Loss-size and Reliability Trade-offs Amongst Diverse Redundant Binary Classifiers

Kizito Salako

The Centre for Software Reliability,
City, University of London, Northampton Sq. EC1V 0HB
The United Kingdom
k.o.salako@city.ac.uk

Abstract. Many applications involve the use of binary classifiers, including applications where safety and security are critical. The quantitative assessment of such classifiers typically involves *receiver operator characteristic* (ROC) methods and the estimation of sensitivity/specificity. But such techniques have their limitations. For safety/security critical applications, more relevant measures of reliability and risk should be estimated. Moreover, ROC techniques do not explicitly account for: 1) inherent uncertainties one faces during assessments, 2) reliability evidence other than the observed failure behaviour of the classifier, and 3) how this observed failure behaviour alters one's uncertainty about classifier reliability. We address these limitations using *conservative Bayesian inference* (CBI) methods, producing statistically principled, conservative values for risk/reliability measures of interest. Our analyses reveals trade-offs amongst all binary classifiers with the same expected loss – the most reliable classifiers are those most likely to experience high impact failures. This trade-off is harnessed by using diverse redundant binary classifiers.

Keywords: reliability assessment, binary classification, diverse redundancy, conservative Bayesian inference

1 Introduction

Numerous applications that society relies upon involve binary classification [21, 22, 26]. Examples include medical diagnosis, autonomous vehicle safety, crime detection/forensic science and IT network protection. The failure of classifiers in such applications can have a significant impact – affecting the well-being, safety and security of those reliant on these technologies. Classifiers *must* be "good enough" to be deployed. But, demonstrating this can be challenging. The primary challenge here is uncertainty: an assessor of such systems is uncertain about if/when the classifiers will fail during operation, the nature of failures should they occur, and the resulting impact failures will have on the wider system. Consequently, the *statistical* assessment of classifiers is necessary. And any serious attempts at quantifying classifier reliability – say, the probability of a classifier's

correct functioning on a sequence of classification tasks – *must* account for these uncertainties. This is not easy to do, because the probability distributions that characterise these uncertainties are often unknown or unknowable.

As an approach to assessing classifiers, *receiver operator characteristic* (ROC) methods are well-suited for comparing certain statistical properties of classifiers [8,11]. But, these methods do not account for all of the aforementioned forms of uncertainty. In this paper, we offer a complementary approach to ROC methods, and the following contributions to the assessment of binary classifiers:

1. We critique the sole use of ROC approaches in the statistical assessment of binary classifiers, particularly for safety/security critical applications;
2. We formalise a statistical model of classifier failure and loss, in terms of loss distributions. We argue, it is loss distributions that classifier assessments should be concerned with – these subsume "point" measures such as sensitivity or specificity.
3. We highlight the statistical challenge an assessor faces – i.e. infinitely many loss distributions, all consistent with a classifier's observed failure behaviour;
4. We show, for a given expected loss, that a trade-off exists between classifier reliability and the size of losses when failures occur. This trade-off has not been reported in the literature before;
5. Using this trade-off we prove the range of those loss distributions – from the most reliable classifiers to the least reliable ones – that are consistent with a given expected loss. Our results allow an assessor to reason conservatively about a classifier's reliability, given it's observed failure behaviour;
6. We demonstrate that convex combinations of diverse classifiers can be used to harness this trade-off. In this way, infinitely many hybrid classifiers may be constructed from a few diverse ones. As a curious aside, we also show how such classifier combinations share striking similarities with the optimal allocation of assets in an investment portfolio – a famous problem in Finance;
7. "Optimal" ways of combining the outputs of diverse classifiers have been argued for in the literature – optimal here means smallest expected loss. These ignore the aforementioned trade-off, and we illustrate how such optimal adjudication schemes do not necessarily produce the most reliable systems;
8. A Bayesian formalisation of an assessor's uncertainty about a classifier's loss distribution. Furthermore, by way of mathematical proof, we show that our Bayesian assessments are guaranteed to be conservative – no other similarly constrained Bayesian prior gives more conservative conclusions than ours.

The outline of the rest of the paper is as follows. Critical context and related work is given in section 2. In section 3, we introduce our statistical model of binary classification. Section 4 then presents analyses of trade-offs between the size of losses (due to classifier failures) and classifier reliability. The consequences of such trade-offs, for optimally combining classifiers, are also explored in some detail. This is followed in section 5 by a novel application of CBI methods, to explicitly account for uncertainties surrounding the assessment of classifiers. This also takes into account the trade-offs in previous sections. The paper concludes with final considerations in section 6.

2 Critical Context and Related Work

During assessment, a classifier's failure propensity and associated risks are estimated by subjecting the classifier to a sequence of statistically representative classification tasks (i.e. operational testing), and averaging over the classifier's observed failure behaviour. Popular statistics computed in this way, and compared using *receiver operator characteristic* (ROC) methods, include estimates of a classifier's false-positive rate (FPR) and true-positive rate (TPR) [32, 33].

A useful graphical tool for comparing classifier performance is ROC-*space* [8, 25], shown in Fig. 1. Each point in the unit-square represents all those *discrete classifiers* with the specific FPR, TPR values at that point. All useful discrete classifiers can be made to lie above the 45° diagonal [10]. The diagonal, itself, represents classifiers that make random or blanket classifications. The best classifiers are located at $(0, 1)$. *Probabilistic classifiers* can have their FPR, TPR values altered by changing the threshold at which they distinguish positive classifications from negative ones. Continuously altering such a threshold produces, in essence, a range of discrete classifiers – a unique ROC curve (e.g. dashed curve) represents the set of discrete classifiers obtained in this way. Given a suitable stochastic process that generates the classification tasks, and given the losses when classifiers fail, an *isocost line* (e.g. line l_1) represents classifiers with the same expected loss. Any parallel line that is closer to the (0,1) point (e.g. line l_2) represents classifiers with smaller expected loss [24]. Consequently, under appropriate invariance assumptions [9, 34], regions **A** and **B** contain classifiers that, respectively, are guaranteed to have expected loss no larger, or smaller, than the expected loss for those classifiers at the point where **A** and **B** meet. For a given ROC curve, the *area under the curve* (AUC) measures the "size" of the set of all classifiers guaranteed to have expected loss at least as small as some classifier on the curve (i.e. the "size" of the union of all **B** regions with north-west corners on the curve). The AUC is also a measure of how likely it is that a probabilistic classifier will rank a randomly chosen positive classification task more highly than a randomly chosen negative task [15].

But, by themselves, these statistics and ROC methods do not explicitly account for an assessor's uncertainty about the accuracy of FPR and TPR estimates, nor the uncertainty about the stochastic process generating the classification tasks. And, they do not explicitly incorporate reliability evidence obtained before the classifier is subjected to operational testing. Further still, they do not provide a means of updating an assessor's uncertainty (about classifier reliability, future failures or losses) upon observing the classifier during operational testing. Moreover, for those safety or security critical applications where any future failure is unacceptable, there are arguably more relevant quantitative measures of reliability to consider, other than FPR and TPR. For instance, the probability that a classifier succeeds on the next n classification tasks, for very large n.

