scientific reports

OPEN

Classification performance assessment for imbalanced multiclass data

Jesús S. Aguilar-Ruiz[1,3]✉ & Marcin Michalak[2,3]

The evaluation of diagnostic systems is pivotal for ensuring the deployment of high-quality solutions, especially given the pronounced context-sensitivity of certain systems, particularly in fields such as biomedicine. Of notable importance are predictive models where the target variable can encompass multiple values (multiclass), especially when these classes exhibit substantial frequency disparities (imbalance). In this study, we introduce the Imbalanced Multiclass Classification Performance (IMCP) curve, specifically designed for multiclass datasets (unlike the ROC curve), and characterized by its resilience to class distribution variations (in contrast to accuracy or F_β-score). Moreover, the IMCP curve facilitates individual performance assessment for each class within the diagnostic system, shedding light on the confidence associated with each prediction—an aspect of particular significance in medical diagnosis. Empirical experiments conducted with real-world data in a multiclass context (involving 35 types of tumors) featuring a high level of imbalance demonstrate that both the IMCP curve and the area under the IMCP curve serve as excellent indicators of classification quality.

Measuring the accuracy of diagnostic systems[1] is an old challenge. In general, a suitable measure of performance must be independent of event frequencies and decision criteria. In machine learning, the classification task deals with assigning a class label to a given observation. To do that, a classification model has been previously learned with training data. On many occasions, the distribution of samples in known classes is skewed, i.e. the proportion of examples that belong to each class is unequal. Class distribution refers to the relative frequency of examples of each class in the dataset, and when these frequencies are not comparable, it is named class imbalance. In the recent scientific literature, we find numerous examples of binary class imbalance (breast cancer prediction[2,3]; cognitive impairment in Parkinson's disease[4]) and multi-class imbalance (31 cancer type classification[5]; identification of Parkinson's disease subtypes[6]; microbiome-based cancer diagnostics[7]; prostate cancer detection[8]; dementia subtype prediction[9]; Gleason grading of prostate cancer[10]; severity of COVID-19 pneumonia[11]; faecal microbiome-based multi-class disease diagnosis[12]).

Class imbalance is another challenge for classification assessment, as most of the evaluation metrics are sensitive to data distribution, i.e. to class sizes. Let D be a dataset with n samples and two classes (Positive and Negative). The confusion matrix is represented by a 2×2 matrix with True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN), as illustrated in Fig. 1 (left: Confusion Matrix). The sum of all positives (negatives) is P (N). The matrix can be normalized by dividing by the number of samples $n = \mathrm{P} + \mathrm{N}$, and all the values can be expressed in terms of the True Positive Rate (TPR), the False Positive Rate (FPR) and the proportion of positives (π), as shown in Fig. 1 (right: Normalized Confusion Matrix). Thus, the sum of all positives is $n\pi$, and of all negatives is $n(1-\pi)$. Both sensitivity (TPR) and specificity (1 − FPR) are not influenced by class distribution (π). However, the accuracy measure contains π in its mathematical expression, i.e. it is highly dependent on class bias, so it is exclusively valid for balanced datasets. Despite the uselessness of accuracy for imbalanced data[13], it is still the most popular evaluation measure for classification. Precision, another well–known measure that calculates the ratio between true positive predictions and all the positive predictions, is also sensitive to class skew (expressed in function of the normalized confusion matrix, it also includes π), and thus F_β–score[14,15], which combines precision and sensitivity, also has the same drawback. Therefore, any measure without the presence of π, like the geometric mean between sensitivity and specificity (G-measure)[16], is not influenced by the imbalance between the classes.

In order to mitigate imbalance in datasets two main techniques are approached: under-sampling and over-sampling. Under-sampling[17–20] means reducing the size of the data to equal the size of the minority (lowest

[1]School of Engineering, Pablo de Olavide University, 41013 Seville, Spain. [2]Department of Computer Networks and Systems, Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland. [3]These authors contributed equally: Jesús S. Aguilar-Ruiz and Marcin Michalak. ✉email: aguilar@upo.es

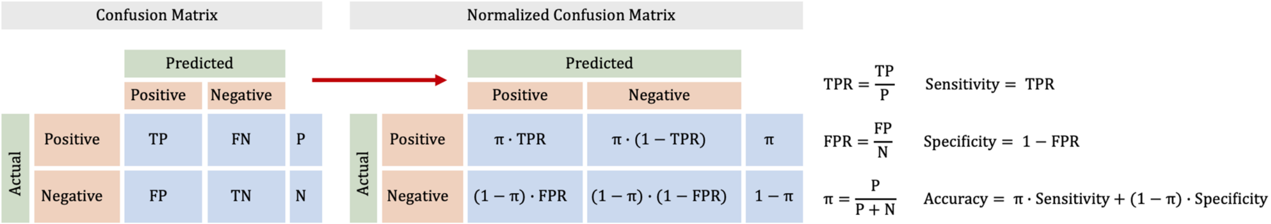

Figure 1. Confusion matrix as a function of true positive rate (TPR), false positive rate (FPR) and the positive class ratio (π).

frequency) class, which is never recommended as it loses information for training the classifier. Instead, oversampling attempts to match the size of all classes to that of the most frequent class, so it does not lose information but it has to generate new samples for the underrepresented classes. This generation can be natural (copies of existing samples) or synthetic (completely new samples from the existing ones). The application of any resampling technique would likely lead to different classification performance, trying to improve the performance before resampling. However, resampling techniques are only focused on the training phase, when the classification model learns. During the testing phase the imbalance will still be present, so the performance metrics must be able to manage it properly. For instance, it would not be correct to apply resampling before cross-validation; the right method would be applying cross-validation and for each training set apply resampling, keeping the test set with the original class distribution. Somehow, modifying the class distribution for the test phase would be equivalent to altering the reality that the classifier is going to face and, therefore, the performance results would be adulterated.

Among the best known sampling techniques, SMOTE[21] has gained popularity. It generates new minority class objects as weighted average of closest neighbors (from the same minority class), and the results during the training phase support its good acceptance in the scientific community[22–26]. Nevertheless, the performance of the classifier—even with better quality learning due to training balance—still depends on the imbalance of the test set. Consequently, classification performance measures should not be sensitive to class bias.

