Taking the Confusion Out of Multinomial Confusion Matrices and Imbalanced Classes

David Lovell[1(✉)], Bridget McCarron[2], Brendan Langfield[3], Khoa Tran[1], and Andrew P. Bradley[1]

[1] Queensland University of Technology, Brisbane, Australia
{David.Lovell,a38.Tran,a6.Bradley}@qut.edu.au
[2] Queensland Health, Brisbane, Australia
Bridget.McCarron@health.qld.gov.au
[3] Services Australia, Brisbane, Australia
Brendan.Langfield@ServicesAustralia.gov.au

Abstract. Classification is a fundamental task in machine learning, and the principled design and evaluation of classifiers is vital to create effective classification systems and to characterise their strengths and limitations in different contexts. Binary classifiers have a range of well-known measures to summarise performance, but characterising the performance of *multinomial classifiers* (systems that classify instances into one of many classes) is an open problem. While *confusion matrices* can summarise the empirical performance of multinomial classifiers, they are challenging to interpret at a glance—challenges compounded when classes are *imbalanced*.

We present a way to decompose multinomial confusion matrices into components that represent the *prior* and *posterior* probabilities of correctly classifying each class, and the intrinsic ability of the classifier to discriminate each class: the *Bayes factor* or *likelihood ratio* of a positive (or negative) outcome. This approach uses the odds formulation of Bayes' rule and leads to compact, informative visualisations of confusion matrices, able to accommodate far more classes than existing methods. We call this method confusR and demonstrate its utility on 2-, 17-, and 379-class confusion matrices. We describe how confusR could be used in the formative assessment of classification systems, investigation of *algorithmic fairness*, and *algorithmic auditing*.

Keywords: Classification · Multiclass · Visualisation · Performance · Fairness · Auditing

1 Introduction

Binary classification systems have a range of performance measures derived from the 2×2 *confusion matrix* produced when a classifier makes predictions about a set of examples whose actual classes are known (Table 1). One dimension of the

Y. Xu et al. (Eds.): AusDM 2021, CCIS 1504, pp. 16–30, 2021.
https://doi.org/10.1007/978-981-16-8531-6_2

confusion matrix (in this paper, the *rows*) relates to the *predicted* class of each example; the other relates to an example's *actual* class. Some performance measures (e.g., *accuracy*, *precision*, *F-score*) depend on the prior abundance of the classes; others—such as the *true* and *false positive rates* from which Receiver Operating Characteristic (ROC) curves are derived—do not [9]. Performance measures that depend on the prior abundance of the classes are especially problematic when classes are *imbalanced* or *skewed*. A trivial example of this is in obtaining an apparent accuracy of 99% from a classifier that always predicts negative, when only 1% of the cases are positive.

Performance measures for binary classifiers are well established in statistics, machine learning and medical decision-making. Not so for *multinomial* classifiers, i.e., systems which classify examples into one of many classes. As organisations seek to develop and deploy these more complex classification systems, there is a growing need for understanding and transparency in model development, as well as a requirement to better understand how the models are operating. This motivates the work that we present here.

Similar to assessment of students' work in educational settings, we can think about performance assessment of classification systems with two ends in mind:

1. *Summative*, in which we wish to have a single measure of performance that we can use to compare and rank different classifiers
2. *Formative*, in which we wish to gain insight into the strengths and limitations of a classification system so we can improve its performance.

We are interested in the latter, noting that in some contexts, understanding and interpretability can trump summative performance: supremely performing models may be blocked from production if they cannot be sufficiently understood. With this in mind, we aim to understand the empirical confusion matrix of a classifier by separating it into components that represent

1. the effect of the prior abundance of different classes in a set of samples presented to the classifier
2. the effect of classifier, i.e., its innate ability to discriminate different classes.

This paper focuses on the visualisation and interpretation of confusion matrices rather than the classification systems that generate them. The design and implementation of multinomial classification systems involves issues such as how to combine the outputs of base classifiers [33], how to set decision thresholds and incorporate misclassification costs [34]. Our hope is that the methods we present here will inform this design and implementation process.

2 Prior Work on Making Sense of Confusion Matrices

We propose a way to visualise the empirical performance of multinomial classification systems using odds ratios to interpret their confusion matrices. Here, we briefly review relevant prior work on these topics.