To complement ROC methods, we choose a Bayesian approach [4, 13, 14]. Classifiers either succeed or fail according to some *unknown* loss distribution – only one of infinitely many plausible distributions that could characterise the classifier's failure behaviour. *Which one* of these is the true loss distribution

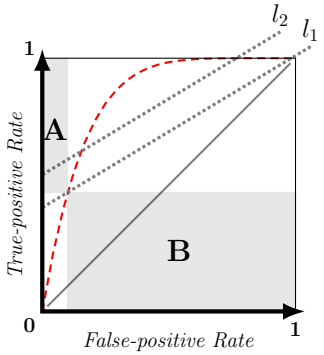


Fig. 1: ROC-space

is an assessor's best guess. In principle, based on evidence gathered prior to operational testing, an assessor's ignorance about a classifier's loss distribution is formalised as a *prior distribution* over the set $\mathcal{L}$ of all such loss distributions. Assessment then proceeds by Bayesian inference, using the classifier's observed failure behaviour during testing to produce a so-called *posterior distribution* over $\mathcal{L}$. And this updated assessor ignorance – this posterior distribution – can be used to compute informed estimates of classifier reliability and risk.

Bayesian inference would typically require the specification of a suitable prior distribution over $\mathcal{L}$ – a daunting task. Perhaps a less daunting task would be to specify the prior only partially – e.g. specify a prior probability that the classifier fails its next classification. Then, amongst all prior distributions over $\mathcal{L}$ that share this value for the probability, one determines which of these priors yields the worst-case value for a posterior reliability measure of interest. With this worst-case value, an assessor gains insight into the range of plausible loss distributions (and thus, classifiers) consistent with their prior evidence and the observed failure behaviour of the classifier. By reasoning conservatively – i.e. reasoning in terms of worst-case values for posterior measures of interest – an assessor is actively avoiding dangerously optimistic assessments resulting from using unjustified priors. In this spirit, *conservative Bayesian inference* (CBI) methods have been used in a number of contexts [2, 18, 30, 35–37]. We develop and apply a new variant of these methods to the problem of producing conservative assessments of binary classifiers. In particular, by performing constrained optimisations over the set $\mathcal{D}$ of all *prior* distributions with sample space $\mathcal{L}$, we produce surprisingly simple expressions for conservative estimates of *posterior* classifier reliability, and identify prior distributions that yield these posteriors.

3 A Statistical Model of Binary Classification

Consider the set $\mathbf{\Omega}$ of all possible classification tasks, i.e. *demands*, that a classifier can be presented with in a given application. When presented with a demand,

a classifier either raises an alarm or not. With ROC methods, one assumes classifiers always give a response – the same response – to a demand. So, imagine the operational environment of a classifier as a black box, spewing forth demands from $\mathbf{\Omega}$ for the classifier to classify. The interplay – of a classifier's deterministic behaviour and the uncertainty about which demand will be presented by the environment next – induces uncertainty about whether a classifier will fail on the next demand. And, uncertainty about whether the next failure will be a *false-positive* (FP) error – the classifier raises an alarm when it should not – or a *false-negative* (FN) error – it does not raise an alarm when it should. These error-types have associated non-zero costs l_{fp} and l_{fn} when they occur. Assume the cost associated with an error-type is always the same whenever the error occurs[1]. Costs are determined by the economic impact of classifier failures; typically, with l_{fp} much smaller than l_{fn}. Correct classification incurs no cost.

A classifier has an associated conditional probability q_{fn} of making an FN error (i.e. 1−TPR), and conditional probability q_{fp} of an FP error (i.e. FPR). Estimates of $(1 - q_{fn})$ and $(1 - q_{fp})$ are referred to as *sensitivity* and *specificity*.

Define the following indicator function:

$$\mathbf{1}_{fn} := \begin{cases} 1\,; \text{ if the classifier commits an FN error} \\ 0\,; \text{ otherwise} \end{cases}$$

and $\mathbf{1}_{fp}$ is similarly defined for FP errors. Then, the loss resulting from a classifier's failure is the random variable $L := l_{fn}\mathbf{1}_{fn} + l_{fp}\mathbf{1}_{fp}$. The *loss distribution* for L is depicted in Fig. 2a, where $\gamma := P(\text{next demand should cause an alarm})$.

4 Loss-size vs Reliability Trade-off

Given the loss values l_{fp} and l_{fn}, and a best estimate of a classifier's expected loss $\mathbb{E}[L]$, what can be conservatively claimed about the classifier's failure behaviour? In theory, there exists a range of possible discrete, 3-point, loss distributions (all with the same $\mathbb{E}[L]$) that could characterise the occurrence of failures and losses for this classifier. And, in practice, specifying which of these distributions conservatively characterises the classifier must ultimately be a judgement call, dependent on the specifics of the situation. There is a trade-off here – between how reliable the classifier is, and how large the losses are when failures occur.

We can elucidate this trade-off. As proved in appendix A, Fig. 2 shows the 3 extremes of the range of possible loss distributions, using a normalised scale for the losses (i.e. the losses have been divided by the largest possible loss, typically l_{fn}). The smallest probability θ of correct classification is either $\theta = \frac{l_{fp} - \mathbb{E}[L]}{l_{fp}}$, when $l_{fp} > \mathbb{E}[L]$ (see Fig. 2b), or it is $\theta = 0$ otherwise (see Fig. 2c). Contrastingly, the largest value of θ is $\theta = 1 - \mathbb{E}[L]$, given by the loss distribution in Fig. 2d.

What do these extremes represent in practice? In Fig. 2b, the only failures are FPs. While, in Fig. 2c, no correct detection ever occurs – only FPs and FNs.

[1] View this as the conditional expected loss, given the occurrence of the relevant error.

