Measures of model interpretability for model selection

André Carrington[1 [0000−0002−0062−5567]], Paul Fieguth[2], and Helen Chen[3]

[1] Systems Design Engineering, University of Waterloo, ON, Canada amcarrin at uwaterloo.ca
[2] Systems Design Engineering, University of Waterloo, ON, Canada pfieguth at uwaterloo.ca
[3] School of Public Health and Health Systems, University of Waterloo, ON, Canada helen.chen at uwaterloo.ca

Abstract. The literature lacks definitions for quantitative measures of model interpretability for automatic model selection to achieve high accuracy and interpretability, hence we define inherent model interpretability. We extend the work of Lipton *et al.* and Liu *et al.* from qualitative and subjective concepts of model interpretability to objective criteria and quantitative measures. We also develop another new measure called simplicity of sensitivity and illustrate prior, initial and posterior measurement. Measures are tested and validated with some measures recommended for use. It is demonstrated that high accuracy and high interpretability are jointly achievable with little to no sacrifice in either.

Keywords: model interpretability · model transparency · support vector machines · kernels

1 Introduction

For machine learning (ML) models, data and results, there is a demand for transparency, ease of understanding and explanations [24] to satisfy a citizen's "right to explanation" in the European Union [20] and to meet health care requirements for justification and explanation [7, 22].

Without quantitative measures of transparency and understandability, doctors (or users) will select models which maximize accuracy but may unnecessarily or unintentionally neglect or sacrifice transparency and understandability, or they will choose models in an ad hoc manner to try and meet all criteria. We refer to the transparency and understandability of models as ***inherent model interpretability***—defined further in Section §3.

We propose criteria and measures of inherent model interpretability to help a doctor select ML models (Table 1 steps 1 and 2) which are more transparent and understandable, in a quantitative and objective manner. More transparent models can offer additional views of results (Table 1 step 3) for interpretation. Our measures facilitate the inclusion of better models as candidates and the selection of better models for use.

Table 1. Measures of inherent model interpretability facilitate model selection (bold text) in steps 1 and 2.

Step	Task	Basis for task
1	The doctor **selects candidate models** for learning and testing based on...	Data types and distributions, Inherent model interpretability (transparency of model)
2	The machine learns model weights for optimal accuracy with various parameters. The doctor **selects the model to use** based on...	Accuracy, Inherent model interpretability (transparency of model and understandability of results)
3	The doctor uses the model to classify new data. The doctor understands and interprets the result and model based on...	Theory, Views of results, Additional views of results
4	The doctor explains the result and model to a patient or peer based on...	Selected interpretations, Theory

Some of our proposed measures are specific to support vector machines (SVM), as one popular ML method. We perform experiments to validate the SVM measures against a set of propositions and evaluate their utility by concordance or matched pair agreement.

Notably, the proposed measures **do not** provide an interpretation or explanation. They also **do not** indicate how useful or meaningful a model is in the context of data. For example, a model that always classifies patient data as belonging to the positive class is very understandable (interpretable). We can easily construct the explanation of the model and result—all patients are classified as positive—but that does not mean that the model is useful, meaningful, appropriate, or unbiased. Accuracy and common sense address the latter issues. The proposed measures only indicate how understandable a model is, i.e., how likely we are **able** to provide an interpretation, as the necessary basis for subsequent explanation.

Making ML more interpretable facilitates its use in health care because there is a perception that ML is a black box [31] lacking interpretability which inhibits its use. Greater use is important because for a good number of health care problems and data, ML methods offer better accuracy in classification [12, 15, 41] than common alternatives among statistical methods, decision trees and rule-based methods and instance-based methods. Interpretable ML also facilitates research on models and model fit.

2 Notation

A machine learning task begins with data in a matrix X consisting of N instances $\underline{x}_i$ which are vectors, each containing n features.

$$X = \left[\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\right]^{\mathrm{T}} \qquad \underline{x}_i \in \mathbb{R}^n \tag{1}$$

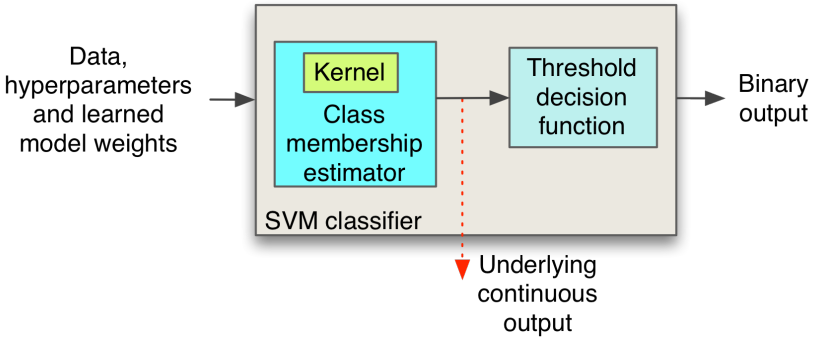

Fig. 1. A model consists of a learning method, SVM in this case, and all of its associated parts as depicted above. Most machine learning and statistical models (or classifiers) have an underlying continuous output that most accurately describes the model's behaviour.

Entry $x_{i,j}$ in the matrix is the j^{th} feature of instance $\underline{x}_i$. We assume real-valued features converting any atomic data type to reals as needed (Appendix B).

A supervised learning task also has N targets (or outcomes) in a vector $\underline{y}$ which are binary in classification,

$$\underline{y} = \left[y_1, y_2, \ldots, y_N \right]^{\mathrm{T}} \qquad y_i \in \{-1, +1\} \tag{2}$$

or continuous in regression:

$$\underline{y} = \left[y_1, y_2, \ldots, y_N \right]^{\mathrm{T}} \qquad y_i \in \mathbb{R} \tag{3}$$

In binary classification there are N^+ instances in the positive class and N^- instances in the negative class.

We refer to a **posterior model** (e.g., Figure 1), or simply **model**, as a learning method (e.g., SVM, neural networks) with all of its associated learning/estimation functions (e.g., kernels and transfer functions), hyperparameters, structure (e.g., layers, connections, components in a composite kernel), constraints and learned model weights, *in the context of specific data*. A model only learns from, and has meaning in, the context of specific data.

We refer to an **initial model** as a model in the context of specific data with initial model weights prior to learning/iteration.

We refer to a **family of models**, or a **prior model**, as the set of models possible when hyperparameters are variables (not specified)—e.g., SVM with a Gaussian RBF kernel with unspecified box constraint and kernel width.

The prior, initial and posterior models are available at different points in the process of machine learning and/or statistical learning process (Figure 2).

Other notation is introduced in the context of discussion.