Table 1. A binary confusion matrix contains the counts of a classifier's predictions in response to a set of examples whose actual classes are known. These counts (TP, FP, FN, TN) are divided by their respective column totals to form the rates TPR, FPR, TNR, FNR. The ratio of true positive rate (TPR) and false positive rate (FPR) is known as the likelihood ratio (or *Bayes factor*) for a positive outcome LR_+), and LR_- is defined similarly. The ratio of LR_+ and LR_- is known as the *diagnostic odds ratio* (DOR) [11].

predicted class	*actual class* positive	*actual class* negative
positive	TP True Positives	FP False Positives
negative	FN False Negatives	TN True Negatives
	Pos	Neg

(a) Binary confusion matrix elements

$$
\begin{aligned}
\text{Pos} &= \text{TP} + \text{FN} \\
\text{Neg} &= \text{FP} + \text{TN} \\
\text{TPR} &= \text{TP}/\text{Pos} \\
\text{FPR} &= \text{FP}/\text{Neg} \\
\text{LR}_+ &= \text{TPR}/\text{FPR} \qquad (1) \\
\text{TNR} &= \text{TN}/\text{Pos} \\
\text{FNR} &= \text{FN}/\text{Neg} \\
\text{LR}_- &= \text{TNR}/\text{FNR} \\
\text{DOR} &= \text{LR}_+/\text{LR}_- \qquad (2)
\end{aligned}
$$

(b) Binary confusion matrix statistics

"The definition of performance measures in the context of multiclass classification is still an open research topic" remarked Jurman *et al.* [15], citing (then) recent reviews [26], empirical comparisons [10] and visualisation strategies [4] before discussing confusion entropy [30] and a multiclass extension of Matthews correlation coefficient [12] as performance measures. That was in 2012. A more recent investigation suggests the topic remains important and unresolved, and highlights issues with performance indices where classes are imbalanced [20] (see also [18]).

In the machine learning domain, single, *summative* measures for comparison or ranking of multinomial classification systems prevail, with micro- and macro-averaging used to combine performance indices for each class versus all others [26,32]. Cohen's Kappa is also used widely, even though it was not originally intended for classification performance measurement, and has a range of problems when used for that purpose [3]. This extensive use of summative classifier performance metrics may reflect the popularity of competitive evaluation in machine learning, with recognition and reward for those whose algorithms outperform all others—noting that such rankings should be interpreted with care [19]. It also probably reflects the inclusion of these metrics in popular machine learning software frameworks (e.g., [16,21]).

Less common are *formative* approaches that seek to understand, and thereby improve the performance of multinomial classification systems. Ren *et al.* [24] tackle this with their carefully designed and evaluated Squares performance visualization system, providing also a comprehensive review of related visualization efforts such as the Confusion Wheel [1]. Squares is designed to be agnostic to performance metrics and focuses on enabling users to explore calibrated probability

scores produced by a classification system in response to test data. Hinterreiter *et al.* [14] propose an interactive system called ConfusionFlow to compare the performances of multinomial classifiers (e.g., during training). In terms of scalability, both Squares and ConfusionFlow were reported to work well with 15–20 classes. Neither approach pays particular attention to class imbalance.

While performance measures like precision, recall, F-score and Area Under the ROC curve (AUC) are popular in machine learning, measures like LR_+ (Eq. 1) and DOR (Eq. 1) are not seen so often, even though they are prominent in medical decision-making and diagnosis [5,11,13]. Next, we show how LR_+ plays a fundamental role in Bayes' rule that cam be applied to interpreting confusion matrices.

3 Factoring the Confusion Matrix Using Class Odds

Sanderson suggests that we can better understand the discriminative performance of a classification model by expressing Bayes' rule in terms of prior odds and Bayes factors [25]. To illustrate this concept, suppose we have a hypothesis (D) that a person actually has a disease, and some evidence (T) about that in the form of a positive test result for that disease. Often we want to know *"if I have a positive test result, what's the chance that I actually have the disease"*; this is known as the positive predictive value, or precision of the test. In terms of probabilities, this is written:

$$\mathrm{P}(\mathrm{D}|\mathrm{T}) = \frac{\mathrm{P}(\mathrm{T}|\mathrm{D})\mathrm{P}(\mathrm{D})}{\mathrm{P}(\mathrm{T}|\mathrm{D})\mathrm{P}(\mathrm{D}) + \mathrm{P}(\mathrm{T}|\overline{\mathrm{D}})\mathrm{P}(\overline{\mathrm{D}})}$$

where T is the event that the test is positive; D is the event that you actually do have the disease, and $\overline{\mathrm{D}}$ is the event that you do not. Sanderson extols the merits of writing this using *odds*:

$$\begin{aligned}\mathrm{O}(\mathrm{D}|\mathrm{T}) &= \mathrm{O}(\mathrm{D})\frac{\mathrm{P}(\mathrm{T}|\mathrm{D})}{\mathrm{P}(\mathrm{T}|\overline{\mathrm{D}})} \\ &= \mathrm{O}(\mathrm{D})\frac{\text{True positive rate}}{\text{False positive rate}}\end{aligned}$$

where the ratio is the *Bayes factor* of the test for a positive result, also known as the *likelihood ratio of a positive outcome*, or LR_+ for short (Eq. 1). This factor represents how our prior odds of having the disease (O(D)) are updated as a result of the test outcome. In other words

$$\begin{aligned}\text{posterior odds} &= \text{prior odds} \times \mathrm{LR}_+ \\ &= \text{prior odds} \times \frac{\text{True positive rate}}{\text{False positive rate}}\end{aligned}$$

In this paper we utilise the realisation that we can visualise these terms on a logarithmic scale, and can exploit the fact that

$$\log(\text{posterior odds}) = \log(\text{prior odds}) + \log(\mathrm{LR}_+) \tag{3}$$

to achieve a graphical presentation in which these odds and LR_+ (the Bayes factor) appear *additively*. This is appealing because these values can be presented in ways that make the most of the human visual system's pre-attentive processing mechanisms [29]. Furthermore, and as we will demonstrate, this strategy can be much more space efficient than the display of 2-dimensional confusion matrices. In reviewing the literature for related work, we learned of Fagan's nomogram [8,13] which is based on similar principles but, as far as we know, has not been used in the interpretation of multinomial confusion matrices.

