A Consolidated Overview of Evaluation and Performance Metrics for Machine Learning and Computer Vision

Tobias Schlosser, Michael Friedrich, Trixy Meyer, and Danny Kowerko

Junior Professorship of Media Computing,

Chemnitz University of Technology,

09107 Chemnitz, Germany,

{firstname.lastname}@cs.tu-chemnitz.de

May 11, 2024

Abstract

Within the rapidly evolving fields of machine learning (ML) and computer vision (CV), the selection and application of appropriate evaluation and performance metrics are paramount to the advancement and validation of innovative models. These metrics not only guide the development and fine-tuning of algorithms but also shape the interpretation of outcomes in diverse applications, ranging from automated disease diagnosis to autonomous driving. This manuscript is therefore also driven by the recent advancements in artificial intelligence (AI) and deep learning (DL) with a focus on big data analysis in areas such as object detection and classification by utilizing learning-based approaches including artificial neural networks (ANN) such as deep neural networks (DNN). For this purpose, we present well-established evaluation metrics, detailing their different advantages, disadvantages, and related origins. By consolidating current knowledge and offering insights into the effective use of these metrics, the goal of this work is to equip researchers and practitioners with the essential tools to critically evaluate and enhance the robustness and reliability of their models. Since this manuscript is designed to be a living resource, it will be systematically updated to incorporate novel evaluation metrics, ensuring its continued development.

Table of Contents

1 Introduction and Motivation **4**
1.1 The need for comprehensive evaluation metrics . . . 4
1.2 Objectives of this manuscript . . . 4

2 Machine Learning **5**
2.1 General . . . 6
2.2 Precision / positive predictive value (PPV) . . . 7
2.2.1 Macro average precision (APmacro) . . . 7
2.2.2 Micro average precision (APmicro) . . . 8
2.2.3 Weighted average precision (APweighted) . . . 8
2.3 Negative predictive value (NPV) . . . 8
2.4 Recall / true positive rate (TPR) / sensitivity / hit rate . . . 9
2.4.1 Macro average recall (ARmacro) . . . 9
2.4.2 Micro average recall (ARmicro) . . . 10
2.4.3 Weighted average recall (ARweighted) . . . 10
2.5 True negative rate (TNR) / specificity / selectivity . . . 10
2.6 Prevalence . . . 10
2.7 Accuracy (A) . . . 11
2.8 Balanced accuracy (BA) . . . 11
2.9 Balanced accuracy weighted (BAW) . . . 11
2.10 Average accuracy (AA) . . . 12
2.11 Average class accuracy (ACA) . . . 12
2.12 Error rate (ER) . . . 12
2.13 Average error rate (AER) . . . 12
2.14 F-score / Fβ-score . . . 13
2.14.1 Macro average F-score . . . 13
2.14.2 Micro average F-score . . . 14
2.14.3 Weighted average F-score . . . 14
2.14.4 F0-score . . . 14
2.14.5 F0.5-score . . . 14
2.14.6 F1-score . . . 15
2.14.7 F2-score . . . 15
2.15 False discovery rate (FDR) . . . 15
2.15.1 Macro average FDR (FDRmacro) . . . 16
2.15.2 Micro average FDR (FDRmicro) . . . 16
2.15.3 Weighted average FDR (FDRweighted) . . . 16
2.16 False omission rate (FOR) . . . 16
2.17 False positive rate (FPR) . . . 17
2.18 False negative rate (FNR) / miss rate . . . 17
2.18.1 Macro average FNR (FNRmacro) . . . 17
2.18.2 Micro average FNR (FNRmicro) . . . 18
2.18.3 Weighted average FNR (FNRweighted) . . . 18
2.19 Positive likelihood ratio (LR+) . . . 18
2.20 Negative likelihood ratio (LR-) . . . 18
2.21 Diagnostic odds ratio (DOR) . . . 19
2.22 Fowlkes–Mallows index (FM) . . . 19

2.23 Informedness / bookmaker informedness (BM) / Youden's J statistic / Youden's index . 19
2.24 Markedness (MK) . . . 20
2.25 Matthews correlation coefficient (MCC) / phi coefficient / Yule phi coefficient . . . 20
2.26 Jaccard index (JI) / threat score (TS) / critical success index (CSI) . . . 20
2.27 Receiver operating characteristic curve (ROC curve) and area under the curve (AUC) . 21
2.28 Average precision (AP) . . . 22
2.29 Mean average precision (mAP) . . . 23
2.30 Cohen's kappa for binary classification . . . 23
2.31 Gini impurity . . . 24
2.32 P4 metric . . . 24
2.33 Skill score (SS) . . . 24
2.34 Relative improvement factor . . . 25
2.35 Examples . . . 26
2.36 Functions . . . 28

3 Computer Vision **29**
3.1 General . . . 30
3.2 Error (E) . . . 30
3.3 Absolute error / sum of absolute errors (AE) . . . 30
3.4 Relative absolute error (RAE) . . . 31
3.5 Mean error (ME) . . . 31
3.6 Mean percentage error (MPE) . . . 31
3.7 Mean absolute error (MAE) . . . 32
3.8 Mean absolute percentage error (MAPE) . . . 32
3.9 Mean absolute scaled error (MASE) . . . 32
3.10 Mean normalized bias (MNB) . . . 32
3.11 Normalized mean bias (NMB) . . . 33
3.12 Squared error / sum of squared errors (SE) . . . 33
3.13 Mean square error (MSE) . . . 33
3.14 Root mean square error (RMSE) . . . 34
3.15 Normalized root mean square error (NRMSE) . . . 34
3.16 Root mean squared logarithmic error (RMSLE) . . . 34
3.17 Peak signal-to-noise ratio (PSNR) . . . 35
3.18 Structural similarity (SSIM) . . . 35
3.19 Structural dissimilarity (DSSIM) . . . 35
3.20 Intersection over union (IoU) . . . 36
3.21 Dice coefficient (DC) . . . 37
3.22 Overlap coefficient (OC) . . . 37
3.23 Functions . . . 38

1 Introduction and Motivation

Within the recent years, the fields of machine learning (ML) and deep learning (DL) have witnessed remarkable growth fueled by advancements in algorithms, data availability, and computational power [1–6]. This evolution has led to significant improvements across a variety of applications including object detection and classification. Fundamental to these developments is the use of learning-based approaches such as deep neural networks (DNN), which are increasingly favored due to their powerful feature extraction and learning capabilities. However, the expansion of these technologies also presents new challenges, particularly in evaluating and optimizing performance under aspects of model interpretability, accountability, and robustness [7–9]. The effectiveness of ML models in real-world scenarios depends on their ability to generalize from training data to unseen data, necessitating robust evaluation metrics that can accurately reflect model performance across diverse conditions and data sets.

1.1 The need for comprehensive evaluation metrics

Evaluation metrics serve as fundamental tools that provide insights into the effectiveness of machine learning models, guiding the selection, tuning, and optimization of models in scientific research and commercial applications. In computer vision (CV), the complexity of tasks such as image segmentation, object recognition, and motion analysis further complicates the assessment of model performance. Traditional metrics such as accuracy or error rate are often insufficient to capture the nuances of model behavior, especially in data sets with imbalanced classes. Moreover, the dynamic nature of ML and CV, with continually evolving models and techniques, demands a standardized yet comprehensive set of metrics that can adapt to new challenges. This need highlights the importance of utilizing well-established metrics that can offer deeper insights and facilitate the effective comparison of state-of-the-art techniques [10].

1.2 Objectives of this manuscript

This manuscript provides a comprehensive overview of evaluation and performance metrics suitable for ML and CV applications. It details established metrics, examining their strengths, weaknesses, and foundational concepts, thereby aiding in their appropriate selection and application. By consolidating an overview of traditional as well as novel metrics, this manuscript therefore serves as a comprehensive resource, supporting researchers and practitioners in accurately assessing and enhancing model performance.