The ROC curve[27–29] represents the relation between the FPR and the TPR, and illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied from the confusion matrix. The curve shows a solid idea of the behavior of a classifier and allows to compare several classifiers in one image for the same dataset. Most importantly, the ROC curve, unlike the accuracy or F-measure, only depends on the FPR and TPR, which are not class biased. Given its usefulness, the ROC curve has been used in many works[30–34], although it has an enormous limitation: it can only be used in binary contexts, i.e. two-class datasets. In case of multi-class datasets, the most common approach is to represent the class of interest (positive) against the others (negative), but it does not contribute to understanding the global performance of the classifier. In case of equal importance of classes, the smoothed solution named *macro-average* is to average all the ROC curves (one-versus-rest for each class), but this compromises its natural insensitivity to class skew. On the contrary, the *micro-average* approach takes into account the proportion of every class (aggregating the contributions of all classes), given more importance to larger classes and, therefore, assigning greater values when the largest class performs better. However when the imbalance is significant, both approaches lead to very different curves with refutable interpretations.

The ROC curve has been proven that is an excellent measure of classification performance[35,36], as it conveys more information than many scoring metrics by visualizing the performance of the classifier by a curve instead of providing a single scalar value. However, the ROC curve also provides a comparable measure of performance: the area under the ROC curve (AU(ROC)). There exist approaches for the AU(ROC) in the context of multi-class classification, based on averaging pairwise comparison of classes[37] or on the use of K-dimensional space to compute the volume of the ROC surface[38–41], being K the number of class labels. However, none of these approaches provide a graphical representation.

In general, when the binary case (positive and negative) is extended to the multi-class case, the class imbalance becomes more prominent. Recently, the Multi-class Classification Performance (MCP) curve solved the problem of showing in a single curve the performance of multi-class datasets for any classifier[42]. However, the MCP curve is also sensitive to class bias. In this work, we introduce the Imbalanced Multi-class Classification Performance (IMCP) curve, as a graphical representation of the performance of the classifier, which is independent of the class distribution, and also the Area Under the IMCP curve (AU(IMCP)), as a scalar measure of prediction quality. Results show the usefulness of both the IMCP curve and the AU(IMCP), firstly with synthetic data to highlight the features of the IMCP curve when analyzing data with class imbalance, and secondly with real–world data to provide an example in which common performance metrics fail at determining the quality of classifiers.

Rationale

The hypothesis is that a consistent assessment metric should calculate classification performance independently of the distribution of the K classes, so that overrepresented classes have no further impact on performance results. To show the impact of the degree of imbalance on classification performance, four synthetic 3-class datasets have been generated, using exactly the same generating functions for each dataset, but varying the distribution of classes exponentially. Figure 2 (left) shows the four datasets, denoted by the letters A, B, C, and D. The datasets have two variables—to simplify visualization on the X and Y axes—and three classes, denoted by the colors blue, red and yellow.

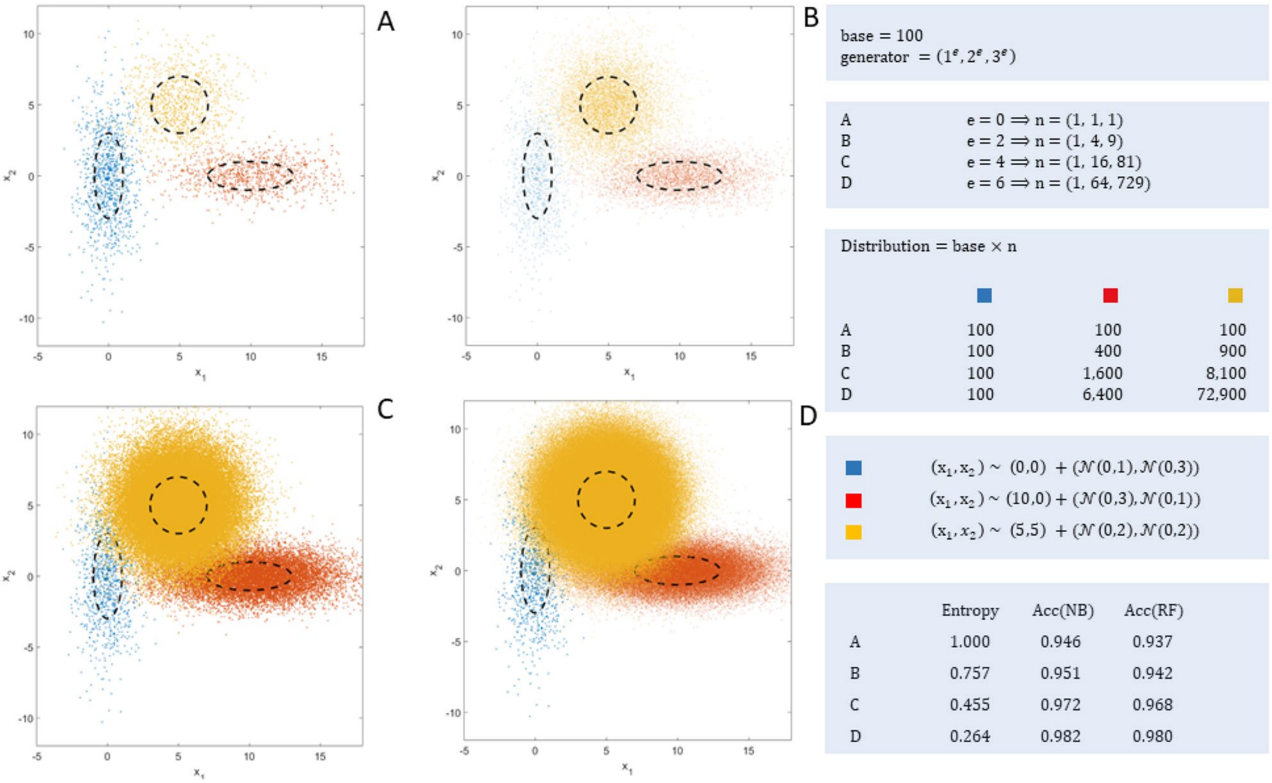

Figure 2. Samples with different distribution of classes.

To provide different degrees of imbalance an exponential methodology supported by a *base* and a *generator* has been designed. The number of samples of each dataset is obtained by multiplying the *base* by the *generator*, which is based on the tuple (1, 2, 3) raised to an exponent *e* that controls the degree of imbalance. The base has been set to 100, and the exponent takes the values 0 (dataset A), 2 (dataset B), 4 (dataset C), and 6 (dataset D). Figure 2 (right) shows the distribution of the tuples for the values of the exponent. To generate the values of each variable a normal distribution has been used, $X, Y \sim \mathcal{N}(\mu, \sigma^2)$, locating the centers at (0,0), (10,0) and (5,5), for blue, red and yellow classes, respectively, and slightly varying the deviations from 1 to 3 to make the shapes somewhat different (ellipses). This method is very reproducible and extensive to more classes, as there is only need to define the base and the exponent for setting the number of examples for each dataset.

