A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives

PETER CHRISTEN, The Australian National University, Australia
DAVID J. HAND, Imperial College London, UK
NISHADI KIRIELLE, The Australian National University, Australia

Methods to classify objects into two or more classes are at the core of various disciplines. When a set of objects with their true classes is available, a supervised classifier can be trained and employed to decide if, for example, a new patient has cancer or not. The choice of performance measure is critical in deciding which supervised method to use in any particular classification problem. Different measures can lead to very different choices, so the measure should match the objectives. Many performance measures have been developed, and one of them is the F-measure, the harmonic mean of precision and recall. Originally proposed in information retrieval, the F-measure has gained increasing interest in the context of classification. However, the rationale underlying this measure appears weak, and unlike other measures, it does not have a representational meaning. The use of the harmonic mean also has little theoretical justification. The F-measure also stresses one class, which seems inappropriate for general classification problems. We provide a history of the F-measure and its use in computational disciplines, describe its properties, and discuss criticism about the F-Measure. We conclude with alternatives to the F-measure, and recommendations of how to use it effectively.

CCS Concepts: • **Computing methodologies** → **Supervised learning by classification**; • **General and reference** → **Measurement**; **Evaluation**; • **Information systems** → Retrieval effectiveness;

Additional Key Words and Phrases: Supervised classification, performance assessment, F1-score, F1-measure, F*-measure, representational measure, pragmatic measure

ACM Reference format:
Peter Christen, David J. Hand, and Nishadi Kirielle. 2023. A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives. *ACM Comput. Surv.* 56, 3, Article 73 (October 2023), 24 pages.
https://doi.org/10.1145/3606367

1 INTRODUCTION

The central challenge of various computational disciplines is the construction of algorithms that automatically improve with experience [32]. An important class of such algorithms is concerned with supervised classification, in which an algorithm is presented with a training set of objects, each of which has a known descriptive vector of measurements, and each of which has a known class membership. The aim is to use this training set to create a *classification method* (also known as a classifier, classification model, classification algorithm, or classification rule) that can classify

Authors' addresses: P. Christen and N. Kirielle, School of Computing, The Australian National University, North Rd, Canberra ACT 2600, Australia; emails: {peter.christen, nishadi.kirielle}@anu.edu.au; D. J. Hand, Department of Mathematics, Imperial College, London SW7 2AZ, United Kingdom; email: d.j.hand@imperial.ac.uk.

0360-0300/2023/10-ART73 $15.00
https://doi.org/10.1145/3606367

future objects based solely on their descriptive feature vectors. In what follows, for convenience, we will assume just two classes of objects (binary classification), labelled 0 and 1 (commonly named the negative and positive class, respectively).

A general high-level strategy for constructing classification methods is to begin by assigning scores to the objects to be classified. In particular, with two class problems, we might score the objects by their estimated probability of belonging to class 1, say. Given such classification scores, the appropriate way to proceed will depend on the problem. If the set of objects to be classified is known and finite (for example, if we wished to classify a given set of patients as having cancer or not), then all objects can be ranked and the top x or the top $y\%$ are assigned to class 1 (having cancer). However, if the set of objects is not fully known, or is of arbitrary and unknown size (for example classifying future patients to have cancer or not), then it is not possible to rank them all and choose the ones with the highest scores.

In such a case, a classification threshold must be chosen, so that those with scores larger than the threshold are assigned to class 1. This threshold could be chosen so as to classify a given percentage to class 1 (based on a known or estimated distribution of scores). Alternatively, it could be chosen in absolute terms: all objects with estimated class 1 probability greater than a threshold of 0.9 could be assigned to class 1, for example. Note an important distinction between these two cases: if the top-scoring $y\%$ are to be assigned to class 1, then whether or not a particular object is assigned to class 1 depends on the scores of other objects. In contrast, if an absolute threshold is used, then objects can be assigned independently of the scores of other objects.

Central to the challenge of creating classification methods is the ability to evaluate them. This is needed so that one can decide if a method is good enough for some purpose, and to choose between competing methods (which possibly have been created using different algorithms, or the same algorithm but different parameter settings). "Choosing between" methods includes notions of algorithm selection, parameter estimation, choice of descriptive features, transformations of features, and so on, as these implicitly select between alternative methods.

Criteria for evaluating classification methods and algorithms are of two kinds: problem-based and "accuracy"-based. Problem-based measures are specific to particular applications, and include aspects such as speed of classification (such as for data streams), speed of adaptation and updating (for example, in spam and fraud detection), ability to handle incomplete data, interpretability (mandatory in certain legal frameworks), how easy they are for a non-expert to use, how effective they are in the hands of an expert, and so on.

"Accuracy"-based measures, on the other hand, are concerned with the accuracy with which objects are assigned to the correct class. In two-class problems, there are two ways in which such assignments may be in error: class 0 objects may be incorrectly assigned to class 1 (known as *false positives* or *Type I* errors), and class 1 objects may be incorrectly assigned to class 0 (*false negatives* or *Type II* errors). To produce a one-dimensional performance measure, which can be used to compare classification methods, the extent of these two types of error must be aggregated into a single value. This can be done in various ways—hence the quotation marks around "accuracy" above when describing this class of criteria: accuracy may mean different things. Various reviews which look at different aspects of classification performance measures have been written, including [11, 15, 19, 23, 27, 50, 57], and others.

In this article, we review and examine one particular measure in detail: the F-measure, which is generally calculated as the harmonic mean of precision and recall. The F-measure was originally proposed in the domain of information retrieval [59] to evaluate the quality of ranked documents retrieved by a search engine. In this discipline, only one class is of real interest (the relevant documents; without loss of generality, call it the positive class), and we are concerned with the proportion of that class misclassified and the proportion classified into this class which are, in fact, from

the other class. These two aspects are captured by recall and precision, respectively. In particular, we are not concerned with the number or proportion correctly classified from the class of no interest. The F-measure is then a way of combining recall and precision, to yield a single number through which classification methods can be assessed and compared.

The appropriateness of the F-measure is illustrated by considering very imbalanced situations [33] in which the uninteresting class is very large (such as all individuals who do not have cancer). In such cases, measures such as error rate will be heavily influenced by a large number of irrelevant correct classifications to the uninteresting class, and therefore can be poor indicators of the usefulness of the classification method used [7, 46, 51, 52]. However, for more general classification problems, misclassifications (and correct classifications) of both classes are of concern. This means that the F-measure is an inappropriate measure of performance for these situations, as we discuss further in Section 4.1.

Apart from the pragmatic benefit of reducing the two measures, recall and precision, to a single number, there does not seem to be a proper justification for why the F-measure is appropriate for evaluating general supervised classification methods (where both classes are of interest). In many publications, no reasons are provided why the F-measure is used for an evaluation of classification methods, or alternatively rationales such as "because it is a commonly used measure in our research discipline", or "because it has been used by previous relevant work" are given. Others have discussed problematic issues with the F-measure earlier [42, 51, 60].

Contributions and outline: In the following section, we trace the use of the F-measure back to its original development in information retrieval, and its increasing use over time in diverse computational disciplines. In Section 3, we then describe the properties of the F-measure, including its generalisation, the F_β measure. From this discussion, we see that the F-measure has characteristics that can be seen as conceptual weaknesses. We discuss these and the resulting criticism of the F-measure in Section 4. In Section 5, we then describe alternatives to the F-measure, and we conclude our work in Section 6 with a discussion and recommendations on how to use the F-measure in an appropriate way.

We have previously explored the shortcomings of the F-measure in the context of record linkage [29], the task of identifying records that refer to the same entities across databases [5, 13]. More recently, we proposed a variant of the F-measure that overcomes some of its weaknesses [30]. Here, we go further and provide a broader discussion of the use of the F-measure and its application in general classification problems. Our work is targeted at a broad audience, where we aim at providing the reader with a deeper understanding of the F-measure. Our objective is also to highlight that using a certain performance measure simply because it is being commonly used in a research community might not lead to an appropriate evaluation of classification methods.

2 A SHORT HISTORY OF THE F-MEASURE

To better understand the ideas and motivation behind the F-measure as developed in the domain of information retrieval, and how it then started to be used in other computational disciplines, we next describe the development of the F-measure, then discuss its use in various disciplines, and end this section with a bibliographic analysis.

2.1 The Development of the F-Measure

The F-measure can be traced to the book "Information Retrieval" by Van Rijsbergen [59] published in 1979. In information retrieval, a document collection is queried using a search term, and a set (usually of fixed size) of documents is returned by a retrieval algorithm, generally in a ranked order based on some measure of relevance [44]. To evaluate the quality of a set of retrieved documents, one needs to know which documents are relevant to a given query (this is similar to having ground

truth data in classification). It is then possible to identify from the set of retrieved documents those that are relevant. The measures of precision and recall are then defined as Reference [44]:

— Precision, P, is the fraction of retrieved documents that are relevant.
— Recall, R, is the fraction of relevant documents that are retrieved.

For classification tasks, as we will discuss in Section 3, precision and recall can be defined similarly. A good retrieval system ideally achieves high precision as well as high recall on a wide range of queries [44].

Van Rijsbergen [59], in chapter 6 (which became chapter 7 in the second edition of his book), extensively discusses the reasoning why precision and recall are useful measures for information retrieval, and the desire to combine them into a single composite number. He also describes various previous attempts, some going back to the early 1960s, to develop single number performance measures for information retrieval. He then develops a measure of retrieval effectiveness, E, defined as

$$E = 1 - \frac{1}{\frac{\alpha}{P} + \frac{(1-\alpha)}{R}}, \tag{1}$$

where $\alpha = 1/(\beta^2 + 1)$, and β is a factor by how much more importance a user gives to recall over precision. Setting $\alpha = 0$ ($\beta = \infty$) corresponds to giving no importance to precision, setting $\alpha = 1$ ($\beta = 0$) corresponds to giving no importance to recall, and setting $\alpha = 1/2$ ($\beta = 1$) corresponds to giving the same importance to precision and recall.

