a decision tree and therefore it always produces complete output knowledge with coverage equal to 1. Conversely, GridEx is based on lists of *if-then* rules that may be non-exhaustive.

[2] https://scikit-learn.org/stable/index.html
[3] https://github.com/psykei/psyke-python

Table 2
Results for the WBC data set (best values in bold). λ stands for *leaves*.

Black box	k-NN				NB			
Extractor	CART		GridEx		CART		GridEx	
Extractor parameters	$\lambda = 3$	$\lambda = 6$	2 splits	3 splits	$\lambda = 3$	$\lambda = 6$	2 splits	3 splits
Classification accuracy	0.92	0.91	0.91	**0.93**	0.90	**0.93**	0.92	0.87
Extracted rules	3	6	5	5	3	6	**2**	14
Coverage	**1.00**	**1.00**	0.94	0.94	**1.00**	**1.00**	0.94	0.93
Q_s score	0.23	0.55	0.48	0.35	0.31	0.42	**0.17**	1.91
FiRe, $\psi = 0.5$	0.49	0.19	0.97	0.72	0.66	0.92	**0.34**	4.08
FiRe, $\psi = 1$	0.24	0.60	0.49	0.36	0.33	0.46	**0.17**	2.04
FiRe, $\psi = 3$	**0.08**	0.20	0.19	0.14	0.11	0.15	0.09	0.73
WInd								
$\varphi = 1, \rho = 1, \xi = 1$	0.90	0.68	0.74	0.75	0.88	0.69	**0.92**	0.13
$\varphi = 3, \rho = 1, \xi = 0$	0.87	0.65	0.72	0.75	0.83	0.67	**0.91**	0.12
$\varphi = 3, \rho = 0.5, \xi = 3$	**0.91**	0.78	0.75	0.79	0.97	0.80	0.86	0.29
$\varphi = 1, \rho = 3, \xi = 0$	0.75	0.33	0.46	0.47	0.74	0.33	**0.87**	0.01
$\varphi = 10, \rho = 0.5, \xi = 3$	**0.80**	0.65	0.63	0.72	0.69	0.72	0.75	0.20

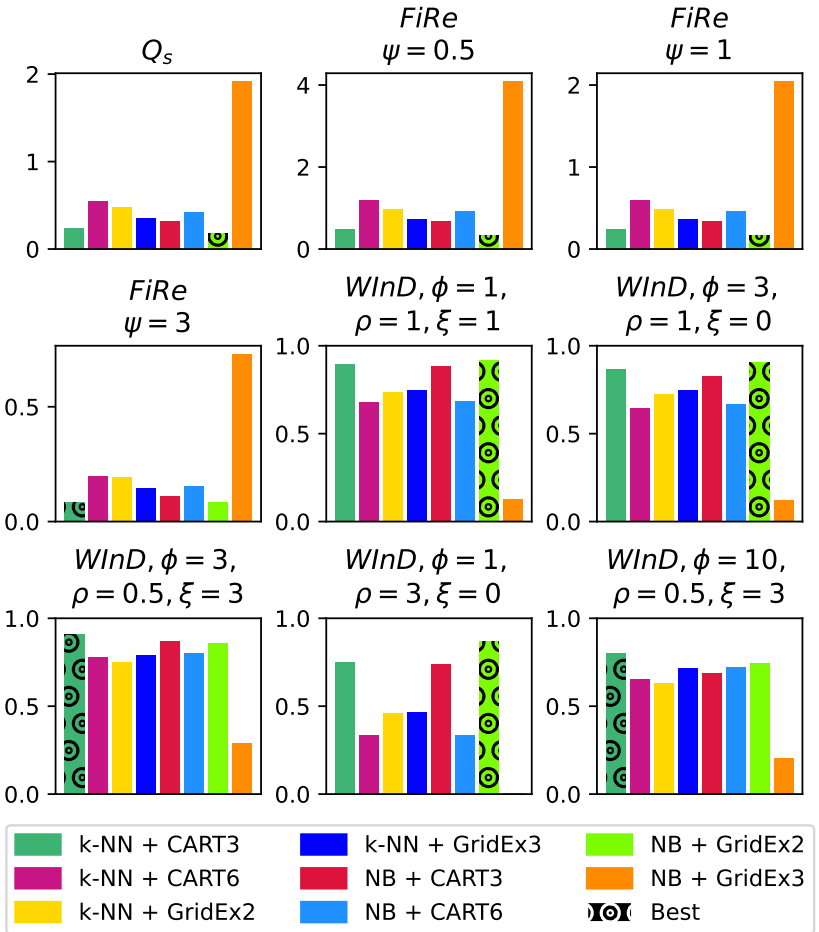

Figure 4: Quality assessments for the SKE techniques.