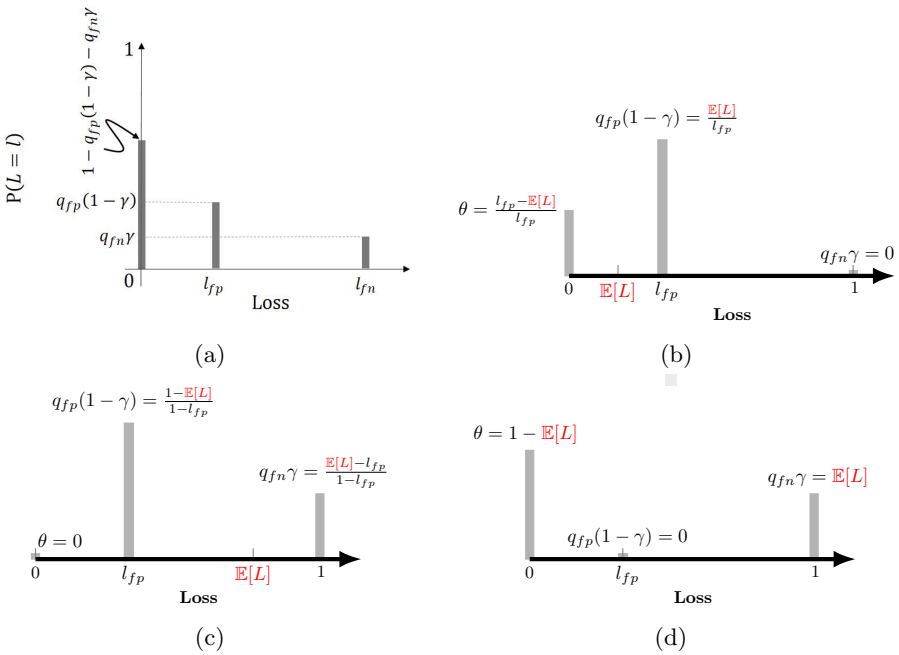


Fig. 2: In **(a)** is the distribution of loss L for a classifier. Also shown, in **(b)**, **(c)** and **(d)**, are the 3 extremes of the range of all loss distributions that share the same expected loss $\mathbb{E}[L]$. All of the extreme distributions are defined over normalised loss values $\{0, l_{fp}, 1\}$. In particular, the distributions in **(b)** and **(c)** give the smallest values for the probability θ of correct detection, depending on whether l_{fp} is larger or smaller than $\mathbb{E}[L]$, respectively. These also give the smallest values for the probability q of an FN error. The largest values for both θ and q are given by the distribution in **(d)**.

In practice, it is fairly easy to determine that one is not in the extreme situation of Fig. 2c, once the classifiers have been observed to correctly classify *some* demands. In contrast, determining that Fig. 2b is not the situation one faces in practice is more challenging, especially since FNs can be very difficult to identify in certain applications (e.g. cybersecurity, medical diagnosis). Moreover, while Fig. 2b is clearly more preferable than Fig. 2c, the choice between Fig. 2b and Fig. 2d is less clear. Of course, lest one get too excited about the most reliable classifier Fig. 2d, it is sobering that this possibility is also easily excluded in practice, once FPs have been observed. But, the usefulness of thinking in terms of these extremes is not that these are "achievable" in practice, but that classifiers worryingly "close" to these are.

There is another way to view this trade-off. Consider when $l_{fp} = \mathbb{E}[L]$. Both distributions in Fig.s 2b and 2c collapse to the same deterministic function – where only failures with associated losses of size $l_{fp} = \mathbb{E}[L]$ occur with probability 1. That is, only demands that should cause no alarms occur, and the classifier

raises alarms on all of these. There is no uncertainty – only FP failures *will* occur with accompanying losses of value $\mathbb{E}[L]$. Unlike the distribution in Fig. 2d, which has *the largest amount of variation amongst all of these distributions*. In this sense, the trade-off between the extreme distributions can be viewed as exchanging the certainty of small losses (i.e. losses due to FPs) for an increase in the reliability of the classifiers, but at the added cost of an increased probability of incurring much larger losses when failures occur (i.e. losses due to FNs).

So far, our discussion has focused on the extremes of a range of 3-point loss distributions, while remarking that the distribution Fig. 2d – for the most reliable classifier – possesses the maximum variation amongst all of these distributions. But the following stronger claim is also true (see appendix A). Amongst *all* loss distributions over *any* (normalised) collection of loss values in the interval $[0, 1]$ (so, not only 3 loss values) – where the distributions all share $\mathbb{E}[L]$ – Fig. 2d is the loss distribution for the most reliable classifiers, and this is also the distribution with the largest possible variation. So, "increased variance" of a loss distribution is the same as "increased reliability and increased losses when failures occur". The proof of this result shares some similarities with the proof of bounds on a system's probability of failure, in [31].[2]

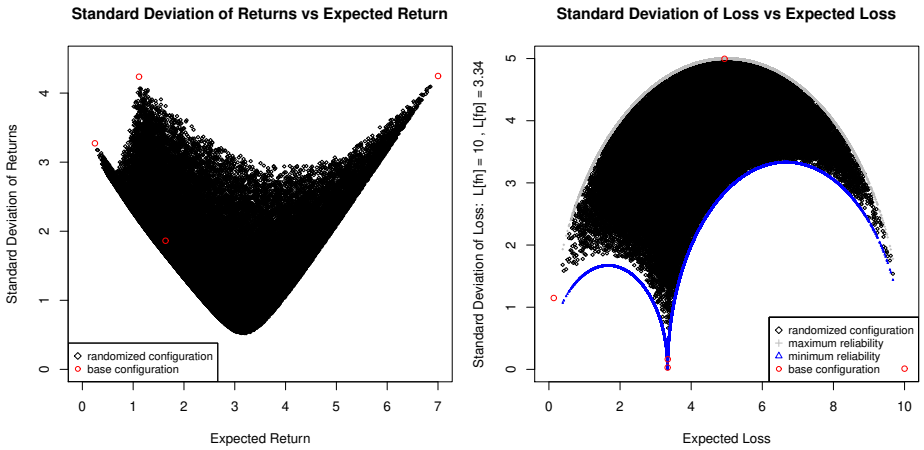
4.1 Trade-off Implications for Randomly Choosing Amongst Diverse Classifiers During Operation

The trade-off becomes more complicated when considering classifiers with different expected losses and variances for their loss distributions. For example, one classifier might have a loss distribution similar to Fig. 2b, while another classifier has a distribution like Fig. 2d, but with a larger expected loss than the first classifier. So, the first classifier has a smaller expected loss, while the second one is noticeably more reliable but perhaps more prone to making FN errors. Consequently, an assessor's preferences for a classifier's failure behaviour may lie somewhere "inbetween" these two classifiers. An "inbetween" classifier can be constructed by a suitable random combination of this pair during operation.

Reducing both expected loss and the probability of failures (i.e. increasing variance) requires multi-objective optimisation techniques. Using expectations and variances together in making multi-objective choices is not a new idea – Markowitz and Sharpe applied this to the finance problem of selecting "efficient" investment portfolios, for which they were awarded the 1990 Nobel prize in Economics [19, 20, 29]. What *is* novel here is the application of these ideas to the problem of choosing optimal configurations of binary classifiers.

Markowitz's modern portfolio theory shows how combinations of diverse risky investments in a portfolio can lower some of the risk associated with the portfolio, possibly leaving only so-called *undiversifiable risk*. So, with a fixed budget to invest, a desirable portfolio is constructed out of varied investments in carefully

[2] In [31] their focus was uncertainty about the value of the probability of failure for a system. Contrastingly, our result applies to the uncertainty about whether a given classifier will fail on its next classification task, and the loss incurred if it does.



(a) (b)

Fig. 3: **(a)** Given 4 distinct risky investments (i.e. "base configurations"), the shaded region contains randomly constructed investment portfolios (i.e. randomised configurations); each point is a convex combination of the investments. Plotted for each portfolio is the standard deviation of returns vs the expected return. All portfolios that lie lower in the region have benefited from diversity reducing uncertainty (i.e. "risk" in this portfolio theory) in the portfolio returns. **(b)** Given 5 distinct classifiers (i.e. "base configurations"), the shaded region contains randomly chosen hybrid classifiers (i.e. randomised configurations). Each hybrid uses a unique distribution for randomly choosing one of the base configurations to exclusively perform the next classification task. Note, one base configuration lies in the bottom right corner of **(b)**. Plotted for each classifier is the standard deviation of losses vs the expected loss, with $l_{fn} = 10$ and $l_{fp} = 3.34$. The classifiers on the boundary have the extreme distributions in Fig. 2, with appropriately scaled losses. All hybrids that lie lower in the region have benefited from diversity reducing the probability of large losses, but also impacting on reliability.

chosen proportions. Here, analogous constructions can be made out of classifiers – a convex combination of classifiers defines a hybrid classifier with properties not wholly possessed by any of the constituent classifiers. For other reasons, the works of Scott *et al.* [28], Provost *et al.* [23, 25] and Gaffney *et al.* [11, 12] have even argued for such convex combinations as a way of creating preferred classifiers out of unsatisfactory ones – these approaches are related to the ROC Convex-Hull theorem (ROCCH) for determining preferred classifiers [8, 25].