As we discuss in more detail in Section 3, the F-measure, or more precisely its weighted generalisation, F_β, corresponds to $F_\beta = 1 - E$ (with α substituted by β), defined as

$$F_\beta = \frac{(\beta^2 + 1)P \cdot R}{\beta^2 P + R}. \tag{2}$$

For $\beta = 1$ this simplifies to $F_1 = 2P \cdot R/(P+R)$, the commonly used default version of the F-measure that is also known as the F_1-measure or the F_1-score [44].

Van Rijsbergen [59] also motivates the use of the harmonic mean to calculate E, and therefore F, through the principle of decreasing marginal relevance, where *at some point a user will be unwilling to sacrifice a unit of precision for an added unit of recall* [44, page 173]. We return to this argument for the use of the harmonic mean in Section 3.

2.2 The Use of the F-Measure in Different Computational Disciplines

From its initial use in information retrieval, the F-measure has subsequently been employed in other computational disciplines. The following discussion is by no means aimed to be a comprehensive analysis of how the F-measure has been used over the past four decades. Rather, our aim is to highlight its diverse use in the context of classification tasks.

Outside information retrieval, one of the first disciplines that employed the F-measure was information extraction (the process of extracting structured information from text), where Chinchor [12] in 1992 proposes to use it for evaluation at the fourth Message Understanding Conference (MUC-4), provides its general formulation according to Equation (2), and names it the F-measure. Sasaki [55] in 2007 notes: *A personal communication with David D. Lewis several years ago revealed that when the F-measure was introduced to MUC-4, the name was accidentally selected by the consequence of regarding a different F function in Van Rijsbergen's book as the definition of the 'F-measure'.*

McCarthy and Lehnert [45] (1995), in their work on using decision trees for coreference resolution (the task of identifying all mentions that refer to the same entity in a given document

or document collection) refer to both the MUC-4 work by Chinchor [12] and the book by Van Rijsbergen [59]. They provide the general formulation from Equation (2), however, besides describing that the F-measure combines precision and recall, they do not justify further why it is a suitable measure for coreference resolution.

In the context of classifying text, Lewis [41] (1995) discusses how to optimise systems when data are time-varying, *such as news feeds or electronic mail*, with the aim of measuring performance without the need to continue collecting true class labels. He suggests using the expected value of the F-measure, assuming the true class labels are binary random variables with probability equal to the estimated class 1 probability. But he then gives an example showing that no fixed probability threshold can be used to optimise the F-measure when the score distribution is unknown.

In their highly influential work on sampling for imbalanced classification problems, Chawla et al. [10] in 2002 interestingly do refer to the book by Van Rijsbergen [59] and describe precision and recall as measures used in information retrieval. However, they do not mention the F-measure, and proceed to employ ROC curves [18] for their experimental evaluation. While the F-measure is calculated for a single specific classification threshold, ROC curves are based on varying such a threshold. Neither of the books by Macmillan and Creelman [43] and Krzanowski and Hand [40], which both cover the topic of ROC curves in detail, mention the F-measure.

The 2009 survey by He and Garcia [33] on learning from imbalanced data also discusses the F-measure, however, without any references. The authors do raise the issue of the F-measure being sensitive to the class imbalance of a given problem. This is because precision considers both the positive and negative classes, while recall is only based on the positive class. Branco et al. [7], in their 2016 survey on predictive modelling of imbalanced data, refer to the book by Van Rijsbergen [59] and provide Equation (2) as the definition of the F-measure. They also state that this measure is *more informative than accuracy about the effectiveness of a classifier on predicting correctly the cases that matter to the user*, and that *the majority of the articles that use F_β for performance evaluation under imbalanced domains adopt $\beta = 1$, which corresponds to giving the same importance to precision and recall.*

Joachims [37] (2005) discusses various performance measures in the context of learning support vector machine classifiers optimised for a certain measure. He mentions the common use of the F-measure for binary classification problems in natural language processing applications such as text classification, and describes how the F-measure is preferable to error rate in imbalanced classification problems. However, he does not refer to any earlier work on the F-measure, nor does he discuss its origin in information retrieval. Similarly, Onan et al. [49] (2016) use the F-measure for their evaluation of keyword extraction methods for text classification, without any discussion why this measure is suitable for this task. In their 2021 survey on deep learning based text classification methods, Minaee et al. [48] briefly mention that the F-measure is mainly used for imbalanced classification problems. However, no references to its origin nor any discussion on why it is a suitable measure in such situations are provided.

In the context of text and Web data mining, Kosala and Blockeel [39] in their 2000 survey mention precision and recall but not the F-measure, while Han and Kamber [21], in the first edition (2000) of their influential data mining text book, discuss precision, recall, and the F-measure in the context of text data mining. They justify the use of the harmonic mean in the F-measure as *the harmonic mean discourages a system that sacrifices one measure for another too drastically* (Section 9.4). Hand et al. [31] in their 2001 data mining book also discuss precision and recall in the context of text retrieval, but they do not mention the F-measure.

In computer vision, the F-measure is commonly used to evaluate image recognition and classification tasks. Achanta et al. [1] (2009), Epshtein et al. [17] (2010), and Brutzer et al. [8] (2011),

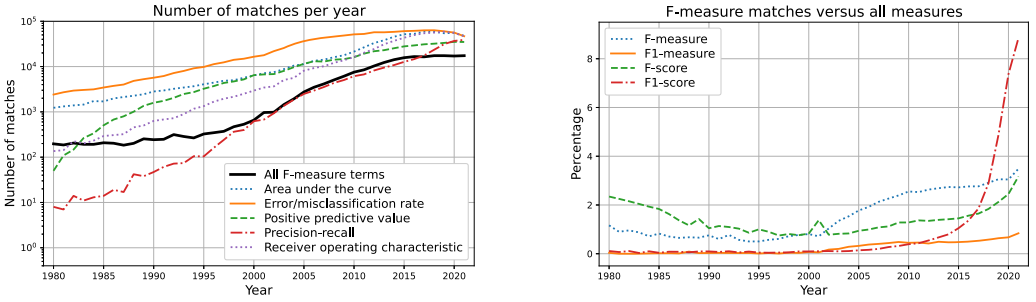

Fig. 1. Google Scholar number of matches (publications) over time for different performance measures (left-side plot), and percentage changes over time for the different variations of F-measure names over all measures from the left-side plot (right-side plot).

are three examples of influential publications that employ the F-measure for such tasks without providing any justifications or references to its origin.

In the context of evaluating diagnostic instruments in medicine, where problems with imbalanced class distributions are common, sensitivity and specificity are widely used measures [27]. Only more recently has the F-measure been employed in the health sciences, as data science based approaches are increasingly being used in this discipline. As an example, Janowczyk and Madabhushi [36] (2016) review deep learning methods for classifying digital pathology images, where they use the F-measure for evaluation without any discussion why this is a suitable measure for this task.

As this discussion has shown, while the F-measure has been used in a variety of computational disciplines in the context of evaluating classification tasks, there seems to be a common lack of justification why it is a suitable performance measure for a given problem domain. In the following we present a bibliographic analysis to better understand how the F-measure has been used over time in comparison to other performance measures.

2.3 Bibliographic Analysis of the Use of the F-Measure

To obtain an overall picture of the popularity of the F-measure, we queried Google Scholar[1] with multiple search terms for individual years of publication, and recorded the number of matches returned (generally shown as "About 123,456 results"). Given the F-measure is known under different names, we queried Google Scholar with the terms "f-measure", "f1-measure', "f-score", and "f1-score", first individually and then as a disjunctive (OR) combination of these four terms. For comparison we queried several other performance measures used in the context of classification, where we limited ourselves to multi-word measures because single-word measures (such as "accuracy", "precision', "recall", "sensitivity", and "specificity") are also used in general text rather than the strict reference to a performance measure. This would likely result in much higher counts of matches for such single-word measures. We queried Google Scholar for counts from 1980 to 2021, noting, however, that the numbers of matches for the past few years are potentially represented less completely in Google Scholar as publications are still being added to its database.

The left-side plot in Figure 1 shows these yearly numbers of matches over time, where the bold black line is for the disjunctive F-measure query. As can be seen, there is a steady increase over time for all performance measures, as is expected given the increase in scientific publications over

[1]See: http://scholar.google.com/, where our code and data of this analysis are available at: https://github.com/nishadi/f-measure.

		True class	
		0	1
Predicted class	0	*TN* (true negatives)	*FN* (false negatives)
	1	*FP* (false positives)	*TP* (true positives)

Fig. 2. Notation for confusion matrix.

time. However, since the late 1990s, the F-measure has seen a stronger increase compared to some of the other measures. Since 1980, the annual numbers of publications referring to the F-measure have increased by two magnitudes, while the numbers referring to precision-recall have increased over three magnitudes. The right-side plot in Figure 1 shows the percentage of matches for the different F-measure names over the total number of matches for all measures we queried (as listed in the left-side plot). A clear increase can be seen since the year 2000 and especially in the past ten years, after an initial drop between the years 1980 and 2000.

To summarise our bibliographic analysis, since its development in information retrieval over forty years ago, the F-measure has been used in a variety of computational disciplines, with a noticeable increase in the past twenty or so years. This coincides with the increased popularity of data based disciplines, including data mining, machine learning (especially deep learning), computer vision, and natural language processing, since the year 2000.

