Hindawi Publishing Corporation
Advances in Artificial Intelligence
Volume 2016, Article ID 8173625, 17 pages
http://dx.doi.org/10.1155/2016/8173625

Research Article

Efficacious Discriminant Analysis (Classifier) Measures for End Users

E. Earl Eiland and Lorie M. Liebrock

Computer Science and Engineering Department, New Mexico Institute of Mining and Technology, 801 Leroy Place, Socorro, NM 87801, USA

Correspondence should be addressed to E. Earl Eiland; eee@nmt.edu

Received 28 March 2015; Revised 3 November 2015; Accepted 8 November 2015

Academic Editor: Elpida Keravnou

Many problem domains utilize discriminant analysis, for example, classification, prediction, and diagnoses, by applying artificial intelligence and machine learning. However, the results are rarely perfect and errors can cause significant losses. Hence, end users are best served when they have performance information relevant to their need. Starting with the most basic questions, this study considers eight summary statistics often seen in the literature and evaluates their end user efficacy. Results lead to proposed criteria necessary for end user efficacious summary statistics. Testing the same eight summary statistics shows that none satisfy all of the criteria. Hence, two criteria-compliant summary statistics are introduced. To show how end users can benefit, measure utility is demonstrated on two problems. A key finding of this study is that researchers can make their test outcomes more relevant to end users with minor changes in their analyses and presentation.

1. Introduction

Artificial intelligence and machine learning are effective discriminant analysis (DA), for example, classification, prediction, and diagnoses, tools [1–3]. These tools are becoming increasingly popular for speeding up applied research and product development. A number of DA tool evaluation measures are seen in the literature, with acceptance varying by research domain. For example, the receiver operating characteristic is common in intrusion detection studies and the F_β-score is common in information retrieval. Researchers use measures because they provide actionable information. The question of whether these measures also provide actionable information to end users is explicitly addressed.

As an example of the problem end users face, two commonly seen summary statistics [4, 5], *total accuracy rate* (TAR) [5] and F_β-score [5], have opposite responses to relative class size (ratio_+) (also known as class imbalance and prevalence) [6]. (In the current context, summary statistics are formulae that take multiple joint probability table (JPT) based values as input and output a single value that represents the target DA tool's composite utility. A key characteristic of summary statistics is that they are not monotonic. When they are plotted against the boundary, they have optima. A common summary statistic, the receiver operating characteristic area under the curve, is boundary invariant. Since it neither increases nor decreases, it is monotonic. Since there is only a single value, it is also an optimum.) The F_β-score monotonically decreases as ratio_+ increases, while TAR monotonically increases.

End users want to know how a specific DA tool will impact their problem. An informative measure for any stakeholder must be sensitive to relevant problem domain characteristics and insensitive to irrelevant (confounding) characteristics. DA tool stakeholders can be partitioned into three groups.

(i) *Basic researchers* focus on developing new DA algorithms. This group expects that an effective new DA algorithm will be useful in many problem domains, so their evaluations need to be application agnostic; specific problem domain characteristics are, in fact, confounding. Basic researchers introduced DA techniques such as k nearest neighbor [7], neural net [8], and support vector machine [9].

(ii) *Applied researchers* use DA algorithms on specific problem domains to create tools and code libraries useful for that domain; specific problem domain characteristics are important. Examples of tools incorporating DA algorithms are anomaly based intrusion detectors for cyber security [10, 11] and document classifiers for enterprise information retrieval systems [12]. In this context, the focus is on the DA tool; data sets are used to develop the tool/library.

(iii) *End users* use DA tools to solve problems in their domain. Domain-specific characteristics are important, as are operational aspects like impact sensitivity to class boundary settings (B, the setting that determines to which class each observation is allocated). Fields of study include medicine [13], molecular biology [14, 15], finance [16], and so forth. The end user's context is the opposite of the other two groups; the DA tool classifies the data, rather than known data being used to evaluate the tool.

The researcher definitions specialize those published by the National Science Board [17].

Jamain and Hand comment on user need in their DA tool meta-analysis:

> *The real question a user generally wants to answer is "which classification methods [are] best for me to use on my problem with my data…" [18].*

In artificial intelligence, Russell and Norvig express a similar sentiment:

> *As a general rule, it is better to design performance measures according to what one actually wants in the environment, rather than according to how one thinks the agents should behave [19].*

In published studies read, proposed measures may be mapped to specific problem domains, but identifying a general means by which end users can quantify DA tool effectiveness in their setting has not been addressed. Indeed, Jamain and Hand generalize Duin's sentiment regarding comparing automated, heavily parametrized DA tools (also known as classifiers):

> *"It is difficult to compare these types of classifiers in a fair and objective way [20]."*

Seemingly, end user needs are viewed as too complex and diverse to address. End user issues, when discussed, have been constrained to specific problem domains. Quantifying end user impact was not found in a literature search.

End users face a daunting challenge, when selecting a DA tool summary statistic. Sokolova et al. comment "…the measures in use now do not fully meet the needs of learning problems in which the classes are *equally important* and where *several algorithms are compared*" [21]. These sentiments are echoed in the text by Japkowicz and Shah:

> *Although considerable effort has been made by researchers in both developing novel learning methods and improving existing models and approaches, these same researchers have not been completely successful in alleviating the users' skepticism with regard to the worth of these developments. This is due, in part, to the lack of both depth and focus in what has become a ritualized evaluation method used to compare different approaches [5].*

These authors also observe that problem domains have *de facto* standard summary statistics. By starting with the most basic questions (the approach used by Artzner et al., addressing a similar situation in the financial risk domain [22]), specifically addressing user's issues, this goes directly to the heart of the matter as perceived by Japkowicz and Shah.

The balance of this paper is organized as follows. Section 2 presents the lexicon. Section 3 discusses relevant work. Section 4 identifies some questions end users have regarding DA tool efficacy. Section 5 explains the research protocol used. Section 6 reports this study's results and summary statistic recommendations. Section 7 assesses the recommended changes. Section 8 summarizes findings and suggests future work.

2. Lexicon

Although this paper applies well-established stochastic concepts, not all discussions use the same terminology [1, 2, 4, 5, 23, 24]. To avoid confusion, a lexicon and alternate terms seen in the literature are proved:

A: it is the population which, when observed, satisfy the characteristics defining A.

$\overline{A}$: it is the population which, when observed, do not satisfy the characteristics defining A.

B: it is boundary, the DA tool vector that defines the partition between Z and $\overline{Z}$.

Class imbalance: see relative class size.

Confidence interval (CI): it is a range within which x% of the observations lie. A 90% CI would exclude the 5% on each pdf tail.

Confusion matrix: see joint probability table.

Contingency table: see joint probability table.

Discriminant analysis (DA): it is a process whereby observations are tagged as being members of a class.

Diagnostic Odds Ratio (DOR): it is a DA tool evaluation measure [25].

Diagnostic Power (DP): it is a DA tool evaluation measure [26].

Error matrix: see joint probability table.

F_β-score: it is a DA tool evaluation measure [27].

Frequency table: see joint probability table.

F_+: it is false positive, class $\overline{A}$ events incorrectly flagged as class A ("Type I error").

F_-: it is false negative, class A events incorrectly flagged as class $\overline{A}$ ("Type II error").

Table 1: Category cardinalities shown in a joint probability table.

		Actual classification (ground truth)		
		Y	$\overline{Y}$	Totals ↓
Test	+: $s_i \in \{Z\}$	t_+	f_+	$\lvert Z\rvert = t_+ + f_+$
Result	−: $s_i \notin \{Z\}$	f_-	t_-	$\lvert\overline{Z}\rvert = f_- + t_-$
Totals		$\lvert Y\rvert = t_+ + f_-$	$\lvert\overline{Y}\rvert = f_+ + t_-$	$\lvert S\rvert = \lvert Y\rvert + \lvert\overline{Y}\rvert = \lvert Z\rvert + \lvert\overline{Z}\rvert$

Ground Truth: it is the actual class to which an observation belongs.

ι: it is the gain or loss (positive or negative) associated with each element output. Each and every element output will affect the end user by the element of I applicable to the category to which the element is binned. (Typically, gains are viewed as positive values and losses are negative values, although there are exceptions.)

ι_{s_n}, ι_{y_n}, $\iota_{\overline{y}_n}$, ι_{z_n}, or $\iota_{\overline{z}_n}$ (n denotes the specific element of the source set, S, Y, $\overline{Y}$, Z, and $\overline{Z}$): they are individual element impacts in the raw data.

$I = (\iota_{T_+}, \iota_{F_+}, \iota_{F_-}, \iota_{T_-})$: it is the vector of JPT category impacts, expressed as statistical expectations (*expected individual element impact*) by category or class.

Joint probability table (JPT): it is an $M \times M$ table in which one axis represents ground truth (Y and $\overline{Y}$) and the other axis represents DA output (Z and $\overline{Z}$). The cells contain the counts for observation's based on their ground truth and DA output. The DA tool under test, configured with boundary vector (B, a "surface" that partitions the problem space), bins S into Z and $\overline{Z}$. The JPT(B) bin counts are snapshots of DA tool labeling versus ground truth at B. Frequently, these counts are presented as proportions of $|S|$. In a JPT of proportions, each cell in Table 1 is divided by $|S|$. The differences between the two table types is that the cell entries in Table 1 are integers, with the total of all four categories equaling $|S|$, whereas the cell entries in a proportional JPT are rational numbers that sum up to one. Additionally, the proportional values represent the probability that, for a given relative class size ($\text{ratio}_+ = Y/\overline{Y}$), any randomly selected DA tool output will be a member of that particular JPT category.

Mathews Correlation Coefficient (MCC): it is a DA tool evaluation measure [28].

Mutual Information Coefficient (IC): it is a DA tool evaluation measure [29].

Probability density function (pdf): it is the probability of a particular value, within the range of all possible values, being observed.

Proportion: see relative class size.

Receiver Operating Characteristic Area Under the Curve (ROC-AUC): it is a DA tool evaluation measure [30–32].

Relative class size: $\text{ratio}_+ = |Y|/|\overline{Y}|$ [5].

S: it is the uncategorized data set.

T_+: it is true positive, correctly identified events in class A.