We construct a convex combination of diverse classifiers as follows. Suppose there are n functionally equivalent classifiers, each with their respective conditional probabilities of FPs, $q^1_{fp}, \ldots, q^n_{fp}$, and FNs, $q^1_{fn}, \ldots, q^n_{fn}$. When a demand from $\mathbf{\Omega}$ occurs, with probability p_i it is classified by classifier i exclusively (where $\sum_{i=1}^{n} p_i = 1$). These define a hybrid configuration of n classifiers, with expected

loss and variance for the hybrid given as

$$\mathbb{E}[L] = \sum_{i=1}^{n} p_i(l_{fp}q^i_{fp}(1-\gamma) + l_{fn}q^i_{fn}\gamma) \tag{1}$$

$$\mathbb{V}[L] = \sum_{i=1}^{n} p_i(l^2_{fp}q^i_{fp}(1-\gamma) + l^2_{fn}q^i_{fn}\gamma) - \left(\sum_{i=1}^{n} p_i(l_{fp}q^i_{fp}(1-\gamma) + l_{fn}q^i_{fn}\gamma)\right)^2 \tag{2}$$

where probability γ is defined in section 3. Note, these formulae are fairly general and do not, for instance, assume failures between classifiers are statistically independent. The set of hybrids implied by (1) and (2) is, in a visually striking sense, the "rotation" of the analogous set of portfolios, where both sets are constructed by convex combinations of classifiers/assets respectively. For instance, see the characteristic "aardvark" (or "bullet") silhouettes in Fig.s 3a and 3b.

4.2 Trade-off Implications for Optimal Adjudication Amongst Diverse Classifiers

If trustworthy estimates of (conditional) expected loss can be obtained for diverse classifiers, then there are ways of combining the classifier responses into responses that minimise expected loss. A given rule for combining classifier responses into single responses is a so-called *adjudication function*. Minimising expected loss by using an "optimal" adjudication function is possible because: 1) the collective responses of a group of classifiers partition $\mathbf{\Omega}$ into disjoint subsets, and 2) on each of these subsets, a response can be chosen that minimizes the conditional expected loss for a subset when demands from this subset occur. In this section, we investigate the relationship between the extreme distributions of Fig. 2 and the loss distribution for a classifier that uses optimal adjudication.

Optimal adjudication has long been advocated in various forms [3, 7, 11]. To illustrate, consider only two classifiers. Their responses partition $\mathbf{\Omega}$ into disjoint subsets (see Fig. 4). Let X_1 be classifier 1's response. Then the two events, "no alarm" $[X_1 = 0]$ and "alarm" $[X_1 = 1]$, divide $\mathbf{\Omega}$ into two disjoint subsets. Similarly, classifier 2's responses split $\mathbf{\Omega}$ into two disjoint subsets. Altogether, (X_1, X_2) divides $\mathbf{\Omega}$ into 4 regions – the subsets of $\mathbf{\Omega}$ that trigger the responses $(1,1), (1,0), (0,1)$ and $(0,0)$, labelled $\boldsymbol{R}_{\mathbf{1},\mathbf{1}}, \boldsymbol{R}_{\mathbf{1},\mathbf{0}}, \boldsymbol{R}_{\mathbf{0},\mathbf{1}}$ and $\boldsymbol{R}_{\mathbf{0},\mathbf{0}}$ respectively.

On each "$\boldsymbol{R}$" the classifiers' responses can be combined into a single response in one of two ways: either issue an alarm or no alarm. An adjudication function is a choice of response on each "$\boldsymbol{R}$". There are 16 possible adjudication functions, $f_1 \ldots f_{16}$ (see Table 1). An "optimal" adjudication function minimises expected loss. Define the expected loss $\mathbb{E}_f[L]$ from adjudication function f as

$$\mathbb{E}_f[L] = l_{fp} \cdot P(f(X_1, X_2) = 1, \text{FP error}) \; + \; 1 \cdot P(f(X_1, X_2) = 0, \text{FN error})$$

To determine an optimal f, one computes two expectations over each "$\boldsymbol{R}$" – the expected loss due to an FP error, and that due to an FN error. In total, over the

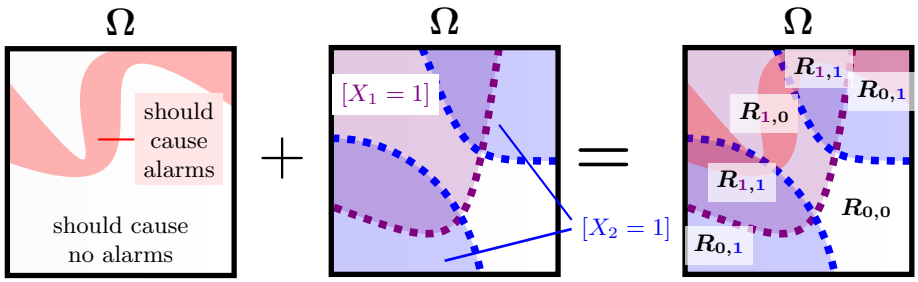


Fig. 4: Two classifiers each give a binary response – alarm "1" or no alarm "0" – upon receiving a demand. There are two ways for $\boldsymbol{\Omega}$ to be partitioned into subsets: 1) The demands divide $\boldsymbol{\Omega}$ into two disjoint subsets; 2) The pair of responses $(X_1(\omega), X_2(\omega))$ that classifiers 1 and 2 give for each demand $\omega \in \boldsymbol{\Omega}$ also partitions $\boldsymbol{\Omega}$, into 4 disjoint subsets labelled $\boldsymbol{R_{i,j}} := \{\omega \in \boldsymbol{\Omega} \,|\, (X_1(\omega), X_2(\omega)) = (i, j) \text{ where } i, j = 0, 1\}$.

Table 1: The 16 adjudication functions for 2 binary classifiers.

$\boldsymbol{R}$	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}
$(\mathbf{1},\mathbf{1})$	0	1	0	0	0	1	1	1	0	0	0	1	1	1	0	1
$(\mathbf{1},\mathbf{0})$	0	0	1	0	0	1	0	0	1	1	0	1	1	0	1	1
$(\mathbf{0},\mathbf{1})$	0	0	0	1	0	0	1	0	1	0	1	1	0	1	1	1
$(\mathbf{0},\mathbf{0})$	0	0	0	0	1	0	0	1	0	1	1	0	1	1	1	1

4 regions, 8 expectations are computed[3]. Using these expectations, the optimal f is simply a choice of responses that give the smallest expected losses for each region. The worst adjudication possible is given by choosing responses which give the largest expected losses instead. This can be done using a table such as Table. 2. Two numerical examples of Table. 2 are illustrated in Table. 3. In either example, $l_{fp} = 0.3$, $l_{fn} = 1$ are the normalised losses. These examples are based on the two probability distributions in Table. 4 for the various $\boldsymbol{R}$ regions in Fig. 4. They show that f_{10} and f_7 are the optimal and worst adjudications for scenario 1, while f_{13} and f_4 are the optimal and worst adjudications for scenario 2. The $\boldsymbol{R}$ distributions were generated from a Dirichlet distribution that randomly assigned probabilities to the regions.