Listing 1 Rules extracted with GridEx2 applied to the NB for the WBC data set.

```
output is 'benign' if
    'Mean concave points' in [0.00, 0.10] and
    'Worst concave points' in [0.00, 0.14].
output is 'malignant' if
    'Mean concave points' in [0.00, 0.20] and
    'Worst concave points' in [0.14, 0.29].
```

Listing 2 Rules extracted with CART3 applied to the k-NN for the WBC data set.

```
output is 'benign' if 'Worst radius' <= 16.30.
output is 'benign' if 'Worst radius' <= 16.79.
output is 'malignant' otherwise.
```

As for the parameters of SKE algorithms, instances of CART with a maximum amount of 3 and 6 leaves (CART3 and CART6, henceforth) have been applied to both the k-NN and the NB classifiers. Analogously, 2 different parametrisations have been adopted for GridEx. Hyper-parameters required by GridEx are the maximum depth (fixed to 3), the predictive error threshold (fixed to 0.1), the minimum amount of data points to consider when building each rule (fixed to 1) and the partitioning strategy to adopt during the knowledge extraction. We opted for an adaptive partitioning based on the input feature relevance, and in particular instances of GridEx performing 2 or 3 slices only along the most relevant input feature have been trained (GridEx2 and GridEx3, henceforth). Examples of extracted knowledge bases can be found in Listings 1 to 3. To assess the feature relevance we relied on the tools provided by the `scikit-learn` Python library [36].

Quality raw indices measured for each possible association of black-box classifier and SKE algorithm amongst those mentioned above have been used to calculate the Q_s, FiRe and WInd scores. As these scoring metrics require the expression of raw quality measurements as losses, we conducted the following calculations. The predictive loss has been calculated as $1 - accuracy$. The knowledge size has been adopted as readability loss. The completeness loss of Q_s has been calculated as suggested by the authors of the metric as $2 - coverage$, in order not to zero the score (regardless of the other losses) for complete knowledge pieces. Since the completeness loss is handled in WInd by an exponential function, this workaround is not necessary and it is calculated as $1 - coverage$, thus representing the rate of test samples that are not covered by the knowledge. This enables a more intuitive definition of the completeness loss, similar to the predictive loss.

The Q_s score accepts no tuning parameters, so it is applied to the raw losses as-is. The FiRe score accepts a user-defined fidelity/readability trade-off parameter (ψ). $\psi \in \{0.5, 1, 3\}$ have been used for the experiments reported here, denoting high, medium and low predominance of the knowledge readability over predictive performance, respectively (high values tend to neglect the readability impact).

For WInd, several different parametrisations have been tested. We identified 5 possible values for the user-defined weights, i.e., 10, 3, 1, 0.5 and 0, corresponding to very high, high, medium, slight and no importance, respectively. Clearly, this customisation is not univocal, but it is suited to reflect interesting knowledge configurations. The tested parametrisation of WInd are reported in Table 2. In the same table, all quality assessments are reported and compared. The best value for each score is highlighted in bold. We recall here that high-quality knowledge is identified by high accuracy and completeness, small amounts of extracted rules, low Q_s and FiRe scores and high WInd scores. A visual comparison of SKE techniques adopted for our experiments is shown in Figure 4, where the hatched bars correspond to the best instances according to each scoring metric.

