




A Semiparametric Approach to Mitigating the Impact of Outliers in ROC Curve Generation for Image Analysis

Regis Cortez Bueno   [Federal Institute of São Paulo | regiscb@ifsp.edu.br]

Renan Gimeniz Marques  [Federal Institute of São Paulo | renan.gimeniz@aluno.ifsp.edu.br]

Raphael Antonio de Souza  [Federal Institute of São Paulo | raphael@ifsp.edu.br]

Sivanilza Teixeira Machado  [Federal Institute of São Paulo | sivanilzamachado@ifsp.edu.br]

 Image Pattern Analysis and Recognition Laboratory (iLab), Federal Institute of São Paulo, Av. Mogi das Cruzes, 1501, Suzano, São Paulo, SP, 08673-010, Brazil.

Received: 04 December 2024 • **Accepted:** 29 July 2025 • **Published:** 31 October 2025

Abstract Artificial intelligence enables the development of machine learning algorithms that can identify and categorize patterns using large amounts of data across various areas. Computational tools were created to analyze these algorithms, allowing for the validation and comparison of their results. The Receiver Operating Characteristic (ROC) is an important statistical technique used for analyzing binary classification models. A ROC curve is commonly utilized in image analysis as a validation metric to compare images generated by a classification model with images created by humans, referred to as Ground Truth (GT). Currently, machine learning algorithms produce ROC curves with a limited number of points, even when trained on large-scale datasets. The result is the presence of outliers which can significantly distort the ROC curve, potentially leading to inaccurate conclusions about the model's performance. This study introduces a novel method for preventing outliers in the creation of ROC curves, guaranteeing a reliable and robust evaluation of image classification models. We implemented our algorithm in Python using a dataset of 1000 grayscale contour images. Performance was compared against Logistic Regression, SVM, Random Forests and SKlearn using ROC curves, AUC, precision, accuracy, and F1-score. Statistical significance was assessed via paired t-tests and Cohen's d for effect size, with outlier detection via Local Outlier Factor. Results demonstrated that SPROC showed a refined curve with more precise AUC values on noisy images in contrast to machine learning approaches.

Keywords: Receiver Operating Characteristic, Roc Curve, ROC Analysis, Images Analysis, Outliers

1 Introduction

Classifiers are computational tools used to bridge the gap in analyzing and recognizing data-driven patterns through a class label defined by a set of attributes (Friedman et al., 1997). Examples of classifiers include random forest [Breiman, 2001], logistic regression [Sachs, 2017], SVM (Support Vector Machine) [Pourghasemi and Rahmati, 2018], convolutional neural networks [Zeiler and Fergus, 2014], and Bayesian classifiers (Friedman et al., 1997), which can generate divergent results when analyzing the same dataset [Khosravi et al., 2018]. In this context, studying receiver operating characteristics (ROC) in images is particularly important as a method for analyzing the overall performance of each classifier.

ROC analysis was first introduced during World War II to quantitatively assess the capabilities of radar operators in differentiating between signals and noise. This approach is a statistical method used to validate results and measure their ability to distinguish between different options, such as diagnosing a condition or predicting an outcome. ROC curves have a wide range of applications in several fields today. They have been utilized in medicine for the purpose of examining and aiding in the diagnosis, particularly with images, as well as for analyzing and comparing the performance of algorithms [Kun-Peng et al., 2018; Hannun et al., 2019; Gao et al., 2020; Zhao et al., 2021]. In geology, they

are used to identify susceptibility to floods and landslides [Termeh et al., 2018; Khosravi et al., 2018; Hong et al., 2018; Pourghasemi and Rahmati, 2018]. Additionally, some studies investigate the impact of biases in ROC analysis [McGowan et al., 2016; Cook, 2017]. They are also used in nuclear technology [Bueno et al., 2018], radiology [Wu et al., 2022], communication [Wang et al., 2022], computing [Sachs, 2017; Moreira, 2020], and medicine [Li et al., 2020; Shkurnikov et al., 2021; Du et al., 2022; Zhao et al., 2022].

ROC curves in image analysis have received considerable attention from the community [He et al., 2010; Bueno et al., 2018; Keidar et al., 2020; Niu et al., 2021] since there is a demand for more accurate metrics to validate proposed algorithms. This study engages a binary classification model that maps cases from the true class to the predicted class, yielding either a positive or negative outcome. The classification is often based on a cutoff value, with each cutoff value represented by a point in the ROC space [Fawcett, 2006]. Binary classification comparison entails the examination of two images: one represents the output of an algorithm, while the other is either manually crafted by a domain expert or generated by an intelligent system to serve as a reference. Images generated are commonly known as Ground Truth (GT) or gold standard. The Ground Truth (GT) is compared with the images generated by the classification algorithms to identify an optimal detection observation for the construction of Re-

ceiver Operating Characteristic (ROC) curves [Bueno *et al.*, 2018].

		Predicted Class	
		Positive (p)	Negative (n)
True Class (GT)	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

Figure 1. Confusion Matrix.

Let a binary set of the true class be labeled as P or N, as illustrated in Figure 1, representing the number of positive and negative instances, respectively. The predictive class, which constitutes the classifier to be analyzed, is represented by p for positive instances and n for negative instances. As illustrated in the confusion matrix in Figure 1, the classification of a set of instances can lead to four distinct situations: True Positive (TP) refers to positive values where the classifier judged them as positive; False Positive (FP) refers to negative values where the classifier judged them as positive; True Negative (TN) refers to negative values where the classifier judged them as negative; and False Negative (FN) refers to positive values where the classifier judged them as negative. The ROC curve is generated by counting the number of TP, FP, FN, and TN, then calculating and plotting the TPR, which represents the true positive rate or the probability of detection, and the FPR, which means the false positive rate (probability of false alarm), according to different detection threshold values [Khawaja *et al.*, 2023]. Equations (1) and (2) define the vectors for the TPR and FPR measures, respectively. Depending on the context, sensitivity may be more or less relevant than specificity [Nahm, 2022]. In terms of object detection or image classification, sensitivity and specificity can vary, but both metrics are fundamental for evaluating and improving image analysis algorithms. Therefore, this present study limits itself to treating sensitivity and specificity as equally relevant, without the necessity of finding the optimal threshold value.