3 THE F-MEASURE AND ITS PROPERTIES

Let $\mathbf{x} = (x_1, \ldots, x_m)$ be the vector of m descriptive characteristics (features) of an object to be classified. From this, a classification method will compute a score $s(\mathbf{x})$ and the object will be assigned to class 0 or class 1 according as

— If $s(\mathbf{x}) > t$ assign the object to class 1.
— If $s(\mathbf{x}) \leq t$ assign the object to class 0.

Here t is the "classification threshold". Clearly by changing t we can shift the proportion of objects assigned to classes 0 and 1. The threshold is thus a control parameter of the classification method. The method's performance is then assessed by applying it to a test dataset of objects with known classifications. Here, we shall assume that the test set is independent of the training set—the reasons for this and the problems arising when it is not true are well-known and have been explored in great depth (see, for example, Hastie et al. [32]).

Application of the classification method to the test set leads to a (mis)-classification table or confusion matrix, as illustrated in Figure 2. Here, FN is the number of test set objects which belong to class 1 but which the classification method assigns to class 0 (that is, which yield a score less than or equal to t), and so on, so that the off-diagonal counts, FP and FN, give the number of test set objects which have been misclassified. This means that $(FP+FN)/n$, with $n = TP+FP+FN+TN$ the total number of objects in the test set, gives an estimate of the overall misclassification or error rate [27], another widely used measure of classification performance.

In what follows, we shall regard class 1 objects as "cases" of relevance or interest (such as exemplars of people with the disease we are trying to detect, of fraudulent credit card transactions, and so on). In terms of this table, recall, R, is defined as the proportion of true class 1 objects which are correctly assigned to class 1, and precision, P, is defined as the proportion of those objects assigned to class 1 which really come from class 1. That is

— Recall $R = TP/(TP + FN)$,
— Precision $P = TP/(TP + FP)$.

The F-measure then combines these two using their harmonic mean, to yield a univariate (single number) performance measure:

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2P \cdot R}{P + R} = \frac{2TP}{2TP + FP + FN}. \tag{3}$$

In numerical taxonomy, this measure is called the Dice coefficient or the Sørensen-Dice coefficient [2, 56, 58] and it also goes under other names. In particular, the F-measure as defined above is commonly known as the F_1-measure (or balanced F-measure), being a particular case of a more general weighted version [54], defined as

$$F_\alpha = \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}}. \tag{4}$$

This can be rewritten as Equation (2), with $\beta^2 = (1 - \alpha)/\alpha$, where with $\alpha = 1/2$ ($\beta = 1$), we get the F_1-measure, short for $F_{\beta=1}$ [44]. As we discussed in Section 2.1, the F_β-measure was derived from the effectiveness measure, E (Equation (1)), as $F_\beta = 1 - E$, developed by Van Rijsbergen [59] who writes that it *measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision* (page 123).[2] However, as we show in Section 4.3, the F-measure can be reformulated as a weighted arithmetic mean. In this reformulation, the weights assigned to precision and recall in the F_β-measure do not only depend upon β but also on the actual classification outcomes.

Precision and recall reflect different aspects of a classification method's performance, so combining them is natural. Moreover, both are proportions, and both have a representational meaning, a topic we return to in Section 4.5. Precision can be seen as an empirical estimate of the conditional probability of a correct classification given predicted class 1 ($Prob(True = 1|Pred = 1)$), and recall as an empirical estimate of the conditional probability of a correct classification given true class 1 ($Prob(Pred = 1|True = 1)$). An average of these, however, has no interpretation as a probability, and unlike many other performance measures also has no representational meaning [24].

The mean of precision and recall does not correspond to any objective feature of classification performance; it is not, for example, an empirical estimate of any probability associated with a classifier method. Formally, and as we discuss further in Section 4.5, the F-measure, as a harmonic mean, is a *pragmatic measurement* [22, 24, 34, 47]: it is a useful numerical summary but does not represent any objective feature of the classifier method being evaluated. This is in contrast with representational measures which correspond to real objective features: precision and recall separately are examples, since they correspond to empirical estimates of probabilities of certain kinds of misclassification.

There is also no straightforward justification for using the harmonic mean to combine precision and recall. A formal argument is sometimes made that for averaging rates the harmonic mean is more natural than, say, the arithmetic mean, but this is misleading. One might argue that the harmonic mean of precision and recall is equivalent to (the reciprocal of) the arithmetic mean of the number of true class 1 cases per class 1 case correctly classified, and the number of predicted class 1 cases per class 1 case correctly classified. But this simply drives home the fact that precision and recall are non-commensurate.

A different argument in favour of the F-measure has been made by Van Rijsbergen [59] using conjoint measurement. The essence of his argument is first to show that there exist non-intersecting isoeffectiveness curves in the (P, R)-space (sometimes called indifference curves:

[2] In the second edition of Van Rijsbergen's book, this quote is given on page 133.

curves showing combinations of P and R which are regarded as equally effective), then to determine the shape of these curves, and hence to decide how to combine P and R to identify which curve any particular (P, R) pair lies on. In particular, he arrives at the conclusion that the harmonic mean (weighted if necessary) determines the shapes of the curves. To explore reasonable shapes for these curves, and noting that P and R are proportions, Van Rijsbergen [59] (pages 122 and 123) makes the assumption of decreasing marginal effectiveness: the user of the system *is willing to sacrifice one unit of precision for an increase of one unit of recall, but will not sacrifice another unit of precision for a further unit increase in recall.* For P and R values near zero, this leads to isoeffectiveness curves which are convex towards the origin. Curves based on the harmonic mean of P and R have this shape.

As we noted above, one way to look at the harmonic mean is that it is the arithmetic mean on the reciprocal of the original scale. That is, it is the reciprocal of the arithmetic mean of $1/P$ and $1/R$, as can be seen in Equation (3). But the reciprocal transformation is not the only transformation of the scale which will produce isoeffectiveness curves of this shape. For example, transforming to $log(P)$ and $log(R)$ will also yield convex isoeffectiveness curves (and results in the geometric mean of P and R, which is known as the Fowlkes-Mallows index [20] of classifier performance). In short, the choice of reciprocal transformation, and hence the harmonic mean, seems arbitrary.

As typically used in numerical taxonomy [56], the F-measure has more justification. Here, it is used as a measure of similarity between two objects, counting the agreement between multiple binary characteristics of the objects. Thus, referring to Figure 2 above, TP is the number of characteristics that are present for both objects, FN is the number of characteristics that are present for object A but absent for B, and so on. Since the number of potential descriptive characteristics of objects can be made as large as you like, the number of characteristics not possessed by either object, that is count TN, should not be included in the measure. But this interpretation seems to be irrelevant in the situation when classification methods are being evaluated, as we discuss below.

4 CRITICISM OF THE F-MEASURE

Over the years, researchers have questioned various aspects of the F-measure and its suitability as an evaluation measure in the context of classification problems [30, 51, 60]. In this section, we summarise and discuss these issues.

4.1 The F-Measure Ignores True Negatives

As can be seen from Equation (3), the F-measure does not take the number of true negatives into account. In its original context in information retrieval, true negatives are the documents that are *irrelevant* to a given query and are correctly classified as irrelevant. Their number can be arbitrarily large (with the actual number even unknown). When comparing the effectiveness of retrieval systems, adding more correctly classified irrelevant documents to a collection should not influence the value of the used evaluation measure. This is the case with precision, recall, and the F-measure [51].

In the context of classification, however, the number of objects in the negative class is rarely irrelevant. Consider a classification method trained on a database of personal health records where some patients are known to have cancer (the class of interest and hopefully also the minority class). While the classification of positives (possible cancer cases for which patients should be offered a test or treatment) is clearly the focus, how many non-cancer patients are correctly classified as not having the disease is also of high importance for these individuals [42]. Therefore, the F-measure would not really be a suitable evaluation measure for such a classification problem.

We illustrate this issue in Figure 3(a) and (b), where the two shown matrices have different counts but yield the same F-measure. Matrix (a) shows nearly 86% (600 out of 700) negative objects

(a)

		True class: 0	True class: 1	
Predicted class	0	$TN = 600$	$FN = 50$	$R = 0.833$
	1	$FP = 100$	$TP = 250$	$F = \mathbf{0.769}$
				$P = 0.714$

(b)

		True class: 0	True class: 1	
Predicted class	0	$TN = 688$	$FN = 105$	$R = 0.650$
	1	$FP = 12$	$TP = 195$	$F = \mathbf{0.769}$
				$P = 0.942$

(c)

		True class: 0	True class: 1	
Predicted class	0	$TN = 552$	$FN = 20$	$R = 0.933$
	1	$FP = 148$	$TP = 280$	$F = \mathbf{0.769}$
				$P = 0.654$

Fig. 3. Three confusion matrices with different numbers of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), all yielding the same F-measure result. All three matrices cover 1,000 objects, with 300 in class 1 and 700 in class 0.

(class 0) correctly classified, while in matrix (b) over 98% of them (688 out of 700) are correctly classified. Note that, however, the classifier that resulted in matrix (a) was able to classify more positive objects (class 1) correctly compared to the classifier that resulted in matrix (b).

It is, therefore, important to understand that the F-measure is only a suitable measure for classification problems when the negative (generally the majority) class is not of interest at all for a given problem or application [51].