Listing 3 Rules extracted with GridEx3 applied to the NB for the WBC data set.

```
output is 'malignant' if
   'Mean concave points' in [0.13, 0.20] and
   'Worst concave points' in [0.19, 0.29].
output is 'malignant' if
   'Mean concave points' in [0.00, 0.13] and
   'Worst concave points' in [0.19, 0.29].
output is 'malignant' if
   'Mean concave points' in [0.06, 0.20] and
   'Worst concave points' in [0.09, 0.19].
output is 'benign' if
   'Mean concave points' in [0.00, 0.04] and
   'Worst concave points' in [0.00, 0.12].
output is 'benign' if
   'Mean concave points' in [0.02, 0.04] and
   'Worst concave points' in [0.12, 0.15].
output is 'benign' if
   'Mean concave points' in [0.05, 0.05] and
   'Worst concave points' in [0.08, 0.10].
output is 'benign' if
   'Mean concave points' in [0.04, 0.05] and
   'Worst concave points' in [0.06, 0.10].
output is 'benign' if
   'Mean concave points' in [0.04, 0.05] and
   'Worst concave points' in [0.10, 0.12].
output is 'malignant' if
   'Mean concave points' in [0.05, 0.06] and
   'Worst concave points' in [0.08, 0.12].
output is 'malignant' if
   'Mean concave points' in [0.04, 0.05] and
   'Worst concave points' in [0.15, 0.17].
output is 'benign' if
   'Mean concave points' in [0.05, 0.05] and
   'Worst concave points' in [0.15, 0.17].
output is 'benign' if
   'Mean concave points' in [0.04, 0.05] and
   'Worst concave points' in [0.12, 0.15].
output is 'malignant' if
   'Mean concave points' in [0.05, 0.06] and
   'Worst concave points' in [0.17, 0.19].
output is 'malignant' if
   'Mean concave points' in [0.05, 0.06] and
   'Worst concave points' in [0.12, 0.17].
```

4.1. Result discussion

The presented experiments produced a set of 8 distinct pieces of knowledge. As shown in Table 2, no single knowledge source exhibits optimum raw quality indices across all aspects. Conversely, it is possible to identify candidates only maximising the predictive performance or the completeness, or only minimising the number of extracted rules. The unique exception is an instance of CART6,

maximising at the same time both completeness and classification accuracy. The identification of the best knowledge should thus be subject to some kind of trade-off between the raw scores.

Depending on the task at hand, it is possible to select more than one candidate with the best knowledge. For instance, GridEx2 applied to the NB classifier extracts only 2 rules (see Listing 1 in the supplementary materials), but these rules are not complete (coverage of 94%) and they have a suboptimum accuracy score (0.92, to be compared with the optimum value of 0.93). If completeness is not mandatory and the human-readability extent of the knowledge is more important than its predictive performance, then this knowledge should be picked as the best one in the set. On the other hand, if completeness is essential, the best knowledge should be selected amongst those provided by CART instances, and in particular those obtained with CART3 applied to the k-NN classifier (3 rules but accuracy of 0.92, see Listing 2 in the supplementary materials) or with CART6 applied to the NB classifier (accuracy of 0.93 but 6 rules). It is reasonable to prefer the CART3 instance given that the very slight worsening in the classification accuracy is balanced with a halving of the knowledge size.

By adopting the Q_s score to compare the knowledge set, the best candidate is GridEx2 applied to the NB model. While this result may be acceptable in some scenarios, the lack of flexibility in Q_s makes it impossible for users to inject the requirement for having complete knowledge into the comparison process.

If evaluated with FiRe, the best knowledge is the same if $\psi \in \{0.5, 1\}$, or the one provided by CART3 applied to the k-NN if $\psi = 3$, given that in this case the knowledge size is less relevant in the overall quality scoring. As mentioned above, both results are acceptable depending on the task at hand, however, when preferring CART, the knowledge completeness is not taken into account by FiRe, given that it does not consider this raw score. Furthermore, the meaning of ψ is not intuitive and tuning its value may be difficult for users. Conversely, WInd appears to be the most flexible, comprehensive and easy-to-customise scoring metric. Indeed, when all the raw quality indices should be evaluated with a medium relevance ($\varphi = 1$, $\rho = 1$, $\xi = 1$), the best knowledge is GridEx2 applied to the NB predictor, in agreement with the assessments of Q_s and FiRe with $\psi = 1$. The same result can be obtained by evaluating with medium relevance the accuracy score and with high relevance the knowledge size, neglecting the contribution of the knowledge completeness ($\varphi = 1$, $\rho = 3$, $\xi = 0$). This WInd customisation corresponds to FiRe with $\psi = 0.5$, indeed they provide the same evaluation, as expected.

The best knowledge is still the same even by augmenting the accuracy score relevance to high and diminishing the knowledge size relevance to medium ($\varphi = 3$, $\rho = 1$, $\xi = 0$).

When assigning slight relevance to the knowledge size and high importance to completeness and predictive performance ($\varphi = 3$, $\rho = 0.5$, $\xi = 3$), the best knowledge is CART3 applied to the k-NN, as expected. By augmenting the predictive performance relevance to very high ($\varphi = 10$) the assessment does not change in favour of the CART instance with optimum classification accuracy because, even with small relevance for the readability score, the very small predictive performance gain is not balanced with the huge human-readability loss. It is worth noting that the knowledge provided by GridEx3 applied to the NB is unanimously deemed the worst according to all the quality indexes adopted in the experiments, given its poor human readability, small completeness and sub-optimum predictive performance.