The entropy of each dataset is a good indicator of the class imbalance, ranging from 1 (dataset A: zero imbalance) to 0.264 (dataset D: huge imbalance). The worst case is when $e = 6$, which provides an imbalance of 0.13% (blue), 8.06% (red), and 91.81% (yellow), visible in Fig. 2 (dataset D). Also, a stratified 10-fold cross-validation was used to compute the accuracy of two well–known techniques: Naïve-Bayes (NB) and Random Forests (RF), which have been chosen to facilitate replicability. These results are also shown in Fig. 2 (lower right corner), and as expected, the accuracy of both methods increases as the imbalance becomes greater (or entropy smaller). In short, there has been an unfair relative improvement in accuracy of 3.8% and 4.6% due to an imbalance for NB and RF, respectively, when transitioning from dataset A to D.

The impact of imbalance on classification performance is obvious when metrics sensitive to class bias are used. To avoid that negative impact, a novel method is introduced: the Imbalanced Multi-class Classification Performance (IMCP) curve. The IMCP curve is based on the MCP curve[42], which was introduced to deal with multi-class datasets, unlike the ROC curve, which can only visualize the performance of binary (2-class) datasets. While the IMCP curve will be described in detail in the next section, Fig. 3 displays the results for the datasets outlined in Fig. 2. These results indicate that the increasing imbalance negatively impacts the classifier's performance, in contrast to accuracy, which actually benefits from the heightened imbalance. In summary, as class imbalance increases, the results of the IMCP curve worsen for both NB and RF. This decline is also evident in the area under the IMCP curve, indicated at the lower left corner of each graph. This is in contrast to the accuracy depicted in Fig. 2, which exhibits remarkable sensitivity to class imbalance. In summary, classifiers that are trained with progressively higher degrees of class imbalance and then tested under the same imbalanced conditions tend to display a misleading increase in classification accuracy. However, when observing the IMCP curve and its corresponding AU(IMCP) under these identical conditions, a decreasing trend becomes apparent. This suggests that the latter metric more accurately captures the diminishing ability to correctly identify samples from the less frequent classes.

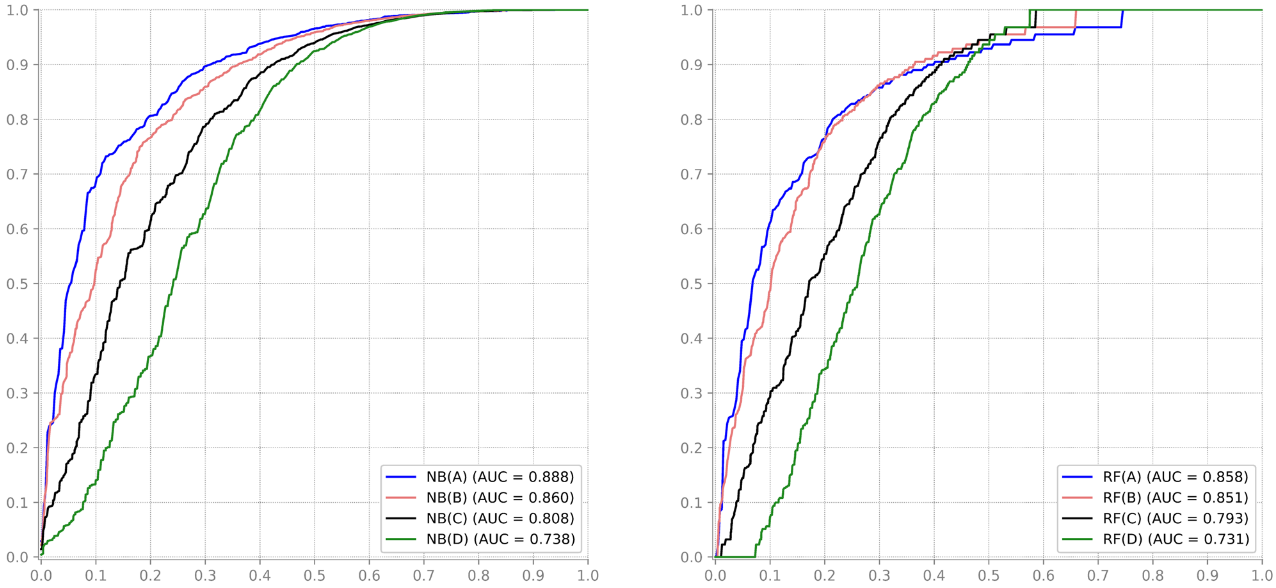

Figure 3. Evolution of the IMCP curve with increasing class imbalance for Naïve-Bayes (left) and Random Forests (right) classifiers, maintaining the data distributions. Class proportions of datasets A, B, C, and D are described in Fig. 2.

Method

For the sake of clarity, the method has been organized into five steps, as depicted in Fig. 4. Let $D = \{e_i \mid e = (\overline{x}, y), \overline{x} \in \mathbb{X}^m, y \in \mathbb{Y}, i = 1, 2, \ldots, n\}$ be a dataset with n samples, which belong to a m-dimensional feature space and have a corresponding outcome y in the space $\mathbb{Y}$. When $|\mathbb{Y}| = 2$ we say that it is a 2-class (binary) classification problem; otherwise, when $|\mathbb{Y}| > 2$, we say that it is a multi–class classification problem. For simplicity, let us assume that $\mathbb{Y} = \{1, 2, \ldots, K\}$ (set of class labels).

The first step is a common evaluation of classifiers by means of any validation technique. In this case, it is visualized a hold-out, but it should ideally be a stratified λ-fold cross-validation (commonly, $\lambda = 10$), which would provide test results of the same size that the original dataset.