Since the diagnostic odds ratio (Eq. 2) has a multiplicative relationship with LR_+ and LR_-, we can visualise that additively on a logarithmic scale using the relationship

$$\begin{aligned} \log(\mathrm{DOR}) &= \log(\mathrm{LR}_+) + \log(1/\mathrm{LR}_-) \\ &= \log(\mathrm{LR}_+) - \log(\mathrm{LR}_-). \end{aligned} \tag{4}$$

So far, we have used binary classification to illustrate the odds formulation of Bayes' rule. We can extend this to multinomial classification by summarising a $C \times C$ confusion matrix as C binary *one-versus-all* confusion matrices and presenting the prior and posterior odds and Bayes factor of each class against all others. We will demonstrate this approach in the next section, but begin by visualising a well known binary classification scenario that has proven challenging to interpret

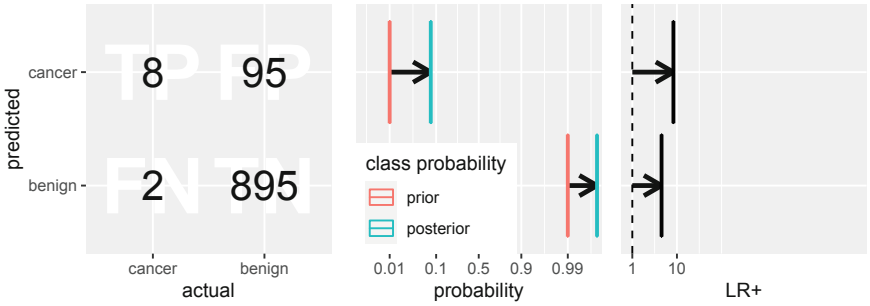
4 Application and Demonstration

We have named our confusion matrix visualisation approach confusR because we have implemented it in R [22], benefiting greatly from the tidyverse suite of packages [31]. To highlight the value of our approach We apply confusR to three increasingly challenging confusion matrices.

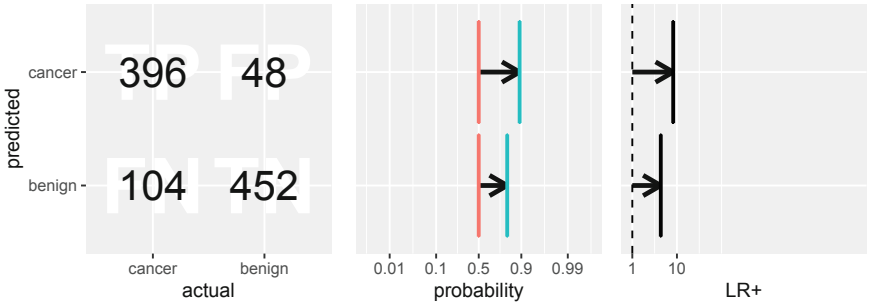
4.1 Eddy's Probabilistic Reasoning Challenge (2 Classes)

Eddy [6] gives an example of a binary diagnostic test for breast cancer where the test had a TPR of 79.2%, an FPR of 9.6% and the prior probability of breast cancer was assumed to be 1%. Eddy found that most physicians (approximately 95 out of 100 in his informal sample) estimated around a 75% probability of someone actually having cancer given that the test predicted that they had. (What is your estimate?)

Figure 1 shows the confusion matrix we would expect if this test was applied to a sample of 1000 people where the prior probability of cancer was 1% and, for comparison, the confusion matrix expected if the same test was applied to a group in whom the prior probability of cancer was much higher (50%). We then use confusR to visualise Eq. 1, the relationship between the prior and posterior odds and LR_+ (the Bayes factor), for both classes of outcome, *cancer* or *benign*.



(a) The confusion matrix (and associated statistics) expected when Eddy's diagnostic test [6] is applied to a sample of 1000 people who have a 1% the prior probability of cancer (the red bar at 0.01 in the top row, middle plot). In this scenario, the probability that someone from this group actually has cancer given that the test predicts that they do is just under 10% (the turquoise bar at 0.096 in the top row, middle plot).



(b) The same test applied to a group of people whose prior probability of cancer is 50% (the red bar at 0.5 in the top row, middle plot). Now the probability that someone from this group actually has cancer given that the test jumps to just under 90% (the turquoise bar at 0.89 in the top row, middle plot).