5. Conclusions

In this paper we introduce the new WInd metric for symbolic knowledge quality assessment. It is based on a set of raw quality indices (i.e., predictive performance, human-readability extent, and completeness) and it accepts user-defined customisation in the form of weights for the raw indices. These characteristics make WInd much more comprehensive and flexible than existing similar scoring functions. A formal definition of WInd is provided and its algebraic properties are demonstrated. Furthermore, we show that our metric may be exploited to compare knowledge rigorously and flexibly, enabling the automated selection of the best knowledge in a set of candidates without renouncing

the capability of tuning the scoring metric according to the task at hand, for instance, by privileging readable knowledge rather than accurate and/or complete alternatives. The WInd metric is particularly suited for agent-based systems where symbolic knowledge underpins decision-making, communication, and coordination. In these contexts, agents benefit from the ability to autonomously assess and prioritise symbolic knowledge depending on their roles, objectives, or environmental constraints. Moreover, WInd opens the way to meta-reasoning capabilities within intelligent agents, by supporting the dynamic selection and refinement of knowledge bases in open and evolving multi-agent environments.

In the future we plan to design a more sophisticated readability function for WInd, enabling the evaluation of readability for individual knowledge items. We also aim to integrate out metric into agent reasoning architectures to support adaptive, self-evaluating agents capable of maintaining high-quality symbolic representations over time.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

Acknowledgments

This work has been supported by PNRR – M4C2 – Investimento 1.3, Partenariato Esteso PE00000013 – "FAIR—Future Artificial Intelligence Research" – Spoke 8 "Pervasive AI", funded by the European Commission under the NextGenerationEU programme and by the European Union's Horizon Europe AEQUITAS research and innovation programme under grant number 101070363.

References

[1] G. Ciatto, F. Sabbatini, A. Agiollo, M. Magnini, A. Omicini, Symbolic knowledge extraction and injection with sub-symbolic predictors: A systematic literature review, ACM Computing Surveys 56 (2024) 161:1–161:35. doi:10.1145/3645103.

[2] E. M. Kenny, C. Ford, M. Quinn, M. T. Keane, Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies, Artificial Intelligence 294 (2021) 103459. doi:10.1016/j.artint.2021.103459.

[3] F. Sabbatini, Four decades of symbolic knowledge extraction from sub-symbolic predictors. A survey, ACM Computing Surveys (2025). doi:10.1145/3749097.

[4] Z. C. Lipton, The mythos of model interpretability, Queue 16 (2018) 31–57. doi:10.1145/3236386.3241340.

[5] A. Rocha, J. P. Papa, L. A. A. Meira, How far do we get using machine learning black-boxes?, International Journal of Pattern Recognition and Artificial Intelligence 26 (2012) 1261001–(1–23). doi:10.1142/S0218001412610010.

[6] N. Barakat, J. Diederich, Eclectic rule-extraction from support vector machines, International Journal of Computer and Information Engineering 2 (2008) 1672–1675. doi:10.5281/zenodo.1055511.

[7] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, Classification and Regression Trees, CRC Press, 1984.

[8] L. A. Castillo, A. González Muñoz, R. Pérez, Including a simplicity criterion in the selection of the best rule in a genetic fuzzy learning algorithm, Fuzzy Sets Syst. 120 (2001) 309–321. URL: https://doi.org/10.1016/S0165-0114(99)00095-0. doi:10.1016/S0165-0114(99)00095-0.

[9] M. W. Craven, J. W. Shavlik, Extracting tree-structured representations of trained networks, in: D. S. Touretzky, M. C. Mozer, M. E. Hasselmo (Eds.), Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference, The MIT Press, 1996, pp. 24–30. URL: http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf.