2 Machine Learning

In the continuously evolving landscape of machine learning, the evaluation of predictive models through various metrics is integral to understanding their performance and guiding their optimization. Evaluation and performance metrics such as precision, recall, and accuracy provide foundational insights, while more nuanced metrics address specific aspects of model behavior and class imbalances.

For example, precision [11, 12], also known as the positive predictive value, gauges the accuracy of positive predictions made by a model, indicating the proportion of true positive results in all positive classifications. This evaluation metric is vital in contexts where the cost of false positives is high, such as in medical diagnostics [13, 14] or fraud detection [15, 16]. To accommodate the characteristics of diverse data sets, precision is often calculated in one of three forms: macro average [17–20], micro average [17, 18], or weighted average precision [21]. Macro average precision treats all classes equally, averaging the precision calculated for each class, which prevents dominant classes from overshadowing smaller ones. Micro average precision aggregates the contributions of all classes to compute the overall precision, which is ideal for handling class imbalances. Weighted average precision, however, assigns a weight to each class's precision based on its representation in the data set, providing a balance between treating all classes equally and recognizing their different sizes.

Recall [22, 23], or the true positive rate, complements precision by measuring the ability to detect all actual positives. It is crucial for applications where missing a positive instance carries severe consequences, such as detecting rare diseases. Like precision, recall can also be categorized into macro [17, 18, 24, 25], micro [17, 18], and weighted averages [21, 26] to adapt to various needs and data set structures. Negative predictive value [11, 12] and true negative rate [22, 23] measure the accuracy of negative predictions and the ability to identify negatives, respectively, which may help in rounding out the model's ability to correctly predict both classes. Accuracy [27, 28] provides an overall measure of model performance but can be misleading in unbalanced data sets. Balanced accuracy [29, 30] and its variations, including weighted balanced accuracy [31, 32], provide a more trustworthy assessment by considering both sensitivity (recall) and specificity (true negative rate) adjusted for prevalence [33, 34]. The F-score [35, 36], including variations such as the F1-score [35, 36] and the F2-score [35, 36], combines precision and recall in a harmonic mean, balancing the trade-off between them. It is especially useful when seeking a model that reliably balances false positives and false negatives. Error metrics including the false discovery rate [37, 38] and false omission rate [39] delve deeper into the types of errors a model may show, providing insights that can be critical for refining model performance. The false discovery rate, for example, focuses on the proportion of false positives among the positive results, which is particularly relevant in scientific testing where validation is costly.

Further sophistication in evaluation metrics is seen in measures such as the diagnostic odds ratio [40, 41], which compares the odds of positive test results, and the Fowlkes–Mallows index [42, 43], which measures the similarity between the predicted and true classifications. These metrics provide additional layers of understanding in scenarios where a simple accuracy is insufficient. Metrics including the positive likelihood ratio [44, 45] and the negative likelihood ratio [44, 45] provide evidence strength that a given condition or feature corresponds to a specific class, which is useful in diagnostic applications. The informedness or bookmaker informedness [46, 47] provides, as opposed to random guessing, a measure of decision effectiveness.

Collectively, these metrics offer a multidimensional framework. They not only assess basic performance but also provide insights into the behavior of models across different scenarios, fostering more informed and effective machine learning applications.

The following evaluation metrics within the context of machine learning are motivated by the contributions of *Metz* [27] and *Fawcett* [48].

2.1 General

Figure 1 and Table 1 give an overview of our general definitions for ML-related metrics.

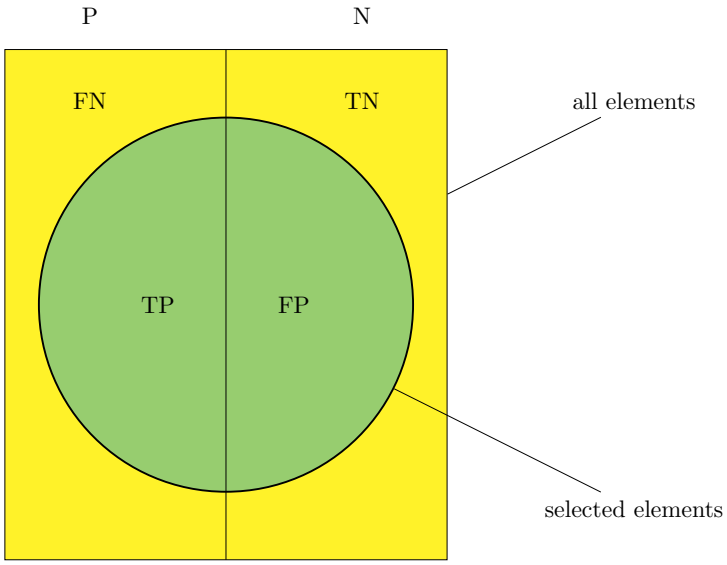

Figure 1: General definitions machine learning.

Abbreviation	Meaning
T	Total
P	Positives
N	Negatives
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
n	Number of values or classes

Table 1: General definitions machine learning.

2.2 Precision / positive predictive value (PPV) [11, 12]

Precision [11, 12] is an essential evaluation metric used to assess the accuracy of classification models, particularly focusing on the correctness of positive predictions. It measures the proportion of predicted positive instances that are truly positive, therefore providing critical insights into the model's performance in contexts where the cost of a false positive is significant, such as in spam detection or legal proceedings. By quantifying how many of the model's positive classifications are accurate, precision determines the reliability of the model's positive predictions. A high precision score indicates that the model is effective in minimizing false positives, thereby ensuring that most of its positive predictions are trustworthy. This metric is particularly valuable when the consequences of erroneous positive predictions are costly or disruptive. However, maximizing precision alone can sometimes lead to a model that is overly conservative in its positive predictions, potentially missing genuine positive cases (low recall). Therefore, in practice, precision is often considered alongside recall to optimize both the accuracy and coverage of the model's predictions, commonly evaluated through the F1-score, which harmonizes the trade-off between precision and recall to provide a more holistic view of model performance.

Fraction of true positives and false positives. (range: $[0, 1]$)

+ Minimizes false positives.
+ Useful for critical false positive sensitive applications.
− Ignores true negatives and false negatives.
− May overlook recall for higher precision.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

2.2.1 Macro average precision ($\mathbf{AP_{macro}}$) [17–20]

Macro average precision calculates the average precision for each class separately and takes the average over all classes. (range: $[0, 1]$)

+ Each class is weighted equally regardless of its frequency in the data set, therefore allowing each class to be evaluated individually.
+ Advantageous when the recognition rate of TP compared to FP is of interest.
− Rare classes have the same weight as frequent classes. This can lead to bias.
− Only partially evaluates a model's classification capabilities.

$$AP_{macro} = \frac{1}{n} \cdot \sum_{i=1}^{n} Precision_i = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{TP_i}{TP_i + FP_i} \tag{2}$$

2.2.2 Micro average precision ($\mathbf{AP_{micro}}$) [17, 18]

Micro average precision calculates the precision across all classes. (range: $[0, 1]$)

+ Use with inconsistent data sets / class distributions. Provides overview of overall model performance.
− Neglects infrequent classes and overestimates frequent classes.

$$AP_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FP_i)} \tag{3}$$

2.2.3 Weighted average precision ($\mathbf{AP_{weighted}}$) [21]

Weighted average precision calculates the weighted precision across all classes. (range: $[0, 1]$)

+ Use with inconsistent data sets / class distributions. Provides overview of overall model performance.
− Neglects under-weighted classes and overestimates over-weighted classes.