$$TPR_i = \frac{\sum_{k=0}^N TP_k}{\sum_{k=0}^N [TP_k + FN_k]} \quad (1)$$

$$FPR_i = \frac{\sum_{k=0}^N FP_k}{\sum_{k=0}^N [FP_k + TN_k]} \quad (2)$$

In addition, there are other metrics that can normally be used together with ROC analysis to evaluate the performance of the models. These include Accuracy (ACC), Precision or Positive Predictive Value (PPV), and F1-score. These metrics can be calculated using Equations (3), (4), and (5) respectively.

$$ACC = \frac{\sum_{k=0}^N [TP_k + TN_k]}{\sum_{k=0}^N [TP_k + TN_k + FP_k + FN_k]} \quad (3)$$

$$PPV = \frac{\sum_{k=0}^N TP_k}{\sum_{k=0}^N TP_k + \sum_{k=0}^N FP_k} \quad (4)$$

$$F1\text{-score} = \frac{2 \cdot PPV \cdot TPR}{PPV + TPR} \quad (5)$$

Accuracy (ACC) consists of the proportion of correctly classified instances of positive and negative classes (TP and TN) over the total population of classes (TP, FN, FP, TN). Generally, this measure is used with other metrics because it can result in distorted measures when the class distribution is imbalanced. In other words, there is a significant difference between the number of samples in each class [Schott *et al.*, 2024]. Precision (PPV) refers to the classifier's ability not to label a negative sample as positive. Its highest value is 1, when there are no false positives ($FP = 0$), and the lowest is zero, when $TP = 0$ and $FP \neq 0$. This measure is also used to calculate the F1-score, which measures classification efficiency and combines the evaluation of TPR and PPV into a single value. If the F1-score is low, this means that either precision or TPR is low, indicating that FN or FP are low. In other terms, the model is not excessively erring in FN or FP. In a hypothetical case where $FN = 0$ and $FP = 0$, the F1-score will be 1. This measure is widely used, especially when class distribution is imbalanced.

The area under the curve (AUC), as described in Equation (6), is a relevant measure in ROC analysis. This metric indicates the probability that a positive class will be considered better than a negative class. The objective is to assess the classification's quality using a single number, which is obtained by combining two measures produced from the ROC curves: the True Positive Rate (TPR) and the False Positive Rate (FPR) values.

$$AUC = \sum_{k=0}^N \frac{1}{2} [(TPR_k + TPR_{k-1}) \cdot (FPR_k - FPR_{k-1})] \quad (6)$$

The AUC is scale-invariant and is generally calculated using the trapezoidal rule, with values ranging from 0 to 1. In Figure 2, the AUC represents the area under the curve, where a value of 0.5 indicates random classification, often depicted by a dotted line. A value greater than 0.5 indicates performance superior to random classification, while a value less than 0.5 shows the opposite behavior. A higher AUC value indicates better performance of the classifier in distinguishing between

classes. As observed in Figure 2, the ROC curve associated with the image "00000001.jpg" from a given classification model achieved an AUC value of 0.6826 (68.26%). The point (0,1.0) on the curve represents perfect classification, corresponding to an AUC value of 1.0 (100%).

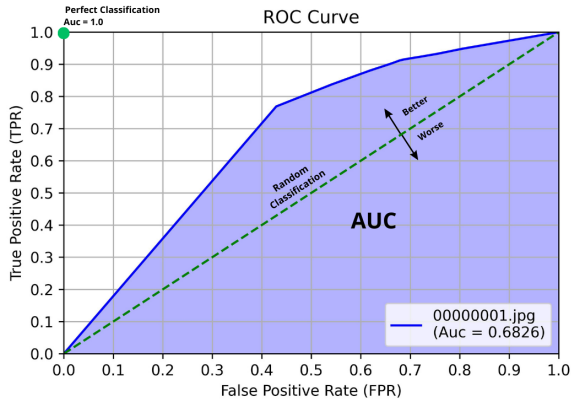


Figure 2. ROC Curve Space.

Current commercial and academic software does not typically employ semiparametric methods. Instead, it generates ROC curves using either empirical or parametric approaches, both of which present inherent limitations [Nahm, 2022]. The empirical approach directly associates the method for generating cutting points with the observed data, thereby eliminating the need to assume a specific data distribution. However, this approach often results in a relatively irregular curve, which can affect the calculation of certain metrics, such as the AUC. When employing the parametric ROC curve approach, it is essential to assess the dataset's data distribution prior to selecting the statistical model. This approach typically involves using machine learning techniques to estimate curve points, which may lead to outliers that can affect the accuracy of the AUC calculation.

This study introduces a novel method, referred to as SPROC (Semi-Parametric ROC), for constructing Receiver Operating Characteristic (ROC) curves, specifically designed to prevent the occurrence of anomalies, such as outliers. By employing a Bayesian methodology independent of machine learning techniques, this approach provides a more robust and reliable evaluation of image classification models.

The rest of the paper is arranged as follows. Section 2 contains a description of the materials and methods used to develop the proposed method, as demonstrated in Section 3. The results of the experiments and discussion are presented in Section 4. The paper closes with a conclusion, some recommendations, and future works.

2 Materials and Methods

The algorithm developed for this study was implemented in the Python 3.11 programming language using IDE Spyder, version 5.3, under the Linux Mint 21.3 operating system. The image dataset was kept on a computer equipped with a 6th generation Intel Core i7 processor, 16 GB of DDR-4 RAM, 500 GB of solid-state drive (SSD) storage, and 2 TB of hard disk

(HD) memory. In addition, we utilized two monitors, specifically a 25-inch Dell monitor and a 34-inch Acer Predator monitor, both boasting a resolution of 3400x1440 to enhance the clarity and detail of the shown images.

The dataset used consists of 1000 images of object contours contained in the scene (Figure 3b) which were generated automatically. It was developed by [Li et al., 2019] and represents the Ground Truth (GT) images of this work, serving as a comparison model to generate metrics between the algorithms analyzed. The Figures 3a and 3b display the original image and its respective contour image.

The original dataset images were acquired on the RGB channels (Red, Green, and Blue) and converted into a single gray-intensity channel to simplify computing complexity. This transformation was applied for each pixel of the image using the ratio $L = (299R + 587G + 114B) \cdot 10^{-3}$, where L is the intensity of gray in R, G, B. Each pixel is characterized by an 8-bit depth number ranging from zero to 255, where zero represents the pixel in black and 255 in white.