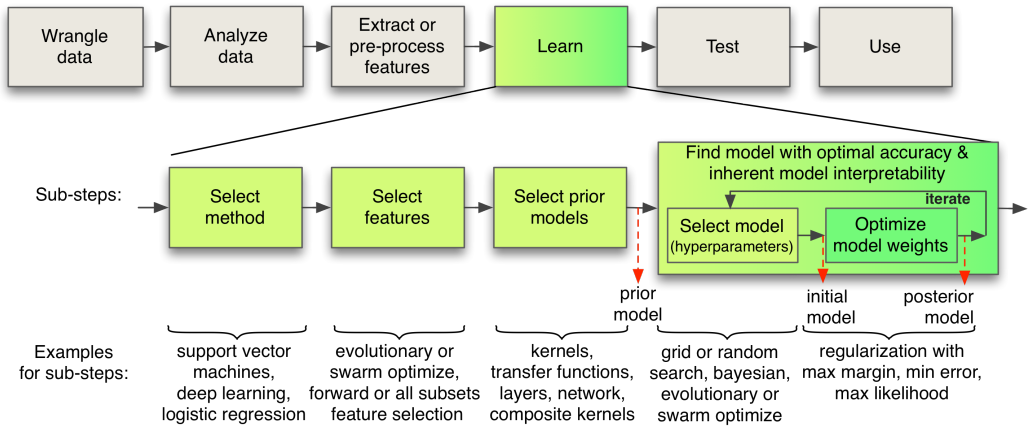

Fig. 2. We measure inherent model interpretability at several points (dashed arrows) in the process of machine learning and/or statistical learning (partially derived from [25]). Note: some steps may not apply to some methods and models.

3 Inherent model interpretability concept and measures

We propose the concept of inherent model interpretability as distinguished from an individual's understanding and we propose two measures for any learning method or model with numeric inputs.

Feynman said that if we understand a concept we must be able to describe it at a freshman level, which often requires simplification or reduction, otherwise we don't really understand it [21]. Badii et al express that complexity is closely related to understanding and that understanding comes from accurate models which use condensed information or reduction schemes[4]. Miller indicates that selection is a key attribute of explanations [38]. Hence, we posit that the simpler a model is, the easier it is to understand, interpret and describe, with all other aspects of the model being equal. This leads to the following general measure.

3.1 A general measure of inherent model interpretability

As stated above, the simpler a model is, the more interpretable it is, inherently. Formally, we propose the following definition.

Definition 1. *Inherent model interpretability (or understandability) U, is a measure with range $[0,1]$ based on either: a measure of model transparency T in the same range, the inverse of semi-infinite model complexity H_∞, or the inverse of finite model complexity H_b, respectively as follows:*

$$U = \begin{cases} T & (i)\ T \in [0,1] \\ \frac{1}{1+(H_\infty - a)} & (ii)\ H_\infty \in [a,\infty) \quad a \in \mathbb{R}^+;\ a < \infty \\ 1 - \left(\frac{H_b - a}{b-a}\right) & (iii)\ H_b \in [a,b] \quad a, b \in \mathbb{R}^+;\ a, b < \infty \end{cases} \tag{4}$$

where:

- *H_∞ and H_b are measures of model complexity based on parts [4] in the categories of information, entropy, code length or dimension [33],*
- *inherent indicates that the measure is independent of an individual, e.g., their specific learning and forgetting curves [44], and*
- *the multiplicative inverse [29] in (4).ii or additive inverse [57] in (4).iii are applied as needed for **absolute** or **relative** measure respectively according to the comparison required. The relative measure is preferred where applicable since it is more intuitive and interpretable (not shown).*
    - *e.g., to compare a set of models where the range $[a, b]$ is known to encompass them all, a relative measure (iii) is fine, however, to compare them to any future model where the maximum b is not known, use an absolute measure (ii), i.e., let $b = \infty$.*

The separation of model interpretability into at least two parts, one part that is inherent to the model (and data) and another part that depends on the individual, aligns with the functionally-grounded approach [17].

In order to use this general measure, one must further define T, H_∞ or H_b, as we do in subsequent sections. We note also that measurement may be performed prior to, initially at, or posterior to, optimizing the model weights (Figure 2).

3.2 A new measure: simplicity of output sensitivity

We consider the continuous underlying output of a classifier (e.g., Figure 1) to be the most accurate representation of a classifier's behaviour. It is available most learning classifiers, in machine learning or statistical learning, such as, neural networks, SVM, logistic regression and naive bayes. It is also facilitated by most implementations, e.g., for SVM it is available in Matlab, R, Python, SPSS, Weka, libsvm and Orange, where the output may be the probability of the positive class or a non-probabilistic value, e.g., "classification score".

Some measure or analyze a classifier's behaviour based on its binary output instead [46]—this approach lacks fine-grained behavioural information. Others measure classifier behaviour by modeling its responses with a separate explanation model that provides a continuous output [46, 5]—this post hoc approach may not meet untested legal, assurance or business requirements.

We use the underlying continuous output, and the logic similar to the previous measure to posit that:

If a model is **uniformly sensitive** in its output to changing values in input features and instances, then its ***sensitivity*** is simple to describe, understand and interpret (as one value). Conversely, a model that is **differently sensitive** to each feature and instance is more difficult to describe, understand and interpret, in those terms or from that perspective. Formally, we propose the following definition:

Definition 2. *The simplicity of output sensitivity U_{Hs} is a measure of inherent model interpretability. It describes the simplicity of the sensitivity of the model's continuous output (e.g., Figures 1) to changes in input. It is specified as the inverse of Shannon entropy H_s with a finite range (4.iii),*

repeated below:

$$U_{Hs} = 1 - \left(\frac{H_s - b}{b} \right) \qquad H_s \in [0, H_{max}] \tag{5}$$

$$H_s = -\sum_i f_i(s) \log f_i(s), \quad i = 1 \ldots N_s \tag{6}$$

$$H_{max} = -\sum_{i=1}^{|s|} \frac{1}{|s|} \log \frac{1}{|s|} \tag{7}$$

where s is the set of sensitivities $S_{j,q}$ of the model's continuous output $\hat{y}_c$ (the value which is underlying for a classifier) to small changes $\varepsilon = (0.1) \cdot 3\sigma$ in each input instance j, one feature q at a time,

$$s = \{S_{j,q}\} \tag{8}$$

$$S_{j,q} = \frac{\hat{y}_c(\underline{x}_j + \underline{\varepsilon}_q) - \hat{y}_c(\underline{x}_j - \underline{\varepsilon}_q)}{2\varepsilon} \tag{9}$$

$$\underline{\varepsilon}_q = \begin{bmatrix} \ldots 0 \ \varepsilon \ 0 \ \ldots \end{bmatrix}^T \qquad \varepsilon in \ q^{th} \ cell$$

and where N_s is the number of bins according to standard binning methods for histograms [47, 18, 53].