T_-: it is true negative, correctly identified events of class $\overline{A}$, the other class.

Total accuracy rate (TAR): it is a DA tool evaluation measure. (TAR has been in use so long; its source is not found cited.)

Y: it is actual class A events in the data set.

$\overline{Y}$: it is actual class $\overline{A}$ events in the data set.

Youden index (J): it is a DA tool evaluation measure [33].

Z: it is events flagged by the DA tool as class A.

$\overline{Z}$: it is events flagged by the DA tool as class $\overline{A}$.

Y and $\overline{Y}$ constitute S as partitioned according to ground truth.

Z and $\overline{Z}$ constitute S as partitioned by the DA tool output.

T_+, F_+, F_-, and T_- cardinalities can be presented in a joint probability table (JPT). These are displayed in lower case: $t_+ = |T_+|$, $t_- = |T_-|$, $f_+ = |F_+|$, and $f_- = |F_-|$.

3. Related Work

Summary statistics are a subset of the measures used for DA evaluation. Measures such as sensitivity, specificity, and true positive predictive rate are often seen in studies using DA. These measures are monotonic (do not have optima by which we identify optimum DA boundary settings); hence, they are not summary statistics. Their value is providing greater detail by quantifying particular aspects of DA tool performance. It should come as no surprise that some summary statistics are functions of monotonic JPT-based measures such as sensitivity, specificity, and true positive predictive rate.

Sokolova and Lapalme tested both monotonic measures and summary statistics used for DA tool evaluation for invariance to various JPT perturbations [34]. Their work's value is based on assuming that measure selection should consider their invariance relative to the problem's need for invariance. This work tightens their constraints by requiring

that measures provide actionable information to the stakeholder, in this case, end users.

It might be fair to state that each end user's need is, in some way, unique. However, uniqueness does not mean that useful common problem characteristics do not exist. Starting with very basic questions, this paper identifies a general DA tool problem framework. Using that framework, this work identifies criteria which end user efficacious measures for DA tool evaluation must satisfy and proposes two summary statistics that meet those criteria. End user efficacious summary statistic optima indicate the best overall DA tool utility across the observed boundary range. Ideally, these summary statistics also quantify some efficacious aspect of DA tool output for end users. Efficacious measure values enable end users to directly estimate how the DA tool will affect their situation.

The state-of-the-art DA evaluation has inspired many classifier evaluation reviews. With the exception of Baldi et. al's narrowly scoped study [35], none address end user interests. In fact, generally, stakeholders were not identified. In some cases, measure characterizations were presented with the expectation that knowledgeable stakeholders would be able to apply the information. Relevant aspects of a few of the many characterization studies in the literature are summarized:

(i) Parker analyzed five measures [36]. His analysis does not identify the stakeholder. However, his recommendation that an integrated measure should be used when possible suggests that his evaluation addresses researchers.

(ii) Japkowicz reviews machine learning evaluation methods and suggests the need for "a framework that would link our various evaluation tools to the different types of problems they address and those that they fail to address" [37]. Japkowicz work focuses on issues within the machine learning domain but does not differentiate the differences between research and user interests.

(iii) Recognizing the large number of comparative studies, Jamain and Hand executed a meta-analysis [18]. By mining relevant studies, they hoped to integrate their findings and gain insight into the problem. Their analysis did provide some insights. However, in closing, they note that the investigation did not shed light on end user issues. Their paper also indicated that they felt the problem was intractable.

(iv) Caruana and Niculescu-Mizil investigate nine measures, partitioned into three different types [38]. Their results lead them to propose a measure suite which outperforms the individual measures. Their study, however, does not specify the stakeholders, nor does it address the end user's potential need to address unequal event impacts.

(v) In the same vein as Caruana and Niculescu-Mizil, Seliya et al. empirically compared twenty-two measures [39]. Their work identified strongly correlated measures with the intent of letting investigators select measure suites with minimal potential redundancies. Their work did not search for root causes or identify stakeholders.

(vi) Baldi et al. assess measures, restricting their scope to two bioinformatics problems [35]. They review eleven measures and conclude that MCC may be the best overall for their domain. Baldi et al. also observed a slight difference in B^* between MCC and IC. For their intended users, impacts are apparently not an issue, as they were not mentioned.

(vii) Sokolova has been actively addressing the classifier performance measure problem [21, 34, 40]. In [34], Sokolova and Lapalme compare eight measures. They conclude that there is a subset of measures that are suitable for problems with restricted access to data, the need to compare several classifiers and equally weighted classes. In [21, 40], she reviews measures for their invariant properties. Thus, potential users can select measures that are suitable for their particular need.

The literature also includes two recent, relevant books. Witten et al. discuss cost-based analysis and present a reasonable case for its use and a means of classifier selection based on ratio_+ [41]. However, they do not identify the root problem and so stop short of identifying and addressing some of the end user impact factors addressed here. Japkowicz and Shah may have the most comprehensive discussion of DA evaluation [5]. Their invaluable book covers many challenges faced by evaluators. Like Witten et al., they "treat the symptoms" but do not identify the root problem.

In a literature review, no common understanding of what constitutes a good DA tool evaluation summary statistic was found; this seemed glaring in its absence. No "good summary statistic" criteria for DA tool evaluation were found. A key contribution of this paper is establishing four criteria for end user efficacious DA tool summary statistics.

4. End User Measure Efficacy Considerations

The focus here is end user interests. For DA tool selection and deployment, texts such as Clemen [42] lay the foundation for three questions informed end users must answer:

(1) *What is the DA tool's impact on my problem?* To answer this question, consider what the summary statistic quantifies and how that relates to end users.

(2) *What is the boundary that provides the optimum impact?* Only a boundary (B) sensitive summary statistic can provide this information.

(3) *How sensitive is the impact to boundary selection?* Only a boundary sensitive summary statistic can provide this information.

Measurement theory provides additional insight on summary statistic efficacy: numbers are used in different ways. These uses constrain their information content and hence their utility. This work uses the scale-type definitions proposed by Stevens [43]. Stevens defined four scale types,

nominal, ordinal, interval, and ratio. Ratio scales have the least functional constraints, so summary statistics using ratio scales are the most information rich ones. Ratio scales have two unique and readily identifiable characteristics.

They Have Meaningful Zeros. A meaningful zero for end users indicates that the statistical expectation of the DA tool's output has no effect on their problem.

They Have a "Standard Unit." This means that $m_x + 1 = m_{x+1}$. One implication of having a standard unit is that there is no upper bound and the lower bound can be either zero or negative infinity. A DA tool's output could negatively impact an end user, so the most generally useful measure's scale range must be $(-\infty, \infty)$.

Reflecting on the end user interests, only a ratio scale summary statistic will satisfy point (1) above. A recurring topic in studies using DA tools is class imbalance or relative class size (a similar concept, prevalence, is seen in medical research [1, 2]). Japkowicz and Shah quantify class imbalance as $\text{ratio}_+ = |Y|/|\overline{Y}|$ [5]. ratio_+ is used to evaluate summary statistics regarding (i) the question answered, (ii) scale type, and (iii) sensitivity to environmental factors, ratio_+, and pdf.

5. Research Protocol

This study develops and tests a mitigation for the end user DA tool assessment gap introduced in Section 1.

5.1. Problem Mitigation Plan. In a literature search, no foundation upon which users can base a measure selection to evaluate DA tools was found. However, the financial domain had a similar situation in assessing market risk. Artzner et al. tackled the problem by first establishing the need, then defining performance criteria, and ending with measurement recommendations [22]. This work applies Artzner et al.'s protocol by the following:

(1) Defining Explicitly the End User's DA Tool Evaluation Problem. This is addressed in Section 4, presented as three questions.

(2) Identifying an Efficacious Measure's Necessary Properties. Measure values that provide end users with actionable information (measure values with which end users can answer the three questions posed) must have certain properties. Section 6.2 proposes criteria necessary for efficacious measures.

(3) Testing Measures for the Existence of These Properties. Section 6.1 shares insights gained evaluating the eight measures relative to Section 4 questions.

(4) Recommending Adjustments to Have Conformant Measures. From Sections 6.4 to 7, this paper applies the insights gained, making specific recommendations regarding end user efficacious measures.

5.2. Mitigation Plan Analysis. DA tool evaluation studies can be partitioned into two groups: those that use "real-world" data and those that use simulated data; both are used here.

Characterizing DA tool evaluation measures requires observing how the summary statistics respond as DA tool output varies. Observing the effect of incremental changes on real-world data is difficult. For this purpose, simulated DA tool output was used.

To preserve generality, distribution insensitive analytic procedures are used here. The analysis is nonparametric; medians are used instead of means and quantiles are used instead of standard deviations. Monte Carlo method-based tests use well-defined pdfs.

DA tool quality is based on supervised tests. This test framework is well established in artificial intelligence circles [4, 5, 19]. Supervised tests provide the ability to compare ground truth (the actual class membership of objects in the test set) to the DA tool's class membership determination of the test set objects, that is, the DA tool's output [44, 45].

One risk with using simulated data is that the situations may be unrealistic. Consequently, studies using solely simulated data suffer from the perception of being unproven. Often DA tests use real-world data from repositories, for example, the University of California, Irvine Machine Learning Repository (UCI). Alas, UCI does not have JPT category impact, which seriously diminishes that analytical approach's sense of realism. Therefore, two published studies are reanalyzed to compare classifiers in two separate domains.

5.3. Study Scope. There is a wide variety of classification problem types. By limiting this study, the risk of obscuring results is removed. This work is limited to problems where

(i) the events mapped to each ground truth class (Y and $\overline{Y}$) are independent;

(ii) each event's impact (ι) is independent;

(iii) the problem is restricted to a 2×2 matrix;

(iv) the end user's problem either treats each input set element individually (as in the case of a medical diagnosis or intrusion detection) or the problem is based on the cumulative effect of the elements in the input stream (as in the case of bank loan application decisions or information retrieval).

6. Results

This section summarizes how well the eight selected measures address the questions posed. Drawing on these findings, four criteria are presented to address end user's needs. This section closes by using the criteria to make recommendations for end user efficacious DA evaluation measures.