In scenario 1 of Table. 3, notice how the optimal response for region $\boldsymbol{R_{11}}$ suggests that one should risk FN failures rather than FP ones. Since, from Table. 4, the chances of demands causing FN errors is very small (approx. 0.008) compared to the chances of demands causing FP errors (approx. 0.112). However, for scenario 2, the probabilities are roughly the same order of magnitude

[3] From these, the expected loss for *any* of the adjudication functions may be computed.

Table 2: The 8 expected losses for the adjudication functions in Table. 1.

$\boldsymbol{R}$	
$(\mathbf{1},\mathbf{1})$	$\frac{P(f(\mathbf{1},\mathbf{1})=1 \text{ \& demand should cause no alarm}) \cdot l_{fp}}{P(f(\mathbf{1},\mathbf{1})=\mathbf{0} \text{ \& demand should cause alarm})}$
$(\mathbf{1},\mathbf{0})$	$\frac{P(f(\mathbf{1},\mathbf{0})=1 \text{ \& demand should cause no alarm}) \cdot l_{fp}}{P(f(\mathbf{1},\mathbf{0})=\mathbf{0} \text{ \& demand should cause alarm})}$
$(\mathbf{0},\mathbf{1})$	$\frac{P(f(\mathbf{0},\mathbf{1})=1 \text{ \& demand should cause no alarm}) \cdot l_{fp}}{P(f(\mathbf{0},\mathbf{1})=\mathbf{0} \text{ \& demand should cause alarm})}$
$(\mathbf{0},\mathbf{0})$	$\frac{P(f(\mathbf{0},\mathbf{0})=1 \text{ \& demand should cause no alarm}) \cdot l_{fp}}{P(f(\mathbf{0},\mathbf{0})=\mathbf{0} \text{ \& demand should cause alarm})}$

Table 3: Two example scenarios of Table 2, with expected losses for each $\boldsymbol{R}$ subset of $\boldsymbol{\Omega}$ (see Fig. 4). In each scenario, for each subset, the adjudication response associated with the smallest expected loss is the optimal response, while the response associated with largest expected loss is the worst response.

$\boldsymbol{R}$	**scenario 1**	**scenario 2**
$(\mathbf{1},\mathbf{1})$	$3.36e{-}2$, worst response $= 1$	$6.781e{-}2$, optimal response $= 1$
	$8.018e{-}3$, optimal response $= 0$	$7.408e{-}2$, worst response $= 0$
$(\mathbf{1},\mathbf{0})$	$1.302e{-}2$, optimal response $= 1$	$2.081e{-}2$, optimal response $= 1$
	$2.836e{-}1$, worst response $= 0$	$3.248e{-}2$, worst response $= 0$
$(\mathbf{0},\mathbf{1})$	$6.436e{-}2$, worst response $= 1$	$9.374e{-}2$, worst response $= 1$
	$4.696e{-}2$, optimal response $= 0$	$8.729e{-}2$, optimal response $= 0$
$(\mathbf{0},\mathbf{0})$	$5.865e{-}2$, optimal response $= 1$	$1.553e{-}2$, optimal response $= 1$
	$9.599e{-}2$, worst response $= 0$	$1.465e{-}2$, worst response $= 0$

(approx. 0.07 and 0.23 respectively). Hence, since l_{fp} is much smaller than l_{fn} in both scenarios, one expects more loss from FN errors than FP errors for $\boldsymbol{R}_{\mathbf{11}}$

in scenario 2. This reversal illustrates how optimal adjudication depends on the likelihood of the various regions in Fig. 4, and the relative sizes of l_{fp} and l_{fn}.

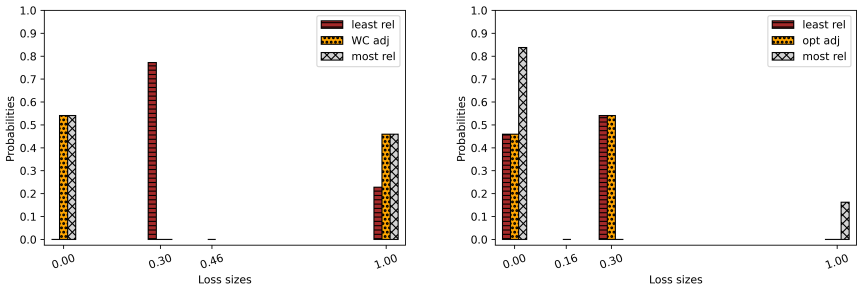
Table 4: Two probability distributions, used respectively to compute the expectations in Table. 3, for the regions in Fig. 4.

$\boldsymbol{R}$		**scenario 1**	**scenario 2**
$(\mathbf{1},\mathbf{1})$	$P(\mathbf{1},\mathbf{1},$ demand should cause no alarm)	0.112005	0.226038
	$P(\mathbf{1},\mathbf{1},$ demand should cause alarm)	0.008017	0.074076
$(\mathbf{1},\mathbf{0})$	$P(\mathbf{1},\mathbf{0},$ demand should cause no alarm)	0.043412	0.069367
	$P(\mathbf{1},\mathbf{0},$ demand should cause alarm)	0.283563	0.032477
$(\mathbf{0},\mathbf{1})$	$P(\mathbf{0},\mathbf{1},$ demand should cause no alarm)	0.214545	0.312464
	$P(\mathbf{0},\mathbf{1},$ demand should cause alarm)	0.046961	0.087287
$(\mathbf{0},\mathbf{0})$	$P(\mathbf{0},\mathbf{0},$ demand should cause no alarm)	0.195507	0.051752
	$P(\mathbf{0},\mathbf{0},$ demand should cause alarm)	0.095986	0.146535

Are optimal configurations very reliable ones, or do they trade-off reliability in favour of making FNs very unlikely? Figure 5 shows a randomly generated example (i.e. a Dirichlet distributed assignment of the probabilities over the regions in Fig. 4) where the worst adjudication is similar to the most reliable classifier with the same expected loss, while optimal adjudication is similar to the least reliable classifier. But how big can the difference be between an optimal adjudicator and the most reliable classifier with the same expected loss? Fig. 6 depicts empirical distributions of the ratio between the accuracy of a randomly chosen extreme adjudication function (over $\boldsymbol{\Omega}$ in Fig. 4) and the accuracy for an extreme loss distribution with the same expected loss. To generate these empirical ratio distributions, 100,000 (Dirichlet distributed) distributions over the $\boldsymbol{R}$ regions in Fig. 4 were sampled. In particular, Fig. 6b shows that in applications where FNs are very rare and $l_{fp} \ll l_{fn}$, the most reliable system can be over two orders of magnitude more reliable than the optimal adjudication configuration with the same expected loss.