We use entropy to measure the global complexity of sensitivities across the space for input data. In the literature, entropy has been applied quite differently to measure the information loss of perturbed features, to indicate their influence — we use entropy instead to measure the complexity of influence with perturbed features.

Our measure uses a first-order central difference (first derivative approximation) as a standard and easy to understand approach to sensitivity that does not require knowing or differentiating the model's formulas. We can generalize this idea to second and third-order differences/derivatives, and so on, like the derivatives in deep Taylor decomposition [39] — but the latter requires a model's formulas and derivatives. Whereas [39] examines the local behaviours of a model, we do that and compute the complexity of the values.

We treat the entries $S_{j,q}$ as a set or random variable s (8) because we are measuring model interpretability overall, across features and instances, not within a feature nor within an instance.

We note that instead of Shannon entropy, it may be possible to apply other types of entropy, such as Renyi entropy, Tsallis entropy, effective entropy or total information [45, 56, 19] and/or Kullback-Leibler (K-L) divergence [14], however such a change would require validation. Prior to this study we experimented with discrete Kullback-Leibler (K-L) divergence as implemented by four measures in the ITK toolkit [55, 54], as an alternative to Shannon entropy, however, our experimental results with K-L divergence did not sufficiently match our expectations, so we focused on Shannon entropy as a more popular and credible measure.

We also implemented differential entropy [14], which is the continuous version of entropy and is defined as the K-L divergence from a uniform probability density function (pdf) to the pdf of interest,

but put that aside based on the previously mentioned K-L divergence results and also because it was more compute intensive as it required a kernel density estimate.

Finally we note that the sensitivity portion of our measure (i.e., entropy aspect aside) differs from how other authors compute sensitivity globally across both instances and features [27].

Table 2. We identify criteria for model interpretability in the literature and translate these into proposed criteria which are objective rather than subjective.

Term	Criteria in the literature	ID	Proposed criteria
Interpretable [34] Decomposable [30]	Each calculation has an intuitive explanation [30].	(a)	The feature space is known/explicit.
		(b)	The feature space has a finite number of dimensions.
	Inputs are interpretable, not anonymous or highly-engineered [30]. Generalized additive models are interpretable [34]	(c)	The model is generalized additive ***with**** known/explicit basis/shape functions.
	Generalized linear models are interpretable [34]. The contributions of individual features in the model, are understandable [34].	(d)	The model is generalized linear [34]
		(e)	The model is multiplicative, e.g., probabilistic, ***with*** known/explicit basis/shape functions.
	n/a	(f)	Model parts are uniform in function.
Transparent algorithm [30]	The training algorithm converges to a unique solution [30].	(g)	Model weights are learned by convex optimization or direct computation.

*Note: Unlike functions of a single variable, basis/shape functions are only available if the kernel is separable.

4 Criteria for model transparency and a measure for SVM

We identify criteria for model transparency from the literature (Table 2) for any model, and propose new criteria in most cases, which are objective, not subjective, and thus suitable for a (quantitative) measure of model transparency.

We apply the proposed criteria (Table 2) for any model, to create a measure specific to kernel methods or support vector machines (SVM).

Using the seven proposed criteria for inherent prior model interpretability (section 4) to define 6 Dirac (binary) measures for SVM (Table 3) meeting each criterion without overlap, except for criterion d (since all SVM kernels are generalized linear models).

We define an overall measure as follows:

$$\check{U}_{\partial} = 1/6\left(\partial_{\text{essep}} + \partial_{\text{fin}} + \partial_{\text{eM}} + \partial_{\times} + \partial_{\text{uni}} + \partial_{\text{adm}}\right)$$

A benefit of this measure is that while independent of the data, it requires little computation and it informs model selection prior to optimization.

Table 3. For kernel methods, e.g., SVM, we propose the following Dirac (binary) measures ∂ of model transparency T. Let $\mathscr{X}_{\mathrm{T}}$ be the space of transparent features derived from simple transforms of the original features $\mathscr{X}$ which are not highly engineered: i.e., given data $\mathscr{X} = \{x\}$, let $\mathscr{X}_{\mathrm{T}} = \left\{x, -x, \frac{1}{x}, \log(x), \tanh(x), \min\left(c_{\mathrm{top}}, x\right), \max\left(c_{\mathrm{bottom}}, x\right)\right\}$.

Name of measure and criterion met	Symbol for measure	Conditions for measure to be true
Explicit symmetric separable (a)	$\partial_{\mathrm{essep}}$	$k(\underline{x}, \underline{z}) = \underline{\phi}(\underline{x})\,\underline{\phi}(\underline{z})$, $\underline{\phi}$ known $x_i, z_i \in \mathscr{X}_0,\ \mathscr{X}_0 \subseteq \mathscr{X}_{\mathrm{T}},\ \underline{\phi} \in \mathscr{F},\ \underline{\phi} : \mathbb{R}^n \to \mathbb{R}$
Finite (b)	∂_{fin}	$\dim(\mathscr{F}) < \infty$
Explicit Mercer (c)	∂_{eM}	$k(\underline{x}, \underline{z}) = \underline{\phi}(\underline{x})^{\mathrm{T}}\,\underline{\phi}(\underline{z})$ $= \sum_q \phi_q(x_q)\,\phi_q(z_q)$, ϕ_q known $x_i, z_i \in \mathscr{X}_0,\ \mathscr{X}_0 \subseteq \mathscr{X}_{\mathrm{T}},\ \phi_q \in \mathscr{F},\ \phi_q : \mathbb{R} \to \mathbb{R}$
Explicit multiplicative (e)	$\partial_{\times}$	$k(\underline{x}, \underline{z}) = \prod_q \phi_q(x_q)\,\phi_q(z_q)$, ϕ_q known $x_i, z_i \in \mathscr{X}_0,\ \mathscr{X}_0 \subseteq \mathscr{X}_{\mathrm{T}},\ \phi_q \in \mathscr{F},\ \phi_q : \mathbb{R} \to \mathbb{R}$
Uniform (f)	∂_{uni}	ϕ_q known and uniform e.g., (c) or (e) with $\phi_q = \phi\ \ \forall q$
Admissible (g)	∂_{adm}	k is positive definite (p.d.) [37] or k is conditionally p.d. (c.p.d.) [8]

5 Creating more measures specific to SVM

In this section we propose measures specific to SVM.

Support vectors: In SVM, a subset of the patients in the data set are key to defining the model. They are known as support vectors since they support the definition of the model's class boundary and decision surface. For example, the decision regarding whether a patient has a disease or not, is determined by a subset of patients, e.g., 5 out of 200 patients, the model learned/picked as positive and negative examples of disease.