6.1. Evaluating the End User of Existing Measures. While investigating the current state of the art, eight commonly seen DA tool evaluation measures appear to be very common: *total accuracy rate*, the *Receiver Operating Characteristic Area Under the Curve*, F_β*-score*, *Youden index*, two related measures, *Diagnostic Odds Ratio* and *Diagnostic Power*, *Mathews*

Correlation Coefficient, and *Mutual Information Coefficient*. Each measures' ability to answer the three questions posed in Section 4 was evaluated. Due to space constraints, findings are summarized here. The full analysis is on-line [46]. Key findings were as follows:

(i) Most are ordinal scale measures and none were ratio scale measures. Answering the questions requires ratio scale values so none could answer the questions.

(ii) The summary statistics measured characteristics that were at best niche problems.

(iii) None included end user impact; however, F_β-score did address category importance.

(iv) Excepting ROC-AUC and TAR, all of the summary statistics reviewed define classifier quality equal to a fair coin as "zero." This may be reasonable in research. However, a fair coin may not have a zero impact on an end user.

Additional measure characterizations are also available in Eiland and Liebrock [6].

6.2. Efficacious End User Summary Statistics. Efficacious end user measures of DA tools must reflect the DA tool's performance in the end users environment and their problem's context. There are many types of performance, but this paper focuses on DA tool output utility, the ability to approach the optimal response for an end user. This section frames insights gained as DA-specific criteria and measurement theory principles. After presenting each criterion, eight commonly seen summary statistics are evaluated for compliance.

6.2.1. Criterion 1, Category Impact. Consider a physician making two treatment decisions, one treating a potential cold and the other treating potential rheumatoid arthritis (RA). Someone with a cold, who is treated for it (T_+), will experience a minimal negative impact on their quality of life (low QoL impact). However, someone with an untreated cold (F_-) can be quite miserable (high negative QoL impact). Someone without a cold and untreated (T_-) will experience no effect (no QoL impact); someone without a cold but treated for it (F_+) will experience a small impact from medication side effects and cost (low QoL impact). In this situation, the best strategy may be to minimize F_-s and accept high F_+, by treating anyone with the slightest indication for a cold.

The situation is different with RA. Someone with RA and treated (T_+ will experience a minimal negative impact on their quality of life (low QoL impact). However, someone with untreated RA (F_-) will be significantly debilitated (high negative QoL impact). Someone without RA and untreated (T_-), will be unaffected (no QoL impact). Someone without RA but treated (F_+) will experience significant disability (high QoL impact). Faced with this decision, the cold treatment strategy is inappropriate: both F_+s and F_-s must be minimized. A conscientious physician must consider category impacts.

An efficacious end user summary statistic must be sensitive to the same factors, to the same degree, as end users are to their respective problems. With regard to utility, the end user's context is defined by the importance, or impact, of elements from each JPT category on the end user. The criterion follows directly from the previous discussion. A small change to any element of I will generate corresponding changes in a compliant measure. For example, TAR and the F_β-score can be interpreted as being the same base ratio, differing only by I (recall that $I = (\iota_{T_+}, \iota_{F_+}, \iota_{F_-}, \iota_{T_-})$):

$$\text{base ratio} = \frac{\iota_{T_+} t_+ + \iota_{T_-} t_-}{\iota_{T_+} t_+ + \iota_{F_+} f_+ + \iota_{F_-} f_- + \iota_{T_-} t_-}. \tag{1}$$

When $I = (1, 1, 1, 1)$, then ratio 1 is TAR. Likewise, when $I = (1 + \beta^2, 1, \beta^2, 0)$, then ratio 1 is F_β-score. The response of these two measures to the same input JPTs is distinctly different. The commonality between TAR and F_β-score leads to the first criterion.

Criterion 1 (category importance). An end user efficacious summary statistic must be a function of problem specific impact set $I = (\iota_{T_+}, \iota_{F_+}, \iota_{F_-}, \iota_{T_-})$, where each element of $I \in \mathbb{Q}$.

A summary statistic that complies with Criterion 1 is sensitive to I. Thus, end users can tune the measure's output to suit their problem. Criterion 1 provides a direct answer to the end user's question "What is the DA tool's impact on my problem?". Since the end user's other two questions also address impact, Criterion 1 also addresses them.

None of the summary statistics reviewed satisfy Criterion 1. The F_β-score, conditioned by β, provides some ability to incorporate impact. However, since $\iota_{T_-} = 0$ regardless of β, it fails. The other summary statistics considered (TAR, Youden index, ROC-AUC, DOR/DP, MCC, and IC) do not have a provision for setting impacts; all have fixed impacts: implicitly, $I = (1, 1, 1, 1)$.

6.2.2. Criterion 2, pdf Sensitivity. When making loan decisions, banks rely heavily on credit scores. In a stable economy, the credit score distribution (pdf) may also remain stable. However, disruptions, such as a reduction in force by a major employer, can cause the pdf to shift. Then, the number of loan defaults by applicants with low but acceptable credit scores may become unacceptable, requiring increasing the acceptable credit score. Finding the new optimum threshold will depend upon the new pdf's shape.

DA problems may exit where the optimum boundary is not sensitive to input class pdfs. However, end users need to identify an appropriate DA tool boundary for their problem. Consider an end user environment with two classes, defined by $f(\mu, \sigma)$ and $g(\mu', \sigma')$, where μ and μ' are the distribution means and σ and σ' are the distribution standard deviations. Let the end user have optimum boundary B^*. Changing both distributions by adding $\Delta\mu$ to μ and μ', then the optimum boundary becomes $B^* + \Delta\mu$ and ι remains constant. There are myriad permutations that can be made to this simple end user environment. Most will affect B^* and/or ι; some will not. The point is that class distribution invariance is not an end-user-efficacious summary statistic characteristic;

end users are better served by summary statistics that are class distribution sensitive.

Criterion 2 (pdf sensitivity). With a change in pdf(Y), (e.g., pdf(Y') = Δ + pdf(Y), pdf(S) = pdf(Y) + pdf($\overline{Y}$), and pdf(S') = pdf(S)+Δ), where Δ describes a perturbation in Y's and S's source population, for all boundaries within Δ, there exists $E(SS(B) \mid S') - E(SS(B) \mid S) \neq 0$. The same is true for a change in $\overline{Y}$ and for any ratio$_+$.

For any boundary within the interval affected by a probability distribution change, an effective summary statistic's expected output will reflect that change. A summary statistic compliant with Criterion 2 will reflect how the target DA tool impacts the end user when provided with different inputs. Criterion 2 addresses the second and third end user questions posed, "what is the boundary that provides the optimum impact?" and "how sensitive is the impact to boundary selection?" which are both related to classifier output pdfs. One end user's optimum boundary may not be optimum for another.

Of the summary statistics considered, only TAR, MCC, and IC comply fully. AUC fails Criterion 2: it is boundary invariant. The Youden index and DOR/DP, being ratio$_+$ invariant, fail for ratio$_+$ sensitive problems. The F_β-score fails because it is T_- invariant.

6.2.3. Criterion 3: DA Tool Output Basis. In the loan decision problem mentioned in Section 6.2.2, after the decision is made, the lender has knowledge of the loans made (T_+, F_+, and Z) but knows only $\overline{Z}$ for the loan applications rejected. Given this information, the lender can verify the quality of their model. By comparing actual results against predictions, the lender can determine if their loan decision process is working as expected.

End users have limited visibility into a DA tool's process. They have knowledge of inputs and outputs but not ground truth. Thus, end users may find DA tool evaluation measures that can be calculated from Z and $\overline{Z}$ more useful than others. Given the proper measure, end users can better assess their DA tool options and monitor DA tool effectiveness. These end user visibility observations lead to the third criterion.

Criterion 3 (DA tool output basis). An end user efficacious summary statistic must be quantifiable with information known and visible to the end user (Z and $\overline{Z}$).

One advantage Criterion 3 confers to end users is the ability to compare predicted outcomes to field observations. Criterion 1 addresses information relevance, and Criterion 3 addresses information availability. Availability is not explicitly mentioned in the three questions but is implicit; none of the questions can be answered, if the information is not available. Hence, Criterion 3 is relevant to all three questions.

The ratio$_+$ invariant measures, ROC-AUC and the Youden index, have as their basis the measure suite $\{t_+/|Y|, f_+/|\overline{Y}|\}$ [6]. Both of these measures are conditioned by ground truth (Y and $\overline{Y}$), not the DA tool outputs visible to the end user (Z and $\overline{Z}$); they do not satisfy Criterion 3. However, the ratio$_+$ invariant measure pair DOR/DP do. Criterion 3 compliance is demonstrated by substituting the four conditional ratios

$$\frac{t_+}{|Z|}, \quad \frac{f_+}{|Z|}, \quad \frac{f_-}{\left|\overline{Z}\right|}, \quad \frac{t_-}{\left|\overline{Z}\right|} \tag{2}$$

for t_+, f_+, f_-, and t_- in the measures. In the DOR equation,

$$\text{DOR} = \frac{t_+/|Z| * t_-/\left|\overline{Z}\right|}{f_+/\left|\overline{Z}\right| * f_-/|Z|}. \tag{3}$$

Multiplying the numerator and denominator by $|Z| * |\overline{Z}|$ results in

$$\text{DOR} = \frac{t_+ * t_-}{f_+ * f_-}, \tag{4}$$

the original DOR equation. DP = $(\sqrt{3}/\pi)\log(\text{DOR})$, and thus DP also satisfies Criterion 3.

Using the same substitution above in the equations for TAR, F_β-score, AUC, MCC, and IC shows none are equivalent to the original equations. Of the commonly seen summary statistics considered, only the DOR/DP satisfies Criterion 3.

6.2.4. Criterion 4: Measure Value Appropriateness. Lenders are interested in loans that maximize profit. Physicians are interested in treatments that maximize patient QoL. Dairymen are interested in breeding cows to maximize milk production. If a dairyman, instead of being given data on pounds of milk produced per cow, was given data on the variation in milk produced per cow, using that information would be difficult to optimize milk production. Outputs from measures such as ROC-AUC and DOR/DP are not mappable to these user's needs. For a measure to be end user efficacious, the end user must be able to map measure output to their problem. The end user wants to avoid making a decision, given unrelated information. An informed end user may know what the values presented quantify. But, if the values are not mappable to their problem, the end user must rely on "soft" evaluations, such as expert opinion, which may incur considerable uncertainty.