Fig. 1. Plots of the confusion matrices (left), prior and posterior odds (middle) and likelihood ratios (right) in two different scenarios inspired by the cancer diagnostic test presented by Eddy [6]. Each "row" relates to the *predicted* classes in this binary decision: *cancer* or *benign*. The x-axes of the middle and right plots show the odds and likelihood ratios on a logarithmic scale. (Ticks on the middle plots refer to probabilities for ease of interpretation.) The arrows emphasise that $\log(\text{posterior}) = \log(\text{prior}) + \log(\text{LR}_+)$. Crucially, in both scenarios, the discriminative ability of the test is the same: the right plots of $\log(\text{LR}_+)$ are identical. What differs between the two scenarios is the prior probability of each class (the red bars). On the logarithmic scale used in the middle and right plots, the discriminative ability of the test (LR_+) adds to the prior class odds (red bars) to yield the odds of each class in light of the test's predictions (turquoise bars). (Color figure online)

Table 2. Confusion matrix from Lu *et al.* [17] (Supplementary Information, Source Data Extended Data Fig. 2.)

	actual																
pred	Lung	Brea	Colo	Panc	Skin	Ovar	Rena	Pros	Head	Esop	Thyr	Blad	Germ	Endo	Live	Adre	Cerv
Lung	180	11	0	3	6	2	3	4	3	3	2	3	0	0	2	0	3
Brea	10	194	1	1	3	4	1	0	0	1	2	0	0	1	1	3	3
Colo	3	4	164	1	1	0	0	0	0	2	0	1	0	1	3	1	0
Panc	21	6	6	114	1	2	0	1	2	6	0	2	1	0	0	1	1
Skin	1	4	0	0	90	0	0	1	1	2	0	0	0	0	0	0	0
Ovar	13	5	0	0	2	92	0	0	0	2	0	0	0	5	0	0	0
Rena	0	1	0	0	0	0	70	0	0	0	1	0	2	0	0	0	0
Pros	3	0	0	0	3	0	0	54	0	2	0	0	0	0	0	0	0
Head	1	0	0	0	1	0	2	1	47	0	0	0	0	0	0	0	0
Esop	0	3	2	1	1	0	1	1	3	32	0	1	2	0	0	0	0
Thyr	0	0	0	0	0	0	0	0	0	0	38	0	0	0	0	0	0
Blad	3	2	0	0	3	1	1	2	0	0	0	35	1	0	0	0	0
Germ	0	0	0	0	0	0	0	0	0	1	0	0	26	0	0	0	0
Endo	1	1	0	1	0	1	0	0	0	0	0	0	0	14	7	0	0
Live	0	0	0	1	0	0	1	0	1	0	0	0	0	0	5	0	0
Adre	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0
Cerv	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	4

(a) Full 17-class confusion matrix showing the distribution of predicted against actual origins of cancers for 1408 examples.

	actual			actual			...		actual	
pred	Lung	n.Lung	pred	Brea	n.Brea	...	pred	Cerv	n.Cerv
Lung	180	45	Brea	194	31	...	Cerv	4	3
n.Lung	56	1127	n.Brea	37	1146	...	n.Cerv	7	1394

(b) Three of the 17 binary one-versus-all confusion matrices derived from the full confusion matrix. Each of these tabulates the count of predicted versus actual for a specific class (e.g., `Lung`) against all other classes (denoted by `n.Lung` meaning "not `Lung`").

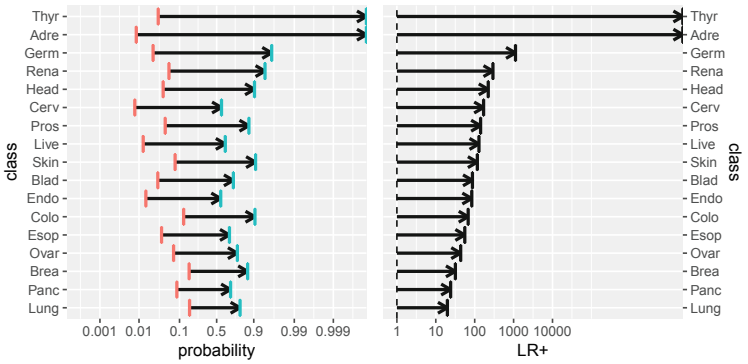
The middle and right panels of Fig. 1 clearly delineate the contribution of the prior probability of each class, the contribution of the classifier's ability to discriminate each class, and the posterior probability of an case actually being from a given class, given that the test's prediction.

Now let us consider a situation with more than two classes.