The more support vectors there are, the more complex the model is, with all other things being equal: $H_{sv} = sv$. SVM models have at least three support vectors in general — at least two to define the line, curve, hyperplane or surface that is the class boundary, and at least one to define the margin, so $sv \geq 3,\ sv \in \mathbb{N}$.

To select a model for one data set, or to compare results between two data sets, we know the maximum number of patients N, so $sv \leq N$, and we apply (4).iii to obtain a relative measure, $U_{sv,r}$. Or

to obtain an absolute measure $U_{sv,a}$, to compare against any current or future data set, we assume $N = \infty$ and apply (4).ii.

Degrees of freedom: Akaike includes all method and kernel hyperparameters and weights as among the degrees of freedom[50]. We calculate the prior complexity measure $\check{H}_{dof}$ with three terms comprised of: the number of SVM hyperparameters, e.g., 1 for C, the number of kernel hyperparameters, e.g., 1 for the kernel width for a Gaussian RBF kernel, the number of independent inputs, e.g., 1 for a Gaussian RBF kernel or stationary kernel, 2 otherwise. We calculate the posterior complexity measure H_{dof} with an additional term for the support vectors and apply the general measure for model interpretability.

$$\check{H}_{dof} = \check{dof} = d_{\text{SVM\_hyp}} + d_{\text{kernel\_hyp}} + d_{\text{input}}$$
$$H_{dof} = dof = d_{\text{SVM\_hyp}} + d_{\text{kernel\_hyp}} + d_{\text{input}} + sv$$

Relevant dimensionality estimate: The relevant dimensionality estimate (rde) [9] provides a way to measure the complexity of the SVM feature space induced by a kernel. There are two complexity measures $H_{rde\text{T}}$ and $H_{rde\text{L}}$ corresponding to two rde methods: the two-component model and the leave-one-out method, respectively.

.

6 Validation of measures

We validate our proposed measures with sanity checks on formulas (not shown) and by agreement with propositions that describe our expectations and knowledge about model complexity and interpretability.

We create propositions based on expected relationships between measures, and check/test the propositions with a statement $\mathbf{P}$ and its inverse $\mathbf{P}^{-1}$ such as the following,

$$\mathbf{P}: \ \check{dof}_1 \leq \check{dof}_2 \ \overset{\text{usually}}{\rightarrow} U^*_{rde1} \geq U^*_{rde2} \tag{10}$$

$$\mathbf{P}^{-1}: \ \check{dof}_1 > \check{dof}_2 \ \overset{\text{usually}}{\rightarrow} U^*_{rde1} < U^*_{rde2} \tag{11}$$

where $\overset{\text{usually}}{\rightarrow}$ is a notation that means "implies the majority of the time". For brevity $\mathbf{P}^{-1}$ is implied but not shown in statements that follow. We measure how much our results agree with these propositions using either Kendall's W coefficient of rank correlation [26] or matched pair agreement [48], where the latter is applied to control for confounding factors.

If a proposition is robust, then the percentage of the concordance coefficient or matched pair agreement indicates how correct and useful the measure is, from that perspective. A measure has some utility, if it is correct the majority of the time, for different models/kernels and data sets, with a confidence interval that does not include 50%.

We validate our propositions using two types of experiments (#1 and #2 as below). We run each experiment five times on each of three data sets from the University of California at Irvine repository: the Statlog Heart, Hepatitis and Bupa Liver data sets. Missing data in the Hepatitis data set are imputed with Stata, taking one of three multiple imputations with Monte Carlo Markov Chains. Bupa Liver is used with the common target [36] rather than the clinically meaningful target.

– Experiment Type #1: For each of 90 points chosen randomly in the hyperparameter space, we choose a pair of models, matched pairs [48], that differ by one hyperparameter/*dof* that is fixed in one and free in the other, and check propositions as the percentage truth of the propositions. We use 3 pairs of kernels that differ by a single *dof*, e.g., a polynomial kernel of varying degree versus a linear kernel, a Gaussian RBF kernel with/without a fixed kernel width and a Mercer sigmoid kernel [11] with/without a fixed horizontal shift.

– Experiment Type #2: From the experiment type #1 we identify three points in the hyperparameter space which perform well for each kernel. For each of 3 fixed points, we choose 30 values of C equally spaced (as logarithms) throughout the range from 10^{-3} to 10^{6} and check propositions as the concordance of the left-hand side with the right-hand side in the propositions, using Kendall's W coefficient of concordance. If the right-hand side should have opposite rank to the left-hand side then we apply a negative to the measure on the right-hand side for concordance to measure agreement of rank. We use the following kernels: linear, polynomial, Gaussian RBF and Mercer sigmoid kernel [11].

6.1 Propositions

Proposition 1.

The majority of the time we expect that a model with less degrees of freedom $\check{dof}_1$, with all other things being equal when compared to another model with $\check{dof}_2$, will be simpler and have a relevant dimensionality estimate (rde) [9] that is less than or equal to the other model and therefore be more interpretable/understandable (U^*_{rde}):

$$\mathbf{1a}: \ \check{dof}_1 \leq \check{dof}_2 \ \overset{usually}{\rightarrow} \ rde_1 \leq rde_2 \tag{12}$$

$$\mathbf{1b}: \ \check{dof}_1 \leq \check{dof}_2 \ \overset{usually}{\rightarrow} \ U^*_{rde1} \geq U^*_{rde2} \tag{13}$$

This applies to rde with the two-component model (rdeT) and the leave-one-out method (rdeL).

Proposition 2.

In SVM, the hyperparameter C is called the box constraint or cost of error. Authors have remarked [49, remark 7.31] that C is not an intuitive parameter, although it has a lower bound for use $C \geq \frac{1}{N}$ and its behaviour suggests $C \doteq \frac{1}{\nu N}$, where ν is a proportion of support vectors. We therefore expect that a model with a higher value C_1 versus a second model with C_2 will have less support vectors (sv) and consequently be more interpretable/understandable (U_{Hs}):

$$\mathbf{2a}: \ C_1 \geq C_2 \ \overset{usually}{\rightarrow} \ sv_1 \leq sv_2 \tag{14}$$

$$\mathbf{2b}: \ sv_1 \leq sv_2 \ \overset{usually}{\rightarrow} \ U_{Hs1} \geq U_{Hs2} \tag{15}$$

$$\mathbf{2c}: \ C_1 > C_2 \ \overset{usually}{\rightarrow} \ U_{sv,a1} \geq U_{sv,a2} \tag{16}$$

$$\mathbf{2d}: \ C_1 > C_2 \ \overset{usually}{\rightarrow} \ U_{Hs1} \geq U_{Hs2} \tag{17}$$

This applies to simplicity of sensitivity U_{Hs} with any binning method.