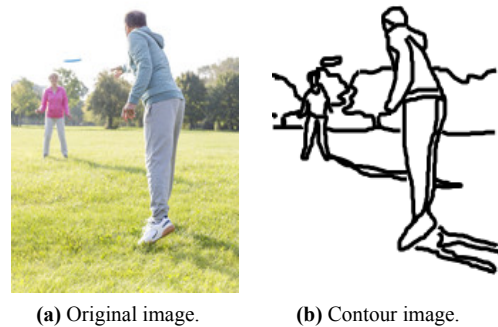


Figure 3. Dataset images developed by [Li et al., 2019].

Logistic Regression [Sachs, 2017], SVM [Pourghasemi and Rahmati, 2018] and Random Forests [Breiman, 2001] and [Friedman et al., 1997] are machine learning algorithms commonly used in the generation of the ROC curve, and have been applied in comparison with the proposed algorithm. These algorithms were implemented with the help of the scikit-learn library for Python. Each of these algorithms generated a number of cutting points obtained by training the dataset images. The SKlearn algorithm was developed to empirically generate ROC curves, using tools provided by the scikit-learn library. In this approach, pixel intensity values from the image are normalized to the range [0, 1] for each pixel, allowing them to be interpreted as continuous score values for binary classification. The normalization process is performed using the equation $score_value = pixel_value/255$. The Scikit-learn library "roccurve" function used these points to create a performance curve for each algorithm of the dataset images.

We created a total of 1000 graphs, each containing five ROC curves, representing the four algorithms and the proposed one. In addition to the ROC curves, the AUC (Area Under the Curve), Precision (PPV), ACC (Accuracy), overall mean AUC and ACC performance across the entire dataset, and F1-score were used as comparison metrics in this study. A paired t-test [Student, 1908] was conducted to assess statis-

tical significance in AUC values per image across the dataset, aiming to evaluate significant differences between each model and the proposed one. The test compares the means of the paired AUCs and determines whether the difference is statistically significant, considering a p-value < 0.05 . A paired-samples t-test was performed using the "ttest_rel" function from the SciPy.stats module (Python). Additionally, Cohen's d [Cohen, 1992] was applied following the paired t-test to assess whether the observed difference has a meaningful practical impact. Equation (7) illustrates the Cohen's d applied in this study, while Table 1 provides its interpretation.

$$d = \frac{\bar{D}}{s_D} \quad (7)$$

where:

\bar{D} = paired mean difference,

s_D = standard deviation of the differences.

Table 1. Interpretation of Cohen's d .

Value	Interpretation
0.2	Small effect
0.5	Medium effect
≥ 0.8	Large effect

Furthermore, the Local Outlier Factor (LOF) was utilized to quantify the presence of outliers in each algorithm as well as in the proposed method.

Outliers detection, also known as anomaly or divergence detection, can be divided into two types: global and local identification. In global identification, all points are considered for analysis. For a local outline, each point is compared to its nearest neighbor. In the ROC curve analysis, the major interest is concentrated on the local identification of the outsiders in the formation of a smooth curve. Many methods are used to detect these types of anomalies present in the data [Alghushairy et al., 2020]. The most commonly used method is the Local Anomaly Factor to identify the density of nearby points [Ghamry et al., 2024]. The Local Anomaly Factor (LOF) is a method for identifying points with anomalous data that analyzes the strangeness of a point in relation to the local density of its neighbors. This measurement is used to detect the presence of outliers accurately in a dataset that presents variations in local densities. Local density refers to the concentration of points around a point of interest. It is used to identify points that differ significantly from the density of their neighbors, indicating that these points may be outliers.

We selected the Canny filter to detect edges in all original images, which were first converted to grayscale. The implementation was carried out using the OpenCV library (version 4.8). This choice was driven by the Canny filter's thresholding process, which effectively determines which pixels correspond to edges, making it suitable for precise edge detection tasks. We employed two thresholds: one to define the upper limit and another for the lower limit. Pixels with gradient magnitudes over the top threshold are considered robust edges, while those falling below the lower threshold are discarded. Pixels with gradient magnitudes between the

two thresholds are considered weak edges and are kept, only if they are connected to strong pixels. The Figure 4 shows the result of the Canny detector applied to the gray-intensity image of the original image of Figure 3a.



Figure 4. Canny filter.

Image analysis and computer vision often employ the Canny filter for its ability to accurately recognize edges with low noise, thereby highlighting TPs. The choice of threshold values (lower and upper) allowed the capture of more detailed and similar edges to those presented by the GT. As a result, this approach effectively highlighted the FNs, as illustrated in Figure 3b and Figure 4. We empirically analyzed the dataset images to select the values for the lower and upper thresholds, resulting in values of 100 and 200, respectively.

3 Proposed Method (SPROC)

SPROC is based on Bayes' theorem to estimate the TPR and FPR values, supported by cutting points, in the generation of ROC curves without relying on machine learning techniques. This approach is classified as semiparametric because it combines a statistical framework with an empirical component, offering users the flexibility to define the number of cutoff points. The Bayesian Theorem used in this work is based on the basic definition found in [Martin, 2024] and can be expressed as:

$$P(X | Y) = \frac{P(Y | X) \cdot P(X)}{P(Y)} \quad (8)$$

where $P(X | Y)$ represents the probability that the X event will occur, given that Y has already occurred; $P(Y | X)$ is the probability of the Y event to occur, considering that X has already taken place; $P(X)$ is the likelihood that the X event will take place, and $P(Y)$ is the probability of the Y event to happen. In this way, the confusion matrix of Figure 1 can be redefined according to the Bayes theorem of the Equation (8) and shown in Figure 5.

The positive and negative classes representing the GT will be described with the symbols (+) and (-), mutually. The symbols (cont) and (no_cont) are predictions of the algorithm for the detection of contour and non-contour, respectively. Be $P(cont)_i$ the threshold values shown in the equation (9), which represent the proportion of pixels considered contours in the image, where N is the amount chosen by the user, $x_0 = 0$, and $\forall i | i \in [0, N - 1]$.