$$AP_{weighted} = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{i=1}^{n} w_i \cdot Precision_i = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{i=1}^{n} w_i \cdot \frac{TP_i}{TP_i + FP_i} \tag{4}$$

where: w_i = samples per class or the relation of samples per class to the total of all samples

2.3 Negative predictive value (NPV) [11, 12]

Fraction of true negatives and false negatives. (range: $[0, 1]$)

+ Emphasizes the accuracy of classifying true negatives in relation to false negatives.
− Neglects true positives and false positives.

$$NPV = \frac{TN}{TN + FN} \tag{5}$$

2.4 Recall / true positive rate (TPR) / sensitivity / hit rate [22, 23]

Recall [22, 23] particularly emphasizes the model's ability to identify all relevant instances within a data set. Commonly used in contexts where the cost of missing a positive case (false negative) is high, such as in medical diagnostics or fraud detection, recall calculates the proportion of actual positives that have been correctly identified by the model. It serves as a critical measure by quantifying the model's sensitivity to detecting positive samples amidst a pool of negatives, thereby highlighting potential weaknesses in capturing all pertinent cases. A high recall score indicates that the model effectively captures the majority of positive cases, which is paramount in scenarios where failing to detect positives could have severe consequences. Consequently, optimizing for recall is essential in fine-tuning the performance of models deployed in high-stakes environments, ensuring that few positive cases go unnoticed. However, focusing solely on recall can sometimes lead to a decrease in precision, whereas the model might also increase the number of false positives. Therefore, recall is often balanced with other metrics such as precision to provide a more comprehensive understanding of a model's overall performance.

Fraction of true positives and false negatives. (range: $[0, 1]$)

+ Ensures minimal false negatives.
+ Critical for identifying rare positive cases.
− Ignores true negatives and false positives.
− May overlook precision for higher recall.

$$Recall = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{6}$$

2.4.1 Macro average recall ($\mathbf{AR_{macro}}$) [17, 18, 24, 25]

Macro average recall calculates the average recall for each class separately and takes the average over all classes. (range: $[0, 1]$)

+ Each class is weighted equally regardless of its frequency in the data set, therefore allowing each class to be evaluated individually.
+ Advantageous when the recognition rate of TP compared to FN is of interest.
− Rare classes have the same weight as frequent classes. This can lead to bias.
− Only partially evaluates a model's classification capabilities.

$$AR_{macro} = \frac{1}{n} \cdot \sum_{i=1}^{n} Recall_i = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{TP_i}{TP_i + FN_i} \tag{7}$$

2.4.2 Micro average recall ($\mathbf{AR_{micro}}$) [17, 18]

Micro average recall calculates the recall across all classes. (range: $[0, 1]$)

+ Use with inconsistent data sets / class distributions. Provides overview of overall model performance.
− Neglects infrequent classes and overestimates frequent classes.

$$AR_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FN_i)} \tag{8}$$

2.4.3 Weighted average recall ($\mathbf{AR_{weighted}}$) [21, 26]

Weighted average recall calculates the weighted recall across all classes. (range: $[0, 1]$)

+ Use with inconsistent data sets / class distributions. Provides overview of overall model performance.
− Neglects under-weighted classes and overestimates over-weighted classes.

$$AR_{weighted} = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{i=1}^{n} w_i \cdot Recall_i = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{i=1}^{n} w_i \cdot \frac{TP_i}{TP_i + FN_i} \tag{9}$$

where: w_i = samples per class or the relation of samples per class to the total of all samples

2.5 True negative rate (TNR) / specificity / selectivity [22, 23]

Fraction of true negatives and false positives. (range: $[0, 1]$)

+ Emphasizes the accuracy of classifying true negatives in relation to false positives.
− Neglects true positives and false negatives.

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} \tag{10}$$

2.6 Prevalence [33, 34]

Fraction of positives with positives and negatives. (range: $[0, 1]$)

+ Straightforward to calculate.
− Neglects TP, TN, FP, and FN.

$$Prevalence = \frac{P}{T} = \frac{P}{P+N} \tag{11}$$

2.7 Accuracy (A) [27, 28]

Fraction of correct classifications. (range: $[0,1]$)

+ Emphasizes the accuracy of classifying true positives and true negatives.
− Misleading for imbalanced data (see also balanced accuracy).

$$A = \frac{TP+TN}{T} = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

2.8 Balanced accuracy (BA) [29, 30]

Balanced fraction of correct classifications. (range: $[0,1]$)

+ Normalizes true positives and true negatives.
− Neglects, e.g., FPR and FNR.

$$BA_{binary} = \frac{TPR+TNR}{2} = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2} \tag{13}$$

$$BA_{multi} = \frac{1}{n} \cdot \sum_{i=1}^{n} Recall_i \tag{14}$$

2.9 Balanced accuracy weighted (BAW) [31, 32]

Building upon the balanced accuracy approach, an additional class weighting is added. (range: $[0,1]$)

+ Suitable for multi-class problems, robust against class imbalance.
− More complex and rarely used metric.

$$BAW = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{i=1}^{n} w_i \cdot Recall_i = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{i=1}^{n} w_i \cdot \frac{TP_i}{TP_i + FN_i} \tag{15}$$

where: w_i = samples per class or the relation of samples per class to the total of all samples

2.10 Average accuracy (AA) [29, 49]

Average accuracy over all classes. (range: $[0, 1]$)

+ Suitable for data sets that consist of balanced classes.
− Can yield poor results when a biased classifier is tested on imbalanced data.

$$AA = \frac{1}{n} \cdot \sum_{i=1}^{n} A_i = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{TP_i + TN_i}{TP_i + FP_i + TN_i + FN_i} \tag{16}$$

2.11 Average class accuracy (ACA) [50, 51]

Calculates a weighted average classification accuracy based on a minority accuracy corresponding to TPR and a majority accuracy corresponding to TNR. (range: $[0, 1]$)

+ Suitable when a significant class imbalance between minority and majority classes exists.
− Difficult to choose a good weighting factor.

$$ACA = w \cdot TPR + (1 - w) \cdot TNR = w \cdot \frac{TP}{TP + FN} + (1 - w) \cdot \frac{TN}{TN + FP} \tag{17}$$

where: w = weight of the positive class in relation to the data set, $0 \leq w \leq 1$

2.12 Error rate (ER) [52, 53]

Complementary metric to accuracy. All faulty classifications are divided by the total number of classifications. (range: $[0, 1]$)

+ Suitable for problems where the evaluation of error recognition is of importance.
− Poor performance for imbalanced data.

$$ER = \frac{FP + FN}{T} = \frac{FP + FN}{P + N} = \frac{FP + FN}{TP + FP + TN + FN} \tag{18}$$

2.13 Average error rate (AER) [54, 55]

Average error rate over all classes. (range: $[0, 1]$)

+ Suitable for problems where the evaluation of error recognition is of importance.
− Poor performance for imbalanced data.

$$AER = \frac{1}{n} \cdot \sum_{i=1}^{n} ER_i = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{FP_i + FN_i}{TP_i + FP_i + TN_i + FN_i} \tag{19}$$

2.14 F-score / Fβ-score [35, 36]

The F-score [35, 36], or Fβ-score, serves as a versatile evaluation metric in classification tasks, allowing for the weighted balancing of precision and recall according to specific application needs. It introduces the parameter β, which adjusts the emphasis placed on recall relative to precision. A β greater than one prioritizes recall, making it particularly useful in contexts where missing a positive instance carries severe repercussions, such as in medical diagnostics or fraud detection. Conversely, a β less than one shifts the focus towards precision, suitable for applications where false positives are more detrimental, such as in spam filtering. By calculating the weighted harmonic mean of precision and recall, the Fβ-score quantifies the trade-offs between both metrics, providing a nuanced view of model performance. High values of the Fβ-score indicate that the model not only performs well in terms of the chosen emphasis on precision or recall but also maintains a reasonable balance according to the specificities dictated by β, thereby ensuring model decisions align closely with the intended requirements.