Our experiment uses three binning methods: Scott $U_{H\text{sc}}$, Freedman-Diaconis $U_{H\text{fd}}$ and and Sturges $U_{H\text{st}}$.

Proposition 3.

The majority of the time we expect that, if a prior measure is useful, then it reflects the same rankings as the posterior measure,

$$\mathbf{3}: \; U^*_{Hs1} \leq U^*_{Hs2} \;\overset{usually}{\rightarrow}\; U_{Hs1} \leq U_{Hs2} \tag{18}$$

Proposition 4.

We expect that the linear kernel is the simplest of all kernels with greater transparency than other kernels such as the polynomial, Gaussian RBF kernel, sigmoid and Mercer sigmoid kernels, whereby,

$$\mathbf{4}: isLinear(k_1) > isLinear(k_2) \;\rightarrow\; \check{U}_{\partial 1} > \check{U}_{\partial 2} \tag{19}$$

7 Results

We summarize the results of our validation tests (Table 4 and 5) as follows: we recommend $\check{U}_{\partial}$ and U_{sv} as good measures. We find that U^*_{rdeT}, U^*_{rdeL} and $U_{H\text{st}}$ are measures which are of limited use, because they may be wrong one third of the time when providing guidance on decisions. $U_{H\text{sc}}$ and $U_{H\text{fd}}$ are not distinguished from chance by our propositions and are therefore not recommended. If $U_{H\text{st}}$ is validated to a greater degree in the future, then the initial measure $U^*_{H\text{st}}$ has been shown to be a good proxy for it, incurring some loss of information.

Table 4. The results from propositions using experiment type #2 validate the support vector measure U_{sv} and simplicity of sensitivity measure with Sturges binning $U_{H\text{st}}$.

Proposition	Measure & Result	Agreement %	Comment
2a	*sv*	82 ± 2.3	*C* validates *sv*, supports B3
2b	$U_{H\text{sc}}$	53 ± 3.3	$U_{H\text{sc}}$ not distinguished by *sv*
	$U_{H\text{fd}}$	48 ± 3.7	$U_{H\text{fd}}$ not distinguished by *sv*
	$U_{H\text{st}}$	62 ± 3.5	*sv* validates $U_{H\text{st}}$
2c	U_{sv}	81 ± 2.3	*C* validates U_{sv}
2d	$U_{H\text{sc}}$	54 ± 3.3	*C* validates $U_{H\text{sc}}$
	$U_{H\text{fd}}$	49 ± 3.7	$U_{H\text{fd}}$ not distinguished by *sv*
	$U_{H\text{st}}$	64 ± 3.2	*C* validates $U_{H\text{st}}$

Legend: Green = affirmative result. Yellow = inconclusive result. Red = contrary result.

Table 5. The results from propositions using experiment #1 validate the relevant dimensionality measures *rde*T and *rde*L, the initial model interpretability measures based on relevant dimensionality U^*_{rdeT} and U^*_{rdeL}, the use of prior measures of simplicity of sensitivity as proxies for posterior measures, and the measure of kernel transparency $\check{U}_\partial$.

Proposition	Measure & Result	Agreement %	Comment
1a	*rde*T	63 ± 5.0	$\check{dof}$ validates *rde*T, supports A2
	*rde*L	59 ± 5.2	$\check{dof}$ validates *rde*L, supports A2
1b	U^*_{rdeT}	62 ± 5.0	$\check{dof}$ validates U^*_{rdeT}
	U^*_{rdeL}	59 ± 5.2	$\check{dof}$ validates U^*_{rdeL}
3	U^*_{Hsc} as a proxy	72 ± 3.1	U_{Hsc} validates U^*_{Hsc} as a proxy
	U^*_{Hfd} as a proxy	76 ± 3.5	U_{Hfd} validates U^*_{Hfd} as a proxy
	U^*_{Hst} as a proxy	80 ± 3.2	U_{Hst} validates U^*_{Hst} as a proxy
4	$\check{U}_\partial$	100 ± 0	k_{Lin} vs. others, validates $\check{U}_\partial$

Legend: Green = affirmative result. Yellow = inconclusive result. Red = contrary result.

Table 6. Result for $\check{U}_\partial$ confirm that the linear kernel is more transparent than other kernels.

Dirac measure	Linear	Polynomial	Gaussian RBF	Sigmoid	Mercer Sigmoid
∂_{essep}	✓	×	✓[13]	×	✓
∂_{fin}	✓	✓	×	×	✓
∂_{eM}	✓	×	×	×	✓
$\partial_{\times}$	✓	×	×	×	×
∂_{uni}	✓	×	×	×	✓
∂_{adm}	✓	✓	✓	×	✓
$\check{U}_\partial$ (%)	100	33	33	0	83

Legend: Green = top result. Light green = second best result.

Our proposed measure of kernel transparency $\check{U}_\partial$, a prior measure, scored 100% agreement. This is a good measure that may be used a priori, but it is high-level and not specific to the match between a model and data. No surprises or complexities arose regarding the attributes of kernels.

The general measure based on the number of support vectors, U_{sv}, scored $81 \pm 2.3\%$ agreement—this is a good measure.

Our proposed simplicity of sensitivity measure with Sturges binning U_{Hst} scored $64 \pm 3.2\%$ and $62 \pm 3.5\%$, which is of limited use—we are interested in agreement that is sufficiently greater than chance (50%), enough to be reliable.

The same measure with Scott binning (U_{Hsc}), however, is barely distinguishable from chance in one test, and not distinguishable in another, and with Freedman-Diaconis binning (U_{Hfd}) it is not distinguishable from chance in both tests. We recommend further validation to examine the role of confounding factors such as kernel width/scale along with C per [16, 6].

If the simplicity of sensitivity measure $U_{H\mathrm{st}}$ can be validated to a greater degree in the future, then the initial measure $U^*_{H\mathrm{st}}$ which scores $80 \pm 3.2\%$ agreement with it, may be used in its place to avoid optimization, or to gain an initial estimate prior to optimization.

The general measure based on the relevant dimensionality of the feature space, U^*_{rdeT} and U^*_{rdeL} scored $62 \pm 5.0\%$ and $59 \pm 5.2\%$ agreement, respectively. These are of some use. We did not include Braun's noise estimate, which in hindsight should improve the measure.

8 Application

We apply model interpretability to results in a toy problem. When we select results for maximum accuracy with the Gaussian RBF kernel, we find that the top result in our sorted list of results achieves 100% accuracy (rounded to no decimal places) with 51 support vectors, while the second best result also achieves 100% accuracy with 40 support vectors and the fifth best result according to the list also achieves 100% accuracy with 25 support vectors.