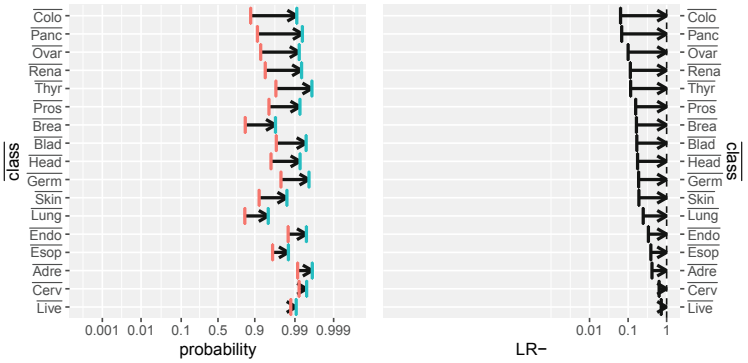
4.2 Cancer of Unknown Primary—CUP (17 Classes)

Lu *et al.* recently developed and evaluated a deep-learning-based system to classify the origin of a cancer primary tumour from histopathology images [17]. Table 2a shows a 17-class confusion matrix from this study and Table 2b shows three of the 17 binary one-versus-all confusion matrices we can derive from it. This number of classes is towards the upper range of what Squares [24] and ConfusionFlow [14] are designed to represent.

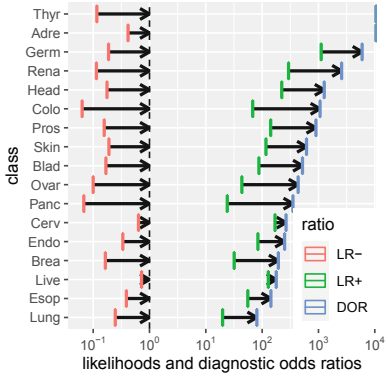
Figure 2a shows a confusR visualisation of the CUP confusion matrix, sorted, by the LR_+, prior and posterior values of the classes. Classes `Thyr` and `Adre` stand out with infinite LR_+ and posterior probabilities, signified by bars at the



(a) Classes sorted by LR_+ (black, right), with prior (red) and posterior (turquoise) probabilities on the left.



(b) Likelihood ratios of a *negative* result for each class, with classes sorted by LR_- (black, right) and prior (red) and posterior (turquoise) probabilities on the left.



(c) Classes sorted by diagnostic odds ratios (blue) using the relationship in Equation 4. The arrows emphasise that $-\log(LR_-)$ is added to $\log(LR_+)$ to get $\log(DOR)$.

Fig. 2. confusR plots of CUP class prior and posterior probabilities, Bayes factors (LR_+, LR_-) and diagnostic odds ratios. (Color figure online)

extreme right of the panels. These two classes had no false positives in the CUP confusion matrix, so their apparent false positive rate and hence denominator of Eq. 1, is 0. This prompts us to look more closely at Table 2a where we can see that there were 12 instances of `Adre` in the data (the `Adre` column total), and none of the 1396 examples from other classes were mistaken for that class. This is a similar outcome for `Thyr` (43 instances) but, in terms of the likelihood ratio of a negative outcome and overall diagnostic odds ratio, we can see that the classification system performs better in discriminating `Thyr` than `Adre`. We will return to the issue of zeros in binary confusion matrices in the next section.

In terms of prior probability, `Lung` is the most abundant class in this dataset (236 instances). (This is more obvious when we sort Fig. 2a by LR_-, which we omit to stay within page limits.) However, it is poorly discriminated by the classification system, having the lowest LR_+ and DOR. This suggests to us that this class merits more attention than other classes in efforts to improve performance.

While it is hard to succinctly describe the multidimensional information encapsulated by an empirical confusion matrix, the confusR visualisations provide a meaningful and accessible visual summary for further consideration. Next we show how this strategy can be extended to much larger numbers of classes, where currently no adequate techniques are available.

4.3 HAndwritten SYmbols—HASY (379 Classes)

Martin Thoma's HASYv2 [27] consists of 32×32 pixel images of HAndwritten SYmbols including "the Latin uppercase and lowercase characters (A-Z, a-z), the Arabic numerals (0–9), 32 different types of arrows, fractal and calligraphic Latin characters, brackets and more", collected from https://detexify.kirelabs.org/classify.html and http://write-math.com/. Thoma has also developed classifiers and published confusion matrices for this 369 class problem available from https://github.com/MartinThoma/algorithms.

Table 3 shows a small section of the training set confusion matrix in numeric form. One way to see all 369 classes is to use a grey scale image or heat map as in Fig. 3a. But with so many classes, it is hard to get a sense of how well the classifier is doing or what the prior abundance of the classes are. Furthermore, humans have difficulty in accurately relating grey scale or colour intensity to quantity [29, p. 168].