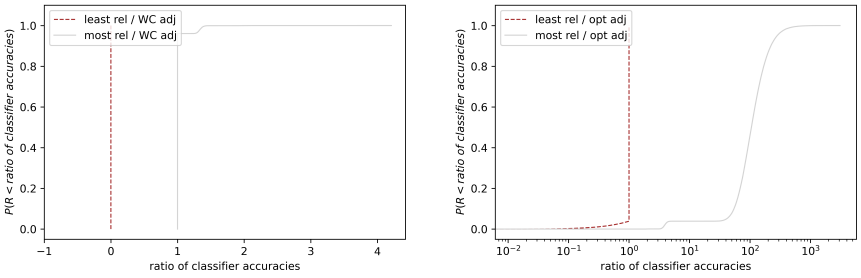
5 Conservative Bayesian Assessment

In this section, we explicitly account for an assessor's ignorance when assessing a classifier. Our proposed approach is Bayesian, and a novel extension of CBI



(a) (b)

Fig. 5: A Dirichlet distributed assignment of probabilities over the $\boldsymbol{R}$ regions (in Fig. 4) produced loss distributions for the worst-case and optimal adjudication functions. These are compared with the extreme loss distributions (in fig.2) that have the same expected losses. In **(a)**, with expected loss 0.46, the worst adjudication is identical to the most reliable classifier with the same expected loss. While **(b)**, with expected loss 0.16, shows the optimal adjudication distribution is identical to the distribution for the least reliable classifier. In both plots, $l_{fp} = 0.3$ and $l_{fn} = 1$.



(a) (b)

Fig. 6: From 100,000 randomly chosen distributions over the $\boldsymbol{R}$ regions in Fig. 4, with FNs unlikely and $l_{fp} \ll l_{fn}$, **(a)** a randomly chosen worst adjudicator (WC adj) gives approximately the same accuracy as the most reliable classifier with the same expected loss. But, **(b)** the most reliable classifier can have "orders of magnitude" greater accuracy than the corresponding, randomly chosen, optimal adjudicator (opt adj).

methods [2, 18, 30, 35–37]. Rather than a completely specified prior distribution, CBI requires only a partial specification of the prior, such as specifying the value of the prior *probability of the classifier passing n tests*. There is a collection of all prior distributions consistent with this probability value. And from this collection, one determines a prior distribution that yields the most undesirable value for a posterior reliability measure of interest (one example measure is the posterior *probability of correctly classifying the next n demands*).

CBI encourages, but does not require, an assessor to be "minimalist" in their application of Bayesian methods. By only providing partial specifications of priors, the assessor can base their assessment only on those properties of a prior they are confident in demonstrating. But, should an assessor completely specify a prior, then CBI becomes identical to traditional Bayesian inference.

Our CBI variant differs from previous ones in two respects. First, it concerns estimating the probability that a known-to-be-imperfect system will have a "perfect run" on the next n demands. Many of the other CBI applications have dealt with systems where *any* failure during testing is unacceptable (i.e. systems with ultra-high reliability requirements). In contrast, classifiers are often known to be imperfect, upon observing them commit FP/FN errors. We determine the worst-case value for the posterior probability of a classifier correctly executing n tasks, after observing the classifier fail "k" out of "$k+r$" tasks.

Secondly, we combine CBI analyses and the extremes implied by the loss-size/reliability trade-off, to obtain conservative measures of reliability and risk that account for this trade-off. Previous applications of CBI do not consider the "impact" of failures explicitly – this new variant does.

More formally, consider estimating the unknown probability U that the classifier fails its next classification (i.e. $u = q_{fp}(1-\gamma) + q_{fn}\gamma$). During operational testing, suppose the classifier fails "k" out of "$k+r$" classification tasks. Furthermore, suppose we have some estimate μ of the probability of seeing these results (e.g. this probability takes the form $\mathbb{E}[U^k(1-U)^r]$ for a sequence of classifications made by the classifier, where these classification tasks arise according to some Bernoulli process with unknown parameter U). Assuming classification tasks follow a Bernoulli process, what is the largest probability of the classifier failing the next classification after operational testing?

Proposition 1. *Consider the set $\mathcal{D}$ of all probability distributions over the unit interval, each distribution representing a prior distribution of U, where U is the unknown probability of a classification failure. For $\mu \in (0,1)$ and $k, r > 0$,*

$$\begin{aligned} \underset{F \in \mathcal{D}}{\textit{maximise}} \quad & \mathbb{E}[\, U \mid k \textit{ failures \& } r \textit{ successes}] \\ \textit{subject to} \quad & \mathbb{E}[U^k(1-U)^r] = \mu \end{aligned}$$

Solution: Appendix B proves the existence of a "single point" prior $F^* \in \mathcal{D}$ – it assigns probability 1 to the unique $u^* \in [\frac{k}{k+r}, 1]$ that satisfies $\mu = (u^*)^k(1-u^*)^r$. Upon using F^* as a prior, $\mathbb{E}[\, U \mid k$ failures & r successes] attains its maximum value, u^*. So, the conservative choice of loss distribution for the classifier would be *any* distribution in $\mathcal{L}$ satisfying $U = u^*$. Note, F^* is unique "*almost everywhere*" [27] (i.e. up to Lebesgue measure zero subsets of $(0,1)$). ■

This result has a number of consequences. The following are limiting cases:

- before any observations (so $k, r = 0$), we must have $\mu = 1$ (i.e. *any* value for the unknown probability of failure U is possible). So, the conservative value u^* must also be 1 – the assessor should conservatively expect the classifier to fail on its next task, in the absence of any evidence to the contrary;

- if no successes are observed (so $r = 0$) then $u^* = 1$: the classifier must be expected to fail its next task, in the absence of any evidence to the contrary;
- if no failures are observed (so $k = 0$) then $u^* = 1 - \mu^{1/r}$. And, as $r \to \infty$, $u^* \to 0$. That is, despite being conservative, with increasing failure free evidence the assessor becomes more convinced that the classifier is "perfect";

Note that the *maximum likelihood estimate* (MLE) for U is $\frac{k}{k+r}$. Since $u^* \in [\frac{k}{k+r}, 1]$, this reassures us that u^* *is* worse than the corresponding MLE. Working backwards, this also means that $\mu \leqslant (\frac{k}{k+r})^k(1-\frac{k}{k+r})^r$ necessarily – a requirement that our assessor can easily use to check the feasibility of their initial μ estimate.

Using this conservative value u^* in the distributions of Fig.s 2b and 2d, we can reason about the possible expected loss values this probability implies. The largest expected loss implied by u^* is u^* (using Fig. 2d). The smallest expected loss is $u^* l_{fp}$ (using Fig. 2b). The case in Fig. 2c applies only if $u^* = 1$; i.e. the classifier *will* fail, the only question is how big the losses will be. Conservatively, one should expect the losses to be as large as possible, so $\mathbb{E}[L] = 1$.

Interestingly, in those cases where we deduced expected losses from Fig.s 2b and 2d, the probability of the classifier being correct is $1-u^*$. This is the smallest plausible value for classifier accuracy, given the assessor's prior evidence μ.