Selecting results for maximum interpretability $U_{sv,r}$, we find the top result uses 9 support vectors for 99% accuracy and the fourth best result uses 10 support vectors for the same accuracy.

We plot the results (Figure 3) of accuracy versus interpretability $U_{sv,r}$ (above 80% in each) and find that there are many results which are highly accurate and highly interpretable, i.e., above 96% in both. These results indicate that there is not a trade-off between accuracy and model interpretability based on support vectors in this data set.

We also plot the results of accuracy versus interpretability $U_{sv,r}$ for other data sets (Figure 4 and Figure 5) and it is clear that there is no trend in all points showing a trade-off between accuracy and model interpretability, although this trend may be present at the pareto front. A trade-off trend would show as an inverse correlation, a trend line running from the top left to the bottom right—instead, high interpretability is consistently achievable with high accuracy, i.e., there are points toward the top right of a bounding box for all points.

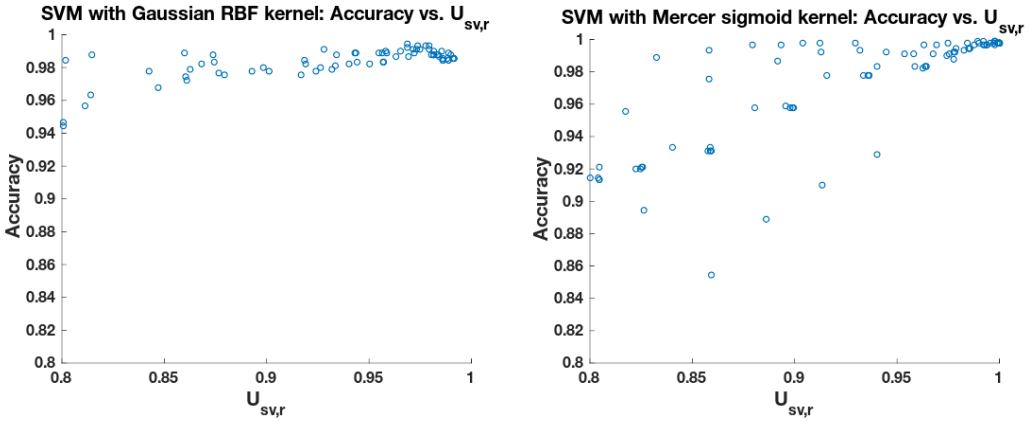

Fig. 3. In classification for the toy problem, there are many results with high accuracy and high model interpretability, with almost no sacrifice in the latter for maximum accuracy.

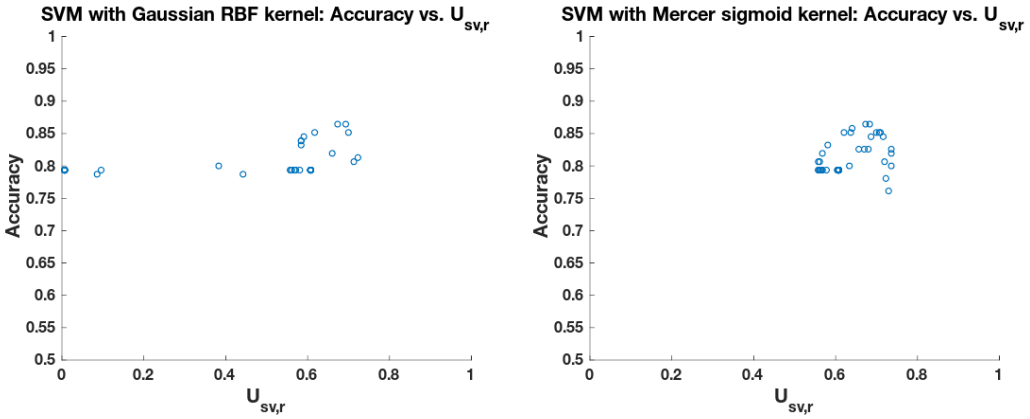

Fig. 4. In classification with the Hepatitis data set there is a less than 5% sacrifice in interpretability for the highest accuracy.

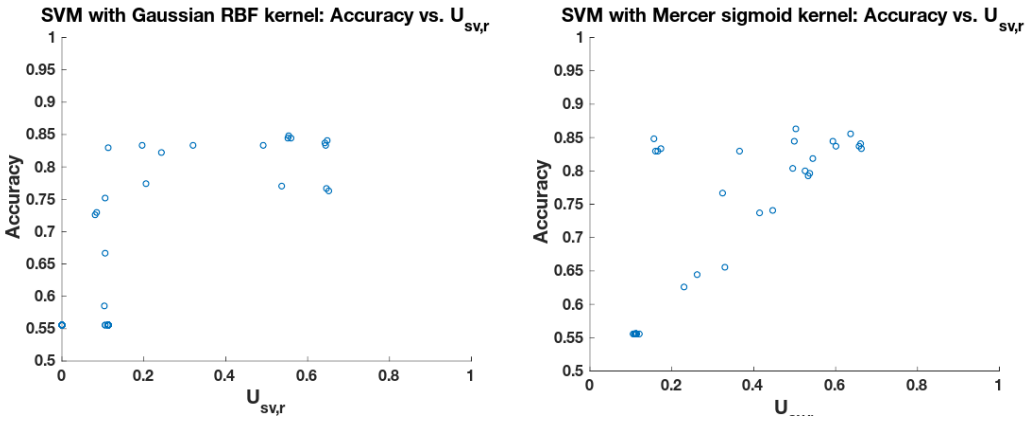

Fig. 5. In classification with Statlog Heart data there are points with high accuracy and interpretability, with minimal sacrifice, 1% and 2%, respectively.

9 Related work

Lipton [30] provides a good taxonomy for **model interpretability** with concepts falling into two broad categories: transparency (the opposite of a black box) and post-hoc interpretability.

Post-hoc interpretability involves an explanatory model separate from the predictive model, or visuals that transform data where the transformation is also a separate explanatory model. Liang [28] cautions against explaining a black box predictive model with another black box explanatory model.

Riberio et al [46] create an external **local linear model** to approximate the prediction model in a post-hoc approach called LIME. They jointly optimize accuracy and model complexity but they do not elucidate much about model complexity as in our work. LIME perturbs features in a separate binary representation of features, which sometimes map to **non-local** features in the original space of data. In their examples they use the binary model output, only referring in passing to the possibility of using a continuous output for classifiers, as we do.

Transparency, on the other hand, focuses on the predictive model itself, and has three aspects: decomposability, simulatability and algorithmic transparency [30].

Decomposability refers to being able to see and understand the parts of the model of the model, e.g., kernels and parameters and the parts of the data, i.e., features and instances—and how they contribute to a result from the predictive model. Some authors refer to the output from decomposition as an **interpretation**, e.g., initial understanding, separate from an **explanation** [24, 39] that may require **analysis**, **selection** or perhaps **synthesis**. Miller adds that explanations are **selected** and **social** [38].