		Predicted Class	
		Positive (cont)	Negative (no_cont)
True Class (GT)	Positive (+)	TP $P(+ cont)$	FN $P(+ no_cont)$
	Negative (-)	FP $P(- cont)$	TN $P(- no_cont)$

Figure 5. Modified Confusion Matrix.

$$P(cont)_i = \sum_{i=0}^{N-1} \left(x_i + \frac{1}{N} \right), \quad (9)$$

$$0 \leq P(cont)_i \leq 1$$

Subsequently, equations (1) and (2) may be reformulated as:

$$TPR_i = \frac{P(+|cont) \cdot P(cont)_i}{P(+)_i} \quad (10)$$

$$FPR_i = \frac{P(-|cont) \cdot P(cont)_i}{P(-)_i} \quad (11)$$

$$P(+)_i = P(+|cont) \cdot P(cont)_i + P(+|no_cont) \cdot [1 - P(cont)_i] \quad (12)$$

$$P(-)_i = P(-|no_cont) \cdot [1 - P(cont)_i] + P(-|cont) \cdot P(cont)_i \quad (13)$$

The equations (3), (4), (5) and (6) produce a single value and remain unaltered. The SPROC algorithm collects the image data (pixels) to generate the classes (TP, FP, TN, and FN) that form the confusion matrix, as shown in Table 2. The user selects the number of cutoff points, and subsequently utilizes equations (9), (10), (11), (12), and (13) to determine the values of TPR and FPR. Table 3 displays the outcome of the computation using $N = 11$ values.

Three significant characteristics should be highlighted. The first refers to the relationships $TP + FN = 1$ and $FP + TN = 1$ that remain valid for the equations (10) and (11), given that:

$$P(cont)_i + [1 - P(cont)_i] = 1,$$

$$P(+|cont) + P(+|no_cont) = 1,$$

$$\forall i \mid i \in [0, N - 1],$$

$$0 \leq P(cont)_i \leq 1,$$

furthermore, the equality is still preserved for $P(+|cont) \cdot P(cont)_i + P(+|no_cont) \cdot [1 - P(cont)_i] = 1$. The second method allows for the calculation of both the lower and upper limits for $P(+)_i$ and $P(-)_i$ as shown in tables 2 and 3, respectively.

$$P(+)_i \in [P(+|cont), P(+|no_cont)],$$

$$P(-)_i \in [P(-|cont), P(-|no_cont)].$$

The third method's characteristic pertains to determining the optimal number of cutoff points required to achieve accurate results. In general, it is unnecessary to compute an excessive number of points since the computation relies on TPR and FPR values derived from the confusion matrix while still maintaining the quality of the results. However, it is possible to determine the minimum number of points for more accurate results. Let N the number of points created to form the ROC curve and its area measure (AUC) to compare two classifiers A and B, which represent two different images of the dataset used in this study. We selected classifiers with AUC values below and above 0.5 for comparison purposes. Table 4 shows the number of points created and their AUCs.

The differences in accuracy of the values are insignificant for $N=100$ in most image-based applications, with variations occurring in the micrometer range (μm) across all tests conducted. This result is crucial for analyzing classifiers with similar ROC curves and AUC values, which can be difficult to distinguish when considering only a few decimal places. For N values exceeding 10,000, the accuracy improves to the nanometer range scale (nm). For values below $N = 15$ cutoff points, the ROC curve becomes less smooth, adversely affecting the accuracy of AUC estimation. It is observed that as the number of cutoff points increases, the computational time complexity rises approximately at the rate of $O(N \log N)$.

4 Results and Discussion

This study introduced a novel approach, termed SPROC, for generating ROC curves to minimize the impact of outliers and ensure an accurate and reliable evaluation of image classification models. The SPROC method generated the ROC curves using $P(cont)_i$ with $N = 100$ cutoff points, applied to a dataset with 1000 images of contours automatically generated by [Li et al., 2019], which correspond to the GTs present in this study (see Fig.3(B)). Machine learning algorithms, including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and the empirical SKlearn (SK), were compared with the proposed method using images both the GTs and images generated by the Canny filter (see Figure 4) across the entire dataset.

The performance of SPROC was evaluated using the visual perception of the ROC curves produced for each method, and the AUC, ACC, PPV, and F1-Score metrics. The Local Anomaly Factor (LOF) of the scikit-learn was used to identify outliers in the results, utilizing the parameter of the nearest neighbors $n = 2$. The low value of n allowed the LOF algorithm to be sensitive to small variations and noises, with the aim of detecting local outliers. Considering a larger number of neighbors can smooth outlier detection and ignore more local and less dense outliers. Additionally, overall mean AUC and ACC values across the entire dataset were computed to provide a comprehensive performance assessment. For a rigorous statistical assessment, this study utilized a paired t-test and Cohen's d to quantify and compare methodological differences.

The results are presented in categories: The first, shown in Figure 6, includes images with significant noise, while the

Table 2. Confusion Matrix of Figure 4.

TP	0.929784894	FN	0.070215106
FP	0.691223951	TN	0.308776049

Table 3. TPR and FPR values from Table 2.

$P(\text{cont})_i$	$P(+)_i$	$P(-)_i$	TPR_i	FPR_i
0	0.070215106	0.308776049	0	0
0.1	0.156172085	0.347020839	0.595359211	0.199188023
0.2	0.242129064	0.38526563	0.768007673	0.358829803
0.3	0.328086043	0.42351042	0.850189987	0.489638922
0.4	0.414043021	0.46175521	0.89824955	0.598779558
0.5	0.5	0.5	0.929784894	0.691223951
0.6	0.585956979	0.53824479	0.952068081	0.770531138
0.7	0.671913957	0.57648958	0.968649956	0.839315717
0.8	0.757870936	0.61473437	0.981470432	0.899541635
0.9	0.843827915	0.652979161	0.991678978	0.952712725
1	0.929784894	0.691223951	1	1

Table 4. AUC associated with cutoff points number.

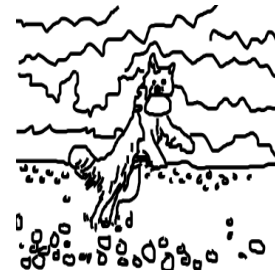
N	AUC (Classifier A)	AUC (Classifier B)
50	0.053239	0.712117
100	0.052743	0.712193
1000	0.052541	0.712216
10000	0.052537	0.712218
100000	0.052537	0.712218

second, illustrated in Figure 7, contains images with minimal noise. The third presents a comprehensive evaluation of the entire dataset, including mean AUC and ACC values, alongside statistical comparisons using a paired t-test and Cohen's d.