Criterion 4 (measure value appropriateness). The summary statistic output must quantify the DA tool's impact on the end user's characteristic of interest.

Criterion 4 may seem self-evident, but not all measures satisfy it. For example, the ROC-AUC quantifies the probability that a randomly selected member of class Y will have a lower test value than a randomly selected member of class $\overline{Y}$. Thus, the ROC-AUC value assumes prior knowledge of ground truth, which, if an end user knew, would mean no DA tool was needed. Criterion 4 may seem similar to Criterion 1, but it addresses the DA tool function, while Criterion 1 addresses the input's information content.

DOR/DP quantify the odds of two randomly selected elements of the test set being one each T_+ and T_-, rather

Table 2: The proposed measure properties mapped well to the end user issues in Section 4.

	End user need		
	Quantify impact	Identify B^*	Identify B sensitivity
Criterion 1: category impacts	Yes	Yes	Yes
Criterion 2: pdf sensitivity	No	Yes	Yes
Criterion 3: perspective	Yes	Yes	Yes
Criterion 4: relevance	Yes	Yes	Yes
Ratio scale	Yes	No	Yes

than one each F_+ and F_-. DOR/DP does not require prior knowledge of ground truth, but it is a very specific scenario. It requires output pairs, rather than considering individual outputs. Secondly, there are ten possible pairings (e.g., two T_+s) and ninety unique ratios. Thus, DOR/DP output is not broadly applicable.

From the end user perspective, the ROC-AUC, DOR/DP, TAR, and F_β-score all share another failing; all have lower bounds of zero and cannot quantify a negative impact.

6.2.5. Preconditions from Measurement Theory. Measurement theory has addressed end user measure efficacy [47]. The scale-type definitions proposed by Stevens [43] are used here. Stevens defined four scale types, nominal, ordinal, interval, and ratio. Ratio scales have the least functional constraints, so measures using ratio scales are preferred. Of the three end user questions asked, measurement theory is relevant to two: "what is the DA tool's impact on my problem?" and "how sensitive is the impact to boundary selection?" Both need measurement scales with meaningful zeros and standard sequences. The remaining question "what is the boundary that provides the optimum impact?" is answerable on an ordinal scale.

Of the measures reviewed (TAR, F_β-score, MCC, Youden index, DOR/DP, IC, and ROC-AUC), none are quantified on a ratio scale.

6.3. The Necessity of the Criteria. First the root problem was identified as a means of identifying measure characteristics that satisfy end user's needs. Table 2 summarizes that mapping. There is a strong relationship between the measure properties proposed and the end user questions posed. In the table, "yes" means the criterion is necessary and "no" means the criterion is not.

The end user questions posed are topics addressed in business management programs and operations research, so they are generally applicable. Hence, the measure properties are generally applicable as well. Each of these criteria addresses at least one end user need. A measure that does not satisfy every criterion fails to provide some information needed by end users. Hence, these criteria are necessary for problems within the scope of this work. Investigating sufficiency will be a topic of future work.

6.4. Two End User Efficacious DA Tool Evaluation Measures. Table 3 recaps how each summary statistic tested conforms to the criteria and ratio scale properties. None satisfy all of the criteria. Sokolova and Lapalme tested invariance to JPT perturbations [34]. Where their tests are relevant to proposed criteria, their results corroborate these results. As noted in Section 1 and supported in Section 6.2, the commonly seen DA tool evaluation measures do not well quantify end user impact. This section proposes suitable measures. Two problem types are considered separately.

When the impact is cumulative, there will be either a gain or loss (impact, ι) associated with each element output. ι can be expressed as a statistical, not necessarily unique, expectation for each JPT category; $(I = (\iota_{T_+}, \iota_{F_+}, \iota_{F_-}, \iota_{T_-}))$. Thus, each and every element output will affect the end user by the element of I applicable to the category to which the element is binned. (The elements of I can be defined in different ways, depending upon the information known about each element of S. For instance, if elements were bank loan applications, then the impact could be quantified per dollar requested. Alternatively, impact could be based on the statistical expectation for the category. Using bank loan applications, the impact could be per loan.) An end user can expect the net gain or loss (ι_{net}) to be the sum of the individual element gains and losses. For problems where impact is cumulative,

$$\iota_I = \iota_{T_+}\frac{t_+}{|S|} + \iota_{F_+}\frac{f_+}{|S|} + \iota_{F_-}\frac{f_-}{|S|} + \iota_{T_-}\frac{t_-}{|S|}. \tag{5}$$

ι_I can also be expressed on Z and $\overline{Z}$, the outputs actually observed by the end user:

$$\iota_I = \iota_Z\frac{|Z|}{|S|} + \iota_{\overline{Z}}\frac{\left|\overline{Z}\right|}{|S|}. \tag{6}$$

For the test set S, *estimated total impact* is $\iota_{\text{tot}} = |S|\iota_I$. "Profit," a measure for customer churn prediction models, was introduced by Verbraken et al. [48]. It differs from ι_I in two ways: (i) profit's costs and benefits must all be positive values and (ii) misclassification costs are deducted, and correct classification gains are added. Intuitively, Verbraken et al.'s constraints seem correct: gains are positive and losses are negative. However, this is not universally true. Recasting a problem to fit profit's requirements could cause the measurement scale to no longer have a meaningful zero resulting in an interval scale and invalidation of analysis such as Verbraken et al.'s proposed cost-benefit ratio. ι_I, as seen in (5), is not susceptible to measure degradation.

There are occasions when the impact is not cumulative, but each output is important individually, for example, a medical diagnosis. In these situations, ratio$_+$ is confounding, so normalized JPTs are used; normalization mathematically balances the relative class sizes, and thus it mitigates any skew resulting from ratio$_+$ (JPT normalization's ratio$_+$ mitigation is limited by the strong law of large numbers; if the minor class sample is too small, then ratio$_+$ skew exerts a significant influence [6]). A normalized JPT is shown in Table 4. To facilitate comparison with nonnormalized JPTs, the sum of

Table 3: None of the summary statistics considered satisfy all four criteria.

	Summary statistic						
	TAR	F_β-score	Youden	MCC	IC	DOR	AUC
Criterion 1: category impacts	No	Partial	No	No	No	No	No
Criterion 2: pdf sensitivity	Yes	Yes	Yes	Yes	Yes	Yes	No
Criterion 3: perspective	No	No	No	No	No	Yes	No
Criterion 4: relevance	No	No	No	No	No	No	No
Meaningful zero	No	No	Yes	Yes	Yes	No	No
Standard sequence	No	No	No	No	No	Yes	No

Table 4: The values in this JPT have been normalized.

		Actual classification		
		Y	$\overline{Y}$	Totals ↓
Test	$+: s_i \in \{Z\}$	$t_{+n} = \dfrac{t_+}{2\lvert Y\rvert}$	$f_{+n} = \dfrac{f_+}{2\lvert\overline{Y}\rvert}$	$\lvert Z_n\rvert$
Result	$-: s_i \notin \{Z\}$	$f_{-n} = \dfrac{f_-}{2\lvert Y\rvert}$	$t_{-n} = \dfrac{t_-}{2\lvert\overline{Y}\rvert}$	$\lvert\overline{Z_n}\rvert$
Normalized totals		0.5	0.5	1

all categories is kept at one ($|S| = 1$) and the individual input class values (Y and $\overline{Y}$) add up to 0.5.

The end user's concern, (a) "given that a result is rendered, how am I affected?," however, can be partitioned into two questions:

(a1) "given that the result is positive, how am I affected?" and

(a2) "given that the result is negative, how am I affected?." (Problems where DA tool results are cumulative can be partitioned in the same way. However, when results are cumulative, question (a) is the most useful for an end user; questions (a1) and (a2) may be of secondary importance. When individual outputs are important, questions (a1) and (a2) are primary.)

These questions indicate that the values of interest are weighted conditional expectations:

$$\iota_Z = \frac{\iota_{T_+} t_{+n} + \iota_{F_+} f_{+n}}{\lvert Z_n\rvert},$$
$$\iota_{\overline{Z}} = \frac{\iota_{T_-} t_{-n} + \iota_{F_-} f_{-n}}{\lvert\overline{Z_n}\rvert}. \tag{7}$$

Output is independent, thus the expected outcome impact equals its average:

$$\iota_\sigma = \frac{\iota_Z + \iota_{\overline{Z}}}{2}. \tag{8}$$

Substituting the normalized expressions from Table 4, the expected impact becomes

$$\iota_\sigma = \frac{1}{2}\left(\frac{\iota_{T_+} t_{+n}}{\lvert Z_n\rvert} + \frac{\iota_{F_+} f_{+n}}{\lvert Z_n\rvert} + \frac{\iota_{T_-} t_{-n}}{\lvert\overline{Z_n}\rvert} + \frac{\iota_{F_-} f_{-n}}{\lvert\overline{Z_n}\rvert}\right). \tag{9}$$

The two problem types (individual impact and cumulative impact) have their unique characteristics, resulting in different sets of relevant measures. For problems where impact is cumulative, the summary statistic,

$$\iota_I = \frac{1}{|S|}\left(\iota_{T_+} t_+ + \iota_{F_+} f_+ + \iota_{F_-} f_- + \iota_{T_-} t_-\right), \tag{10}$$

provides actionable information to the end user. The monotonic measures upon which it is based and which provide more insight into DA tool impact are the four individual category impacts, $I = (\iota_{T_+}, \iota_{F_+}, \iota_{F_-}, \iota_{T_-})$, where

$$\iota_{T_+} = \frac{\sum_{s\in T_+} \iota_s}{t_+},$$
$$\iota_{F_+} = \frac{\sum_{s\in F_+} \iota_s}{f_+},$$
$$\iota_{F_-} = \frac{\sum_{s\in F_-} \iota_s}{f_-},$$
$$\iota_{T_-} = \frac{\sum_{s\in T_-} \iota_s}{t_-}. \tag{11}$$

Both ι_I and ι_σ have optima, so they are summary statistics. In the case of ι_I, the associated measure suite (measures that provide insight into a specific aspect of the DA tool output) consists of ι_{Z_I} and $\iota_{\overline{Z}_I}$, the two outputs observable by end users. ι_I differs from the usual summary statistic in that it directly quantifies the characteristic of interest to end users. ι_σ's measure suite consists of the conditional expectations of the two outputs observable by an end user, ι_{Z_σ} and $\iota_{\overline{Z}_\sigma}$. For problem domains where ι_σ is appropriate, the summary statistic does contain less information. Consider, for instance, the example where a person receives a medical diagnosis. If the test result is positive, then that person's impact will be *either* ι_{T_+} *or* ι_{F_+}. Likewise, if the test result is negative, then that person's impact will be *either* ι_{T_-} *or* ι_{F_-}. The three composite measures, ι_σ, ι_{Z_σ}, and $\iota_{\overline{Z}_\sigma}$, may have little utility for the patient. The values are, however, useful to diagnosticians in assessing diagnostic and treatment strategies.