Since the social and synthesis tasks are more suitable to a person than a computer—it is reasonable for our work to focus on inherent measures of interpretability, rather than explanations.

[34] express that some types of models are more **intelligible** (i.e., decomposable) than others. We include categories for generalized linear and generalized additive models in our measures as a result of their work.

Simulatability, as another aspect of transparency, refers to a model that a person can mentally simulate or manually compute in reasonable time [30] and is correlated, for example, with the number of features in a linear model, or the depth of the tree in a decision tree. **Model complexity** is implied Lipton's examples but the term is not invoked although other authors refer to it [35, 42, 10].

Ockham's razor, also called the principle of **parsimony** [50], is a well known principle related to model complexity. Regarding models, it says that among sufficient explanations (e.g., equally accurate[4] models), the simplest[5] should be preferred. A quick note on sufficiency: for multiple equally accurate models, none are necessary, because any one of them is sufficient. Model accuracy is sought first, then simplicity. Using our proposed measure one can search for the model with highest interpretability among equally accurate models.

Backhaus et al propose a quantitative measure of model interpretability [3]—but that is for a different meaning or definition—the ability for a model to interpret data, with relevance in relevance vector machines as the context.

Related to our work, **sensitivity analysis of model outputs** (SAMO) [2, 23] describe how sensitive a model output is to a change in feature values, one at a time—which is the approach of our proposed general measure.

In variance-based sensitivity analysis, Sobol [51] finds the variance in the output explained by an input feature. Liu et al [32] performs entropy-based sensitivity analysis, called global response probabilistic sensitivity analysis (GRPSA), to find the influence of input features—where entropy is used

[4] where accuracy cannot be distinguished with statistical significance

[5] [Sober] refers to [Akaike]'s definition of the simplest model as the model with the least degrees of freedom, i.e., least number of (independent) coefficients.

to compute the effect as information loss. Lemaire *et al.* [27] apply sensitivity analysis but their perturbations are non-local and could easily create points outside of any known clusters of instances and true states of nature. Poulin *et al.* [43] provides effective visualization and analysis tools but for SVM they only apply their method to linear SVM and its binary output.

Automatic model selection methods have been proposed for accuracy [1, 40]—these are based on rules computed from many data sets. The rule-based approach is brittle in comparison to our measures, since it only works with a fixed set of candidate kernels.

Conclusions

We developed and validated measures for inherent model interpretability to enable automatic model selection and ongoing research. Two measures are recommended: our proposed kernel transparency measure $\check{U}_\partial$ which is an inexpensive prior measure, and a posterior measure based on support vectors U_{sv}. Three other measures, U^*_{rdeT}, U^*_{rdeL} and $U_{H\text{st}}$ were found to be of limited use but may be further validated by future work.

We also contributed ideas as a foundation for these measures: the concept of inherent model interpretability, a general measure, a simplicity of sensitivity measure, and measurement of interpretability at different points in the learning process, i.e., via prior, initial and posterior models.

We applied our measure to model selection and demonstrated that choosing a model based on a sorted list of accuracy alone can result in models with substantively less inherent model interpretability despite the consistent availability of models with high accuracy and high interpretability in multiple data sets. The notion of a trade-off between accuracy and interpretability does not hold for these data sets.

Appendix A: Treating features of any atomic data type as continuous

Table 7. Atomic data types are based on Steven's scales of measurement

Atomic data type	Steven's scale	Summary of key attributes			
		Continuous	Discrete	Ordered	Fixed zero
Real	Ratio	✓		✓	✓
Integer	Ratio		✓	✓	✓
Datetime	Interval	✓		✓	
Date	Interval		✓	✓	
Ordinal	Ordinal		✓	✓	
Binary	Nominal		✓		
Nominal	Nominal		✓		

Assuming that we are not given a fixed pre-trained model, but can instead the machine learning method and model, we can select one that handles continuous values, and we can treat features of

any atomic data type (defined below) as continuous. This treatment requires three steps — and most of the content in these steps are standard practice, with a few exceptions denoted by an asterix*.

We define **atomic data types** (Table 7) as the following set of data types which are fundamental building blocks for all electronic data[6]: reals, integers, datetimes, dates, ordinals, binary and nominals. These atomic data types are based on Steven's scales of measurement [52], but are specified at a level that is more interpretable and useful.

Although binary values may also be considered nominals, we identify them separately because there are methods in the literature specific to binary data (e.g., for imputation and similarity measurement) and the data type is specifically defined in programming languages, machine learning platforms, database schema and data extraction tools.

1. Treat missing data. Assuming data are missing completely at random (MCAR) do the following, otherwise refer to [cite].
   (a) Impute missing data for reals, integers, datetimes, dates and ordinals, using whichever method meets requirements—e.g., multiple imputation with Monte Carlo Markov chain, expectation maximization, hot-deck imputation or mean imputation.
   (b) Impute missing data for nominals using the mode, i.e., the most frequent level.
   (c) Impute missing binary data with a method that will produce continuous values and which is appropriate for binary distributions—e.g., multiple imputation or expectation maximization. We refer to the output as continuously-imputed binary data.
2. Convert nominals to binary indicators, one for each level.
3. Center and normalize data
   (a) For continuously-imputed binary data, bottom-code and top-code the data to the limits, then min-max normalize the data to the range [-1,+1] for SVM or [0,1] for neural networks and logistic regression.
   (b) For binary data, min-max normalize the data to the set {-1,+1} for SVM or {0,1} for neural networks and logistic regression. This data will be treated as reals by the methods/models, but {-1,+1} makes more sensible use of the symmetric kernel geometry in SVM than {0,1}.
   (c) For all other data types, center and normalize each feature using z-score normalization (or scalar variations based on 2 or 3 sigma instead of 1 sigma).

Now all of the data are ready to be treated as reals by the methods/models.

[6] e.g., a combination of atomic data types can make up a complex data type—e.g., a combination of letters or symbols (nominals) make up a string as a complex data type.

Bibliography

[1] Shawkat Ali and Kate A Smith. On learning algorithm selection for classification. *Applied Soft Computing*, 6(2):119–138, 2006.

[2] Benjamin Auder and Bertrand Iooss. Global sensitivity analysis based on entropy. In *Safety, reliability and risk analysis-Proceedings of the ESREL 2008 Conference*, pages 2107–2115, 2008.

[3] Andreas Backhaus and Udo Seiffert. Quantitative measurements of model interpretability for the analysis of spectral data. In *Computational Intelligence and Data Mining (CIDM), 2013 IEEE Symposium on*, pages 18–25. IEEE, 2013.