The second step provides two matrices: the prediction probability and the true probability matrices. The true probability of sample e_i, denoted by $t(e_i)$, can be encoded as a K-dimensional one-hot vector, in which all the values are 0 except for one 1 at the position k, that satisfies $y_i = k$ (following the *indicator function*, for all k, $t(e_i)_k = 1_{y_i = k}$), being $k \in \mathbb{Y}$. The true probability matrix is in fact the binarization of the target variable (class). Consequently, the sum of all the values of the vector is 1. The prediction probability of sample e_i, denoted by $p(e_i)$, is also a K-dimensional vector, whose values are provided by the classifier as output probabilities of belonging to class $k \in \mathbb{Y}$. Most classification techniques (e.g. Naïve-Bayes, Decision Trees, Random Forests, Neural Networks, etc.) automatically provide the prediction probability vector for each test sample. This vector contains the probabilities that the test sample belongs to each class, so assuming there are K classes, the dimensionality is K, and

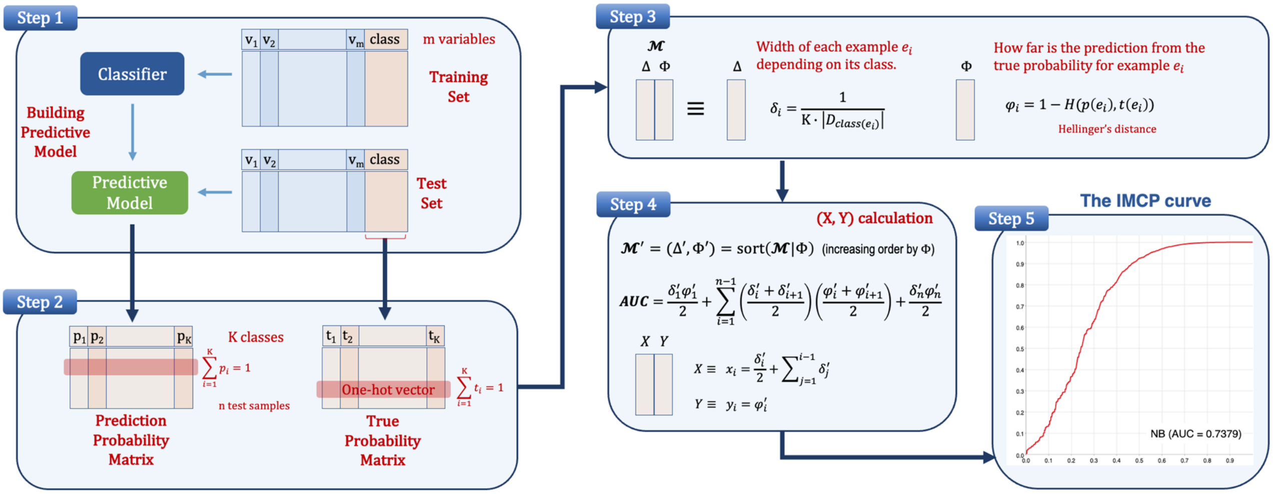
Step 1
Classifier
Building Predictive Model
Predictive Model
v_1 v_2 v_m class
m variables
Training Set
v_1 v_2 v_m class
Test Set
Step 2
p_1 p_2 p_K
K classes
$\sum_{i=1}^{K} p_i = 1$
n test samples
Prediction Probability Matrix
t_1 t_2 t_K
One-hot vector
$\sum_{i=1}^{K} t_i = 1$
True Probability Matrix
Step 3
$\mathcal{M}$
Δ Φ
$\equiv$
Δ
Width of each example e_i depending on its class.
$\delta_i = \frac{1}{K \cdot |D_{class(e_i)}|}$
Φ
How far is the prediction from the true probability for example e_i
$\varphi_i = 1 - H(p(e_i), t(e_i))$
Hellinger's distance
Step 4
(X, Y) calculation
$\mathcal{M}' = (\Delta', \Phi') = \mathrm{sort}(\mathcal{M}|\Phi)$ (increasing order by Φ)
$AUC = \frac{\delta'_1 \varphi'_1}{2} + \sum_{i=1}^{n-1} \left(\frac{\delta'_i + \delta'_{i+1}}{2}\right)\left(\frac{\varphi'_i + \varphi'_{i+1}}{2}\right) + \frac{\delta'_n \varphi'_n}{2}$
X Y
$X \equiv x_i = \frac{\delta'_i}{2} + \sum_{j=1}^{i-1} \delta'_j$
$Y \equiv y_i = \varphi'_i$
Step 5
The IMCP curve
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
NB (AUC = 0.7379)

Figure 4. Description of the methodology for any classifier: from data to the IMCP curve.

the sum of all the probabilities is 1. From these two matrices that contain the prediction and true probabilities of each test sample, we could apply probability distribution distances in order to know how far is the prediction probability from the true probability.

The third step calculates two values for each test sample, located at columns Δ and Φ in matrix $\mathscr{M}$: Δ is associated with the X-axis and Φ with the Y-axis. Column Δ calculates the width δ_i associated to each test sample e_i in the X-axis, that depends on the class distribution (examples belonging to overrepresented class will have smaller values). The sum of all δ_i is 1, but the sum of all δ_i associated with each class is the same, i.e., it is exactly $\frac{1}{K}$, being K the number of classes. In order to observe the quality of the classifier prediction, a distance function d between the two discrete probability distributions must be applied to each sample $e \in D$. Let d be a distance function between two distributions t and p, such as $d : \{0,1\}^K \times [0,1]^K \rightarrow [0,1]$. The Hellinger distance[43,44] is a metric oriented to probability distributions, and has been chosen due to its interesting properties[42]. Let $P = [p_1, \ldots, p_K]$ and $Q = [q_1, \ldots, q_K]$ be two discrete probability distributions, the Hellinger distance is defined as:

$$H(P,Q) = \frac{1}{\sqrt{2}} \left(\sum_{j=1}^{K} \left(\sqrt{p_j} - \sqrt{q_j} \right)^2 \right)^{\frac{1}{2}} \tag{1}$$

Column Φ first calculates the Hellinger distance $H(p(e_i), t(e_i))$ between the prediction probability and the true probability of each sample e_i. If the distance between the true probability $t(e_i)$ and the prediction probability $p(e_i)$ of a sample e_i is close to 0, then the classifier is making a good prediction for e_i; otherwise, when it is close to 1, the prediction is bad. Good or bad predictions are not necessarily correct and incorrect, respectively. The distance is a quality measure of the classification performance, so it would be possible to provide at the same time a not small distance associated with a correct prediction (e.g., for a 3-class problem, when t=[1,0,0] and p=[0.35,0.33,0.32], H(t,p)=0.639—very high value—and the prediction is correct). This fact provides interesting insights to address the uncertainty of classifiers, since even when they correctly assign the class the value of the highest probability might not be far from the others.