ι_I and ι_σ are suitable for many DA problems and extensible to DA problems with more than two classes. Since they are additive, all that is needed is to sum up the impact adjusted JPT values. For ι_σ, the JPT must be normalized; then, each category is conditioned by its DA tag. Then, the impact adjusted values can be summed up.

Table 5: ι_I and ι_σ satisfy the four criteria and measurement theory criteria.

Criterion	Summary statistic								
	ι_I	ι_σ	TAR	F	J	MCC	IC	DOR	AUC
(1) Category impact	Yes	Yes	No	Not t_-	No	No	No	No	No
(2) pdf sensitivity	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
(3) Perspective	Yes	Yes	No	No	No	No	No	Yes	No
(4) Relevance	Yes	Yes	No	No	No	No	No	No	No
Meaningful zero	Yes	Yes	No	No	Yes	Yes	Yes	No	No
Standard sequence	Yes	Yes	No	No	No	No	No	Yes	No

6.4.1. Comparing ι_I and ι_σ against the Criteria. Commonly seen summary statistics do not satisfy these four criteria. How well do the impact measures ι_I and ι_σ satisfy the criteria?

Criterion 1 (category importance). A measure that reflects category importance will exhibit sensitivity to change in category impact. If ι_{T_+} in (5) is replaced by $\iota_{T_-} + \epsilon$, where ϵ represents a small change, then ι_I becomes $(\iota_I + \epsilon)t_+$, as ϵ changes, with a corresponding change in ι_I. (The correspondence is true for $t_+ > 0$. If $t_+ = 0$, then ϵ has no effect. If T_+ elements do not occur, then they have no impact.) A similar situation exists for ι_σ. $\iota_{T_{+n}}$ in (9) is replaced by $\iota_{T_{+n}} + \epsilon$, and then ι_σ becomes $(\iota_\sigma + \epsilon)\epsilon t_{+n}/4|Z_n|$. Other than the trivial case when a JPT cell cardinality equals zero, an ϵ change in the corresponding element of I will change the value of the ratio. ι_I and ι_σ satisfy Criterion 1.

Criterion 2 (pdf sensitivity). The discussion in Section 6.2 shows that JPT category cardinalities are sensitive to changes in pdf. Hence, Criterion 2 compliant measures must be sensitive to changes in JPT category cardinalities. ι_I and ι_σ are sensitive to changes in all four JPT category cardinalities, and thus they satisfy Criterion 2.

Criterion 3 (DA tool output basis). ι_σ is based on the weighted conditional expectations (Equation (7)) on Z and $\overline{Z}$, the outputs visible to the end user. Thus, ι_σ can be calculated from data available to end users and ι_σ can be deconstructed into Z and $\overline{Z}$ relevant components. ι_I, as written in (6), is the sum of the estimated impact of Z and $\overline{Z}$; ι_I can be deconstructed into Z and $\overline{Z}$ relevant components. Thus both satisfy Criterion 3.

Criterion 4 (scale appropriateness). For Criterion 4, I needs to be quantified in a unit appropriate for end user impact, and then ι_I and ι_σ satisfy Criterion 4.

Measurement Theory: Outputting Ratio Scale Values. When either $\iota_I = 0$ or $\iota_\sigma = 0$, then the DA tool has no noticeable overall effect on the end user; the measures have meaningful zeros.

The discussion for Criterion 1 shows that ι_I and ι_σ values have standard intervals. Regardless of the value of any $\iota_{\{T_+,F_+,F_-,T_-\}}$, a change in its value will cause a corresponding change in ι_I and ι_σ.

Table 5 summarizes this study's results. From the end user perspective, ι_I and ι_σ have all of the desired characteristics.

7. Discussion

Starting from a well-defined problem, impact measures for end user problems have been devised. Here, measure usage and differences in the fundamental nature of the impact measures and the commonly seen ratio measures are discussed. Following this, measure outputs are compared on simulated classifier output. Finally a reanalysis of published studies shows additional insights end users can gain by using impact measures.

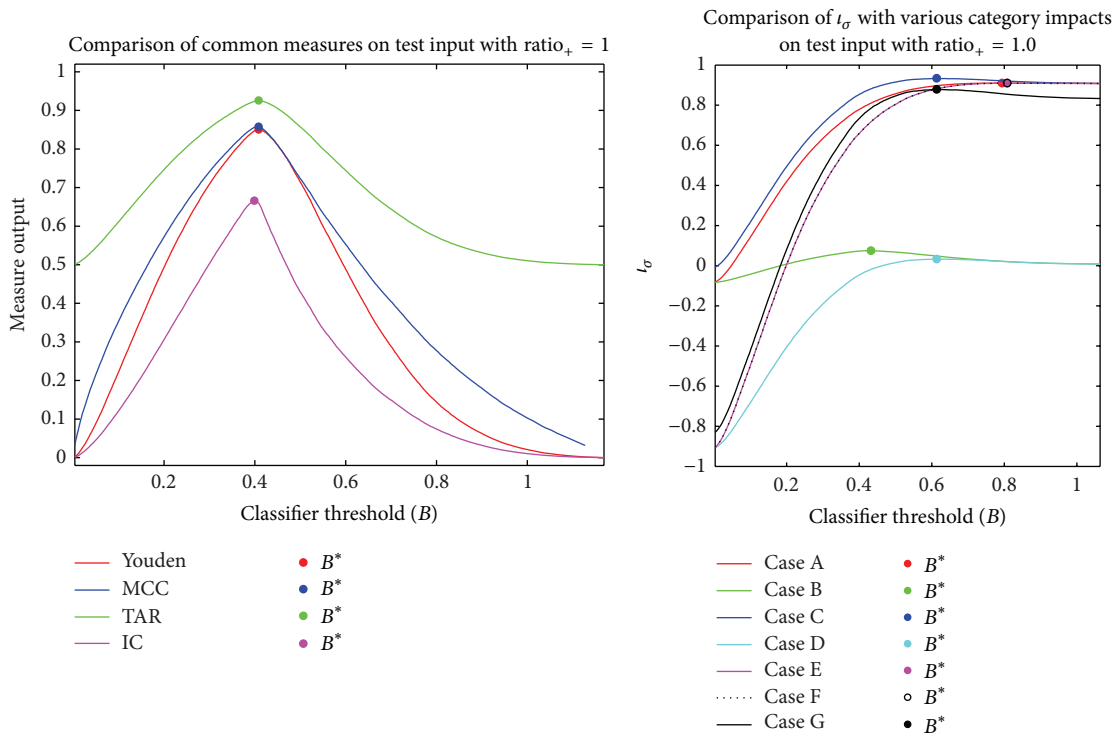
7.1. Measure Comparisons. There are two key differences between the additive measures ι_I and ι_σ and the commonly seen measures (all of which are ratio based):

(i) The ratio-based measures are either unitless or use units with weak utility for end users. ι_I and ι_σ have units defined in I. For end user efficacy, the units must be relevant to the problem and provide end users actionable information.

(ii) The ratio-based measures are mostly measured on ordinal scales, which limits comparison to rank ordering. ι_I and ι_σ are measured on a ratio scale, where DA impacts are ordered, but difference and magnitude are also valid comparisons. An end user can determine not only which DA tool is better, but also *how much better*.

As noted in Section 1, measures differ in their response to changes in ratio$_+$ and accuracy. Generally, the differences are monotonically consistent, so rank ordering would not change. There are exceptions: (i) the ratio$_+$ sensitive measure TAR monotonic increase, as ratio$_+$ increases, MCC and IC monotonically decrease; (ii) ι_I's response to ratio$_+$ is subject to I: it can either monotonically increase or decrease. When DA tools are evaluated with the same test data (for ι_I, I must remain the same as well), this will not cause a change in rank ordering. If test sets with different ratio$_+$s are used to evaluate DA tool performances, TAR would rank them differently from MCC or IC.

7.2. Simulation Tests. Some use cases illustrate the value of the impact measures. Symmetrical class probability distributions can mask some differences, so for all measures class samples have the beta (1.5, 5.0) distribution.



(a) Each of these commonly seen measures identify similar optimum boundaries. However, the values differ significantly

(b) ι_σ is tunable to the end user's situation. End users can readily identify B^*, B sensitivity, and the expected impact the classifier will have on their problem

FIGURE 1: B^* for each of the common measures are similar. ι_σ, however, shows that end user results are sensitive to I. An end user using ι_σ receives much more actionable information than if one of the common measures is used.

Regarding ROC-AUC, J, DOR/DP, TAR, MCC, and IC; testing took into consideration the following:

(i) They all implicitly use $I = (1, 1, 1, 1)$.

(ii) ROC-AUC, J, and DOR/DP are ratio_+ invariant, so they can be used on test sets with any ratio_+. Their invariance, however, forces $\text{ratio}_+ = 1$. These measures are compared in use cases with $\text{ratio}_+ = 1$.

(iii) TAR, MCC, and IC are considered inappropriate for data with $\text{ratio}_+ \neq 1$. These measures are compared in use cases with $\text{ratio}_+ = 1$.