4.2 The Same F-Measure can be Obtained for Different Pairs of Precision and Recall

A common aspect of all performance measures that combine the numbers in a confusion matrix into a single value is that different counts of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), can result in the same value for a certain measure. This is the rationale behind isoeffectiveness curves [59]. For example, for a given $n = TP + FP + FN + TN$, any pair of TP and TN that sum to the same value will lead to the same accuracy result.

Specifically for the F-measure, from the right-hand side of Equation (3), we can see that for any pair of triplets (TP_a, FP_a, FN_a) and (TP_b, FP_b, FN_b) arising from classifiers applied to the same dataset (so that $TP + FN = n_1$) any pair (TP, FP) satisfying $FP = k \cdot TP - n_1$, will yield the same F-measure, even though precision and recall may differ between the classifiers. Here k is a constant, where all three matrices in Figure 3 have $k = 1.6$. This means that potentially classification methods that achieve very different results when evaluated using precision and recall will provide the same F-measure result.

An example can be seen in Figure 3(b) and (c), where the confusion matrix (b) results in $P = 0.942$ and $R = 0.650$, matrix (c) results in $P = 0.654$ and $R = 0.933$, while for both these matrices $F = 0.769$.

The F-measure should, therefore, never be reported in isolation without also reporting precision and recall results. In situations where a single measure is evaluated, such as for hyperparameter tuning for automated machine learning [38], using the F-measure can be dangerous because the performances of classification methods are being compared in a way that potentially can provide very different outcomes (of course, this is true whenever measures are summarised into a single number).

4.3 The Weights Assigned to Precision and Recall Depend Not Only on Alpha (or Beta)

As we discussed in Section 3, a generalised version of the F-measure allows assigning weights to precision and recall using the parameter α, see Equation (4), or equivalently β, see Equation (2).

In an effort to understand the use of the F-measure in the context of record linkage (also known as entity resolution) [5, 13], the process of identifying records that refer to the same entities within or across databases, Hand and Christen [29] showed that the harmonic mean representation of the F-measure can be reformulated as a weighted arithmetic mean of precision and recall as $F = pR + (1-p)P$, where $p = (TP+FN)/(2TP+FP+FN) = P/(R+P)$. In this weighted arithmetic mean reformulation, however, the value of the weight p given to recall depends upon the outcome of the evaluated classification method. When several classification methods are compared, the weight

p assigned to recall can be different if the numbers of false positives and false negatives obtained by these methods differ. From the example confusion matrices in Figure 3, we can calculate $p = 0.462$ for matrix (a), $p = 0.592$ for matrix (b), and $p = 0.412$ for matrix (c).

As a result, in this weighted arithmetic mean reformulation of the F-measure, the weights assigned to precision and recall do not only depend upon the values of α or β, but also upon the actual classification outcomes. We describe this property of the F-measure, including an extension of the work by Hand and Christen [29] for the generalised F_β-measure from Equation (2), in more detail in Appendix A.

4.4 The F-Measure has an Asymmetric Behaviour for Varying Classification Thresholds

In Section 3, we have discussed how a specific confusion matrix for a given classification method can be obtained by setting a "classification threshold" t to a certain value. For a given classification problem (a set of objects to be classified), by modifying this threshold the individual counts of TP, FP, FN, and TN will likely change, while their total number $n = TP + FP + FN + TN$, and the numbers of actual positive (class 1) objects, $n_1 = TP + FN$, and negative (class 0) objects, $n_0 = TN + FP$ (with $n = n_0 + n_1$), are fixed. Generally, lowering the threshold t means more objects are being classified as positives, with the numbers of TP and FP increasing and the numbers of TN and FN decreasing. Conversely, increasing t generally results in more objects to be classified as negatives, with the numbers of TP and FP decreasing and the numbers of TN and FN increasing.

Therefore, as we lower the classification threshold t, recall (R) either stays the same (no new objects in class 1 have been classified to be in class 1 with a lower t) or it increases (more objects in class 1 have been classified to be in class 1 with a lower t). Recall, however, can never decrease as t gets lower.

Precision (P), on the other hand, can increase, stay the same, or decrease, both when the classification threshold t is increased or decreased. A change in the value of precision depends upon the distributions of the scores of objects in the two classes, as well as the class imbalance. For example, for some decrease of the threshold t more class 1 objects might be newly classified as being in class 1 compared to class 0 objects, while for another decrease of t more class 0 objects might be newly classified as being in class 1 compared to class 0 objects. With large class imbalances, where $n_1 < n_0$, precision generally decreases as t gets lower because more class 0 objects are classified to be in class 1 (as false positives) compared to class 1 objects (as true positives). We show how precision changes for real datasets in Appendix B.

If we assume the scale of scores, $s(\mathbf{x})$, assigned to objects is standardised into the range 0 to 1, we can accordingly set the threshold $0 \leq t \leq 1$. Further assuming that in the extreme case $t = 0$ all objects are classified as positives (classified as class 1) and in the case $t = 1$ all are classified as negatives (classified as class 0), then we will have the following:

- If $t = 0$ then $TP = n_1$, $FP = n_0$, $TN = 0$, and $FN = 0$, and therefore $P = n_1/(n_1 + n_0) = n_1/n$ and $R = n_1/n_1 = 1$.
- If $t = 1$ then $TP = 0$, $FP = 0$, $TN = n_0$, and $FN = n_1$, and therefore $P = 0$ (for convenience we define that $P = 0/0 = 0$)

With $t = 0$ precision, therefore, becomes $P = n_1/n = 1/(ci + 1)$, a ratio which depends upon the *class imbalance* of the given classification problem, $ci = n_0/n_1$. Here, we assume that $n_0 \geq n_1$ and therefore $ci \geq 1$ (the negative class, 0, is the majority class and the positive class, 1, the minority class).

For a balanced classification problem with $ci = 1$, for $t = 0$ we obtain $P = 1/2$, $R = 1$, and therefore $F = 2/3$. For an imbalanced problem where, for example, 20% of all objects are positive and 80% are negative ($ci = 4$), for $t = 0$ we obtain $P = 1/5$, $R = 1$, and therefore $F = 1/3$. For $t = 1$, for both problems, we obtain $F = 0$ because $TP = 0$.

4.5 The F-Measure is not a Representational Measure

Performance measures can be categorised into *representational* and *pragmatic* measures [22, 24, 34, 47]. Measures in the former category quantify some property of the attributes of real objects, while measures in the latter category assign some numerical value to objects where these values may not represent any attributes of these objects. Examples of representational measures are the height and weight of people, while a pragmatic measure would be their university GPA (grade point average) scores. GPA is a construct, without objective empirical existence.

Unlike precision and recall, which are both representational measures, the harmonic mean formulation of the F-measure is a pragmatic measure. It is a useful numerical summary but it does not represent any objective feature of the classification method being evaluated. In the quest to develop an intuitive interpretable transformation of the F-measure, Hand et al. [30] recently proposed the F^* (F-star) measure, which we describe in the following section.

A criticism raised by Powers [51] is that averaging F-measure results is nonsensical because it is not a representational measure. Averaging the results of classification methods is commonly conducted in experimental studies. For example, to identify the best performing method from a set of classification methods, for each method, its results obtained on different datasets can be summarised using the arithmetic mean and standard deviation. Because recall and precision are proportions, averaging over multiple precision or recall results, respectively, will yield a value that is also a proportion and therefore has a meaning (the average recall or average precision). However, given the F-measure is a pragmatic measure, averaging several F-measure results is akin to *comparing apples with pears* [51].

5 ALTERNATIVES TO THE F-MEASURE

Based on the properties we discussed in the previous section, researchers have been investigating alternatives to the use of the F-measure in the context of evaluation of classification problems.

5.1 The F* (F-star) Measure

In an attempt to provide an interpretable combination of precision and recall, Hand et al. [30] have proposed the F^*-measure (F-star), defined as:

$$F^* = \frac{F}{2 - F} = \frac{P \cdot R}{P + R - P \cdot R} = \frac{TP}{TP + FP + FN}. \tag{5}$$

It is an empirical estimate of the probability that an object will belong to the intersection of the class 1 objects and the objects predicted to be class 1. Clearly, the bigger this intersection, the better the classifier. It can be interpreted as: *F^* is the proportion of the relevant classifications which are correct, where a relevant classification is one which is either really class 1 or classified as class 1* [30].

Researchers may recognise the F^*-measure as the Jaccard coefficient [35], as widely used in domains where true negatives may not be relevant, such as numerical taxonomy and fraud analytics [3, 56]. They may also recognise it as the intersection over union statistic, as used, for example, in image recognition [53].

The F^*-measure is a monotonic transformation of the F-measure [30], which means that any conclusions reached by assessing which F^* value is larger will be the same as when assessing which value of F is larger. As a result, selecting the "best" among a set of classification methods

based on the highest value of F will give the same as when selecting using F^*. It should be noted, however, that the F^*-measure is no longer an average of precision and recall, because it holds that $F^* \leq min(P, R)$.

5.2 The Area under the Precision-Recall Curve

In all our discussions so far, we assumed that a single confusion matrix has been used, as generated from the evaluation of a specific classification method and a certain classification threshold t. However, in practice, it is often not possible to identify a suitable value for t, and instead, the performance of a classification method is evaluated for a range of values for t. As we discussed in Section 4.4, for different values of t there will likely be different values of TP, FP, FN, and TN. For a sequence of classification thresholds in the range $0 \leq t \leq 1$ (assuming classification scores also in the range $0 \leq s \leq 1$), connecting the corresponding pairs of precision and recall values will result in a precision-recall curve [44]. We show examples of such curves in Appendix B.