[4] Remo Badii and Antonio Politi. *Complexity: Hierarchical structures and scaling in physics*, volume 6. Cambridge University Press, 1999.

[5] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert MÃžller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.

[6] Asa Ben-Hur and Jason Weston. A user's guide to support vector machines. In *Data mining techniques for the life sciences*, pages 223–239. Springer, 2010.

[7] Eta S Berner. *Clinical Decision Support Systems*. Springer Science+ Business Media, LLC, 2007.

[8] Sabri Boughorbel, J-P Tarel, and Nozha Boujemaa. Conditionally positive definite kernels for svm based image recognition. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 113–116. IEEE, 2005.

[9] Mikio L Braun, Joachim M Buhmann, and Klaus-Robert MÃžller. On relevant dimensions in kernel feature spaces. *Journal of Machine Learning Research*, 9(Aug):1875–1908, 2008.

[10] Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

[11] André M Carrington, Paul W Fieguth, and Helen H Chen. A new mercer sigmoid kernel for clinical data classification. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 6397–6401. IEEE, 2014.

[12] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

[13] Andrew Cotter, Joseph Keshet, and Nathan Srebro. Explicit approximations of the gaussian kernel. *arXiv preprint arXiv:1109.4603*, 2011.

[14] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[15] Joseph A Cruz and David S Wishart. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2, 2006.

[16] Olivier Devos, Cyril Ruckebusch, Alexandra Durand, Ludovic Duponchel, and Jean-Pierre Huvenne. Support vector machines (svm) in near infrared (nir) spectroscopy: Focus on parameters optimization and model interpretation. *Chemometrics and Intelligent Laboratory Systems*, 96 (1):27–33, 2009.

[17] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. 2017.

[18] David Freedman and Persi Diaconis. On the histogram as a density estimator: L 2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(4):453–476, 1981.

[19] Murray Gell-Mann and Seth Lloyd. Information measures, effective complexity, and total information. *Complexity*, 2(1):44–52, 1996.

[20] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". In *1st Workshop on Human Interpretability in Machine Learning, International Conference of Machine Learning*, 2016.

[21] David L Goodstein and Judith R Goodstein. *Feynman's lost lecture: the motion of planets around the sun*, volume 1. WW Norton & Company, 1996.

[22] Robert A Greenes. *Clinical decision support: the road ahead*. Academic Press, 2011.

[23] Kenneth M Hanson and François M Hemez. *Sensitivity Analysis of Model Output: Proceedings of the 4th International Conference on Sensitivity Analysis of Model Output (SAMO 2004): Santa Fe, New Mexico, March 8-11 2004*. Los Alamos National Laboratory, 2005.

[24] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[25] Marvin Ed Jernigan and Paul Fieguth. *Introduction to Pattern Recognition*. University of Waterloo, 2004.

[26] Maurice G Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.

[27] Vincent Lemaire, Raphael Féraud, and Nicolas Voisine. Contact personalization using a score understanding method. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 649–654. IEEE, 2008.

[28] Percy Liang. Provenance and contracts in machine learning. In *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, 2016.

[29] Dekang Lin. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304, 1998.

[30] Zachary C Lipton, David C Kale, Charles Elkan, Randall Wetzell, Sharad Vikram, Julian McAuley, Randall C Wetzell, Zhanglong Ji, Balakrishnan Narayaswamy, Cheng-I Wang, et al. The mythos of model interpretability. *IEEE Spectrum*, 2016.

[31] Paulo JG Lisboa. Interpretability in machine learning–principles and practice. In *International Workshop on Fuzzy Logic and Applications*, pages 15–21. Springer, 2013.

[32] Huibin Liu, Wei Chen, and Agus Sudjianto. Relative entropy based method for probabilistic sensitivity analysis in engineering design. *Journal of Mechanical Design*, 128(2):326–336, 2006.

[33] Seth Lloyd. Measures of complexity: a nonexhaustive list. *IEEE Control Systems Magazine*, 21(4):7–8, 2001.

[34] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM, 2012.

[35] David Martens and Bart Baesens. Building acceptable classification models. In *Data Mining*, pages 53–74. Springer, 2010.

[36] James McDermott and Richard S Forsyth. Diagnosing a disorder in a classification benchmark. *Pattern Recognition Letters*, 73:41–43, 2016.

[37] James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, pages 415–446, 1909.

[38] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum. In *IJCAI-17 Workshop on Explainable AI (XAI)*, page 36, 2017.

[39] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[40] Jesmin Nahar, Shawkat Ali, and Yi-Ping Phoebe Chen. Microarray data classification using automatic svm kernel selection. *DNA and cell biology*, 26(10):707–712, 2007.

[41] Randal S Olson, William La Cava, Patryk Orzechowski, Ryan J Urbanowicz, and Jason H Moore. Pmlb: a large benchmark suite for machine learning evaluation and comparison. *BioData mining*, 10(1):36, 2017.

[42] Pedro Santoro Perez, Sérgio Ricardo Nozawa, Alessandra Alaniz Macedo, and José Augusto Baranauskas. Windowing improvements towards more comprehensible models. *Knowledge-Based Systems*, 92:9–22, 2016.

[43] Brett Poulin, Roman Eisner, Duane Szafron, Paul Lu, Russell Greiner, David S Wishart, Alona Fyshe, Brandon Pearcy, Cam MacDonell, and John Anvik. Visual explanation of evidence with additive classifiers. In *Proceedings Of The National Conference On Artificial Intelligence*, volume 21, page 1822. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[44] Martin V Pusic, Kathy Boutis, Rose Hatala, and David A Cook. Learning curves in health professions education. *Academic Medicine*, 90(8):1034–1042, 2015.

[45] Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[46] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[47] David W Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

[48] Steve Selvin. *Statistical analysis of epidemiologic data*. Oxford University Press, 2004.

[49] John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[50] Elliott Sober. Parsimony and predictive equivalence. *Erkenntnis*, 44(2):167–197, 1996.

[51] Ilya M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1):271–280, 2001.

[52] Stanley Smith Stevens. On the theory of scales of measurement, 1946.

[53] Herbert A Sturges. The choice of a class interval. *Journal of the american statistical association*, 21(153):65–66, 1926.

[54] Zoltán Szabó, Barnabás Póczos, and András Lőrincz. Undercomplete blind subspace deconvolution. *Journal of Machine Learning Research*, 8:1063–1095, 2007.

[55] Zoltán Szabó, Barnabás Póczos, and András Lőrincz. Separation theorem for independent subspace analysis and its consequences. *Pattern Recognition*, 45:1782–1791, 2012.

[56] Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1):479–487, 1988.

[57] Alan Tussy and R Gustafson. *Elementary Algebra*. Nelson Education, 2012.