In contrast, the confusR visualisation of Fig. 3b makes the most of our visual system's ability to compare point positions pre-attentively, especially when ordering is used to reduce uninformative variation. In this Figure, classes are ordered from bottom to top by LR_+ then prior probability, allowing us to discern some interesting relationships on this training data. As LR_+ increases (from classes `\sum` up to `\n`) we see a rough but noticeable decrease in the prior abundance of classes (red dots). This indicates to us that many of these relatively

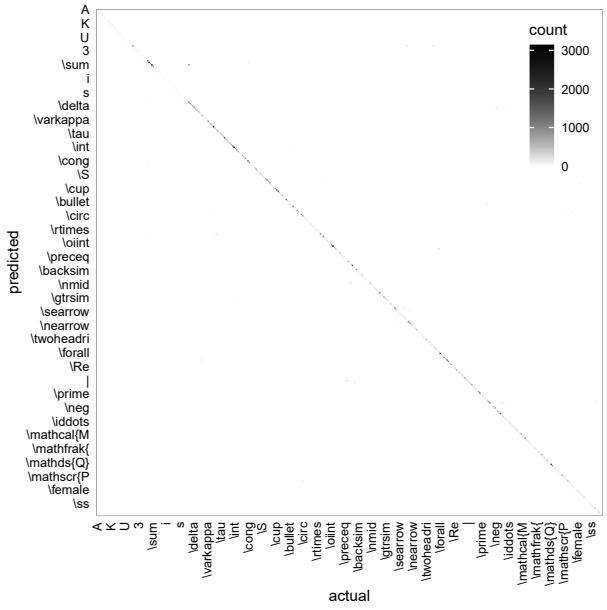
Table 3. Section of HASYv2 training set confusion matrix.

	actual									
predict	\\nu	\\xi	\\Xi	\\Pi	\\rho	\\varrho	\\tau	\\phi	\\Phi	\\varphi
\\nu	344	0	0	0	0	0	0	0	0	0
\\xi	0	2309	0	0	0	0	0	0	0	0
\\Xi	0	0	350	0	0	0	0	0	0	0
\\Pi	0	0	0	451	0	0	0	0	0	0
\\rho	0	0	0	0	622	1	0	0	0	0
\\varrho	0	0	0	0	0	198	0	0	0	0
\\tau	0	0	0	0	0	0	369	0	0	0
\\phi	0	0	0	0	0	0	0	561	13	2
\\Phi	0	0	0	0	0	0	0	25	532	0
\\varphi	0	0	0	0	0	0	0	2	0	1366

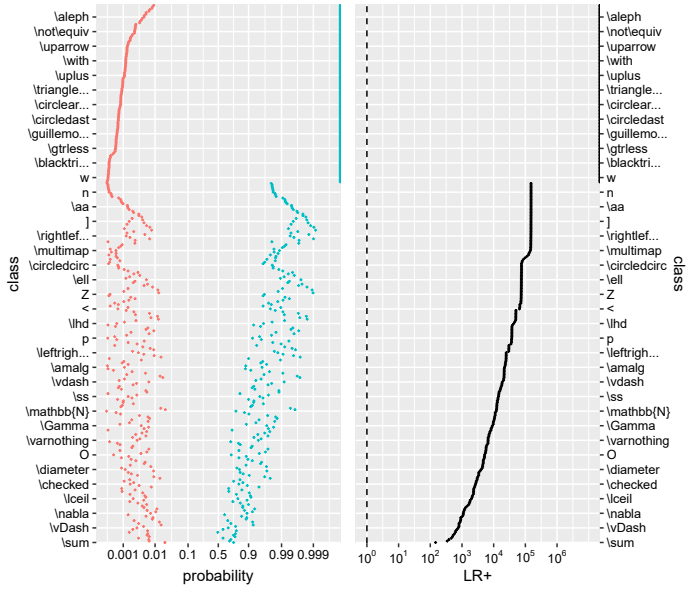
infrequent classes are distinctive to the classifier. There is also a band of classes from `\multmap` to `\n` in which we can clearly see the impact of prior abundance on posterior probabilities while the classifier's ability to discriminate these classes (LR_+) stays steady.

This is a good point to stress that we see these confusR visualisations as *complementary to* the traditional confusion matrix representation rather than a replacement for it. The confusR representation does not show what classes are being confused, but it does give rapid insight into the extent to which classes are being confused, as well as meaningfully factoring apart the role of prior abundance from the classifier's intrinsic ability to discriminate a particular class. We think that this has potential to help developers focus in on more manageable *subsets* of the confusion matrix or data for further attention, perhaps using interactive strategies like Squares [24] or ConfusionFlow [14].

We also see potential for confusR visualisations to compare the empirical performance of a classifier on different datasets, as shown in Fig. 4a which compares confusion matrices from HASYv2 training and test data. The test set confusion matrix has a number of classes for that show zero true positives and/or zero false positives (Fig. 4b), in which case LR_+ is off the scale of Fig. 4a or undefined (when both TP and FP are zero). Figure 4a visualises confusion matrices on two different data sets; this approach could easily be extended to show the distribution of LR_+ values observed across many data sets, e.g., as would occur in cross-validation of classifier performance.



(a) HASYv2 training set confusion matrix as an image in which the count of each element is shown in grey scale.