Measures the user-defined classification effectiveness. (range: $[0, 1]$)

+ Balances precision and recall.
+ Adjustable for precision or recall emphasis.
− Harder to interpret than individual components.
− Not sensitive to data distribution changes.

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall} = \frac{(1 + \beta^2) \cdot TP}{(1 + \beta^2) \cdot TP + \beta^2 \cdot FN + FP} \tag{20}$$

2.14.1 Macro average F-score [56, 57]

Averaged F-score over all classes. (range: $[0, 1]$)

+ Values all classes equally.
− Valuing all classes equally can be detrimental in case of class imbalance.

$$Macro\ average\ F\text{-}score = \frac{1}{n} \cdot \sum_{i=1}^{n} F_i \tag{21}$$

where: F_i = F-score of class i with a chosen β

2.14.2 Micro average F-score [57, 58]

Averaged F-score based on the micro average of precision and recall. (range: $[0, 1]$)

+ Useful when the performance of larger classes is of importance.
− Prone to issues with class imbalance.

$$\textit{Micro average F-score} = 2 \cdot \frac{AP_{micro} \cdot AR_{micro}}{AP_{micro} + AR_{micro}} \tag{22}$$

2.14.3 Weighted average F-score [59, 60]

Weighted averaged F-score over all classes. (range: $[0, 1]$)

+ More robust against class imbalance.
− Not widely used within the literature.

$$\textit{Weighted average F-score} = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{i=1}^{n} (w_i \cdot F_i) \tag{23}$$

where: w_i = samples per class or the relation of samples per class to the total of all samples
F_i = F-score of class i

2.14.4 F0-score [35, 36]

The recall is weighted 0 times as important as the precision. (range: $[0, 1]$)

+ The importance of the recall is maximized.
− The importance of the precision is minimized.

$$F_0 = (1 + 0^2) \cdot \frac{\textit{Precision} \cdot \textit{Recall}}{(0^2 \cdot \textit{Precision}) + \textit{Recall}} = \frac{\textit{Precision} \cdot \textit{Recall}}{\textit{Recall}} \tag{24}$$

2.14.5 F0.5-score [35, 36]

The recall is weighted 0.5 times as important as the precision. (range: $[0, 1]$)

+ The importance of the recall is increased.
− The importance of the precision is decreased.

$$F_{0.5} = (1 + 0.5^2) \cdot \frac{Precision \cdot Recall}{(0.5^2 \cdot Precision) + Recall} = 1.25 \cdot \frac{Precision \cdot Recall}{(0.25 \cdot Precision) + Recall} \tag{25}$$

2.14.6 F1-score [35, 36]

Measures the harmonic mean of the precision and the recall. (range: $[0, 1]$)

+ Equal importance of precision and recall.
− Does not consider true negatives.

$$F_1 = (1 + 1^2) \cdot \frac{Precision \cdot Recall}{(1^2 \cdot Precision) + Recall} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{26}$$

2.14.7 F2-score [35, 36]

The recall is weighted 2 times as important as the precision. (range: $[0, 1]$)

+ The importance of the recall is decreased.
− The importance of the precision is increased.

$$F_2 = (1 + 2^2) \cdot \frac{Precision \cdot Recall}{(2^2 \cdot Precision) + Recall} = 5 \cdot \frac{Precision \cdot Recall}{(4 \cdot Precision) + Recall} \tag{27}$$

2.15 False discovery rate (FDR) [37, 38]

The false discovery rate (FDR) [37,38] is an evaluation metric used predominantly in multiple hypothesis testing to measure the proportion of false positives among the rejected hypotheses. This is particularly crucial in fields such as genomics, where large numbers of simultaneous tests are conducted and distinguishing between truly significant results and those due to chance. FDR quantifies the expected proportion of erroneous rejections (false discoveries) in a set of claimed findings. This focus on controlling the rate of false positives means that FDR is particularly useful when the cost of a false positive is significant. A lower FDR value indicates a more reliable rejection of null hypotheses, implying that a higher proportion of identified significant results are likely to be truly significant. FDR provides a balance between discovery and error, which allows to maximize true discoveries while maintaining a controlled rate of false positives. This balance is especially important in exploratory research where the ability to detect genuine effects without being overwhelmed by spurious findings is crucial.

Fraction of false positives and true positives. (range: $[0, 1]$)

+ Focuses on proportion of false positives.
+ Useful in multiple comparison scenarios.
− Ignores true negatives and false negatives.
− Sensitive to class imbalance.

$$FDR = \frac{FP}{FP + TP} \tag{28}$$

2.15.1 Macro average FDR (FDRmacro)

$$FDR_{macro} = \frac{1}{n} \cdot \sum_{i=1}^{n} FDR_i = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{FP_i}{FP_i + TP_i} \tag{29}$$

2.15.2 Micro average FDR (FDRmicro)

$$FDR_{micro} = \frac{\sum_{i=1}^{n} FP_i}{\sum_{i=1}^{n} (FP_i + TP_i)} \tag{30}$$

2.15.3 Weighted average FDR (FDRweighted)

$$FDR_{weighted} = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{i=1}^{n} w_i \cdot FDR_i = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{i=1}^{n} w_i \cdot \frac{FP_i}{FP_i + TP_i} \tag{31}$$

2.16 False omission rate (FOR) [39]

Fraction of false negatives and true negatives. (range: $[0, 1]$)

+ Emphasizes the accuracy of classifying false negatives in relation to true negatives.
− Neglects true positives and false positives.

$$FOR = \frac{FN}{FN + TN} \tag{32}$$

2.17 False positive rate (FPR) [61]

Fraction of false positives and true negatives. (range: $[0, 1]$)

+ Emphasizes the accuracy of classifying false positives in relation to true negatives.
− Neglects true positives and false negatives.

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \tag{33}$$

2.18 False negative rate (FNR) / miss rate [61]

The false negative rate (FNR) [61], which is an important metric in the field of statistical classification, measures the proportion of positive instances that a model incorrectly classifies as negative. This rate is critical in scenarios where the consequences of missing a positive detection are severe, such as in medical diagnosis or security surveillance systems [62, 63]. FNR is particularly valuable in contexts where ensuring the capture of all positive cases is paramount to prevent harmful outcomes. By quantifying the likelihood that true positives are overlooked by a model, the FNR provides essential insights into the sensitivity of a classification system. A high FNR indicates that a significant number of positive cases are being missed, potentially leading to hazardous oversights. Consequently, minimizing FNR is crucial in high-stakes environments where the cost of a false negative is high, such as failing to diagnose a serious illness or missing a security threat. It helps in tuning the thresholds of detection algorithms to better suit specific operational requirements, ensuring a more reliable performance in critical applications. Therefore, the FNR is integral to developing robust predictive models that effectively manage the risks associated with incorrect negative classifications.