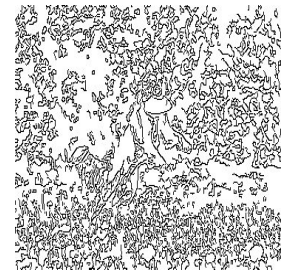
Figure 6 illustrates the first category. The images in this category are characterized by significant noise, making it difficult to clearly distinguish the objects of interest. The ROC curves produced by machine learning methods, empirical method, and SPROC are similar, as shown in Figure 6(c). It can identify a great deal of noise in the images in this category due to many FPs. The number of FPs and TPs found is nearly equal, resulting in all curves displaying an AUC value close to 0.5.

Figure 7 illustrates the second category of images, characterized by minimal noise. These images preserve the objects of interest, thereby facilitating their visual identification. Figures 7(c) and 7(f) show a significant difference between ROC curves created by SPROC and those produced by standard machine learning and empirical algorithms. The images in this category exhibit minimal noise due to the high number of true positives (TPs). It is evident that the number of false positives (FPs) is lower than the identified TPs, although the number of FPs remains significant. SPROC generated a curve with a larger area, demonstrating the method's superior accuracy in identifying objects within acceptable quality images.

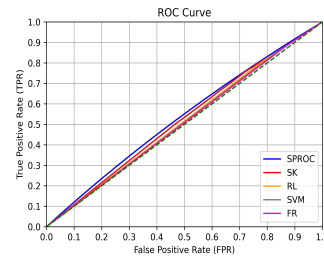
The AUC value estimator considers the curve's structure, and the number of points generated for its formation is critical in determining a precise AUC value. In contrast to the proposed approach in this work, which generated 100 cutting points for curve creation, traditional machine learning and empirical algorithms produced only a limited number



(a) Objects contour (GT).



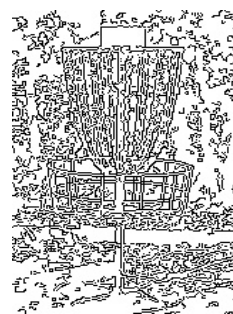
(b) Canny filter.



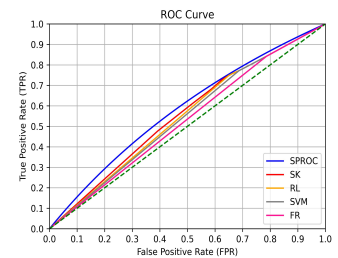
(c) ROC curve.



(d) Objects contour (GT).



(e) Canny filter.



(f) ROC curve.

Figure 6. Noisy images.

of connecting points. This resulted in less refined curves, significantly impacting the accuracy of the AUC calculation. Despite the high number of false positives (FPs) associated with the Canny filter, SPROC was able to estimate a curve

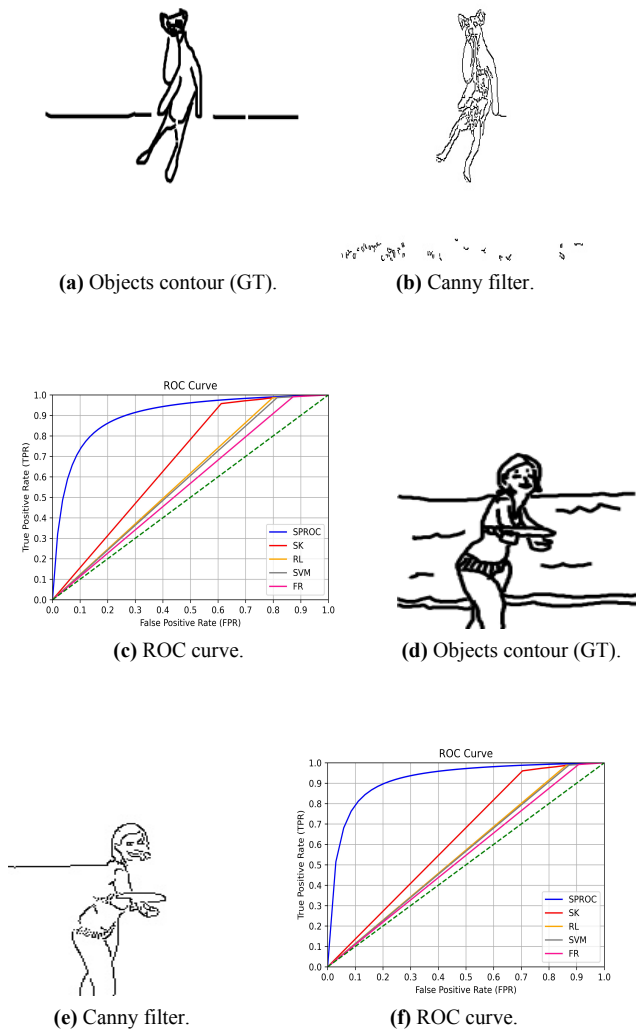


Figure 7. Images showing minimal noise.

more closely aligned with visual perception in terms of object detection.

The Tables 5 and 6 present the estimates for Figure 6 and 7, respectively. As shown in Table 5, the values of the AUC, ACC, and PPV estimator values are very similar across all the methods analyzed. This occurs because the number of FPs is close to the number of TPs, resulting in noisy images and difficult visual perception of objects. The low PPV values, approximately 0.5 across all methods, indicate a high occurrence of FPs. The F1-score, with values around 0.6, reflects the combination of TPR and PPV assessments, highlighting relatively low efficiency. In the case of the images in Figure 6, these values suggest that the model is committing a high number of FP errors, but not for false negatives (FNs).

In Table 6, the ACC, PPV and F1-score estimator values are similar, indicating that the ROC curves generated by the machine learning and empirical algorithms are comparable. However, the AUC values for SPROC are higher, as the images shown in Figures 7(b) and 7(e) show significant differences from Figure 6. The images enable the visual identification of objects of interest with greater ease due to the high number of points generated by SPROC. The values obtained by SPROC for the AUC of 0.8993 in Figure 7(b) and 0.9135 in Figure 7(e) indicate excellent performance of the

classifier in the distinction between classes, even when the model contains a high number of FPs. Machine learning and empirical algorithms produced low AUC values by generating a limited number of points. These values do not accurately represent the visual perception of images in the context of object detection. Other techniques may be used to improve the performance of machine learning-based methods. One such approach is the estimation of the maximum likelihood (EVM), which statistically increases the number of points and results in a more accurate graph. In this study, we applied EVM to the dataset; however, it did not produce significantly different results from those presented in the Figures 6 and 7.