For ι_I, ι_σ, and F-score, there are a limited number of variations. For category impacts, JPTs are bilaterally symmetrical, so there are seven cases. The following list shows the category impact relationships and I used:

(i) One impact $\gg$ three impacts:

(a) $\max(|I|) = \iota_{T_{\{+,-\}}}$, $I = (1, -0.1, -0.1, 0.1)$ (Case A);

(b) $\max(|I|) = \iota_{F_{\{+,-\}}}$, $I = (0.1, -1, -0.1, 0.1)$ (Case B).

(ii) Two impacts $\gg$ two impacts:

(a) $\max(|I|) = \{\iota_{T_+}, \iota_{T_-}\}$, $I = (1, -0.1, -0.1, 1)$ (Case C);

(b) $\max(|I|) = \{\iota_{F_+}, \iota_{F_-}\}$, $I = (0.1, -1, -1, 0.1)$ (Case D);

(c) $\max(|I|) = \{\iota_{T_+}, \iota_{F_+}\}$, $I = (1, -1, -0.1, 0.1)$ (Case E);

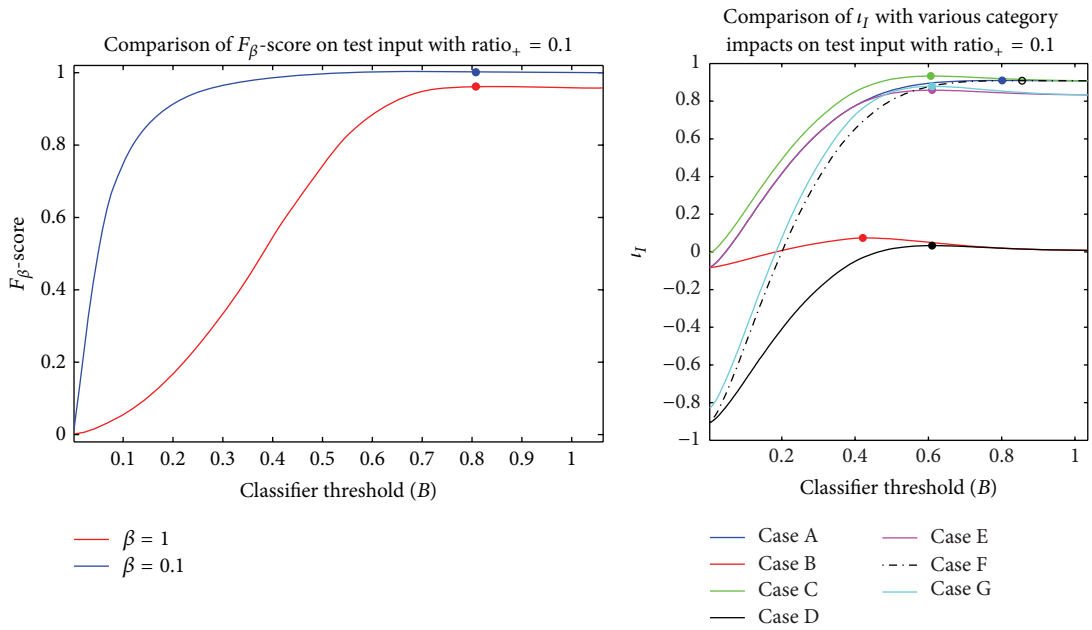
(d) $\max(|I|) = \{\iota_{T_+}, \iota_{F_-}\}$, $I = (1, -0.1, -1, 0.1)$ (Case F);

(e) $\max(|I|) = \{\}$, $I = (1, -1, -1, 1)$ (Case G: this is the additive measure equivalent of $I = (1, 1, 1, 1)$ for ratio measures).

This gives a total of twenty use cases, seven each for the two *impact* measures and six for F_β-score (since $\iota_{T_-} = 0$ for F_β-score, cases A and C are identical).

Mapping I to β is not exact. However, from β's somewhat subjective definition, it may be viewed as ι_{F_-}/ι_{F_+}. This reduces F_β-score's test cases to three: $\beta = 10$, $\beta = 1$, and $\beta = 0.1$. Interestingly, $\beta = 10$ and $\beta = 0.1$ result in the same F_β-score, so the seven distinct I cases resolve to two distinct F_β-scores.

Figure 1 compares the common measures suitable for $\text{ratio}_+ = 1$ and ι_σ. The dots on each curve indicate the optimum boundary, B^*. One important difference between the measures in Figures 1(a) and 1(b) is that the Y-axis units in Figure 1(b) are meaningful to end users. For the dairyman's problem, perhaps the units would be tons of milk per day per herd. For the banker, perhaps annual profit. The Y-axis values in Figure 1(a) have no such relevance to the user. The variation in the optima and curve shapes in Figure 1(b) show



(a) Varying β does impact the F-score. However, B^* are essentially the same

(b) ι_I is tunable to the end user's situation. End users can readily identify B^*, B sensitivity, and the expected impact the classifier will have on their problem. In some cases, the end user could experience a negative result. This is reflected in ι_I; F-score cannot show this information

FIGURE 2: In contrast to ι_I shows that B^*, B sensitivity, and expected impact can vary considerably. F-score does not reveal this information.

ι_I's sensitivity to I. This is a benefit for end users, as they can determine B^* and B sensitivity. Cases A and E appear identical left of B^*(Case E), otherwise all of the cases tested are significantly different. Alas, I sensitivity is confounding for DA research. Fortunately, there is commonality between one common measure and ι_I: TAR and ι_I, Case G, (the additive measure equivalent of $I = (1, 1, 1, 1)$ for ratio measures) are extremely similar. The two measure's output similarities may mean that ι_I and TAR are functionally equivalent (in an operational, not mathematical sense).

DOR/DP and ROC-AUC are not shown in these figures. DOR/DP is measured on an interval scale, so to show it would require a separate figure. Its characteristics, as defined by B^* and curve shape, were far different from the other six measures evaluated. Its graph is included in the on-line technical report [46]. ROC-AUC is boundary invariant; its "curve" is a horizontal line, so ROC-AUC is not included in Figure 1(a).

Figure 2 shows the comparison between the F_β-score and ι_σ for ratio$_+$ = 0.1. As with the previous comparisons, the important difference between the measures in Figures 2(a) and 2(b) is that the Y-axis units in Figure 2(b) are meaningful to end users. Although β varies by an order of magnitude in the two tests, B^* is essentially the same. Users for whom the importance of F_+ and F_- is equal will be less satisfied than users for whom there is an order of magnitude difference in F_+ and F_-'s importance. ι_I exhibits a much greater sensitivity to I. In every case, there are differences in the curves and B^*. In all cases, the impact measures provide information on B^*, B sensitivity, and expected impact that are not available using the other measures.

7.3. Real-World Problem 1: Evaluating Rheumatoid Arthritis Diagnostic Tests. Nishimura et al. [13] published a meta-analysis [49] of evaluations of two rheumatoid arthritis (RA) diagnostic tests. The meta-analysis is quite thorough and accounts for many potential variations between studies. The team concludes that one test is better than the other, however, they do so without using a summary statistic. This reanalysis adds ι_σ, the appropriate impact measure identified in Section 6.4. Using ι_σ, end users can identify B^*, impact, and B sensitivity. Analysis is limited as results at the pre-defined B^* are given, but not the underlying data. Information difference is shown between ι_σ and the original results, but additional insights are not gained.

Nishimura et al.'s study uses two measures, positive likelihood ratio (LR$_+$) and negative likelihood ratio (LR$_-$). These measures are not summary statistics, but they allow calculation of the underlying normalized JPT. The authors observe that RA treatment is harmful to and costly for persons with false positive results. Regardless of the diagnosis, a correct diagnosis maximizes the subject's quality of life. Accordingly, the meaningful zero is defined as the cost associated with a correct diagnosis: $\{\iota_{T_+} = 0, \iota_{T_-} = 0\}$. An incorrect diagnosis results in reduced quality of life, so Lajas et al.'s reported costs [50] are rounded to two significant digits, setting the misdiagnosis costs at $\iota_{F_+} = -\$7{,}900$ and $\iota_{F_-} = -\$13{,}000$. Hence, $\iota_\sigma = 0$ means "the test results have

Table 6: (a) and (b) compare the summary likelihood ratios originally reported [13] (a) and the corresponding ι_Z, $\iota_{\overline{Z}}$, and ι_σ (b). Both measure suites show that the anti-CCP test is better, as does the summary statistic ι_σ.

(a)

	Normalized odds ratio measures	
Test	LR_+	LR_-
Anti-CCP	12.46 (9.72–15.98)	0.36 (0.31–0.42)
RF	4.86 (3.95–5.97)	0.38 (0.33–0.44)

(b)

	Expected (annual economic) impact ($\$10^3$)		
Test	ι_Z	$\iota_{\overline{Z}}$	ι_σ
Anti-CCP	−0.55 (−0.56–−0.44)	−9.6 (−9.7–−9.5)	−5.1 (−5.1–−5.0)
RF	−1.4 (−1.6–−1.3)	−9.5 (−9.6–−9.4)	−5.5 (−5.5–−5.5)

no negative effect on patient QoL" and the lower the value, the worse the impact on the patient. Since $\iota_\sigma = 0$ has a meaningful zero, a practitioner using ι_σ would seem to have a direct mapping from test result to the patient's expected experience.

In contrast, $LR_+ \in (0, \infty)$ means the expected test result is somewhere between "a positive test result is never correct" and "a negative test result is never correct." Similarly, $LR_- \in (0, \infty)$ means the expected test result is somewhere between "a positive test is always correct" and "a negative test is always correct." Both LR_+ and LR_- are related to patient QoL, but the mapping is not clear.

In an extension to Nishimura et al.'s report, calculate $\iota_{Z(\sigma)}$, $\iota_{\overline{Z}(\sigma)}$, and ι_σ on the pooled test data. Table 6 shows the original likelihood ratios reported by Nishimura et al. and the proposed new measures. (The parenthesized range is the 95% confidence interval. On a single tailed test, only one bound is relevant, and thus the bound indicates a 97.5% confidence.) The conclusions are reached using the proposed measures and the original measures match: the anti-CCP test is better than the RF test. Comparing ι_σ for each test given that the end user context requires ratio_+ invariance, the anti-CCP test estimated annual economic impact on patients is four hundred dollars less than the RF test's estimated annual economic impact.