A precision-recall plot shows the detailed performance of one or more classification methods and allows selection of a suitable method. An optimal classifier, correctly classifying all objects, would go through the (1,1) top-right corner of such a plot. If the curve for one classification method is always below the curve of another method then this indicates a consistent lower performance of the former method compared to the latter in the precision-recall space across all values of the classification threshold t. More commonly, however, are situations where one curve is above another for a certain range of t but below for another range.

To summarise the performance of a method over the full range of the classification threshold into a single number, the area under the precision-recall curve (AUC-PR) can be calculated [6, 14, 44]. This measure averages the performance of a classification method, with an AUC-PR value of 1 indicating perfect classification (no false positives and no false negatives for any value of t). The AUC-PR has a deep connection with the area under the Receiver Operator Characteristic (ROC) curve (AUC-ROC) [14], where the latter was shown to have a fundamental conceptual weakness (of being equivalent to taking average performance over some distribution, where the distribution varies from classification method to classification method) and should never be used to compare classification methods [26, 28]. Further research is needed to investigate if the AUC-PR measure exhibits this weakness as well.

6 DISCUSSION AND RECOMMENDATIONS

Duin [16] and Hand [23, 25] have pointed out some of the realities of evaluating the performance of classification methods. In addition to the problem-based aspects mentioned in Section 1, they include the fact that empirical evaluation has to be on some datasets, and these datasets may not be similar to the data that the classification method is being applied to. Moreover, there are many different kinds of accuracy measures: ideally a measure should be chosen which reflects the objectives. So, for example, we might wish to minimise the overall proportion of objects which are misclassified (perhaps leading to misclassifying all cases from the smaller class if the class sizes are very unequal), to minimise a cost-weighted overall loss, to minimise the proportion of class 1 objects misclassified subject to misclassifying no more than a certain percentage of the other class (for example, 5%), to minimise the proportion of objects classified as class 1, which, in fact, come from class 0 (the false discovery rate), and so on, effectively endlessly.

While the fact that there are only four counts in a confusion matrix (and, indeed, summing to a fixed total) means that these measures are related, it does not determine which measure is suitable for a particular problem. But the choice can be critical. Benton [4, Figure 4.1] gave a real-data example in which optimising two very widely used measures of performance led to linear

combinations of the variables which were almost orthogonal: the "best" classification methods obtained using these two measures could hardly have been more different. Things are complicated yet further by the need to choose a threshold to yield a confusion matrix or misclassification table. This has led to measures such as the AUC-PR [14], the area under the ROC curve [18, 40] (with its known conceptual weakness [28]) and the H-measure [26] (which we discuss in Appendix C). All these measures average over a distribution of possible thresholds. And yet further issues are described by Hand [25].

The implication of all of this is that just as much thought should be given to how classifier performance is to be measured as to the choice of what classification method(s) is/are to be employed for a given classification problem. Software tools which provide a wide variety of classification methods and their easy use are now readily available, but far less emphasis has been placed on the choice of measure of classification performance. And yet a classification method which appears to be good under one measure may be poor under another. The critical issue is to match the measure to the objectives of a given classification problem.

The F-measure, as widely used, is based on an ad hoc notion of combining two aspects of classifier performance, precision and recall, using the harmonic mean. This results in a pragmatic measure that has a poor theoretical base: it seems not to correspond to any fundamental aspect of classifier performance. With the aim at helping researchers improve the evaluation of classification methods, we conclude our work with a set of recommendations of how to use the F-measure in an appropriate way:

- The first aspect to consider is if the F-measure is really a suitable performance measure for a given classification problem. Specifically, is there clear evidence that incorrect classifications of one of the classes are irrelevant to the problem? Only if this question can be answered affirmatively should the F-measure be considered. Otherwise, a performance measure that considers both classes should be employed.
- As we discussed in Section 4.2, different pairs of precision and recall values can yield the same F-measure result. It is, therefore, important to not only report the F-measure but also precision and recall when evaluating classification methods [42]. Only when assessing and comparing the values of all three measures can a valid picture of the comparative performance of classification methods be made.
- If a researcher prefers an interpretable measure, then the F^*-measure [30] discussed in Section 5.1, a monotonic transformation of the F-measure, can be used.
- If a researcher knows what importance they want to assign to precision and recall, the general weighted F_β or F_α versions from Equations (2) or (4), respectively, can be used. For the weighted arithmetic mean reformulation of the F-measure we discussed in Section 4.3 (and also Appendix A), we recommend to specify the weight p assigned to recall in Equations (7) and (11), and correspondingly set the individual classification thresholds, t, for each classification method being compared such that the required same number $TP + FP$ of objects are classified to be in class 1. Alternatively, as we discuss in Appendix A.2, a researcher can set the weight for recall as w and for precision as $(1 - w)$, and instead of the F-measure explicitly calculate the weighted arithmetic mean of precision and recall, $wR + (1 - w)P$, for all classification methods being compared.
- If it is not possible to specify a specific classification threshold, then we recommend using a measure such as the H-measure [26, 28], which averages the performance of classification methods over a distribution of threshold values, t, as we discuss in Appendix C. Alternatively, precision-recall curves should be provided which illustrate the performance of a classification method over a range of values for the threshold t.

The critical issue for any classification problem is to decide what aspects of classification performance matter, and to then select an evaluation measure which reflects those aspects.

APPENDICES

A AN ARITHMETIC MEAN FORMULATION OF THE F-MEASURE

As we discussed in Section 3, given that the justification for the harmonic mean interpretation of the F-measure seems weak, perhaps a more intuitive strategy would be to use the arithmetic mean. Fortunately, it is possible to interpret the F-measure in this way, however, doing so reveals a property of the measure which might be regarded as a weakness.

This appendix builds on our earlier work [29], where we explored the weaknesses of the F-measure in the context of record linkage (also known as entity resolution) [5, 13]. Here, we extend this work to general classification problems and also cover the more general weighted version of the F_β measure.

A.1 Reformulating the F-Measure

As we discussed in Section 4.3, the standard definition of the F-measure as a harmonic mean can be rewritten as

$$\begin{aligned} F &= \frac{2}{\frac{1}{P}+\frac{1}{R}} = \frac{2TP}{2TP+FP+FN} \\ &= \frac{TP+FN}{2TP+FP+FN}\cdot\frac{TP}{TP+FN} + \frac{TP+FP}{2TP+FP+FN}\cdot\frac{TP}{TP+FP} \\ &= pR + (1-p)P, \end{aligned} \tag{6}$$

where:

$$p = \frac{TP+FN}{2TP+FP+FN} = \frac{P}{R+P}. \tag{7}$$

This is a weighted arithmetic mean, with recall, R, being weighted by p and precision, P, by $(1-p)$. It leads to an immediate and natural interpretation: in using the F-measure, we are implicitly saying that we wish to use a weighted sum of recall and precision, with relative importance p and $(1-p)$, respectively. Such weighted sums, with the weights representing importance, are very familiar. One interpretation is that R is regarded as $p/(1-p)$ times as important as P, since an increase of ε in R has the same impact on the overall F as an increase of $\varepsilon p/(1-p)$ in P.

However, this reformulation also reveals something more unsettling. This is that p depends on FP, FN, and TP, which means that the weights assigned to precision and recall depend on the output of the classification method being evaluated. This is inappropriate. If precision and recall are regarded as the relevant descriptive characteristics of a method's performance, then the relative importance accorded to them must depend on the problem and aims, not on the particular method being evaluated. We would not say "I regard recall as twice as important as precision if a neural network is used, but only half as important if a random forest is used." In other contexts, we have likened this sort of situation to using an elastic ruler [26], it contravenes a fundamental principle of performance evaluation.

For the general weighted version of the F-measure, F_β, as per Equation (2), we can similarly calculate:

$$F_\beta = \frac{(\beta^2+1)P\cdot R}{\beta^2 P+R} = \frac{(\beta^2+1)TP}{(\beta^2+1)TP+FP+\beta^2 FN} = p_\beta R + (1-p_\beta)P, \tag{8}$$

where:

$$p_\beta = \frac{\beta^2(TP+FN)}{(\beta^2+1)TP+FP+\beta^2 FN}. \tag{9}$$

Small FN / Large FP		Large FN / Small FP		Random	
$TN = 30$	$FN = 5$	$TN = 45$	$FN = 20$	$TN = 25$	$FN = 25$
$FP = 20$	$TP = 45$	$FP = 5$	$TP = 30$	$FP = 25$	$TP = 25$
$p = 0.435$		$p = 0.588$		$p = 0.5$	
$p_{\beta=2} = 0.755$		$p_{\beta=2} = 0.851$		$p_{\beta=2} = 0.8$	
$p_{\beta=1/2} = 0.161$		$p_{\beta=1/2} = 0.263$		$p_{\beta=1/2} = 0.2$	

Small FN / Large FP		Large FN / Small FP		Random	
$TN = 62$	$FN = 1$	$TN = 77$	$FN = 6$	$TN = 40$	$FN = 10$
$FP = 18$	$TP = 19$	$FP = 3$	$TP = 14$	$FP = 40$	$TP = 10$
$p = 0.351$		$p = 0.541$		$p = 0.286$	
$p_{\beta=2} = 0.684$		$p_{\beta=2} = 0.825$		$p_{\beta=2} = 0.615$	
$p_{\beta=1/2} = 0.119$		$p_{\beta=1/2} = 0.227$		$p_{\beta=1/2} = 0.091$	

Fig. 4. Weights, p, assigned to recall for balanced ($ci = 1$, top row) and imbalanced ($ci = 4$, bottom row) classes for three different classification outcomes assuming 100 data objects (with 50 in class 1 and 50 in class 0 for the balanced classes in the top row, and 20 in class 1 and 80 in class 0 for the imbalanced classes in the bottom row), for $\beta = 1$, $\beta = 1/2$, and $\beta = 2$.