[10] J. Huysmans, B. Baesens, J. Vanthienen, ITER: An algorithm for predictive regression rule extraction, in: Data Warehousing and Knowledge Discovery (DaWaK 2006), Springer, 2006, pp. 270–279. doi:10.1007/11823728_26.
[11] K. Saito, R. Nakano, Extracting regression rules from neural networks, Neural Networks 15 (2002) 1279–1288. doi:10.1016/S0893-6080(02)00089-8.
[12] R. Setiono, J. Y. L. Thong, An approach to generate rules from neural networks for regression problems, Eur. J. Oper. Res. 155 (2004) 239–250. URL: https://doi.org/10.1016/S0377-2217(02)00792-0. doi:10.1016/S0377-2217(02)00792-0.
[13] R. Setiono, W. K. Leow, J. M. Zurada, Extraction of rules from artificial neural networks for nonlinear regression, IEEE Transactions on Neural Networks 13 (2002) 564–577. doi:10.1109/TNN.2002.1000125.
[14] A. S. d'Avila Garcez, K. Broda, D. M. Gabbay, Symbolic knowledge extraction from trained neural networks: A sound approach, Artificial Intelligence 125 (2001) 155–207.
[15] S. N. Tran, A. S. d'Avila Garcez, Knowledge extraction from deep belief networks for images, in: IJCAI-2013 workshop on neural-symbolic learning and reasoning, 2013.
[16] J. Fan, A. Kalyanpur, D. C. Gondek, D. A. Ferrucci, Automatic knowledge extraction from documents, IBM Journal of Research and Development 56 (2012) 5–1.
[17] D. Demner-Fushman, W. J. Rogers, A. R. Aronson, Metamap lite: an evaluation of a new java implementation of metamap, Journal of the American Medical Informatics Association 24 (2017) 841–844.
[18] C. A. Smith, S. Hetzel, P. Dalrymple, A. Keselman, Beyond readability: investigating coherence of clinical text for consumers, Journal of medical Internet research 13 (2011) e1842.
[19] S. K. Karmaker, M. M. Hassan, M. J. Smith, L. Xu, C. Zhai, K. Veeramachaneni, Automl to date and beyond: Challenges and opportunities, ACM Computing Surveys (CSUR) 54 (2021) 1–36.
[20] F. Sabbatini, R. Calegari, Symbolic knowledge-extraction evaluation metrics: The FiRe score, in: K. Gal, A. Nowé, G. J. Nalepa, R. Fairstein, R. Rădulescu (Eds.), Proceedings of the 26th European Conference on Artificial Intelligence, ECAI 2023, Kraków, Poland. September 30 – October 4, 2023, 2023. URL: https://ebooks.iospress.nl/doi/10.3233/FAIA230496. doi:10.3233/FAIA230496.
[21] F. Sabbatini, R. Calegari, On the evaluation of the symbolic knowledge extracted from black boxes, AI Ethics 4 (2024) 65–74. doi:https://doi.org/10.1007/s43681-023-00406-1.
[22] F. Sabbatini, R. Calegari, ICE: An evaluation metric to assess symbolic knowledge quality, in: AIxIA 2024 – Advances in Artificial Intelligence, volume 15450 of *Lecture Notes in Computer Science*, Springer, Cham, Switzerland, 2025, pp. 241–256. doi:10.1007/978-3-031-80607-0_19, XXIII International Conference of the Italian Association for Artificial Intelligence, AIxIA 2024, Bolzano, Italy, November 25–28, 2024, Proceedings.
[23] G. Bologna, C. Pellegrini, Three medical examples in neural network rule extraction, Physica Medica 13 (1997) 183–187. URL: https://archive-ouverte.unige.ch/unige:121360.
[24] B. Baesens, R. Setiono, C. Mues, J. Vanthienen, Using neural network rule extraction and decision tables for credit-risk evaluation, Management Science 49 (2003) 312–329. doi:10.1287/mnsc.49.3.312.12739.
[25] Y. Hayashi, R. Setiono, K. Yoshida, A comparison between two neural network rule extraction techniques for the diagnosis of hepatobiliary disorders, Artificial intelligence in Medicine 20 (2000) 205–216. doi:10.1016/s0933-3657(00)00064-6.
[26] A. Hofmann, C. Schmitz, B. Sick, Rule extraction from neural networks for intrusion detection in computer networks, in: 2003 IEEE International Conference on Systems, Man and Cybernetics, volume 2, IEEE, 2003, pp. 1259–1265. doi:10.1109/ICSMC.2003.1244584.
[27] M. T. A. Steiner, P. J. Steiner Neto, N. Y. Soma, T. Shimizu, J. C. Nievola, Using neural network rule extraction for credit-risk evaluation, International Journal of Computer Science and Network Security 6 (2006) 6–16.
[28] M. G. Augasta, T. Kathirvalavakumar, Reverse engineering the neural networks for rule extraction in classification problems, Neural Process. Lett. 35 (2012) 131–150. URL: https://doi.org/10.1007/s11063-011-9207-8. doi:10.1007/s11063-011-9207-8.