Of course, an assessor might also look to more optimistic assessments. Nevertheless, CBI still offers a useful check against dangerously optimistic assessments. CBI reliability estimates – when compared with estimates from alternative, similarly constrained, approaches – reveal how optimistic these alternatives are. In fact, appendix C proves the following "most optimistic" posterior expected value for U, in a Bernoulli process of classifications.

Proposition 2. *Consider the set $\mathcal{D}$ of all probability distributions over $[0, 1]$, each distribution representing a potential prior distribution of U, where U is the unknown probability of a classification failure. For $\mu \in (0, 1)$ and $k, r > 0$,*

$$\begin{aligned} \underset{F \in \mathcal{D}}{\textit{minimise}} \quad & \mathbb{E}[\, U \mid k \textit{ failures } \& \textit{ } r \textit{ successes}] \\ \textit{subject to} \quad & \mathbb{E}[U^k(1-U)^r] = \mu \end{aligned}$$

Solution: Appendix C proves the existence of a "single point" prior $F_* \in \mathcal{D}$ – it assigns probability 1 to the unique $u_* \in [0, \frac{k}{k+r}]$ that satisfies $\mu = (u_*)^k(1-u_*)^r$. Upon using F_* as a prior, $\mathbb{E}[\, U \mid k$ failures & r successes$]$ attains its minimum value, u_*. So, the optimistic choice of loss distribution for the classifier would be *any* distribution in $\mathcal{L}$ satisfying $U = u_*$. Estimates of U that are close in value to u_* should be used with caution. Note, F_* is unique "*almost everywhere*". ■

Finally, using u^* and u_*, appendix D proves the worst-case (i.e. smallest) probability of a classifier correctly classifying the next n tasks, having already observed the classifier correctly classify "r" out of "$k + r$" tasks. That is,

Corollary 1. *Consider the previous propositions (proved in appendices B and C) which give the largest and smallest values for $\mathbb{E}[\, U \mid k$ failures & r successes$]$, u^* and u_* respectively. Then,*

$$(1-u^*)^n \leqslant \mathbb{E}[\, (1-U)^n \mid k \textit{ failures } \& \textit{ } r \textit{ successes}] \leqslant 1 - u_*$$

In particular, the lower bound is attained with the prior distribution $F^* \in \mathcal{D}$. So, any loss distribution in $\mathcal{L}$ with $U = u^*$ conservatively characterises the long-run failure behaviour of the classifier. Note that $1 - u^*$ is the smallest plausible (i.e. consistent with the evidence) value for the classifier's unknown accuracy, θ. Consequently, as $n \to \infty$, $(1-u^*)^n$ tends to zero faster than any other plausible value for accuracy. Given the well-known unsuitability of accuracy as a measure of classifier performance in imbalanced dataset settings [5, 16, 17], conservative long-run success probabilities provide an increasingly stringent (as n increases) alternative measure of classifier performance.

6 Conclusions

This work has two main focuses: 1) to explicitly account for various uncertainties inherent in assessing binary classifiers, and 2) to highlight trade-offs between classifier reliability and the size of losses when a classifier fails. These have consequences for deciding which, amongst a collection of classifiers, is most desirable.

Assessment is fraught with uncertainty. And, while ROC techniques address some of these, they do not go far enough. Classifiers fail and incur losses according to some unknown loss distribution, and our work bounds the possible range of such distributions. For example, for classifiers with the same expected loss, the more reliable a classifier is, the larger the losses when it fails – a trade-off.

A consequence of such trade-offs is that hybrid classifiers – i.e. convex combinations of diverse classifiers – can have a reduced risk of a sequence of high-impact failures, but at the expense of reliability. This is akin to how, in modern portfolio theory, diverse risky assets can be combined into one investment portfolio, to reduce investment risk. There are visually arresting parallels (Fig.s 3a and 3b) between the sets in these two scenarios. The trade-offs also have implications for "optimal adjudication" schemes that combine the outputs of diverse classifiers to reduce expected loss. Such schemes can be significantly less reliable than the most reliable classifier with the same "optimal" expected loss.

These trade-offs are a strong argument for using multi-objective optimisations during assessment – it is not enough to only consider the well-known trade-off between sensitivity and specificity. In this sense, our approach furthers the benefits of traditional ROC approaches. It also turns out that the loss distributions for the most reliable classifiers (amongst classifiers with a common expected loss) are those with the largest variation. This strongly suggests parallels with mean-variance optimisation methods used in modern portfolio theory. So, our work also complements techniques employed in modern portfolio theory.

Of course, there are significant differences between assets and classifiers that limit the analogy. For example, assets can be leveraged against each other – borrowing on the one hand to buy an investment on the other hand. Such leveraging does not make sense when randomising amongst classifiers. Also, as of this writing, classifiers are typically "indivisible" and cannot be broken up into smaller functionally equivalent artefacts, unlike many investment assets.

Unlike either ROC approaches or modern portfolio theory, with CBI it is fundamental that an assessor explicitly models their uncertainty, about the occurrence of classifier failures. This must be based on justifiable evidence gathered prior to operational testing. Classifiers are then assessed by carrying out "worst-case" inference, using prior distributions in conjunction with observed reliability.

A number of this paper's results (e.g. propositions 1 and 2) apply more widely (e.g. to multiclass classifiers). But, outstanding challenges remain. For instance, how best to quantify and gain sufficient confidence in prior estimates, such as the prior probability $\mathbb{E}[U^k(1-U)^r]$ of observing classifier failures? Or, investigating whether analogous trade-offs hold for metrics other than $\mathbb{E}[L]$ and reliability, such as F-measures or surrogate estimates of loss [1, 6]. Investigating settings where classifications and likelihoods arise according to processes more general than Bernoulli ones. And, explicitly accounting for how classifier performance can evolve and change over time (e,g, due to "learning" or patching) – currently, the techniques we have outlined apply inbetween changes in classifier performance.

7 Appendices And Supplementary Material

For all of the proofs, please see this paper's appendices online, in Springer's Electronic Supplementary Materials (ESM) system.

Acknowledgment

This work was supported by the European Commission through the H2020 programme under grant agreement 700692 (DiSIEM). My thanks to the anonymous reviewers for their helpful suggestions for improving the presentation.