Fraction of false negatives and true positives. (range: $[0, 1]$)

+ Emphasizes identifying missed positive cases.
+ Important for recall-focused assessments.
− Ignores true negatives and false positives.
− May lead to overemphasis on recall.

$$FNR = \frac{FN}{P} = \frac{FN}{FN + TP} \tag{34}$$

2.18.1 Macro average FNR (FNRmacro)

$$FNR_{macro} = \frac{1}{n} \cdot \sum_{i=1}^{n} FNR_i = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{FN_i}{FN_i + TP_i} \tag{35}$$

2.18.2 Micro average FNR (FNRmicro)

$$FNR_{micro} = \frac{\sum_{i=1}^{n} FN_i}{\sum_{i=1}^{n}(FN_i + TP_i)} \tag{36}$$

2.18.3 Weighted average FNR (FNRweighted)

$$FNR_{weighted} = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{i=1}^{n} w_i \cdot FNR_i = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{i=1}^{n} w_i \cdot \frac{FN_i}{FN_i + TP_i} \tag{37}$$

2.19 Positive likelihood ratio (LR+) [44, 45]

In medicine, LR+ describes the probability of, e.g., a positive result in a sick person (true positive) in relation to the probability of a positive result in a healthy person (false positive). (range: $[0, \infty)$)

+ Emphasizes the accuracy of the TPR in relation to the FPR.
− Neglects, e.g., TNR and FNR (see also diagnostic odds ratio).

$$LR+ = \frac{Sensitivity}{1 - Specificity} = \frac{TPR}{FPR} \tag{38}$$

2.20 Negative likelihood ratio (LR−) [44, 45]

In medicine, LR− describes the probability of, e.g., a negative result in a sick person (false negative) in relation to the probability of a negative result in a healthy person (true negative). (range: $[0, \infty)$)

+ Emphasizes the accuracy of the FNR in relation to the TNR.
− Neglects, e.g., TPR and FPR (see also diagnostic odds ratio).

$$LR- = \frac{1 - Sensitivity}{Specificity} = \frac{FNR}{TNR} \tag{39}$$

2.21 Diagnostic odds ratio (DOR) [40, 41]

In medicine, DOR measures the classification effectiveness of a diagnostic test. [40] (range: $[0, \infty)$)

+ Balances sensitivity and specificity.
+ Independent of disease prevalence.
− Requires true positive, true negative, false positive, false negative counts.
− Increased calculation and interpretation complexity.

$$DOR = \frac{LR+}{LR-} = \frac{TP \cdot TN}{FP \cdot FN} \tag{40}$$

2.22 Fowlkes–Mallows index (FM) [42, 43]

Measures the geometric mean of the precision and the recall. In clustering, FM measures the similarity of two clusters. (range: $[0, 1]$)

+ Increased robustness to noise.
− Less well known and used.

$$FM = \sqrt{PPV \cdot TPR} = \sqrt{Precision \cdot Recall} \tag{41}$$

2.23 Informedness / bookmaker informedness (BM) / Youden's J statistic / Youden's index [46, 47]

Quantifies the probability of an "informed decision". (range: $[-1, 1]$)

+ Takes all predictions into account.
− Its result ranges from -1 to 1.

$$Informedness = Sensitivity + Specificity - 1 = TPR + TNR - 1 \tag{42}$$

2.24 Markedness (MK) [64]

Quantifies the "markedness", i.e., the state of being irregular or uncommon. (range: $[-1, 1]$)

+ Takes all predictions into account.
− Its result ranges from -1 to 1.

$$MK = PPV + NPV - 1 \tag{43}$$

2.25 Matthews correlation coefficient (MCC) / phi coefficient / Yule phi coefficient [65–67]

A balanced measure of TPR, TNR, PPV, and NPV. (range: $[0, 1]$)

+ Takes all predictions into account.
− Increased calculation and interpretation complexity.

$$MCC = \sqrt{TPR \cdot TNR \cdot PPV \cdot NPV} \tag{44}$$

2.26 Jaccard index (JI) / threat score (TS) / critical success index (CSI) [68, 69]

Fraction of TP with TP, FN, and FP. (range: $[0, 1]$)

+ Emphasizes the accuracy of classifying TP in relation to FN and FP.
− Neglects TN.

$$JI = \frac{TP}{TP + FN + FP} \tag{45}$$

2.27 Receiver operating characteristic curve (ROC curve) and area under the curve (AUC) [48, 70, 71]

The receiver operating characteristic curve (ROC curve) and its associated metric [48, 70, 71], the area under the curve (AUC), are pivotal evaluation tools used in the analysis of classification models [72–74]. The ROC curve is a graphical representation that plots the true positive rate (sensitivity) against the false positive rate (specificity) at various threshold settings, providing a comprehensive view of a model's ability across a spectrum of conditions. The AUC, a scalar value derived from the area under the ROC curve, quantifies the overall ability of the model to discriminate between classes irrespective of any specific threshold. A higher AUC value indicates better model performance, with a value of 1.0 representing perfect discrimination and 0.5 denoting no discriminative ability (equivalent to random guessing). These metrics are particularly valuable in environments where distinguishing between classes is crucial and the costs of different types of classification errors vary significantly. By evaluating the trade-offs between sensitivity and specificity without committing to a specific classification threshold, ROC and AUC facilitate an objective comparison of model performance when robust and reliable predictions are of central importance.

Calculates the area under the so-called receiver operating characteristic curve by adjusting the confidence threshold for, e.g., classification, detection, or segmentation. (range: $[0, 1]$)

+ Evaluates model performance across various decision thresholds.
+ Suitable for imbalanced data sets.
− Insensitive to class distribution changes.
− Does not directly address practical performance thresholds.

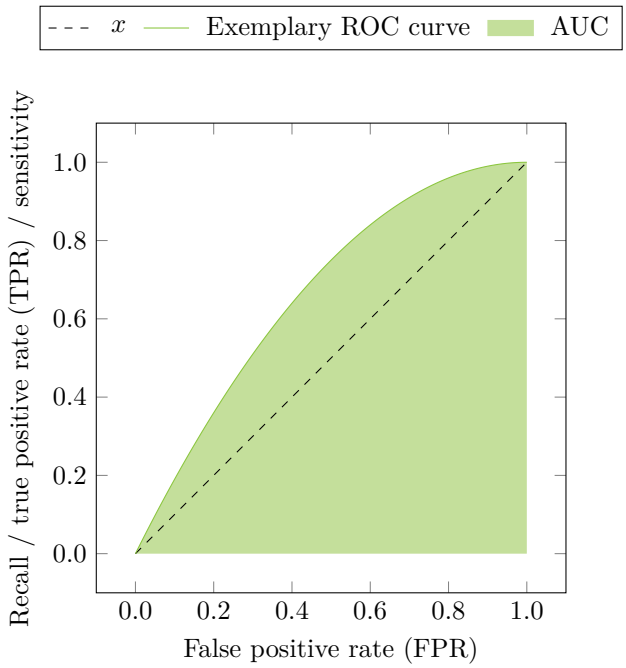

Figure 2: Receiver operating characteristic curve (ROC curve) and area under the curve (AUC).

$$AUC = \int_{x=0}^{1} TPR(FPR^{-1}(x))\,dx = \frac{1}{n} \cdot \sum_{i=1}^{n} TPR(FPR_i) \tag{46}$$

2.28 Average precision (AP) [75, 76]

Calculates the area under the precision-recall curve by adjusting the confidence threshold for, e.g., classification, detection, or segmentation [77–79]. (range: $[0, 1]$)

+ Evaluates model performance across various decision thresholds.
+ Suitable for imbalanced data sets.
− Insensitive to class distribution changes.
− Does not directly address practical performance thresholds.