The distance value is very important because it informs about how far the prediction is from the true observation for each sample. The value $\varphi_i = 1 - H(p(e_i), t(e_i))$ can be interpreted as the probability that the classifier correctly assigns the observed class label to the test sample e_i. If φ_i is zero, then the distance H between prediction and true probabilities is maximum (one). On the contrary, if φ_i is one, then the distance H is zero, which means that prediction and true probabilities are equal, i.e., the prediction probability vector is a one–hot vector similar to the true probability vector.

The fourth step prepares the values of columns Δ and Φ to be visualized as X-axis and Y-axis, respectively. First, the matrix $\mathscr{M}$ is increasingly sorted according to column Φ, providing Φ'. Once obtained the matrix $\mathscr{M}' = (\Delta', \Phi')$, the AU(IMCP) can be easily calculated following the expression provided in Fig 4 (Step 4).

The new values in the Δ' column are then used to calculate the values for X-axis of the IMCP curve, while the values of Y-axis are the same as in Φ'.

The fifth step draws the points of the (X, Y) matrix in the unit square, as well as visualizes the area under the IMCP curve. The first point $(x = 0)$ and the last point $(x = 1)$ retain the first and last values of the Φ' column, i.e. these will be the points $(0, \varphi_1)$ and $(1, \varphi_n)$.

In order to highlight the influence of the class distribution in the dataset, Fig. 5 shows a synthetic example in which the importance of each class is balanced, while varying proportionally the importance of each sample. The role of δ', φ', x and y in the method is illustrated in Fig. 5 (e.g., for position $i = 12$ in (X, Y)).

The features of the IMCP curve provide interesting aspects related to the interpretability and explainability of predictive models:

- Upper bound: the ideal classifier should display the function $y = 1$, that has AU(IMCP) $= 1$, which means that for each test sample the distance between prediction probability and true probability is zero, i.e. the probability of assigning the true class is always one.
- Lower bound: the worst case is the function $y = 0$, which has AU(IMCP) $= 0$, which means that for each test sample the prediction probability assigned to the true class is zero.
- Correct classification threshold (η): the Hellinger distance can be expressed as a function of the symmetric Bhattacharyya coefficient (BC)[45], where $\mathrm{BC}(P,Q) = \sum_{j=1}^{K} \sqrt{p_j q_j}$ and $\mathrm{H}(P,Q) = \sqrt{1 - \mathrm{BC}(P,Q)}$; if $\mathrm{BC} > \sqrt{0.5}$ then the predicted class will be the true class. Therefore, if $\varphi > 1 - \eta = 1 - \sqrt{1 - \sqrt{0.5}} = 0.459$ then the test sample will be correctly satisfied (the opposite is not true).
- Incorrect classification threshold (θ): the threshold for a correct prediction for a K-classification problem $(K > 1)$ is $\theta = \sqrt{1 - \frac{1}{\sqrt{K}}}$. Therefore, if $\varphi < 1 - \theta$ then the test sample will not be correctly classified.
- Uncertainty: correct and incorrect classification can be defined in terms of the φ values. However, there is an interval of uncertainty $[1 - \theta, 1 - \eta]$ in which all the prediction probabilities for each class are below 0.5, so the difference between the best and the second best values could be very close to zero.

For the ROC curve, only curves above the diagonal line joining the points (0,0) and (1,1) make sense (better than random), since below this it is interpreted as "worse than random". The IMCP curve uses completely the domain of X and Y, i.e., the unit square. Therefore, neither the ROC and the IMCP curves, nor the areas under such curves, are comparable.

The potential of the approach and its features will be discussed in detail in the case of study.

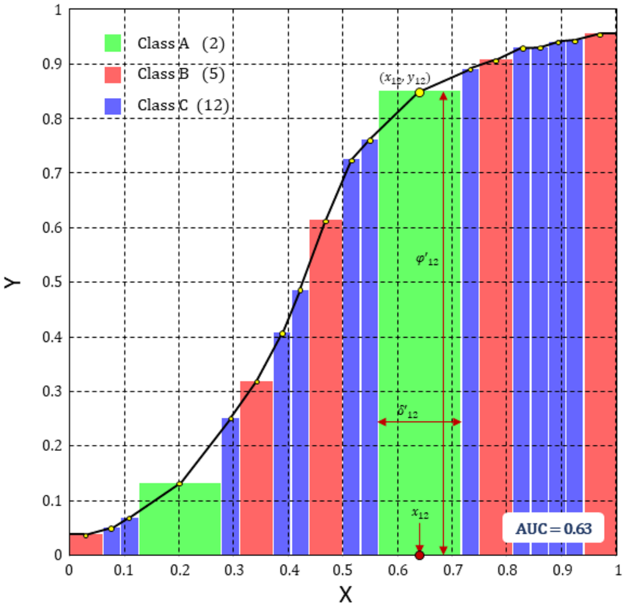

Figure 5. The IMCP curve for a simple example with three imbalanced classes (2 samples of class A, 5 of class B, and 12 of class C.

A case of study: tumor type prediction

To showcase the impact of imbalance on multi-class datasets, recent leading research on cancer prediction has been selected[46]. The work focus on the application of several machine learning techniques to predict the tumor tissue of origin in cases of advanced-stage metastatic tumors. This is a very important issue as a significant number of patients have no conclusive diagnoses (typically via histopathology), and they suffer from very limited therapeutic options as primary cancer type classification is a key factor in guiding treatment decisions.

The original dataset collects information from a wide range of molecular methods, and includes 4131 features extracted for classifying 35 cancer types based on driver/passenger and simple/complex mutations. A total of 6756 samples from the Hartwig Medical Foundation (metastatic tumors) and the Pan-Cancer Analysis of Whole Genomes consortium (primary tumors) were gathered. Univariate feature selection was performed to provide 463 variables, and 48 new features were generated for regional mutational density profiles by means of non-negative matrix factorization. In summary, the dataset used henceforth has 6756 samples, 511 features, and a target variable (tumor type) with 35 possible discrete values (biliary, breast, cervix, liver, pancreas, thyroid, etc.).

The multi-class nature of the dataset limits the use of some classification performance measures (e.g., the ROC curve), and its large imbalance impacts the results. Out of the 35 tumor types, the most frequent is *Breast*, with 996 samples (14.7%), and the least frequent is *Skin_Carcinoma*, with 25 samples (0.4%). The seven most frequent tumors account for more than 58%, and the seven least frequent tumors do not together reach 3.5%. In fact, 16 tumor types have less than 1% of the total, and 23 tumor types have less than they should to become a balanced dataset (2.86%).