Generally, β is seen as a way to assign weights (importance) to recall and precision [54, 59], where values $\beta < 1$ are seen to give higher weight to precision and $\beta > 1$ give higher weight to recall [44]. The reformulation in Equations (8) and (9), however, shows that the weights assigned to precision and recall not only depend on β but also upon the outcome of a classification method with regard to the values of FP, FN, and TP, when given the arithmetic mean interpretation.

From Equation (7) we can see that if $FN = FP$, then $p = 1/2$ because the equation can be rewritten as:

$$p = \frac{TP + FN}{(TP + FP) + (TP + FN)}. \tag{10}$$

From this one can immediately see that $p < 1/2$ if $FP > FN$ and $p > 1/2$ if $FP < FN$. As a result, for the special (and most commonly used) case $\beta = 1$, the F_1-measure or F_1-score, with the arithmetic mean formulation of the F-measure, the weights given to recall and precision directly depend upon the ratio of the number of false positives (FP) to the number of false negatives (FN). If there are more false positives than false negatives, then more weight is given to precision than recall ($p < 1/2$); conversely if there are more false negatives than false positives, then more weight is given to recall than precision ($p > 1/2$). Note that this holds independent of any class imbalance and is only affected by the values of FP and FN in the confusion matrix.

In Figure 4, we illustrate this issue with several example confusion matrices for both a balanced and imbalanced class distribution ci, for $\beta = 1$ (p), $\beta = 2$ ($p_{\beta=2}$), and $\beta = 1/2$ ($p_{\beta=1/2}$) for three situations with different numbers of false positives and false negatives. As can clearly be seen, depending upon the outcomes of a classification method, different weights are assigned to recall, and therefore to precision, when the arithmetic mean formulation of the F-measure is being used. The values of p obtained on the datasets and classifiers we used in our experimental evaluation in Appendix B are shown in Figure 6 for varying classification thresholds.

Below we explore the consequences of this problem of different weights being assigned to precision and recall when the arithmetic mean formulation of the F-measure is used, and how to resolve it. First, we examine some other properties of the F-measure.

It is straightforward to see that the harmonic mean of two positive values lies closer to the smaller than the larger (if the two positive values are different the harmonic mean is less than the arithmetic mean). This means that the smaller of precision and recall dominate the F-measure. An

alternative, even more extreme, measure of this kind would be $min(P, R)$. The value P can vary between 0 (if $TP = 0$ and $FP \neq 0$) and 1 (if $FP = 0$ while $TP \neq 0$). However, when $R = 1$, then P lies between the proportion of objects belonging to class 1 and 1. The value of R can also vary between 0 and 1, taking the value 0 if $TP = 0$ and the value 1 if $FN = 0$ (assuming $TP + FN \neq 0$ of course).

By the property of the harmonic mean noted above, this means that if $R = 1$ then the value for F is close to the one of precision, P. This means that the F-measure is asymmetric with regard to the extreme values of P and R, as we have also discussed in Section 4.4.

A.2 Modifying the F-Measure

In general, the relative weight to give to precision and recall should depend on the problem and the researcher's aims; that is, on which they regard as the more critical in a particular context, and just how much more critical one is than the other. Let w be the weight the researcher considers appropriate for R and $(1 - w)$ the weight for P (so that the relative weight given to R relative to P is $w/(1 - w)$. This yields the overall measure $wR + (1 - w)P$ when combining recall and precision using the arithmetic mean.

Choosing weights one considers appropriate can be difficult. It is possible to develop betting approaches, similar to those for the elucidation of priors in Bayesian statistics, but such methods are not straightforward and they are not without their weaknesses. Similar problems arise if, instead of precision and recall, one summarises the misclassification table in Figure 2 using the proportions of class 0 and class 1 correctly classified, as is typically done in medical applications, or using precision and the proportion of the overall population classified as class 1, as is often done in consumer credit applications. However one looks at it, determining suitable weights can be difficult.

For this reason, as well as encouraging researchers to think about what weights might be appropriate, we also recommend using a conventional standard. In particular, we recommend regarding precision and recall as equally important and using $w = 1/2$.

A.3 Adapting the F-Measure

Adapting the F-measure so that the arithmetic mean interpretation allows meaningful comparison of classification methods requires choosing the classification threshold so that recall receives the same weight, p, for all methods being compared, as we discuss next.

A.3.1 Choosing Thresholds to Permit Legitimate Comparisons. So far, we have assumed that the classification method is completely defined, including the choice of classification threshold, so that precision and recall are given. Use of the F-measure for evaluation means that one is using an implied weight $w = p$ for the importance of recall, and under the arithmetic mean interpretation, since this is a function of P and R, it means that using the F-measure is equivalent to weighting P and R differently for different classification methods (at least, if they have different P and/or R values). However, we also note that the classification threshold t is a control parameter of the method: change it and one obtains different P and R values. We could therefore change the classification thresholds so that p is the same for all classification methods. This would make comparison legitimate: comparisons made using the arithmetic mean interpretation of the F-measure if the thresholds were chosen so that p was the same for all classification methods would be legitimate.

Equality of p values is equivalent to equality of the ratio $p/(1 - p)$ and from the definition of p we see that:

$$\frac{p}{1-p} = \frac{TP + FN}{TP + FP}. \tag{11}$$

Since $TP + FN$ is fixed (at the number of class 1 objects in the test set) we can ensure equality of the p values by choosing individual thresholds for all classification methods so that all methods being compared classify the same number of objects, $TP+FP$, to class 1. That is, if we arrange things so that each classification method assigns the same number of objects to class 1, R (and hence also P) is being given the same weight p by each classification method, so that the F-measure is a legitimate measure of performance. Note, however, that this is more restricted than the situation described in Appendix A.2. There we assumed that different classification methods generally assign different numbers of test objects to class 1, with the weight w being chosen independently of this number.

Assigning the same number of test objects to class 1 ensures that all classification methods are using the same p value. As discussed previously, the choice of an appropriate p should be made on the basis of the problem and aims, but this can be difficult. If, instead, we wish to adopt the convention of weighting recall and precision equally ($p = 1/2$) then the relationship $p/(1 - p) = (TP + FN)/(TP + FP)$ shows that we must choose the classification thresholds so that each classification method being compared assigns $TP + FN = TP + FP$ objects to class 1: that is, all methods must assign to class 1 the same number of objects as there are in class 1.

A.3.2 Calibrated Scores. Classification performance measures based on the values of TP, FN, FP, and TN are invariant to monotonic increasing transformations of the score continuum, since these counts are obtained by comparing score values, with a particular score value chosen as the threshold t. In particular, this means that we can transform the score so that it is calibrated. That is, we can transform the score so that it is the probability of a data object to belong to class 1. Technically, of course, we should speak of estimated probabilities and also deal with potential overfitting issues arising from using limited data to calibrate [32], as well as estimate parameters and evaluate performance. Here, however, we are concerned with conceptual issues of classifier performance rather than practicalities, so we shall ignore these complications. Calibration means that classification has the attractive interpretation that an object is assigned to class 1 if its (estimated) probability of belonging to class 1 is greater than the threshold t. This leads to two possible strategies:

(1) We might choose a probability threshold (the calibrated score being interpretable as a probability) t, which is the same for all classification methods being compared. This is conceptually attractive: we would hardly want to say that we will classify an object to class 1 if its estimated class 1 probability is greater than 0.9 when using a neural network, but greater than 0.7 when using a random forest.
(2) We might choose a weight w that is the same for all classification methods being compared. This is also conceptually attractive. We would hardly want to say that we will weight recall as 0.7 (and precision as 0.3) when using a neural network, but weight recall as 0.4 (and precision as 0.6) when using a random forest. As noted above, the weights represent an aspect of the problem, not the classification method.

These two strategies can be simultaneously applied if the weights applied to precision and recall are chosen independently of the threshold, as described in Appendix A.2. That is, a probability threshold t, the same for all classification methods being compared, is chosen, and this yields R and P values for each of the methods. These are then weighted using w, again the same for all methods, to yield the overall measure $wR + (1 - w)P$.

However, from the weighted arithmetic mean perspective, using the F-measure implies a fundamental contradiction between the two strategies (1) and (2). This is easily seen by supposing that we used strategy (1) and chose a probability threshold t so that all objects with estimated class 1 probability greater than t are assigned to class 1. Now, if the classification methods do not have identical performance, then, in general, they will produce different values of TP, FN, FP, and TN

Table 1. Details of the three Datasets Used in the Empirical Evaluation Discussed in Appendix B

Data set	Number of records, n	Number of features, m	Class imbalance ci	Results with $t = 0$		
				P	R	F
Wisconsin Breast Cancer	699	9	1.90	0.345	1.0	0.513
German Credit	1,000	24	2.33	0.300	1.0	0.462
PIMA Diabetes	768	8	1.87	0.348	1.0	0.516

for any given threshold t. This means that the weights p and $1 - p$ implied by the F-measure will be different for different methods.

The converse also applies. If we chose the thresholds so that p was the same for all classification methods (corresponding to strategy (2) of using F-measure values where all methods give the same weight to recall), it is likely that this would correspond to different estimated probabilities of belonging to class 1; that is, different thresholds t.