Inspecting the raw JPT values in Table 7, anti-CCP has a substantially lower F_+ rate than the RF test. This might lead an end user to place substantially more trust in the RF test's negative result than in a negative result from the anti-CCP test. However, that trust does not result in a better outcome for the patient; both tests have statistically equivalent negative impacts. Because ι_Z and $\iota_{\overline{Z}}$ are measured in the same units, with the same meaningful zero, comparisons can be made between them. For example, the end user can see that, in contrast to a positive test result, a negative result can have a substantial negative annual cost: around 9,600 USD per year. An end user may not want to conclude a patient with a negative test result is RA-free without strong corroboration.

Assessing the RA test's impacts is useful to the researcher but is substantially more valuable to the end user. Table 8 shows the JPTs for both raw proportions and impacts.

Table 7: These *normalized JPTs* for pooled anti-CCP and RF test data were calculated from Nishimura et al.'s reported sensitivities and specificities [13]. A person without RA is far less likely to be misdiagnosed than one with the disease when the anti-CCP test is used.

(a)

		Actual RA condition	
		Diseased	Not diseased
Anti-CCP test	Positive	0.67 (0.65–0.68)	0.05 (0.04–0.06)
Result	Negative	0.33 (0.32–0.35)	0.95 (0.94–0.95)
	Totals	1	1

(b)

		Actual RA condition	
		Diseased	Not diseased
RF test	Positive	0.69 (0.68–0.7)	0.15 (0.14–0.16)
Result	Negative	0.31 (0.3–0.32)	0.85 (0.84–0.86)
	Totals	1	1

The raw test data were not available. If the raw data were available, it would be interesting to see how the curves for ι_Z and $\iota_{\overline{Z}}$ compare to LR_+ and LR_-. Likelihood ratios are not summary statistics, so they cannot directly provide B^*; it would also be interesting to compare B^* derived from the likelihood ratios with that predicted by ι_σ.

7.4. Real-World Problem 2: Bank Loan Decisions. Optimizing bank loan decisions is a "cumulative output" DA problem type; thus end user impact is best quantified by ι_I. There is a body of credit scoring algorithm tests; none with sufficient data was available for a full reanalysis. The Abdou work [16] included sufficient details to compare peak outputs identified by the algorithms tested.

Abdou provides normalized JPTs and reports a misclassification cost ratio (MCR) of 5 : 1; MCR = cost of Type II errors (f_+)/cost of Type I errors (f_-). (MCR considers direct costs only. It does not include the opportunity cost, the lost income attributable to qualified applicants not being funded.) Abdou

Table 8: (a) and (b) show $\iota_{Z(\sigma)}$, $\iota_{\overline{Z}(\sigma)}$, and ι_σ as well as the proportional contribution of each JPT category. Values are in thousands of dollars and indicate the unnecessary annual economic cost resulting from an incorrect diagnosis.

(a)

Anti-CC test	Actual RA condition		
Result	Diseased	Not diseased	ι
Positive	0	−0.55 (−0.56–−0.44)	$\iota_{Z(\sigma)} = -0.55$ (−0.56–−0.44)
Negative	−9.6 (−9.7–−9.5)	0	$\iota_{\overline{Z}(\sigma)} = -9.6$ (−9.7–−9.5)
			$\iota_\sigma = -5.1$ (−5.1–−5.0)

(b)

RF test	Actual RA condition		
Result	Diseased	Not diseased	ι
Positive	0	−1.4 (−1.5–−1.3)	$\iota_{Z(\sigma)} = -1.4$ (−1.6–−1.3)
Negative	−9.5 (−9.6–−9.4)	0	$\iota_{\overline{Z}(\sigma)} = -9.5$ (−9.6–−9.4)
			$\iota_\sigma = -5.5$ (−5.5–−5.5)

does not provide loan amount information, so a "standard loan unit" was defined as some arbitrary number of Egyptian pounds (EGP) and impact was calculated per loan unit. The JPT categories are defined as follows.

Good Applicants ($\iota_{T_+} = 1.0$). These are loans that are made and pay-off as expected.

Known Deadbeats ($\iota_{T_-} = -0.01$). These are applicants rejected where ground truth is a known default. ι_{T_-} all applicants, including those rejected, incur an application processing cost; hence, ι_{T_-} is negative.

Unknown Deadbeats ($\iota_{F_+} = -0.05$). These are loans made that defaulted. The value is based on MCR = 5 : 1.

Unknown Good Applicants ($\iota_{F_-} = -0.01$). These are rejected applicants that would have proven to be good.

In order to limit complexity, the standard loan unit has a defined annual profit expectation. Intuitively, a user could expect $\iota_{T_-} > 0$. In Abdou's scenario, ι_{T_-} has a slight negative impact. This is due to application processing costs incurred for all applications.

Abdou also normalizes his data $\text{ratio}_+ = 1.0$. However, the problem is ratio_+ sensitive, so JPT tuning is used to adjust to the reported value, $\text{ratio}_+ = 2.1$. Abdou ran a sensitivity analysis on EMC; JPT tuning use illustrates how an end user can run a ratio_+ sensitivity analysis. (Such a sensitivity analysis can test results at the identified boundary; however, ratio_+ causes the optimum boundary to shift. So, without the actual data, JPT tuning cannot be used to estimate the peak impact [6].) JPT tuning is used for two other relative class sizes: $\text{ratio}_+ = 1.5$ and $\text{ratio}_+ = 2.5$. Table 9 compares ι_I results with the estimated misclassification cost, EMC, reported by Abdou.

Abdou concludes that the WOE model performs best, based on EMC. ι_I shows that genetic programming performs best. Abdou does not report confidence intervals, but since probit analysis (PA) and GP_p results are so similar, the difference is likely statistically insignificant. Based on ι_I, WOE seems to perform worst. WOE has a substantial negative impact on the lender compared to either PA or GP. This reanalysis indicates that GP is at least equivalent to the best nonartificial intelligence method tested; this is consistent with other tests comparing artificial intelligence (AI) and non-AI methods.

Using the Egyptian banking assumptions presented here, sensitivity analysis of GP shows that, for $\text{ratio}_+ = [1.5, 2.5]$, the annual profit per loan unit would range from fifty-six to sixty-seven percent of the amount that would be received if loan decisions were perfect. Thus, by using ι_I, the bank decision-makers receive valuable information that can be used to define loan application scoring policy and procedures. The banking environment assumptions used are probably not extensible to a wide bank pool. Thus, ι_I will be most useful when each institution tunes the values to their specific environment.

If Abdou's raw data were available, it would be interesting to test B^* identification and B sensitivity. This problem's meaningful zero is the loan portfolio with a net zero profit; B^* is almost certainly different from those determined by Abdou. The hypothetical profits are probably greater than those determined using Abdou's results.

8. Conclusion

An important characteristic for DA tool end user efficacious summary statistics has been identified: impact (ι). First, An impact vector sets the end user impact of each joint probability table category: $I = (\iota_{T_+}, \iota_{F_+}, \iota_{F_-}, \iota_{T_-})$. Next, four criteria for end user efficacious summary statistics and measurement theory were evaluated and applied.

Criterion 1 (category importance). An end user efficacious summary statistic must be a function of problem specific, rational number impacts for each JPT category.

Table 9: This table compares Abdou's results [16] using estimated misclassification cost (EMC) and ι_I as measures. The highlighted values indicate the best results. EMC is a cost, and thus the lower the value, the better; WOE is best. ι_I estimates the net impact; the higher the number, the better; GP_t is best. The ι_I results are consistent with other studies; generally, AI-derived algorithms outperform manual algorithms.

Test	Original and reanalysis results				
	EMC	ι_I at $\text{ratio}_+ =$			
	5 : 1	$\text{ratio}_+ = 1$	$\text{ratio}_+ = 1.5$	$\text{ratio}_+ = 2.1$	$\text{ratio}_+ = 2.5$
WOE_{T2}	**0.4627**	0.3484	0.4223	0.4764	0.5068
PA	0.6232	0.4377	0.5299	0.5975	0.6354
GP_p	0.5679	0.4315	0.5555	0.5896	0.6259
GP_t	0.6964	**0.4590**	**0.5561**	**0.6271**	**0.6670**

Criterion 2 (pdf sensitivity). An end user efficacious summary statistic must be sensitive to differences between end user environments and changes in an end user's environment, as expressed in the DA tool input population ratio_+ and pdf.

Criterion 3 (DA tool output basis). An end user efficacious summary statistic must be quantifiable with information known and visible to the end user (Z and $\overline{Z}$).

Criterion 4 (measure value appropriateness). *An end user efficacious summary* statistic must quantify the DA tool's impact on the characteristic of interest.

Measurement Theory. Ratio scales allow the most extensive analysis for end users.

Eight commonly seen DA tool summary statistics, total accuracy rate, F-score, Youden index, Diagnostic Odds Ratio (and associated measure, discriminant power), ROC area under the curve, Mathews Correlation Coefficient, and Mutual Information Coefficient, fail to satisfy these criteria. Two criterion compliant end user efficacious summary statistics were identified along with their measure suites.

(i) For cumulative DA tool output impact, the summary statistic is

$$\iota_I = \frac{1}{|S|}\left(\iota_{T_+} t_+ + \iota_{F_+} f_+ + \iota_{F_-} f_- + \iota_{T_-} t_-\right). \tag{12}$$

The end user efficacious measure suite consists of

$$\iota_{Z(I)} = \iota_Z \frac{\iota_{T_+} t_+ + \iota_{F_+} f_+}{|S|},$$
$$\iota_{\overline{Z}(I)} = \iota_{\overline{Z}} \frac{\iota_{F_-} f_- + \iota_{T_-} t_-}{|S|}. \tag{13}$$

(ii) For noncumulative DA tool output impact, the summary statistic is

$$\iota_\sigma = \frac{1}{2}\left(\frac{\iota_{T_+} t_{+n}}{|Z_n|} + \frac{\iota_{F_+} f_{+n}}{|Z_n|} + \frac{\iota_{T_-} t_{-n}}{|\overline{Z_n}|} + \frac{\iota_{F_-} f_{-n}}{|\overline{Z_n}|}\right). \tag{14}$$

The end user efficacious measure suite consists of

$$\iota_{Z(\sigma)} = \frac{\iota_{T_+} t_{+n} + \iota_{F_+} f_{+n}}{|Z_n|},$$
$$\iota_{\overline{Z}(\sigma)} = \frac{\iota_{T_-} t_{-n} + \iota_{F_-} f_{-n}}{|\overline{Z_n}|}. \tag{15}$$

Generally, the intent of publishing DA tool performance data is to inform a broad readership, including potential end users. Using specific I would be too restrictive, unless the report is for a specific audience, such as the rheumatoid arthritis study. Otherwise, end users would be better served if DA tool performance reports provide data that enables them to calculate their own impacts.