(b) HASY as a confusR one-vs-all plot

Fig. 3. Two ways of visualising HASYv2 training set confusion matrix from https://github.com/MartinThoma/algorithms. For legibility, only every tenth class label is printed on the y-axes. Classes sorted by LR_+ (black, right), then prior (red) probability with posterior probability in turquoise. (Color figure online)

(a) The performance of Martin Thoma's classifier on the HASYv2 training set and test set (from `https://github.com/MartinThoma/algorithms`). Classes are sorted in order of the empirical LR_+ observed on the training set, highlighting the familiar tendency of classifiers to fit the training data best.

	TP == 0	TP > 0
FP == 0	3 $LR_+ = NaN$	34 $LR_+ = \infty$
FP > 0	12 $LR_+ = 0$	320

(b) Confusion matrices may contain classes for which no true positives (TP) or false positives (FP) are observed, giving rise to LR_+ values that are zero, infinite, or not defined (NaN: Not a Number). This table shows the number of classes where these LR_+ values occurred in the confusion matrix derived from the HASYv2 test set.

Fig. 4. confusR visualisations can be used to compare a classifier's performance on different datasets.

5 Discussion and Conclusions

The odds formulation of Bayes' rule has been around for long time [7], but it does not seem to have been used in visualising the relationship between prior and posterior odds as we have described here with our confusR approach. By putting prior and posterior class odds and Bayes factors onto a logarithmic scale, we

provide a 1-dimensional, compact and readily interpretable representation of a 2-dimensional confusion matrix which allows us to separate the innate ability of the classifier to discriminate different classes from the prior abundance of imbalanced classes. This allows us to deal with much larger confusion matrices than existing visualisation methods [14,24] which are effective with up to 15–20 classes; confusR could serve as a practical way to identify subsets of classes that could be further explored with these approaches.

In addition to enabling greater insight into multinomial classifier performance, we believe confusR may have useful application in the evaluation of *algorithmic fairness*. In their survey of fairness definitions, Verma and Rubin [28] describe *equalized odds* as the situation in which two different groups have the same true positive rates, and the same false positive rates with respect to a predicted outcome, i.e., the same LR_+ (Eq. 1). Equalised odds has been discussed in relation to a binary classifier; the confusR approach could extend this measure of fairness to scenarios where there are more than two classes at play. We see the potential to use confusR in the analysis of classifier performance on different subgroups (e.g., the performance of a medical diagnostic for classifying skin lesions across different skin tones [35]) and as part of algorithmic auditing processes [23].

In terms of future work, we see opportunities to investigate the incorporation of uncertainty (e.g., through simulation or theoretical approaches [2]) into the confusR visualisation approach. Incorporation of decision costs [34] would be a valuable advance, but we suspect this would require more information (i.e., calibrated class probability estimates) than empirical confusion matrices alone could provide.

We are developing an R package that implements the methods presented here, however, the underlying computations are simple and we hope that this paper will provide sufficient information for others to use the confusR concept in making sense of confusion matrices and imbalanced classes.

References

1. Alsallakh, B., Hanbury, A., Hauser, H., Miksch, S., Rauber, A.: Visual methods for analyzing probabilistic classification data. IEEE Trans. Visual Comput. Graphics **20**(12), 1703–1712 (2014). https://doi.org/10.1109/TVCG.2014.2346660
2. Caelen, O.: A Bayesian interpretation of the confusion matrix. Ann. Math. Artif. Intell. **81**(3), 429–450 (2017). https://doi.org/10.1007/s10472-017-9564-8
3. Delgado, R., Tibau, X.A.: Why Cohen's Kappa should be avoided as performance measure in classification. PLoS ONE **14**(9), e0222916 (2019). https://doi.org/10.1371/journal.pone.0222916
4. Diri, B., Albayrak, S.: Visualization and analysis of classifiers performance in multi-class medical data. Expert Syst. Appl. **34**(1), 628–634 (2008). https://doi.org/10.1016/j.eswa.2006.10.016
5. Dujardin, B., Van den Ende, J., Van Gompel, A., Unger, J.P., Van der Stuyft, P.: Likelihood ratios: a real improvement for clinical decision making? Eur. J. Epidemiol. **10**(1), 29–36 (1994). https://doi.org/10.1007/BF01717448