It is important to note that the aim of this research is not to analyze the methodology proposed in[46], but to show that, in general, good accuracy, precision, or sensitivity results for imbalanced multi-class datasets should be studied in more depth because they could lead to optimistic conclusions. In this study, the random forest (RF) technique with default parameters was chosen, as in[46]. However, isotonic regression to calibrate the probabilities produced by the RF ensemble was not necessary, as the one-versus-rest strategy to discriminate one cancer type versus other cancer types transforms the multi-class classification into multiple binary classifications, which is inappropriate for evidencing weaknesses in imbalanced multi-class classification performance measures.

Stratified 10-fold cross-validation was used to compute several classification performance measures. The accuracy, which has been shown to be class-biased, was 0.924, which would indicate that the diagnostic system is of sufficient quality. However, an analysis of the confusion matrix revealed that for 12 tumor types the precision (also class skewed) was less than 0.90 (for the tumor type *Sarcoma_Other* was 0.68), and for 13 tumor types the sensitivity was less than 0.90 (for the tumor type *Lung_SmallCell* was 0.32). In tumor prediction, which guides subsequent medical treatments, some of these low precision and sensitivity values are not acceptable.

More revealing than accuracy, precision or sensitivity, is the IMCP curve (Fig. 6a), which is based on the diagnostic system's ability to make predictions based on the probability that the prediction is close to reality, i.e., that it does not exhibit uncertainty that has to be decided by some voting method. While a perfect diagnostic system would achieve an area under the IMCP curve of 1, the system based on RF covers only 57.8% of the ideal classification performance. The IMCP curve measures, in essence, how close these values are to 1, i.e. to assigning a cancer type with a probability close to 1. The larger the area under the IMCP curve, the lower the uncertainty of the predictive system. Therefore, what appeared to be a sound diagnostic system (accuracy = 0.924) may not be so reliable (AU(IMCP)= 0.578).

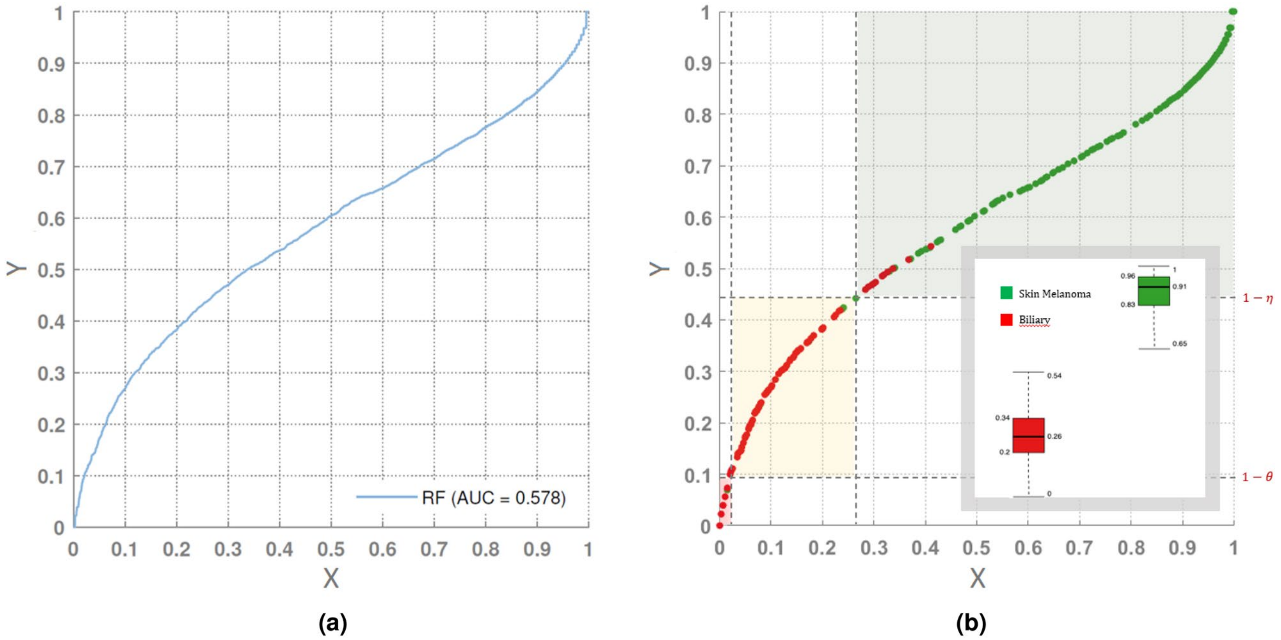

Figure 6. The IMCP curves for the tumor dataset.

The incorrect (θ) and correct (η) classification thresholds provide the landscape of certainty for the dataset: [0, 0.088) for incorrect classifications (72 samples: 1.06%), [0.088, 0.459] for uncertainty (1023 samples: 15.14%), and (0.459, 1] for correct classifications (5661 samples: 83.79%). These values, related to the percentage of samples, are very illustrative because are delimiting the proportion of samples that could be correctly classified ([83.79%, 98.93%]). However, confidence in prediction can vary substantially. This fact is of extreme importance in some predictive systems, such as in the case of tumor diagnosis.

The IMCP curve also allows to analyze in depth the behavior of the classifier for each class, which is very useful for experts, as they could re-evaluate the prediction provided by the system, and possibly supplement it with other factors in case of low reliability. To explain the nature of the predictions, Fig. 6b illustrates the best (*Skin_Melanoma*) and the worst (*Biliary*) classification performances in a graph showing all samples for only both classes, in green and red, respectively (the same values displayed in the IMCP curve in Fig. 6a). Most of the predictions for *Skin_Melanoma* are above $1 - \eta$ within the green area (correct classification). However, most of the predictions for *Biliary* are between $1 - \theta$ and $1 - \eta$, within the yellow area (uncertainty). This means that most of the predictions were unreliable, in the sense that the classifier could assign to other tumor types prediction probabilities lower but very close to the predicted tumor.