Another important aspect to consider in parametric curves is the knowledge of data distribution, as the accuracy of the curve relies on the proper specification of the underlying model. Incorrect model selection can pose significant challenges and lead to biased conclusions. The Canny filter does not inherently produce a known statistical distribution; rather, it generates a distribution that is highly dependent on the content of the analyzed image. Table 7 displays the advantages and disadvantages of generating the empirical, parametric, and SPROC ROC curves. Furthermore, in parametric ROC curves, the dataset goes through the supervised machine learning algorithm's training phase. Both the empirical and parametric methods calculate the class values simultaneously with the TPR and FPR values during the dataset collection process.

Tables 5 and 6 show that SPROC has only two outlier points in Figure 7, out of a total of 100 cutting points. When SPROC generates more than 100 cutting points, the number of outlier points remains constant. The SK algorithm outperformed machine learning algorithms in terms of cutting points, but it had an excess of LOF points. The FR algorithm produced no LOF points, but only three for the ROC curve. SVM generated only five points for curve construction and two LOF points, whereas the RL algorithm produced LOFs for all images.

The third category of results presents the mean AUC and ACC metrics for the entire dataset, along with statistical comparisons performed using the paired t-test and Cohen's d effect size measure. Table 8 presents the mean ACC and AUC values for the entire dataset, considering all algorithms used in this study.

The combination of ACC and AUC enables the assessment of model quality in classification tasks. SPROC algorithm demonstrates the best overall performance. SK exhibits ACC similar to the SPROC; however, its lower AUC (0.6265) suggests greater difficulty in correctly distinguishing classes. RL shows ACC within the same range as the SK algorithm, but its AUC of 0.5743 suggests less efficient class separation. SVM presents inferior performance, with ACC and AUC lower than those of other algorithms, indicating low classification quality. FR shows the worst overall performance, with ACC and AUC lower than those of other algorithms, which may indicate that this model is failing to learn useful patterns for class separation.

The high paired t-test values presented in Table 9 indicate significant differences among the SPROC algorithm compared to SK, RL, SVM, and FR. SPROC with FR exhibited the highest t-value ($t = 58.5294$), suggesting a statistically

Table 5. Estimators for Figure 6

Fig. 6 (c)	SPROC	SK	RL	SVM	FR
AUC	0.5342	0.5219	0.5120	0.5092	0.5069
ACC	0.5194	0.5180	0.5123	0.5069	0.5069
PPV	0.5130	0.5120	0.5076	0.5040	0.5040
F1-score	0.6148	0.6136	0.6414	0.6414	0.6414
LOF	0	3	2	2	0
Number of points (N)	100	32	7	5	3
Fig. 6 (f)	SPROC	SK	RL	SVM	FR
AUC	0.5834	0.5569	0.5482	0.5421	0.5277
ACC	0.5520	0.5497	0.5491	0.5277	0.5277
PPV	0.5369	0.5356	0.5351	0.5170	0.5170
F1-score	0.6284	0.6243	0.6241	0.6407	0.6407
LOF	0	12	5	2	0
Number of points (N)	100	12	7	5	3

Table 6. Estimators for Figure 7

Fig. 7 (c)	SPROC	SK	RL	SVM	FR
AUC	0.8993	0.6746	0.5940	0.5863	0.5595
ACC	0.5995	0.5940	0.5940	0.5595	0.5595
PPV	0.5559	0.5527	0.5527	0.5320	0.5320
F1-score	0.7119	0.7083	0.6923	0.6923	0.6923
LOF	2	11	7	2	0
Number of points (N)	100	11	8	5	3
Fig. 7 (f)	SPROC	SK	RL	SVM	FR
AUC	0.9135	0.6290	0.5625	0.5581	0.5423
ACC	0.5680	0.5624	0.5625	0.5423	0.5423
PPV	0.5367	0.5337	0.5338	0.5223	0.5223
F1-score	0.6973	0.6930	0.6930	0.6842	0.6842
LOF	2	10	5	2	0
Number of points (N)	100	10	6	5	3

more pronounced advantage. These results demonstrate that SPROC significantly improves model performance, with statistical significance ($p < 0.05$) confirming this effect is not due to random variation. The comparative analysis (Table 9) demonstrates that the SPROC outperforms the SK algorithm by 1.3687 standard deviations, as indicated by Cohen's d metric. According to the criteria presented in Table 1, this difference represents a substantially high effect, highlighting a considerable statistical superiority. Furthermore, the p -value, which is close to zero, strengthens the robustness of the statistical analysis of the classified data, providing compelling evidence against the null hypothesis. These findings indicate that the observed disparity between the algorithms is not attributable to chance but rather reflects a significant impact on the performance of the tested algorithms, with a particular emphasis on the advantage demonstrated by the SPROC.

5 Conclusion

ROC analysis is a statistical technique commonly used to evaluate image classifier models. Three main approaches are used to generate points in the construction of the curve: empirical, parametric, and semiparametric. Few studies use a semiparametric approach to create points on the ROC curve.

The empirical analysis uses all points, resulting in an irregular curve. Currently, machine learning algorithms generate the curve of the parametric approach, making it the most widely used method. These algorithms generally generate a limited number of points to trace the curve, which can result in outliers that affect the accuracy of determining the exact value of the area under the curve (AUC). The aim of this study was to introduce a novel semiparametric technique based on Bayes' theorem to avoid the impact of divergent points in curve generation without relying on machine learning.

The programming language Python was selected, and a dataset with 1000 images was utilized as a reference for the tests. We applied the Canny filter to all images to verify the proposed algorithm (SPROC) with the machine learning algorithms Logistic Regression, SVM, Random Forest, and an empirical approach referred to in this study as SKlearn. SPROC was evaluated against these algorithms using the metrics AUC, PPV, F1-score, paired t -tests, Cohen's d and Local Outlier Factor (LOF) estimators, categorized into three groups. The first refers to images with minimal noise, which makes it possible to identify the objects present. In this case, the metric values were similar across all the evaluated images, as the number of FPs was close to the number of TP. The second refers to images with significant noise. SPROC produced a smooth curve with a significantly higher AUC than the other algorithms. The third category reports mean

Table 7. Comparative analysis of the methodologies employed in the generation of ROC curves.