6. Eddy, D.M.: Probabilistic reasoning in clinical medicine: Problems and opportunities. In: Tversky, A., Kahneman, D., Slovic, P. (eds.) Judgment under Uncertainty: Heuristics and Biases, pp. 249–267. Cambridge University Press, Cambridge (1982). https://doi.org/10.1017/CBO9780511809477.019
7. Etz, A., Wagenmakers, E.J.: J. B. S. Haldane's contribution to the bayes factor hypothesis test. Stat. Sci. **32**(2), 313–329 (2017)
8. Fagan, T.: Nomogram for bayes's theorem. N. Engl. J. Med. **293**(5), 257–257 (1975). https://doi.org/10.1056/NEJM197507312930513
9. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006). https://doi.org/10.1016/j.patrec.2005.10.010
10. Ferri, C., Hernández-Orallo, J., Modroiu, R.: An experimental comparison of performance measures for classification. Pattern Recogn. Lett. **30**(1), 27–38 (2009). https://doi.org/10.1016/j.patrec.2008.08.010
11. Glas, A.S., Lijmer, J.G., Prins, M.H., Bonsel, G.J., Bossuyt, P.M.M.: The diagnostic odds ratio: a single indicator of test performance. J. Clin. Epidemiol. **56**(11), 1129–1135 (2003). https://doi.org/10.1016/S0895-4356(03)00177-X
12. Gorodkin, J.: Comparing two K-category assignments by a K-category correlation coefficient. Comput. Biol. Chem. **28**(5), 367–374 (2004). https://doi.org/10.1016/j.compbiolchem.2004.09.006
13. Grimes, D.A., Schulz, K.F.: Refining clinical diagnosis with likelihood ratios. The Lancet **365**(9469), 1500–1505 (2005). https://doi.org/10.1016/S0140-6736(05)66422-7
14. Hinterreiter, A., et al.: ConfusionFlow: a model-agnostic visualization for temporal analysis of classifier confusion. IEEE Trans. Visualization Comput. Graph., 1 (2020). https://doi.org/10.1109/TVCG.2020.3012063
15. Jurman, G., Riccadonna, S., Furlanello, C.: A comparison of MCC and CEN error measures in multi-class prediction. PLoS ONE **7**(8) (2012). https://doi.org/10.1371/journal.pone.0041882
16. Kuhn, M.: Building predictive models in r using the caret package. J. Stat. Softw. Articles **28**(5), 1–26 (2008). https://doi.org/10.18637/jss.v028.i05
17. Lu, M.Y., et al.: AI-based pathology predicts origins for cancers of unknown primary. Nature **594**(7861), 106–110 (2021). https://doi.org/10.1038/s41586-021-03512-4
18. Luque, A., Carrasco, A., Martín, A., de las Heras, A.: The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recogn. **91**, 216–231 (2019). https://doi.org/10.1016/j.patcog.2019.02.023
19. Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., et al.: Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. **9**(1), 5217 (2018). https://doi.org/10.1038/s41467-018-07619-7
20. Mullick, S.S., Datta, S., Dhekane, S.G., Das, S.: Appropriateness of performance indices for imbalanced data classification: an analysis. Pattern Recogn. **102** (2020). https://doi.org/10.1016/j.patcog.2020.107197
21. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
22. R Core Team: R: A language and environment for statistical computing. Technical report, Vienna, Austria (2020). https://www.R-project.org/, R Foundation for Statistical Computing

23. Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., et al.: Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 33–44. ACM, Barcelona (2020). https://doi.org/10.1145/3351095.3372873
24. Ren, D., Amershi, S., Lee, B., Suh, J., Williams, J.D.: Squares: supporting interactive performance analysis for multiclass classifiers. IEEE Trans. Visual Comput. Graphics **23**(1), 61–70 (2017). https://doi.org/10.1109/TVCG.2016.2598828
25. Sanderson, G.: The medical test paradox: can redesigning Bayes rule help? (2020). https://www.youtube.com/watch?v=lG4VkPoG3ko
26. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. **45**(4), 427–437 (2009). https://doi.org/10.1016/j.ipm.2009.03.002
27. Thoma, M.: The HASYv2 dataset. arXiv:1701.08380 [cs] (2017)
28. Verma, S., Rubin, J.: Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness, FairWare 2018, pp. 1–7. ACM, New York (2018). https://doi.org/10.1145/3194770.3194776
29. Ware, C.: Information visualization: perception for design. Interactive technologies, 3rd edn. Morgan Kaufmann, Waltham (2013)
30. Wei, J.M., Yuan, X.J., Hu, Q.H., Wang, S.Q.: A novel measure for evaluating classifiers. Expert Syst. Appl. **37**(5), 3799–3809 (2010). https://doi.org/10.1016/j.eswa.2009.11.040
31. Wickham, H., et al.: Welcome to the tidyverse. J. Open Source Softw. **4**(43), 1686 (2019). https://doi.org/10.21105/joss.01686
32. Wu, X.Z., Zhou, Z.H.: A unified view of multi-label performance measures. arXiv:1609.00288 [cs] (2017)
33. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2002, pp. 694–699. Association for Computing Machinery, New York (2002). https://doi.org/10.1145/775047.775151
34. Zhou, Z.H., Liu, X.Y.: On multi-class cost-sensitive learning. In: Proceedings of the 21st National Conference on Artificial Intelligence, AAAI 2006, vol. 1, pp. 567–572. AAAI Press, Boston (2006)
35. Zicari, R.V., Ahmed, S., Amann, J., Braun, S.A., Brodersen, J., et al.: Co-design of a trustworthy AI system in healthcare: deep learning based skin lesion classifier. Front. Hum. Dyn. **3**, 40 (2021). https://doi.org/10.3389/fhumd.2021.688152