A practical comment is worth making. Except for certain pathological situations, as the threshold t increases, so $TP + FP$ will decrease (while $TP + FN$ remains constant). This means that if all test object scores are different it will usually be possible to choose a threshold t such that $TP + FN = TP + FP$. In other situations, it might not be possible (for example, if groups of test objects have the same score), but then a simple averaging can be adopted. To do this, find the threshold t_1 for which $TP+FP$ is closest to $TP+FN$ while $TP+FP > TP+FN$, and also the threshold t_2 for which $TP+FP$ is closest to $TP+FN$ while $TP+FP < TP+FN$. Calculate the weights p_1 and p_2 for these two thresholds, and also the corresponding precision and recall values, P_1, P_2, R_1, and R_2, respectively. Then, calculate a weighted average of the recall values for these two thresholds, and similarly the precision values as $\overline{R} = wR_1+(1-w)R_2$ and $\overline{P} = wP_1+(1-w)P_2$, respectively. The weight, w, is chosen such that $wp_1 = w(1 - p_2) = 1/2$. That is $w = (1/2 - p_2)/(p_1 - p_2)$.

We have shown that the F-measure can be reinterpreted as a weighted arithmetic mean, which does have a straightforward intuitive interpretation. In particular, it allows a researcher to specify the relative importance they accord to precision and recall. When used in this way, with a user specified choice of the weight p assigned to recall, the F-measure makes sense. Alternatively, it might be better to abandon the F-measure altogether, and go directly for a weighted combination of precision and recall, with weights reflecting the researcher's attitudes to their relative importance, with equal weights as an important special case.

B REAL-WORLD EXAMPLES

To illustrate the properties of the F-measure we discussed in Section 4 and Appendix A, in this appendix we show actual classification results obtained when applying different classification methods on different datasets, where we change the classification threshold from $t = 0$ to $t = 1$. Specifically, we selected three datasets from the UCI Machine Learning Repository,[3] as detailed in Table 1. We applied four classification methods (decision tree, logistic regression, random forest, and support vector machine) as implemented in the Python sklearn machine learning library.[4] We randomly split the datasets into training and testing sets, in proportions of 80% / 20%, and estimated the parameters for the four classification methods by minimising misclassification rate using a grid search over a variety of settings for the relevant parameters. Note that our experiments are aimed at illustrating issues with the F-measure, we are not focused on obtaining the highest classifier performance.

[3]See: https://archive.ics.uci.edu/ml/index.php.
[4]See: https://scikit-learn.org.

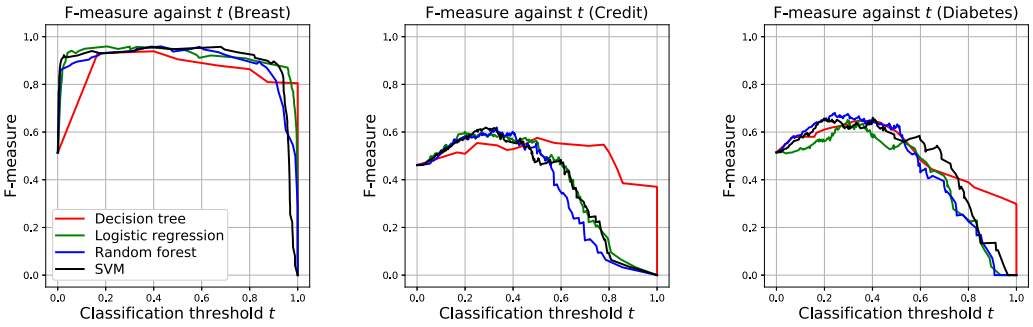

Fig. 5. Four different classification methods applied on three datasets from the UCI Machine Learning repository, showing F-measure results for a classification threshold varied from $0 \leq t \leq 1$, as described in Appendix B.

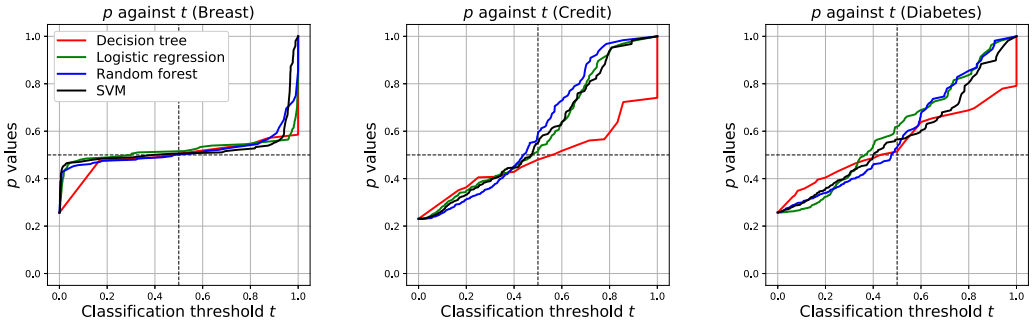

Fig. 6. Weights, p, assigned to recall for a classification threshold varied from $0 \leq t \leq 1$ for the four classification methods and three datasets described in Appendix B.

In Figure 5, we illustrate the asymmetric behaviour of the F-measure, calculated using Equation (3), with varying classification thresholds, t. As we discussed in Section 4.4, for all three datasets the discussed asymmetric behaviour is visible, even for the Wisconsin Breast Cancer data set which seems to be easy to classify. For all classification methods used (and on all datasets), with $t = 1$ the values for precision (P), recall (R), and, therefore, the F-measure, are 0, while the corresponding values for $t = 0$ are shown in Table 1.

As can also be seen in Figure 5, for the two more difficult to classify data sets (German Credit and PIMA Diabetes) the best F-measure results are not obtained with the $t = 0.5$ classification threshold that is often used as the default threshold, but rather with much lower thresholds [42] of around $t = 0.25$ in our examples. With such thresholds, recall is much higher than precision (in the range of 0.683 to 0.867 for German Credit and 0.736 to 0.906 PIMA Diabetes), which might not be a desirable classification outcome. While the precise shapes of F-measure curves depend both upon the classification method employed as well as the distribution of the scores of the classified objects, as these examples illustrate, the asymmetric behaviour of the F-measure needs to be considered when it is used to evaluate classification methods.

Figure 6 shows the values of the weight p, as calculated using Equation (7), for the four classification methods and varying thresholds shown in Figure 5. As can be seen, the different classification methods lead to different values of p as the threshold t is changing. This holds even for different classification methods applied on the same dataset. With low thresholds the value of p is below 0.5, which means less weight is assigned to recall compared to precision. For higher thresholds the

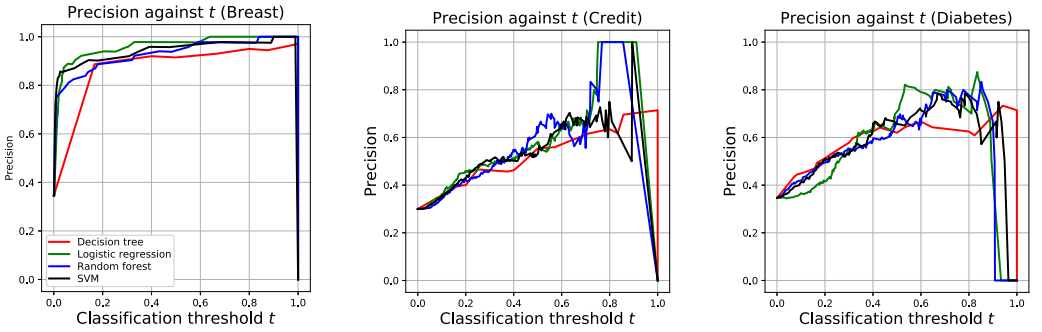

Fig. 7. Precision results for a varying classification threshold for the three datasets and four classification methods.

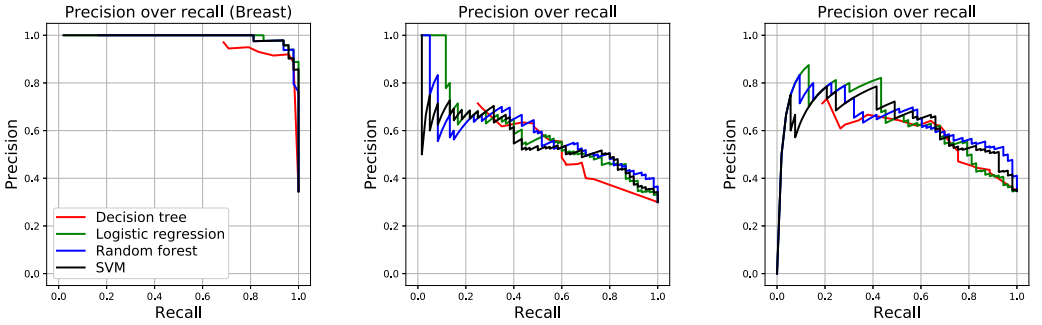

Fig. 8. Precision-recall curves for a varying classification threshold for the three datasets and four classification methods.

value of p increases to above 0.5, resulting in higher weights being assigned to recall compared to precision. This general pattern means that low classification thresholds give more weight to precision (if $FP > FN$) while high classification thresholds give more weight to recall (if $FP < FN$), as we discussed in Appendix A.1.

In Figure 7, we show precision results for varying classification thresholds, as we discussed in Section 4.4. As can clearly be seen, precision does not necessarily increase or decrease monotonically as the classification threshold changes. These plots illustrate the actual distribution of classification scores of objects in the positive and negative classes as obtained with the four different classifiers.

Finally, Figure 8 shows precision-recall curves for varying classification thresholds, as we discussed in Section 5.2. As clearly visible, the resulting **areas under the precision-recall curves** (**AUC-PR**) are much larger for the Wisconsin Breast Cancer dataset which seems to be easier to classify accurately compared to the other two datasets. What can also be seen is that for different ranges of the classification threshold one classification method shows higher performance than the other methods, given their precision-recall curves are above the ones of the other methods.