The values of φ' for each class are also depicted in box-plots, illustrating that the medians are 0.26 (*Biliary*) and 0.91 (*Skin_Melanoma*). The Q3 value for *Biliary* (0.34) indicates that 75% of the predictions are extremely low. In contrast, the Q1 value for *Skin_Melanoma* (0.83) indicates that 75% of the predictions are excellent. In summary, the reliability of the diagnostic system is not similar for all the tumor types. The expert may have high confidence in a prediction for *Skin_Melanoma*, but very low confidence in a prediction of *Biliary*.

In total, 7 out of 37 tumor types had medians below $1 - \eta$ (more than 50% of samples under the correct classification threshold), and only 19 out of 35 had the Q1 value above $1 - \eta$ (more than 75% of samples over the correct classification threshold). An overall conditional classification performance approach is depicted in Fig. 7, in which it is easy to detect the diagnostic ability of the system for each type of tumor. The information provided by the figure is very valuable from a diagnostic system quality perspective, as it not only highlights the tumor types for which the system is reliable but also identifies the tumor types for which predictions should be examined more carefully by experts in the field. In addition, the conditional classification performance results identify to which classes greater efforts should be applied, from a data analytics perspective, to improve the quality of predictions.

In summary, a thorough analysis of the quality of the diagnostic system, in principle with excellent indicators (accuracy = 0.924, average precision = 0.908, average sensitivity = 0.866, mean F_1-score = 0.877), reveals that these measures are very sensitive to imbalance, as the classification performance provided by the area under the IMCP curve is 0.578, mainly due to low confidence on predictions for some tumor types.

The analysis performed with this dataset is extensible to any dataset. It extracts interesting aspects about the overall quality of the diagnostic system and also about the classification performance for each class. The method is useful for multi-class datasets, and especially suitable for more than two classes (where the ROC curve is not applicable). Furthermore, it is insensitive to class imbalance, which adds further benefit to the examination of the performance of predictive systems.

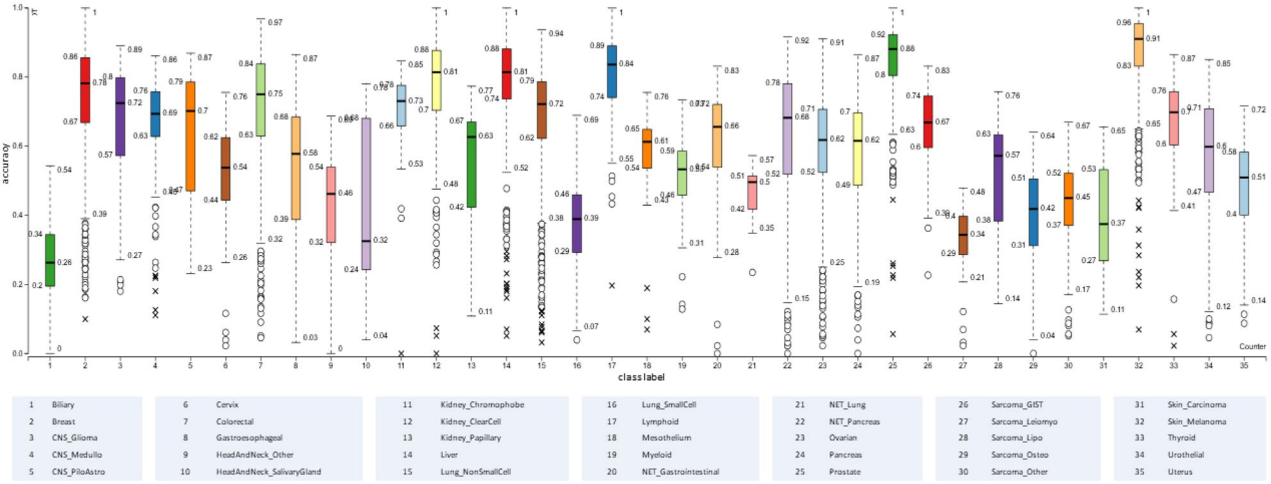

Figure 7. Conditional classification performance for each tumor type.

Conclusion

Classification models are designed to predict unseen events from what has been learned during training. The quality of the model indicates how good the prediction would be in the future, and can be measured in different ways. Several important performance measures are not valid for multi-class datasets (e.g., the ROC curve), and many of them fail at analyzing imbalanced datasets (e.g., the accuracy or the F_β-score). The aim of this work is to provide a novel method, named the IMCP curve, that graphically represents the classification performance for both multi-class and imbalanced datasets. The usefulness is broad, particularly in biomedicine, as most of the gene expression datasets are imbalanced (e.g., proportion of cancerous versus healthy samples) and many of them include more than two target-categories (e.g., subtypes of tumors).

The IMCP reveals that some quality measures could be optimistic, as they provide overall values of classification performance of the diagnostic system without taking into account the class bias. Moreover, the approach facilitates an in-depth conditional analysis of the diagnostic system, firstly by identifying which classes perform better (high confidence), and secondly by allowing further research on specific target classes that perform worse (high risk).

The promising scenario opened by the IMCP curve requires further analysis and study of the properties in relation to the most commonly used metrics in the scientific literature, since prediction probabilities have much more potential than the confusion matrix.

Data availability

The datasets used in the "Rationale" section are publicly available at https://github.com/adaa-polsl/imcp/tree/main/tests with the following names: Dataset A: A_exp-0.base-100.csv Dataset B: B_exp-2.base-100.csv Dataset C: C_exp-4.base-100.csv Dataset D: D_exp-6.base-100.csv. A "Case study" data are taken from: Nguyen et al.[46]. This dataset is public, as indicated in the "Data availability" section of the aforementioned reference and is freely accessible for academic purposes (Supplementary Information). To generate the IMCP curves, all experiments used the Python package publicly available at https://github.com/adaa-polsl/imcp.

Received: 25 September 2023; Accepted: 6 May 2024
Published online: 10 May 2024

References

1. Swets, J. A. Measuring the accuracy of diagnostic systems. *Science* **240**, 1285–1293 (1988).
2. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
3. Shen, Y. *et al.* Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. *Nat. Commun.* **12**, 5645 (2021).
4. Harvey, J. et al. Machine learning-based prediction of cognitive outcomes in de novo Parkinson's disease. *npj Parkinson's Dis.* **8**, 150 (2022).
5. Lee, K., Jeong, H.-O., Lee, S. & Jeong, W.-K. CPEM: Accurate cancer type classification based on somatic alterations using an ensemble of a random forest and a deep neural network. *Sci. Rep.* **9**, 16927 (2019).
6. Zhang, X. *et al.* Data-driven subtyping of Parkinson's disease using longitudinal clinical records: A cohort study. *Sci. Rep.* **9**, 797 (2019).
7. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).
8. Perincheri, S. *et al.* An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. *Mod. Pathol.* **34**, 1588–1595 (2021).
9. Asanomi, Y. *et al.* Dementia subtype prediction models constructed by penalized regression methods for multiclass classification using serum microrna expression data. *Sci. Rep.* **11**, 20947 (2021).
10. Bulten, W. *et al.* Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge. *Nat. Med.* **28**, 154–163 (2022).