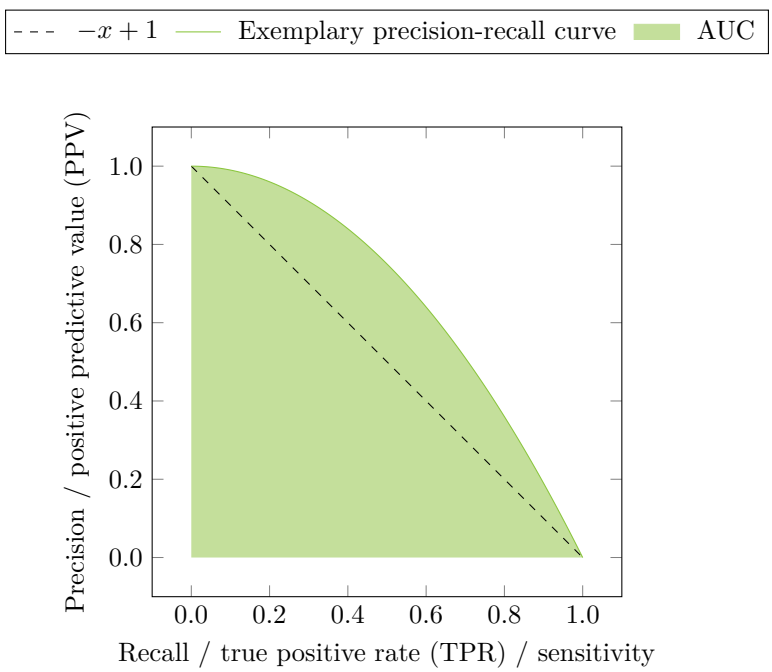

Figure 3: Precision-recall curve and area under the curve (AUC).

$$AP = \int_{x=0}^{1} Precision(Recall^{-1}(x))\,dx = \frac{1}{n} \cdot \sum_{i=1}^{n} Precision(Recall_i) \tag{47}$$

2.29 Mean average precision (mAP) [75, 76]

Averages the area under the precision-recall curve over multiple classes. (range: $[0, 1]$)

+ Allows for an even more meaningful evaluation over multiple classes.
− Increased calculation complexity.

$$mAP = \frac{1}{n} \cdot \sum_{i=1}^{n} AP_i \tag{48}$$

2.30 Cohen's kappa for binary classification [80–82]

Cohen's kappa is a metric for evaluating inter-rater reliability. It is applicable to both binary and categorical data by providing a quantitative measure. It is especially useful for the training of raters and adjustments in methodologies, particularly in complex settings with multiple raters. Yet, the main argument against Kappa centers on the problematic interpretations it can foster. Critics highlight that Kappa can be misleading due to its dependence on the prevalence of the attribute being measured and the possible bias of the raters. For instance, even when multiple different raters are in substantial agreement, kappa scores can be unexpectedly low if the prevalence of a certain categorization is particularly high or low. This has led to concerns that kappa may not always provide a robust or accurate measure of agreement. Hence, it could be misleading when forming a decisions based on its result. [83–85]

Ranges from -1 for a fully wrong prediction to $+1$ for a completely correct prediction. (range: $[-1, 1]$)

+ Accounts for chance agreement.
+ Useful for inter-rater agreement assessment.
− Sensitivity to class imbalance.
− Interpretation complexity with multiple raters.

$$\begin{aligned} \kappa &= \frac{P_o - P_e}{1 - P_e} \\ P_e &= \left(\frac{TP + FP}{T}\right) \cdot \left(\frac{TP + FN}{T}\right) + \left(\frac{TN + FN}{T}\right) \cdot \left(\frac{TN + FP}{T}\right) \end{aligned} \tag{49}$$

where: κ = Cohen's kappa
P_e = expected accuracy
P_o = observed accuracy

2.31 Gini impurity [86–88]

The Gini impurity indicates the probability that two randomly selected samples from a data set have different original label types. A lower value indicates a higher purity of the data set. (range: $[0, 1]$)

+ Well-suited for application in decision tree algorithms.
− Favors binary decisions, which can result in decision trees of reduced complexity.

$$Gini\ impurity(D) = \sum_{i=1}^{n} \sum_{i' \neq i} p_i p_{i'} = 1 - \sum_{i=1}^{n} p_i^2 \tag{50}$$

where: D = data set
p = probability of samples

2.32 P_4 metric [89]

The P_4 metric covers four probabilities, precision, recall, specificity, and NPV, at once, forming their harmonic mean. (range: $[0, 1]$)

+ Tends to zero when any of the four conditional probabilities tends to zero. Tends to one when all four conditional probabilities tend to one.
+ It is symmetrical with respect to data set labels swapping.
− Does not take weighting.
− Currently barely used metric.

$$P_4 = \frac{4}{\frac{1}{Precision} + \frac{1}{Recall} + \frac{1}{Specificity} + \frac{1}{NPV}} = \frac{4 \cdot TP \cdot TN}{4 \cdot TP \cdot TN + (TP + TN) \cdot (FP + FN)} \tag{51}$$

2.33 Skill score (SS) [90]

Measures the quality of a prediction against a reference. (range: $(-\infty, 1]$)

+ Intuitive way to rank model performance.
− Scales indefinitely into the negative for larger variations.

$$SS = 1 - \frac{Metric_{GT}}{Metric_P} \tag{52}$$

where: $Metric_{GT}$ = best possible expectation for a model based on a given metric

$Metric_P$ = actual prediction of a model based on a given metric

2.34 Relative improvement factor [91]

Measures the relative quality of a prediction against a reference. (range: $[0, \infty)$)

+ Intuitive way to rank model performance.

− Scales indefinitely into the positive for larger variations.

$$Relative\ improvement\ factor = \frac{1 - Metric_{GT}}{1 - Metric_P} \tag{53}$$

where: $Metric_{GT}$ = best possible expectation for a model based on a given metric

$Metric_P$ = actual prediction of a model based on a given metric

2.35 Examples

Figure 4 shows an example result for a classification problem with 3 imbalanced classes, while Fig. 5 illustrates an example result for a classification problem with 3 balanced classes created from the Iris flower data set of *Fisher* [92]. Both examples show an exemplary confusion matrix on the top left together with the number of samples per class. The metrics recall and false negative rate (FNR) with their equations and example calculations are aligned to the rows, whereas precision and false discovery rate (FDR) and their respective equations and calculations are aligned with the columns of the confusion matrix. Accuracy as well as micro, macro, and weighted precision, recall, and F1-scores are shown on the bottom right. Note that micro and macro average recall, precision, and F1-score remain constant for the balanced data set. This is not the case for the imbalanced data set. Since the examples discuss 3-class problems, we set $n = 3$ in the equations for, e.g., macro average precision, recall, and F1-score (Eq. 2, 7, and 21, respectively).

Ground truth class \ Predicted class	1: *cat*	2: *fox*	3: *dog*	$\sum$	$Recall_i$	FNR_i
1: *cat*	41	11	12	64	64.1%	35.9%
2: *fox*	62	22	23	107	20.6%	79.4%
3: *dog*	31	0	52	83	62.7%	37.4%
$\sum$	134	33	87	254		
$Precision_i$	30.6%	66.7%	59.8%			
FDR_i	69.4%	33.3%	40.2%			

Eq. {6}	$\frac{TP}{TP+FN}$	Eq. {34}	$\frac{FN}{FN+TP}$
$Recall_1$	$\frac{41}{41+11+12}$	FNR_1	$\frac{11+12}{11+12+41}$
$Recall_2$	$\frac{22}{22+62+23}$	FNR_2	$\frac{62+23}{62+23+22}$
$Recall_3$	$\frac{52}{52+31+0}$	FNR_3	$\frac{31+0}{31+0+52}$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{41+22+52}{41+22+52+11+12+62+23+31+0} = 45.28\%$$

Eq. 1	$Precision_1$	$Precision_2$	$Precision_3$
$\frac{TP}{TP+FP}$	$\frac{41}{41+62+31}$	$\frac{22}{22+11+0}$	$\frac{52}{52+12+23}$
Eq. 28	FDR_1	FDR_2	FDR_3
$\frac{FP}{FP+TP}$	$\frac{62+31}{62+31+41}$	$\frac{11+0}{11+0+22}$	$\frac{12+23}{12+23+52}$

Class	*Precision* [%]	*Recall* [%]	F_1 [%]	$\sum$
1	30.60	64.06	41.41	64
2	66.67	20.56	31.43	107
3	59.77	62.65	61.18	83
Macro average	52.34	49.09	44.67	
Micro average	45.28	45.28	45.28	
W. average	55.32	45.28	43.67	
	Eq. {1,2–4}	Eq. {6,7–9}	Eq. {26,21–23}	

Figure 4: Classification result example (1) showing a 3-class problem with imbalanced classes.