C THE H-MEASURE

Another widely used measure of classifier performance is the Area Under the Receiver Operating Characteristic Curve (AUC). Instead of being based on a single choice of classification threshold, the AUC is equivalent to an average misclassification loss, where the average is taken over a distribution of classification thresholds. Since any choice of classification threshold is equivalent to a

choice of relative severities of the two kinds of misclassification (class 0 objects to class 1, or class 1 objects to class 0), the AUC is equivalent to averaging the misclassification loss over a distribution of the ratio of severities of the two kinds of misclassification.

However, it has been shown that for the AUC, this distribution depends on the score distribution produced by the classifier [26, 28]. This means it is fundamentally incoherent: it is equivalent to using a stretchy ruler when measuring length, with the measuring instrument differing between the classifiers being assessed. Put another way, the AUC implies that researchers using different classifiers have different distributions of severity-ratios for the two kinds of misclassification. This is nonsensical: misclassifying a class 0 object as a class 1 object cannot be twice as serious as the reverse when using a random forest but three times as serious when using deep learning. The relative seriousness depends on the problem and the aims, not the classifier.

The H-measure overcomes this fundamental incoherence by specifying a fixed distribution for the severity-ratio. In its most common form, a beta distribution is used [28].

Another way of describing the problem with the AUC is that it is equivalent to using different probability scoring functions when evaluating the accuracy of the estimates of the probability of belonging to class 1 produced by different classifiers. Buja et al. [9] show that the H-measure, using the beta distribution, overcomes this.

REFERENCES

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. 2009. Frequency-tuned salient region detection. In *Conference on Computer Vision and Pattern Recognition*. IEEE, Miami, 1597–1604.

[2] Brian Austin and Rita R. Colwell. 1977. Evaluation of some coefficients for use in numerical taxonomy of microorganisms. *International Journal of Systematic and Evolutionary Microbiology* 27, 3 (1977), 204–210.

[3] Bart Baesens, Veronique Van Vlasselaer, and Wouter Verbeke. 2015. *Fraud Analytics using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection*. John Wiley and Sons, Hoboken, New Jersey.

[4] Thomas Benton. 2001. *Theoretical and empirical models*. Ph. D. Dissertation. Department of Mathematics, Imperial College, London.

[5] Olivier Binette and Rebecca C. Steorts. 2022. (Almost) all of entity resolution. *Science Advances* 8, 12 (2022), eabi8021.

[6] Kendrick Boyd, Kevin H. Eng, and C. David Page. 2013. Area under the precision-recall curve: Point estimates and confidence intervals. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 451–466.

[7] Paula Branco, Luís Torgo, and Rita P. Ribeiro. 2016. A survey of predictive modeling on imbalanced domains. *Computing Surveys* 49, 2 (2016), 1–50.

[8] Sebastian Brutzer, Benjamin Höferlin, and Gunther Heidemann. 2011. Evaluation of background subtraction techniques for video surveillance. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 1937–1944.

[9] Andreas Buja, Werner Stuetzle, and Yi Shen. 2005. *Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications*. Technical Report. The Wharton School, University of Pennsylvania.

[10] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.

[11] Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21, 1 (2020), 6.

[12] Nancy Chinchor. 1992. MUC-4 evaluation metrics. In *Fourth Message Understanding Conference*. ACL, 22–29.

[13] Peter Christen, Thilina Ranbaduge, and Rainer Schnell. 2020. *Linking Sensitive Data*. Springer.

[14] Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and ROC curves. In *International Conference on Machine Learning*. ACM, 233–240.

[15] Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7, Jan (2006), 1–30.

[16] Robert P. W. Duin. 1996. A note on comparing classifiers. *Pattern Recognition Letters* 17, 5 (1996), 529–536.

[17] Boris Epshtein, Eyal Ofek, and Yonatan Wexler. 2010. Detecting text in natural scenes with stroke width transform. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2963–2970.

[18] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 8 (2006), 861–874.

[19] César Ferri, José Hernández-Orallo, and R. Modroiu. 2009. An experimental comparison of performance measures for classification. *Pattern Recognition Letters* 30, 1 (2009), 27–38.

[20] Edward B. Fowlkes and Colin L. Mallows. 1983. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association* 78, 383 (1983), 553–569.
[21] Jiawei Han and Micheline Kamber. 2000. *Data Mining: Concepts and Techniques* (1st ed.). Morgan Kaufmann, San Francisco.
[22] David J. Hand. 1996. Statistics and the theory of measurement. *Journal of the Royal Statistical Society: Series A* 159, 3 (1996), 445–473.
[23] David J. Hand. 1997. *Construction and Assessment of Classification Rules*. John Wiley and Sons, Chichester.
[24] David J. Hand. 2004. *Measurement Theory and Practice: The World Through Quantification*. Edward Arnold, London.
[25] David J. Hand. 2006. Classifier technology and the illusion of progress. *Statistical Science* 21, 1 (2006), 1–14.
[26] David J. Hand. 2009. Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning* 77, 1 (2009), 103–123.
[27] David J. Hand. 2012. Assessing the performance of classification methods. *International Statistical Review* 80, 3 (2012), 400–414.
[28] David J Hand and Christoforos Anagnostopoulos. 2022. Notes on the H-measure of classifier performance. *Advances in Data Analysis and Classification* 17, 1 (2022), 109–124.
[29] David J. Hand and Peter Christen. 2018. A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing* 28, 3 (2018), 539–547.
[30] David J. Hand, Peter Christen, and Nishadi Kirielle. 2021. F*: An interpretable transformation of the F-measure. *Machine Learning* 110, 3 (2021), 451–456.
[31] David J. Hand, Heikki Mannila, and Padhraic Smyth. 2001. *Principles of Data Mining*. MIT Press.
[32] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer, New York.
[33] Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. *Transactions on Knowledge and Data Engineering* 21, 9 (2009), 1263–1284.
[34] David Torres Irribarra. 2021. *A Pragmatic Perspective of Measurement*. Springer.
[35] Paul Jaccard. 1908. Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles* 44, 163 (1908), 223–270.
[36] Andrew Janowczyk and Anant Madabhushi. 2016. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics* 7, 1 (2016), 29.
[37] Thorsten Joachims. 2005. A support vector method for multivariate performance measures. In *International Conference on Machine Learning*. ACM, 377–384.
[38] Shubhra Kanti Karmaker, Md Mahadi Hassan, Micah J. Smith, Lei Xu, Chengxiang Zhai, and Kalyan Veeramachaneni. 2021. AutoML to date and beyond: Challenges and opportunities. *Computing Surveys* 54, 8 (2021), 1–36.
[39] Raymond Kosala and Hendrik Blockeel. 2000. Web mining research: A survey. *ACM SIGKDD Explorations Newsletter* 2, 1 (2000), 1–15.
[40] Wojtek J. Krzanowski and David J. Hand. 2009. *ROC Curves for Continuous Data*. CRC Press, New York.
[41] David D. Lewis. 1995. Evaluating and optimizing autonomous text classification systems. In *Conference on Research and Development in Information Retrieval*. ACM, 246–254.
[42] Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. 2014. Optimal thresholding of classifiers to maximize F1 measure. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Springer, 225–239.
[43] Neil A. Macmillan and C. Douglas Creelman. 2005. *Detection Theory: A User's Guide* (2nd ed.). Lawrence Erlbaum Associates, New York.
[44] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge.
[45] Joseph F. McCarthy and Wendy G. Lehnert. 1995. Using decision trees for coreference resolution. In *International Joint Conference on Artificial Intelligence*. AAAI, 1050–1055.
[46] Giovanna Menardi and Nicola Torelli. 2014. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery* 28, 1 (2014), 92–122.
[47] Ave Mets. 2019. A philosophical critique of the distinction of representational and pragmatic measurements on the example of the periodic system of chemical elements. *Foundations of Science* 24, 1 (2019), 73–93.
[48] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: A comprehensive review. *Computing Surveys* 54, 3 (2021), 1–40.
[49] Aytuğ Onan, Serdar Korukoğlu, and Hasan Bulut. 2016. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications* 57 (2016), 232–247.
[50] David M. W. Powers. 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology* 2, 1 (2011), 37–63.

[51] David M. W. Powers. 2019. What the F-measure doesn't measure: Features, flaws, fallacies and fixes. https://arxiv.org/abs/1503.06410v2
[52] Troy Raeder, George Forman, and Nitesh V. Chawla. 2012. Learning from imbalanced data: Evaluation matters. In *Data Mining: Foundations and Intelligent Paradigms*. Springer, Berlin, 315–331.
[53] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 658–666.
[54] Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, Singapore.
[55] Yutaka Sasaki. 2007. The truth of the F-measure. University of Manchester, MIB-School of Computer Science.
[56] Robert R. Sokal and Peter H. A. Sneath. 1963. *Numerical Taxonomy*. W.H. Freeman and Co., San Francisco.
[57] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing and Management* 45, 4 (2009), 427–437.
[58] Thorvald Julius Sørensen. 1948. *A Method of Establishing Groups of Equal Amplitude in Plant Sociology based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*. I Kommission Hos E. Munksgaard, Copenhagen.
[59] Cornelius J. Van Rijsbergen. 1979. *Information Retrieval*. Butterworth and Co., London.
[60] Adam Yedidia. 2016. Against the F-score. Retrieved from https://adamyedidia.files.wordpress.com/2014/11/f_score.pdf. Blogpost, accessed 12 April 2023.

Received 9 June 2022; revised 4 April 2023; accepted 2 June 2023