References

1. Bartlett, P., Jordan, M., McAuliffe, J.: Convexity, classification, and risk bounds. Journal of the American Statistical Association **101**, 138–156 (02 2006). https://doi.org/10.1198/016214505000000907
2. Bishop, P., Bloomfield, R., Littlewood, B., Povyakalo, A., Wright, D.: Toward a formalism for conservative claims about the dependability of software-based systems. IEEE Transactions on Software Engineering **37**(5), 708–717 (2011)
3. Blough, D.M., Sullivan, G.F.: A comparison of voting strategies for fault-tolerant distributed systems. In: Proceedings Ninth Symposium on Reliable Distributed Systems. pp. 136–145 (Oct 1990). https://doi.org/10.1109/RELDIS.1990.93959
4. Box, G.E., Tiao, G.C.: Nature of Bayesian Inference, chap. 1, pp. 1–75. John Wiley & Sons, Ltd (2011). https://doi.org/10.1002/9781118033197.ch1
5. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks **106**, 249 – 259 (2018). https://doi.org/https://doi.org/10.1016/j.neunet.2018.07.011, `http://www.sciencedirect.com/science/article/pii/S0893608018302107`
6. Dembczyński, K., Kotłowski, W., Koyejo, O., Natarajan, N.: Consistency analysis for binary classification revisited. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine

Learning Research, vol. 70, pp. 961–969. PMLR, International Convention Centre, Sydney, Australia (06–11 Aug 2017), http://proceedings.mlr.press/v70/dembczynski17a.html
7. Di Giandomenico, F., Strigini, L.: Adjudicators for diverse-redundant components. In: Proceedings Ninth Symposium on Reliable Distributed Systems. pp. 114–123 (Oct 1990). https://doi.org/10.1109/RELDIS.1990.93957
8. Fawcett, T.: An introduction to roc analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006), http://dx.doi.org/10.1016/j.patrec.2005.10.010
9. Fawcett, T., Flach, P.A.: A response to webb and ting's on the application of roc analysis to predict classification performance under varying class distributions. Mach. Learn. **58**(1), 33–38 (Jan 2005). https://doi.org/10.1007/s10994-005-5256-4, https://doi.org/10.1007/s10994-005-5256-4
10. Flach, P., Shaomin, W.: Repairing concavities in roc curves. In: Unknown. pp. 702 – 707. IJCAI (Aug 2005), conference Proceedings/Title of Journal: Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05)
11. Gaffney, J.E., Ulvila, J.W.: Evaluation of intrusion detectors: A decision theory approach. In: Proceedings of the 2001 IEEE Symposium on Security and Privacy. pp. 50–61. IEEE (2001), http://dl.acm.org/citation.cfm?id=882495.884438
12. Gaffney, J.E., Ulvila, J.W.: Evaluation of intrusion detection systems. J. Res. Natl. Inst. Stand. Technol. **108**(6), 453–473 (2003)
13. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis. Chapman and Hall/CRC, 2nd edn. (2004)
14. Gelman, A., Shalizi, C.R.: Philosophy and the practice of bayesian statistics. British Journal of Mathematical and Statistical Psychology **66**(1), 8–38 (2013). https://doi.org/10.1111/j.2044-8317.2011.02037.x
15. Hand, D.J., Till, R.J.: A simple generalisation of the area under the roc curve for multiple class classification problems. Mach. Learn. **45**(2), 171–186 (Oct 2001). https://doi.org/10.1023/A:1010920819831, https://doi.org/10.1023/A:1010920819831
16. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intell. Data Anal. **6**(5), 429–449 (Oct 2002)
17. Koyejo, O.O., Natarajan, N., Ravikumar, P.K., Dhillon, I.S.: Consistent binary classification with generalized performance metrics. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 2744–2752. Curran Associates, Inc. (2014)
18. Littlewood, B., Salako, K., Strigini, L., Zhao, X.: On reliability assessment when a software-based system is replaced by a thought-to-be-better one. Reliability Engineering & System Safety **197**, 106752 (2020). https://doi.org/https://doi.org/10.1016/j.ress.2019.106752
19. Markowitz, H.M.: Portfolio selection. The Journal of Finance **7**(1), 77–91 (1952)
20. Markowitz, H.M.: Portfolio Selection, Efficient Diversification of Investments. John Wiley and Sons (1959)
21. Nayak, J., Naik, B., Behera, D.H.: A comprehensive survey on support vector machine in data mining tasks: Applications and challenges. International Journal of Database Theory and Application **8**, 169–186 (01 2015). https://doi.org/10.14257/ijdta.2015.8.1.18
22. Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M.P., Shyu, M.L., Chen, S.C., Iyengar, S.S.: A survey on deep learning: Algorithms, techniques, and applications. ACM Comput. Surv. **51**(5) (Sep 2018). https://doi.org/10.1145/3234150, https://doi.org/10.1145/3234150

23. Provost, F., Fawcett, T.: Robust classification systems for imprecise environments. In: Proc. AAAI-98. pp. 706–713. AAAI press (1998)
24. Provost, F., Fawcett, T.: Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In: Proceedings of the Third International Conference on Knowledge Discovery and Data Mining. p. 43–48. KDD'97, AAAI Press (1997)
25. Provost, F., Fawcett, T.: Robust classification for imprecise environments. Mach. Learn. **42**(3), 203–231 (Mar 2001), https://doi.org/10.1023/A:1007601015854
26. RavinderReddy, R., Kavya, B., Yellasiri, R.: A survey on svm classifiers for intrusion detection. International Journal of Computer Applications **98**, 34–44 (07 2014). https://doi.org/10.5120/17294-7779
27. Schilling, R.: Measures, Integrals and Martingales. Cambridge University Press, 2nd edn. (2017)
28. Scott, M.J.J., Niranjan, M., Prager, R.W.: Realisable classifiers: Improving operating performance on variable cost problems. In: Proceedings of the British Machine Vision Conference. pp. 31.1–31.10. BMVA Press (1998). https://doi.org/10.5244/C.12.31
29. Sharpe, W.F.: Capital asset prices: A theory of market equilibrium under conditions of risk. The Journal of Finance **19**(3), 425–442 (1964)
30. Strigini, L., Povyakalo, A.A.: Software fault-freeness and reliability predictions. In: International Conference on Computer Safety, Reliability and Security. vol. 8153, pp. 106–117. Springer (2013)
31. Strigini, L., Wright, D.: Bounds on survival probability given mean probability of failure per demand; and the paradoxical advantages of uncertainty. Reliability Engineering and System Safety **128**, 66–83 (Aug 2014). https://doi.org/10.1016/j.ress.2014.02.004
32. Swets, J., Dawes, R., Monahan, J.: Better decisions through science. Scientific American **283**, 82–7 (11 2000). https://doi.org/10.1038/scientificamerican1000-82
33. Swets, J.A.: Measuring the accuracy of diagnostic systems. Science **240**(4857), 1285–93 (1988)
34. Webb, G., Ting, K.: On the application of roc analysis to predict classification performance under varying class distributions. Machine Learning **58**, 25–32 (01 2005). https://doi.org/10.1007/s10994-005-4257-7
35. Zhao, X., Littlewood, B., Povyakalo, A., Strigini, L., Wright, D.: Modeling the probability of failure on demand (pfd) of a 1-out-of-2 system in which one channel is 'quasi-perfect'. Reliability Engineering & System Safety **158**, 230–245 (2017)
36. Zhao, X., Littlewood, B., Povyakalo, A., Strigini, L., Wright, D.: Conservative claims for the probability of perfection of a software-based system using operational experience of previous similar systems. Reliability Engineering & System Safety **175**, 265 – 282 (2018). https://doi.org/https://doi.org/10.1016/j.ress.2018.03.032
37. Zhao, X., Robu, V., Flynn, D., Salako, K., Strigini, L.: Assessing the safety and reliability of autonomous vehicles from road testing. In: the 30th Int. Symp. on Software Reliability Engineering (ISSRE). IEEE, Berlin, Germany (2019), in press