Ground truth class \ Predicted class	1: *setosa*	2: *versicolor*	3: *virginica*	$\sum$	$Recall_i$	FNR_i
1: *setosa*	50	0	0	50	100%	0%
2: *versicolor*	0	47	3	50	94%	6%
3: *virginica*	0	2	48	50	96%	4%
$\sum$	50	49	51	150		
$Precision_i$	100%	95.9%	94.1%			
FDR_i	0%	4.1%	5.9%			

Eq. 6	$\frac{TP}{TP+FN}$	Eq. 34	$\frac{FN}{FN+TP}$
$Recall_1$	$\frac{50}{50+0+0}$	FNR_1	$\frac{0+0}{0+0+50}$
$Recall_2$	$\frac{47}{47+0+3}$	FNR_2	$\frac{0+3}{0+3+47}$
$Recall_3$	$\frac{48}{48+0+2}$	FNR_3	$\frac{0+2}{0+2+48}$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{50+47+48}{50+47+48+0+0+0+3+0+2} = 96.7\%$$

Eq. 1	$Precision_1$	$Precision_2$	$Precision_3$
$\frac{TP}{TP+FP}$	$\frac{50}{50+0+0}$	$\frac{47}{47+0+2}$	$\frac{48}{48+0+3}$
Eq. 28	FDR_1	FDR_2	FDR_3
$\frac{FP}{FP+TP}$	$\frac{0+0}{0+0+50}$	$\frac{0+2}{0+2+47}$	$\frac{0+3}{0+3+48}$

Class	*Precision* [%]	*Recall* [%]	F_1 [%]	$\sum$
1	100.00	100.00	100.00	50
2	95.92	94.00	94.95	50
3	94.12	96.00	95.05	50
Macro average	96.68	96.67	96.67	
Micro average	96.67	96.67	96.67	
W. average	96.68	96.67	96.67	
	Eq. {1,2–4}	Eq. {6,7–9}	Eq. {26,21–23}	

Figure 5: Classification result example (2) showing a 3-class problem with balanced classes.

2.36 Functions

The following table gives an overview of machine learning metrics commonly used with the Python programming language machine learning libraries scikit-learn[1], TensorFlow[2] (and Keras[3]), and PyTorch[4,5].

Equation	scikit-learn	TensorFlow	PyTorch
True Positives	/	TruePositives	/
True Negatives	/	TrueNegatives	/
False Positives	/	FalsePositives	/
False Negatives	/	FalseNegatives	/
Precision (1)	precision_score	Precision	Precision
Recall (6)	recall_score	Recall	Recall
True negative rate (TNR) (10)	/	/	Specificity
Macro average precision (AP_{macro}) (2)	average_precision_score	/	AveragePrecision
Micro average precision (AP_{micro}) (3)	average_precision_score	/	AveragePrecision
Weighted average precision ($AP_{weighted}$) (4)	average_precision_score	/	AveragePrecision
Accuracy (A) (12)	accuracy_score	Accuracy	Accuracy
Balanced accuracy (BA) (13,14)	balanced_accuracy_score	/	/
F-score (20)	fbeta_score	FBetaScore	FBetaScore
F1-score (26)	f1_score	F1Score	F1Score
Positive likelihood ratio (LR+) (38)	class_likelihood_ratios	/	/
Negative likelihood ratio (LR−) (39)	class_likelihood_ratios	/	/
Fowlkes–Mallows index (FM) (41)	fowlkes_mallows_score	/	/
Matthews correlation coefficient (MCC) (44)	matthews_corrcoef	/	MatthewsCorrCoef
Jaccard index (JI) (45)	jaccard_score	/	JaccardIndex
Receiver operating characteristic curve (ROC curve)	roc_curve	/	ROC
Area under the curve (AUC) (46)	auc	AUC	AUROC
Average precision (AP) (47)	average_precision_score	/	AveragePrecision
Cohen's kappa (49)	cohen_kappa_score	/	CohenKappa

Table 2: Selection of function calls for the available metrics in scikit-learn, TensorFlow, and PyTorch. The call for the respective metrics follows the corresponding scheme: scikit-learn – sklearn.metrics.<metric>, TensorFlow – tf.keras.metrics.<metric>, and PyTorch – torchmetrics.<metric>.

[1]scikit-learn project page, https://scikit-learn.org/stable/
[2]TensorFlow project page, https://www.tensorflow.org/
[3]Keras project page, https://keras.io/
[4]PyTorch project page, https://pytorch.org/
[5]TorchMetrics project page, https://torchmetrics.readthedocs.io/en/stable/

3 Computer Vision

In the realm of computer vision, the evaluation of algorithms through various error and similarity metrics is essential for ensuring the reliability and accuracy of models applied in areas such as image processing, object recognition, and scene understanding. These metrics serve as fundamental tools for quantifying the performance of vision systems across a range of real-world tasks.

The core of these evaluations begins with the concept of error measurement. The error, abbreviated as E, provides a basic indication of the deviation of predicted values from the ground truth. Based on E, the absolute error (AE) [93] quantifies the absolute difference between each predicted and actual value. By aggregating these errors, a sense of overall error magnitude without accounting for the direction of errors can be obtained. This metric is particularly useful in applications where errors of different orientations, or signs, are equally detrimental. Complementing AE, the relative absolute error (RAE) [94–96] normalizes the absolute error by the magnitude of the true value, providing a scale-independent measure. The mean error (ME) [97, 98] and the mean absolute error (MAE) [99, 100] extend these concepts by averaging errors across all predictions, with MAE providing a robust measure against outliers by using the absolute values of errors. For a percentage-based perspective, the mean percentage error (MPE) [101, 102] and the mean absolute percentage error (MAPE) [94, 100] express errors as a percentage of the actual values, which is valuable for comparing performance across data sets of varying scales. The mean absolute scaled error (MASE) [100, 103] offers an advanced normalization by scaling MAE against the in-sample MAE. It is especially suitable for time-series predictions in vision tasks involving motion or tracking.

Expanding on squared errors, the squared error (SE) [104] and the mean squared error (MSE) [105] provide a measure where larger errors are exponentially penalized, making these metrics sensitive to outliers but valuable in applications where large errors are particularly undesirable. Following MSE, the root mean square error (RMSE) [99, 100, 106] brings these scales back to the original units of measurement, while the normalized root mean square error (NRMSE) [107, 108] offers a relative measure by normalizing RMSE against the range of observed data, potentially enhancing comparability across different scales and data sets. The root mean squared logarithmic error (RMSLE) [109] addresses scenarios where proportional differences are more significant than absolute differences. Therefore, it is more frequently used in depth and logarithmic scale predictions. In terms of assessing the quality of visual outputs, the peak signal-to-noise ratio (PSNR) [110, 111] is pivotal in image processing, especially in lossy compression, by comparing the level of desired signals to the background noise. Likewise, the structural similarity (SSIM) [112, 113] and its counterpart, the structural dissimilarity (DSSIM) [112, 113], evaluate the visual impact of changes to, e.g., luminance, contrast, and structure, providing a comprehensive measure of image quality degradation due to compression or other distortions.

In tasks including image segmentation, metrics such as the intersection over union (IoU) [68, 69, 114, 115], the Dice coefficient (DC) [116, 117], and the overlap coefficient (OC) [118–121] are indispensable. They assess the overlap between predicted and ground truth segments, which is crucial for evaluating the accuracy of boundary detection algorithms within domains such as medical imaging [122, 123], video surveillance [124, 125], and autonomous driving systems [126, 127].