One suggestion is for researchers to publish ι_I and ι_σ with balanced I, on normalized JPT(B^*), and the normalized JPT(B) for a range of boundaries. Researchers can use the published ι_I and ι_σ values and end users have the values necessary to calculate actionable information for their specific problem.

The three-step process for end users is as follows:

(1) Identify the appropriate measure, ι_I or ι_σ.

(2) If ι_I is appropriate, then use JPT tuning to compensate for their problem domain's ratio_+. If ι_σ is appropriate, then condition the published JPTs by Z and $\overline{Z}$.

(3) Calculate the selected measure and select the JPT(B^*) with the best impact.

With these values, end users can also determine the DA output's sensitivity to B.

Future work includes loosening problem constraints to include unsupervised tests, exploring the potential relationship between F_β-scores β and I and studying the relationship between ratio_+ and DA tool output. The cost curves discussed by Witten et al. [41] may provide a means to avoid publishing multiple JPTs for end users with cumulative DA tool type problems.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The authors gratefully acknowledge Dr. Andrew Barnes (General Electric Global Research, Niskayuna, NY). His insights during this study's formative stage were invaluable.

References

[1] D. G. Altman, *Practical Statistics for Medical Research*, Chapman & Hall, London, UK, 1st edition, 1997.

[2] B. S. Everitt and C. R. Palmer, *Encyclopaedic Companion to Medical Statistics*, John Wiley & Sons, Chichester, UK, 2nd edition, 2011.

[3] R. C. James and G. James, *Mathematics Dictionary*, Van Nostrand Reinhold, New York, NY, USA, 5th edition, 1992.

[4] S. David and J. Hand, *Constructin and Assessment of Classification Rules*, John Wiley and Sons, Chichester, UK, 1997.

[5] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms A Classification Perspective*, Cambridge University Press, New York, NY, USA, 2014.

[6] E. E. Eiland and L. M. Liebrock, "Efficacious end user measures; part 1: relative class size and end user problem domain," *Advances in Artificial Intelligence*, vol. 2013, Article ID 427958, 22 pages, 2013.

[7] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[8] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, 1943.

[9] V. N. Vapnik and A. Lerner, "Pattern recognition using generalized portrait method," *Automation and Remote Control*, vol. 24, pp. 774–780, 1963.

[10] M. Schonlau, W. DuMouchel, W.-H. Ju, A. F. Karr, M. Theus, and Y. Vardi, "Computer intrusion: detecting masquerades," *Statistical Science*, vol. 16, no. 1, pp. 58–74, 2001.

[11] A. D. Solis and S. Rackovsky, "Information and discrimination in pairwise contact potentials," *Proteins: Structure, Function and Genetics*, vol. 71, no. 3, pp. 1071–1087, 2008.

[12] B. A. Almquist, *Mining for evidence in enterprise corpora [Ph.D. thesis]*, Univrsity of Iowa, Ames, Iowa, USA, 2011.

[13] K. Nishimura, D. Sugiyama, Y. Kogata et al., "Meta-analysis: diagnostic accuracy of anti-cyclic citrullinated peptide antibody and rheumatoid factor for rheumatoid arthritis," *Annals of Internal Medicine*, vol. 146, no. 11, pp. 797–808, 2007.

[14] M. S. Cline, K. Karplus, R. H. Lathrop, T. F. Smith, R. G. Rogers Jr., and D. Haussler, "Information-theoretic dissection of pairwise contact potentials," *Proteins*, vol. 49, no. 1, pp. 7–14, 2002.

[15] C. S. Miller and D. Eisenberg, "Using inferred residue contacts to distinguish between correct and incorrect protein models," *Bioinformatics*, vol. 24, no. 14, pp. 1575–1582, 2008.

[16] H. A. Abdou, "Genetic programming for credit scoring: the case of Egyptian public sector banks," *Expert Systems with Applications*, vol. 36, no. 9, pp. 11402–11417, 2009.

[17] National Science Board, "Globalization of science and engineering research science and engineering indicators 2010 (nsb 10-3)," February 2015, http://www.nsf.gov/statistics/nsb1003/definitions.htm.

[18] A. Jamain and D. J. Hand, "Mining supervised classification performance studies: a meta-analytic investigation," *Journal of Classification*, vol. 25, no. 1, pp. 87–112, 2008.

[19] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 3rd edition, 2009.

[20] R. P. W. Duin, "A note on comparing classifiers," *Pattern Recognition Letters*, vol. 17, no. 5, pp. 529–536, 1996.

[21] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation," in *AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4–8, 2006. Proceedings*, vol. 4304 of *Lecture Notes in Computer Science*, pp. 1015–1021, Springer, Berlin, Germany, 2006.

[22] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent measures of risk," *Mathematical Finance*, vol. 9, no. 3, pp. 203–228, 1999.

[23] K. P. Murphy, *Machine Leariing A Probabilistic Prespective*, MIT Press, Cambridge, Mass, USA, 1st edition, 2012.

[24] L. R. Ott and M. Longnecker, *An Introduction to Statistical Methods and Data Analysis*, Duxbury Press, Pacific Grove, Calif, USA, 5th edition, 2001.

[25] A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. M. Bossuyt, "The diagnostic odds ratio: a single indicator of test performance," *Journal of Clinical Epidemiology*, vol. 56, no. 11, pp. 1129–1135, 2003.

[26] D. D. Blakeley, E. Z. Oddone, V. Hasselblad, D. L. Simel, and D. B. Matchar, "Noninvasive carotid artery testing: a meta-analytic review," *Annals of Internal Medicine*, vol. 122, no. 5, pp. 360–367, 1995.

[27] C. J. Van, "Information Retrieval," 1979, http://www.dcs.gla.ac.uk/Keith/Preface.html.

[28] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975.

[29] B. Rost and C. Sander, "Prediction of protein secondary structure at better than 70% accuracy," *Journal of Molecular Biology*, vol. 232, no. 2, pp. 584–599, 1993.

[30] J. A. Swets, "Form of emperical ROCs in discrimination and diagnostic tasks: implications for theory and measurement performance," *Psychological Bulletin*, vol. 99, no. 2, pp. 181–198, 1986.

[31] J. A. Swets, "Indices of discrimination or diagnostic accuracy: their ROCs and implied models," *Psychological Bulletin*, vol. 99, no. 1, pp. 100–117, 1986.

[32] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, no. 4857, pp. 1285–1293, 1988.

[33] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[34] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, no. 4, pp. 427–437, 2009.

[35] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[36] C. Parker, "An analysis of performance measures for binary classifiers," in *Proceedings of the 11th IEEE International Conference on Data Mining (ICDM '11)*, pp. 517–526, Vancouver, Canada, December 2011.

[37] N. Japkowicz, "Why question machine learning evaluation methods?" in *Proceedings of the AAAI-06: Evaluation Methods for Machine Learning Workshop*, pp. 6–11, 2006.

[38] R. Caruana and A. Niculescu-Mizil, "Data mining in metric space: an empirical analysis of supervised learning performance criteria," in *Proceedings of Knowledge Discovery and Data Mining*, pp. 859–864, Seattle, Wash, USA, August 2004.

[39] N. Seliya, T. M. Khoshgoftaar, and J. Van Hulse, "A study on the relationships of classifier performance metrics," in *Proceedings of the 21st IEEE International Conference on Tools with Artificial Intelligence (ICTAI '09)*, pp. 59–66, IEEE, Newark, NJ, USA, November 2009.

[40] M. Sokolova, "Assessing invariance properties of evaluation measures," in *Proceedings of the 19th Neural Information Processing Systems Conference Workshop on Testing of Deployable Learning and Decision Systems*, 2006.

[41] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Burlington, Mass, USA, 3rd edition, 2011.

[42] R. T. Clemen, *Making Hard Decisions: An Introduction to Decision Analysis*, Duxbury Press, Pacific Grove, Calif, USA, 1996.

[43] S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, no. 2684, pp. 677–680, 1946.

[44] S. Krig, *Computer Vision Metrics, Survey, Taxonomy, and Analysis*, Apress, Oxford, UK, 2014.

[45] K. H. Zou, W. M. Wells III, M. R. Kaus, R. Kikinis, F. A. Jolesz, and S. K. Warfield, "Statistical validation of automated probabilistic segmentation against composite latent expert ground truth in MR imaging of brain tumors," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2002*, T. Dohi and R. Kikinis, Eds., vol. 2488 of *Lecture Notes in Computer Science*, pp. 315–322, Springer, Berlin, Germany, 2002.

[46] E. Earl and L. M. Liebrock, "Efficaceous end user measures for discriminant analysis tools," Tech. Rep., Institute of Mining and Technology, Socorro, NM, USA, 2015.

[47] D. J. Hand, *Measurement Theory and Practice: The World Through Quantification*, Oxford University Press, New York, NY, USA, 2004.

[48] T. Verbraken, W. Verbeke, and B. Baesens, "A novel profit maximizing metric for measuring classification performance of customer churn prediction models," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 5, pp. 961–973, 2013.

[49] I. K. Crombie and H. T. O. Davies, *What is Meta-Analysis? "What is ... ?" Series*, NPR09/1112, Hayward Medical Communications, London, UK, 2009.

[50] C. Lajas, L. Abasolo, B. Bellajdel et al., "Costs and predictors of costs in rheumatoid arthritis: a prevalence-based study," *Arthritis Care & Research*, vol. 49, no. 1, pp. 64–70, 2003.