[29] E. W. Saad, D. C. Wunsch II, Neural network explanation using inversion, Neural Networks 20 (2007) 78–93. URL: https://doi.org/10.1016/j.neunet.2006.07.005. doi:10.1016/j.neunet.2006.07.005.

[30] G. P. J. Schmitz, C. Aldrich, F. S. Gouws, ANN-DT: an algorithm for extraction of decision trees from artificial neural networks, IEEE Transactions on Neural Networks 10 (1999) 1392–1401. doi:10.1109/72.809084.

[31] Z. Zhou, Y. Jiang, S. Chen, Extracting symbolic rules from trained neural network ensembles, AI Commun. 16 (2003) 3–15. URL: http://content.iospress.com/articles/ai-communications/aic272.

[32] I. Czarnowski, A. M. Caballero, R. J. Howlett, L. C. Jain, Intelligent Decision Technologies 2016: Proceedings of the 8th KES International Conference on Intelligent Decision Technologies (KES-IDT 2016)–Part I, volume 56, Springer, 2016.

[33] I. Steinwart, A. Christmann, Estimating conditional quantiles with the help of the pinball loss, Bernoulli 17 (2011) 211–225. URL: https://doi.org/10.3150/10-BEJ267. doi:10.3150/10-BEJ267.

[34] C. J. van Rijsbergen, Information Retrieval, Butterworth, 1979.

[35] F. Sabbatini, G. Ciatto, A. Omicini, GridEx: An algorithm for knowledge extraction from black-box regressors, in: D. Calvaresi, A. Najjar, M. Winikoff, K. Främling (Eds.), Explainable and Transparent AI and Multi-Agent Systems. Third International Workshop, EXTRAAMAS 2021, Virtual Event, May 3–7, 2021, Revised Selected Papers, volume 12688 of *LNCS*, Springer Nature, Basel, Switzerland, 2021, pp. 18–38. doi:10.1007/978-3-030-82017-6_2.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research (JMLR) 12 (2011) 2825–2830. URL: https://dl.acm.org/doi/10.5555/1953048.2078195.

[37] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, On the design of PSyKE: A platform for symbolic knowledge extraction, in: R. Calegari, G. Ciatto, E. Denti, A. Omicini, G. Sartor (Eds.), WOA 2021 – 22nd Workshop "From Objects to Agents", volume 2963 of *CEUR Workshop Proceedings*, Sun SITE Central Europe, RWTH Aachen University, 2021, pp. 29–48. 22nd Workshop "From Objects to Agents" (WOA 2021), Bologna, Italy, 1–3 September 2021. Proceedings.

[38] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, Symbolic knowledge extraction from opaque ML predictors in PSyKE: Platform design & experiments, Intelligenza Artificiale 16 (2022) 27–48. URL: https://doi.org/10.3233/IA-210120. doi:10.3233/IA-210120.

[39] F. Sabbatini, G. Ciatto, A. Omicini, Semantic Web-based interoperability for intelligent agents with PSyKE, in: D. Calvaresi, A. Najjar, M. Winikoff, K. Främling (Eds.), Explainable and Transparent AI and Multi-Agent Systems, volume 13283 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 124–142. URL: http://link.springer.com/10.1007/978-3-031-15565-9_8. doi:10.1007/978-3-031-15565-9_8.

[40] R. Calegari, F. Sabbatini, The PSyKE technology for trustworthy artificial intelligence 13796 (2023) 3–16. URL: https://doi.org/10.1007/978-3-031-27181-6_1. doi:10.1007/978-3-031-27181-6_1, xXI International Conference of the Italian Association for Artificial Intelligence, AIxIA 2022, Udine, Italy, November 28 – December 2, 2022, Proceedings.

[41] F. Sabbatini, R. Calegari, Unlocking insights and trust: The value of explainable clustering algorithms for cognitive agents, in: R. Falcone, C. Castelfranchi, A. Sapienza, F. Cantucci (Eds.), Proceedings of the 24th Workshop "From Objects to Agents", Roma, Italy, November 6–8, 2023, volume 3579 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 232–245. URL: https://ceur-ws.org/Vol-3579/paper18.pdf.

[42] W. N. Street, W. H. Wolberg, O. L. Mangasarian, Nuclear feature extraction for breast tumor diagnosis, in: R. S. Acharya, D. B. Goldgof (Eds.), Biomedical Image Processing and Biomedical Visualization, volume 1905, International Society for Optics and Photonics, SPIE, 1993, pp. 861 – 870. URL: https://doi.org/10.1117/12.148698. doi:10.1117/12.148698.