Empirical	Parametric	SPROC
No machine learning.	Uses machine learning.	No machine learning.
Exhibits a non-uniform curvature.	Shows a smooth curve.	Shows a smooth curve.
No need for knowledge of the data distribution.	Knowledge of the data distribution is necessary.	No need for knowledge of the data distribution.
All points are considered for plotting the curve.	Few points are used in plotting the curve.	The curve shape and results remain unaffected by the number of points chosen, provided that they exceed 100.
No need for specifying an underlying model, making it more flexible.	A better fit occurs: When the selected parametric model fits the data well, the parametric ROC curve can provide a more accurate representation of the relationship between true and false positives.	No need for specifying an underlying model, making it more adaptable and better representing the relationship between true and false positives.
Intensive computing: The curve is constructed using all the image's points.	Intensive computing: It requires a dataset to train the chosen machine learning model.	Computation is simple: No need for the use of machine learning.
Low accuracy when it involves high spatial resolution imagery. It may have limited accuracy due to the need to estimate the probability density in high-dimensional regions.	More accuracy when it involves a high-resolution image to train the machine learning algorithm.	Same accuracy in images with low and high resolution.

Table 8. Mean ACC and AUC metrics for the entire dataset.

Algorithm	ACC	AUC
SPROC	0.5762	0.7141
SK	0.5744	0.6265
RL	0.5743	0.5743
SVM	0.5479	0.5687
FR	0.5478	0.5478

AUC/ACC across the dataset, with statistical comparisons (paired t-test, Cohen's d). This study clearly demonstrated that the AUC calculation considers the curve's structure to provide a more accurate value, and SPROC exhibited the highest t-value (paired t-test), suggesting a statistically robust advantage with Cohen's $d > 0.8$. Furthermore, the machine learning algorithms generated a limited number of points for plotting the curve, which impacted the AUC calculation and resulted in more potential outliers.

It is important to note that the precision-recall curve may be the best option for models with high false positive (FP) rates because it accurately shows how well the model is doing and gives better insights for improving the model, especially for models where $FP \geq TP$. SPROC can be easily adapted to accommodate the Precision-Recall curve, simply by changing from FPR to Precision. For further research, it is recommended to use Youden's statistic, Euclidean distance, and the maximum product of sensitivity and specificity to determine the optimal cutoff value for the SPROC method. Other studies can be conducted to determine alternative methods for calculating the AUC with SPROC, such as the DeLong method, and compare it with the trapezoidal method used in

this study.

Declarations

Acknowledgements

The authors would like to thank [Li et al., 2019] for providing the necessary image resources to run the experiments.

Funding

This research was funded by the Brazilian agency of the National Council for Scientific and Technological Development (CNPq) through the CNPq-SETEC/MEC Call (2014) and by the Federal Institute of Education, Science and Technology of São Paulo (IFSP).

Authors' Contributions

RB contributed to the conception of this study. RM, RB and RS performed the experiments. SM and RB analyzed the results. SM and RM review the Materials and Methods. RB and RM are the main contributors and writers of this manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they don't have competing interests.

Availability of data and materials

Generated data can be made available upon request.

Table 9. Performance comparison between SPROC and machine learning models (full dataset).

	SPROC with SK	SPROC with RL	SPROC with SVM	SPROC with FR
Paired t-test	43.2816	53.3679	54.5543	58.5294
p-value	2.4667×10^{-231}	9.4272×10^{-295}	7.4018×10^{-302}	0.0000
Cohen's d	1.3687	1.6876	1.7252	1.8509

References

- Alghushairy, O., Alsini, R., Soule, T., and Ma, X. (2020). A review of local outlier factor algorithms for outlier detection in big data streams. DOI: 10.3390/bdcc5010001.
- Breiman, L. (2001). Random forests. 45:5–32. DOI: 10.1023/A:1010933404324.
- Bueno, R. C., Masotti, P. H., Justo, J. F., Andrade, D. A., Rocha, M. S., Torres, W. M., and de Mesquita, R. N. (2018). Two-phase flow bubble detection method applied to natural circulation system using fuzzy image processing. *Nuclear Engineering and Design*, 335:255–264. DOI: 10.1016/j.nucengdes.2018.05.026.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1):155–159. DOI: 10.1037/0033-2909.112.1.155.
- Cook, J. A. (2017). Roc curves and nonrandom data. *Pattern Recognition Letters*, 85:35–41. DOI: 10.1016/j.patrec.2016.11.015.
- Du, Z. X., Chang, F. Q., Wang, Z. J., Zhou, D. M., Li, Y., and Yang, J. H. (2022). A risk prediction model for acute kidney injury in patients with pulmonary tuberculosis during anti-tuberculosis treatment. *Renal Failure*, 44:625–635. DOI: 10.1080/0886022X.2022.2058405.
- Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874. DOI: 10.1016/j.patrec.2005.10.010.
- Friedman, N., Geiger, D., Provan, G., Langley, P., and Smyth, P. (1997). Bayesian network classifiers *. 29:131–163. Available at: <https://link.springer.com/article/10.1023/A:1007465528199>.
- Gao, Y., Li, T., Han, M., Li, X., Wu, D., Xu, Y., Zhu, Y., Liu, Y., Wang, X., and Wang, L. (2020). Diagnostic utility of clinical laboratory data determinations for patients with the severe covid-19. *Journal of Medical Virology*, 92:791–796. DOI: 10.1002/jmv.25770.
- Ghamry, F. M., El-Banby, G. M., El-Fishawy, A. S., El-Samie, F. E., and Dessouky, M. I. (2024). A survey of anomaly detection techniques. *Journal of Optics (India)*, 53:756–774. DOI: 10.1007/s12596-023-01147-4.
- Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., and Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25:65–69. DOI: 10.1038/s41591-018-0268-3.
- He, X., Gallas, B. D., and Frey, E. C. (2010). Three-class roc analysis toward a general decision theoretic solution. *IEEE Transactions on Medical Imaging*, 29:206–215. DOI: 10.1109/TMI.2009.2034516.
- Hong, H., Liu, J., Bui, D. T., Pradhan, B., Acharya, T. D., Pham, B. T., Zhu, A. X., Chen, W., and Ahmad, B. B. (2018). Landslide susceptibility mapping using j48 decision tree with adaboost, bagging and rotation forest ensembles in the guangchang area (china). *Catena*, 163:399–413. DOI: 10.1016/j.catena.2018.01.005.
- Keidar, D., Yaron, D., Goldstein, E., Shachar, Y., Blass, A., Charbinsky, L., Aharon, I., Lifshitz, L., Lumelsky, D., Neeman, Z., Mizrahi, M., Hajouj, M., Eizenbach, N., Sela, E., Weiss, C. S., Levin, P., Benjaminov, O., Shabshin, N., Elyada, Y. M., and Eldar, Y. C. (2020). Covid-19 classification of x-ray images using deep neural networks. DOI: 10.1007/s00330-021-08050-1/Published.
- Khawaja, A. M., Asayesh, B. M., Hainzl, S., and Schorlemmer, D. (2023). Towards improving the spatial testability of aftershock forecast models. *Natural Hazards and Earth System Sciences*, 23:2683–2696. DOI: 10.5194/nhess-23-2683-2023.
- Khosravi, K., Pham, B. T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Prakash, I., and Bui, D. T. (2018). A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at haraz watershed, northern iran. *Science of the Total Environment*, 627:744–755. DOI: 10.1016/j.scitotenv.2018.01.266.
- Kun-Peng, Z., Xiao-Long, M., and Chun-Lin, Z. (2018). Overexpressed circpvt1, a potential new circular rna biomarker, contributes to doxorubicin and cisplatin resistance of osteosarcoma cells by regulating abcb1. *International Journal of Biological Sciences*, 14:321–330. DOI: 10.7150/ijbs.24360.
- Li, K., Fang, Y., Li, W., Pan, C., Qin, P., Zhong, Y., Liu, X., Huang, M., Liao, Y., and Li, S. (2020). Ct image visual quantitative evaluation and clinical classification of coronavirus disease (covid-19). DOI: 10.1007/s00330-020-06817-6/Published.
- Li, M., Lin, Z., Mech, R., Yumer, E., and Ramanan, D. (2019). Photo-sketching: Inferring contour drawings from images. DOI: 10.1109/wacv.2019.00154.
- Martin, O. (2024). *Bayesian Analysis with Python - Third Edition: A Practical Guide to Probabilistic Modeling*. Packt Publishing. Book.
- McGowan, L. D., Bullen, J. A., and Obuchowski, N. A. (2016). Location bias in roc studies. *Statistics in Biopharmaceutical Research*, 8:258–267. DOI: 10.1080/19466315.2016.1173583.
- Moreira, D. (2020). Comparing empirical roc curves using a java application: Cercus. DOI: 10.1007/978-3-030-24302-9_3.
- Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Korean Journal of Anesthesiology*, 75:25–36. DOI: 10.4097/kja.21209.
- Niu, M., Song, K., Huang, L., Wang, Q., Yan, Y., and Meng, Q. (2021). Unsupervised saliency detection of rail surface defects using stereoscopic images. *IEEE Transactions on Industrial Informatics*, 17:2271–2281. DOI:

- 10.1109/TII.2020.3004397.
- Pourghasemi, H. R. and Rahmati, O. (2018). Prediction of the landslide susceptibility: Which algorithm, which precision? *Catena*, 162:177–192. DOI: 10.1016/j.catena.2017.11.022.
- Sachs, M. C. (2017). Plotroc: A tool for plotting roc curves. *Journal of Statistical Software*, 79. DOI: 10.18637/jss.v079.c02.
- Schott, S. M. C., da Silva, M. C. B., de Andrade, D. A., and de Mesquita, R. N. (2024). Convolutional neural network-based pattern recognition in natural circulation instability images. *Concilium*, 24:267–288. DOI: 10.53660/clm-2919-24d10.
- Shkurnikov, M., Nersisyan, S., Jankevic, T., Galatenko, A., Gordeev, I., Vechorko, V., and Tonevitsky, A. (2021). Association of hla class i genotypes with severity of coronavirus disease-19. *Frontiers in Immunology*, 12. DOI: 10.3389/fimmu.2021.641900.
- Student (1908). The probable error of a mean. *Biometrika*, 6(1):1–25. DOI: 10.2307/2331554.
- Termeh, S. V. R., Kornejady, A., Pourghasemi, H. R., and Keesstra, S. (2018). Flood susceptibility mapping using novel ensembles of adaptive neuro fuzzy inference system and metaheuristic algorithms. *Science of the Total Environment*, 615:438–451. DOI: 10.1016/j.scitotenv.2017.09.262.
- Wang, D., Fan, G., Wu, S., Yang, T., Xu, J., Yang, L., Zhao, J., Zhang, X., Bai, C., Kang, J., Ran, P., Shen, H., Wen, F., Huang, K., Chen, Y., Sun, T., Shan, G., Lin, Y., Xu, G., Wang, R., Shi, Z., Xu, Y., Ye, X., Song, Y., Wang, Q., Zhou, Y., Li, W., Ding, L., Wan, C., Yao, W., Guo, Y., Xiao, F., Lu, Y., Peng, X., Zhang, B., Xiao, D., Wang, Z., Bu, X., Zhang, H., Zhang, X., An, L., Zhang, S., Zhu, J., Cao, Z., Zhan, Q., Yang, Y., Liang, L., Dai, H., Cao, B., He, J., and Wang, C. (2022). Development and validation of a screening questionnaire of copd from a large epidemiological study in china. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 19:118–124. DOI: 10.1080/15412555.2022.2042504.
- Wu, J.-p., Ding, W.-Z., Wang, Y.-L., Liu, S., Zhang, X.-q., Yang, Q., Cai, W.-J., Yu, X.-l., Liu, F.-y., Kong, D., et al. (2022). Radiomics analysis of ultrasound to predict recurrence of hepatocellular carcinoma after microwave ablation. *International Journal of Hyperthermia*, 39(1):595–604. DOI: 10.1080/02656736.2022.2062463.
- Zeiler, M. D. and Fergus, R. (2014). Lncs 8689 - visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901. DOI: 10.48550/arXiv.1311.2901.
- Zhao, S., Pan, H., Guo, Q., Xie, W., and Wang, J. (2022). Platelet to white blood cell ratio was an independent prognostic predictor in acute myeloid leukemia. *Hematology (United Kingdom)*, 27:426–430. DOI: 10.1080/16078454.2022.2055857.
- Zhao, W., Lu, M., Wang, X., and Guo, Y. (2021). The role of sarcopenia questionnaires in hospitalized patients with chronic heart failure. *Aging Clinical and Experimental Research*, 33:339–344. DOI: 10.1007/s40520-020-01561-9.