11. Shiri, I. *et al.* High-dimensional multinomial multiclass severity scoring of COVID-19 pneumonia using CT radiomics features and machine learning algorithms. *Sci. Rep.* **12**, 14817 (2022).
12. Su, Q. *et al.* Faecal microbiome-based machine learning for multi-class disease diagnosis. *Nat. Commun.* **13**, 6818 (2022).
13. 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, KDD'97*. 43–48 (AAAI Press, 1997).
14. van Rijsbergen, C. J. Foundation of evaluation. *J. Docum.* **30**, 365–373 (1974).
15. Chinchor, N. A. & Sundheim, B. M. Muc-5 evaluation metrics. In *MUC* (1993).
16. Kubat, M., Holte, R. C. & Matwin, S. Machine learning for the detection of oil spills in satellite radar images. *Mach. Learn.* **30**, 195–215 (1998).
17. Tomek, I. Two modifications of CNN. *IEEE Trans. Syst. Man Cybern.* **6**, 769–772 (1976).
18. Kubat, M. & Matwin, S. Addressing the curse of imbalanced training sets: One-sided selection. In Proceedings of the Fourteenth International Conference on Machine Learning. 179–186 (Morgan Kaufmann, 1997).
19. Zheng, M. *et al.* UFFDFR: Undersampling framework with denoising, fuzzy c-means clustering, and representative sample selection for imbalanced data classification. *Inf. Sci.* **576**, 658–680 (2021).
20. Ng, W. W. Y. et al. Hashing–based undersampling ensemble for imbalanced pattern classification problems. IEEE Trans. Cybern. **52** (2022).
21. Chawla, N., Bowyer, K., Hall, L. & Kegelmeyer, W. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **9**, 321–357 (2002).
22. Chawla, N. V., Lazarevic, A., Hall, L. O. & Bowyer, K. W. SMOTEBoost: Improving prediction of the minority class in boosting. In *Knowledge Discovery in Databases: PKDD 2003. 107–119* (eds Lavrač, N. *et al.*) (Springer, 2003).
23. Han, H., Wang, W.-Y. & Mao, B.-H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing. 878–887* (eds Huang, D.-S. *et al.*) (Springer, 2005).
24. Zhang, H. & Li, M. RWO-sampling: A random walk over-sampling approach to imbalanced data classification. *Inf. Fusion* **20**, 99–116 (2014).
25. Chen, Q., Zhang, Z.-L., Huang, W.-P., Wu, J. & Luo, X.-G. PF-SMOTE: A novel parameter-free SMOTE for imbalanced datasets. *Neurocomputing* **498**, 75–88 (2022).
26. Dangut, M. D., Skaf, Z. & Jennions, I. K. Handling imbalanced data for aircraft predictive maintenance using the BACHE algorithm. *Appl. Soft Comput.* **123**, 108924 (2022).
27. Peterson, W., F,W., Birdsall, T.: The theory of signal detectability. Trans. *IRE Profess. Group Inf. Theory* **4**, 171–212 (1954).
28. Swets, J. A. The relative operating characteristic in psychology. *Science* **182**, 990–1000 (1973).
29. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
30. Zweig, M. H. & Campbell, G. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clin. Chem.* **39**, 561–577 (1993).
31. Lu, L. *et al.* A language-matching model to improve equity and efficiency of COVID-19 contact tracing. *Proc. Natl. Acad. Sci. USA* **118** (2021).
32. Bury, T. M. *et al.* Deep learning for early warning signals of tipping points. *Proc. Natl. Acad. Sci. USA* **118** (2021).
33. Mueller, H., Groeger, A., Hersh, J., Matranga, A. & Serrat, J. Monitoring war destruction from space using machine learning. *Proc. Natl. Acad. Sci. USA* **118** (2021).
34. Groh, M., Epstein, Z., Firestone, C. & Picard, R. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proc. Natl. Acad. Sci. USA* **119** (2022).
35. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).
36. Ferri, C., Hernández-Orallo, J. & Modroiu, R. An experimental comparison of performance measures for classification. *Pattern Recognit. Lett.* **30**, 27–38 (2009).
37. Hand, D. J. & Till, R. J. A simple generalisation of the area under the roc curve for multiple class classification problems. *Mach. Learn.* **45**, 171–186 (2001).
38. Mossman, D. Three-way ROCs. *Med Decis. Mak.* **1**, 78–89 (1999).
39. Ferri, C., Hernández-Orallo, J. & Salido, M. A. Volume under the ROC surface for multi–class problems. In *Machine Learning: ECML 2003*. 108–120 (Springer, 2003).
40. He, X. & Frey, E. C. The meaning and use of the volume under a three-class ROC surface (VUS). *IEEE Trans. Med. Imaging* **27**, 577–588 (2008).
41. Landgrebe, T. C. & Duin, R. P. Efficient multiclass ROC approximation by decomposition via confusion matrix perturbation analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**, 810–822 (2008).
42. Aguilar-Ruiz, J. S. & Michalak, M. Multiclass classification performance curve. *IEEE Access* **10**, 68915–68921 (2022).
43. Hellinger, E. *Die Orthogonalvarianten Quadratischer Formen von Unendlich Vielen Variablelen*. Ph.D. thesis, University of Göttingen (1907).
44. Hellinger, E. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *J. Reine Angewandte Math.* **1909**, 210–271 (1909).
45. Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Calcutta Math. Soc. Bull.* **35**, 99–109 (1943).
46. Nguyen, L., Van Hoeck, A. & Cuppen, E. Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features. *Nat. Commun.* **13**, 4013 (2022).

Acknowledgements

This work was supported by Grant PID2020-117759GB-I00 funded by MCIN/AEI/10.13039/ 501100011033, the Andalusian Plan for Research, Development and Innovation and the Department of Computer Networks and Systems (RAu9) at Silesian University of Technology.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-61365-z.

Correspondence and requests for materials should be addressed to J.S.A.-R.