Together, these metrics provide a robust framework for diagnosing and enhancing the performance of computer vision systems, ensuring their efficiency and reliability in both controlled and erratic environments.

The following evaluation metrics within the context of machine learning are motivated by the contributions of *Wang et al.* [128] and *Hore and Ziou* [129].

3.1 General

Table 3 gives an overview of our general definitions for CV-related metrics.

Abbreviation	Meaning
GT	Ground truth
P	Prediction
n	Number of values

Table 3: General definitions computer vision.

3.2 Error (E)

The amount by which a prediction differs from the ground truth. (range: $(-\infty, \infty)$)

+ Intuitive and straightforward to apply and interpret.
− Does not distinguish between types of errors.

$$E = GT - P \tag{54}$$

3.3 Absolute error / sum of absolute errors (AE) [93]

Calculates the sum (total) of all absolute errors. (range: $[0, \infty)$)

+ Intuitive and straightforward to apply and interpret.
+ Accounts for absolute magnitude of errors.
− No differentiation depending on the number of compared values is made.
− Large individual differences equal to many small ones (distribution problem).

$$AE = \sum_{i=1}^{n} |E_i| = \sum_{i=1}^{n} |GT_i - P_i| \tag{55}$$

3.4 Relative absolute error (RAE) [94–96]

Normalization of the absolute error by dividing the total absolute error of the simple predictor. (range: $[0, \infty)$)

+ Comparison of models that differ significantly.
− Not sensitive to outliers and scaling.

$$\begin{aligned} RAE &= \frac{AE}{\sum_{i=1}^{n} |GT_i - \overline{GT}|} = \frac{\sum_{i=1}^{n} |GT_i - P_i|}{\sum_{i=1}^{n} |GT_i - \overline{GT}|} \\ \overline{GT} &= \frac{1}{n} \cdot \sum_{i=1}^{n} GT_i \end{aligned} \tag{56}$$

where: $\overline{GT}$ = average of the ground truth

3.5 Mean error (ME) [97, 98]

The average over all error measurements. (range: $(-\infty, \infty)$)

+ Intuitive and straightforward to apply.
− Positive and negative error values can cancel each other out.

$$ME = \frac{1}{n} \cdot \sum_{i=1}^{n} E_i = \frac{1}{n} \cdot \sum_{i=1}^{n} (GT_i - P_i) \tag{57}$$

3.6 Mean percentage error (MPE) [101, 102]

The average over all error measurements in percentage. (range: $(-\infty\%, \infty\%)$)

+ Intuitive overview of the underlying situation.
− Undefined as soon as a single ground truth value is zero.

$$MPE = \frac{100}{n} \cdot \sum_{i=1}^{n} \frac{E_i}{GT_i} = \frac{100}{n} \cdot \sum_{i=1}^{n} \frac{GT_i - P_i}{GT_i} \tag{58}$$

3.7 Mean absolute error (MAE) [99, 100]

Calculates the mean of the sum (total) of all absolute errors. (range: $[0, \infty)$)

+ Partially solves the distribution problem.
− No differentiation depending on the maximum assumable error is made.

$$MAE = \frac{AE}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} |GT_i - P_i| \tag{59}$$

3.8 Mean absolute percentage error (MAPE) [94, 100]

The average over all absolute error measurements in percentage. (range: $[0\%, \infty\%)$)

+ Intuitive and scale independent.
− Undefined as soon as a single ground truth value is zero.

$$MPE = \frac{100}{n} \cdot \sum_{i=1}^{n} \frac{|E_i|}{|GT_i|} = \frac{100}{n} \cdot \sum_{i=1}^{n} \frac{|GT_i - P_i|}{|GT_i|} \tag{60}$$

3.9 Mean absolute scaled error (MASE) [100, 103]

Mean absolute error of the measurements scaled by the mean absolute error of the ground truth. (range: $[0, \infty)$)

+ Scale invariant.
− Less sensitive to outliers.

$$MASE = \frac{MPE}{\left(\frac{1}{n} - 1\right) \cdot \sum_{i=2}^{n} |GT_i - GT_{i-1}|} = \frac{\frac{100}{n} \cdot \sum_{i=1}^{n} \frac{GT_i - P_i}{GT_i}}{\left(\frac{1}{n} - 1\right) \cdot \sum_{i=2}^{n} |GT_i - GT_{i-1}|} \tag{61}$$

3.10 Mean normalized bias (MNB) [130, 131]

Calculates the variance between the predicted values and the ground truth values. Divides by the reference variable, subsequently calculating the mean. (range: $(-\infty, \infty)$)

+ Enables the specific evaluation of systematic errors across the entire model.
− Does not detect specific errors in individual parts of the model.

$$MNB = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{E_i}{P_i} = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{GT_i - P_i}{P_i} \tag{62}$$

3.11 Normalized mean bias (NMB) [130, 132]

Calculates the average of the variances between the prediction and the reference variable, subsequently normalizing it by the reference variable. (range: $(-\infty, \infty)$)

+ Comparison of models independently of scaling.
− Sensitive to outliers.

$$NMB = \frac{\sum_{i=1}^{n} E_i}{\sum_{i=1}^{n} P_i} = \frac{\sum_{i=1}^{n} (GT_i - P_i)}{\sum_{i=1}^{n} P_i} \tag{63}$$

3.12 Squared error / sum of squared errors (SE) [104]

Calculates the sum (total) of all squared errors. (range: $[0, \infty)$)

+ Emphasizes the contribution of large errors.
− No differentiation depending on the number of compared values is made.

$$SE = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (GT_i - P_i)^2 \tag{64}$$

3.13 Mean square error (MSE) [105]

Calculates the mean of the sum (total) of all squared errors. (range: $[0, \infty)$)

+ Partially solves the distribution problem.
− No differentiation depending on the maximum assumable error is made.

$$MSE = \frac{SE}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} (GT_i - P_i)^2 \tag{65}$$

3.14 Root mean square error (RMSE) [99, 100, 106]

Calculates the root of the mean of the sum (total) of all squared errors. (range: $[0, \infty)$)

+ Provides a result in the range of the compared values.
− No differentiation depending on the maximum assumable error is made.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (GT_i - P_i)^2} \tag{66}$$

3.15 Normalized root mean square error (NRMSE) [107, 108]

Normalization of RMSE. (range: $[0, \infty)$)

+ Comparison of models that differ significantly.
− Not sensitive to outliers and scaling.

$$\begin{aligned} NRMSE &= \frac{RMSE}{\overline{GT}} = \frac{\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (GT_i - P_i)^2}}{\overline{GT}} \\ \overline{GT} &= \frac{1}{n} \cdot \sum_{i=1}^{n} GT_i \end{aligned} \tag{67}$$

where: $\overline{GT}$ = average of the ground truth

3.16 Root mean squared logarithmic error (RMSLE) [109]

Calculates the mean squared error of the logarithmized ground truth in comparison to the logarithmized predictions. (range: $[0, \infty)$)

+ Robust to outliers.
− Biased penalty. Underestimation is penalized more than overestimation.

$$RMSLE = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (\ln(GT_i + 1) - \ln(P_i + 1))^2} \tag{68}$$

3.17 Peak signal-to-noise ratio (PSNR) [110, 111]

Calculates the MSE in relation to the maximum assumable error. (range: $[0, \infty)$)

+ A differentiation depending on the maximum assumable error is made.
− No differentiation depending on structural similarities is made.

$$PSNR = 10 \cdot \log_{10} \frac{E_{max}^2}{MSE} = 10 \cdot \log_{10} \frac{E_{max}^2}{\frac{1}{n} \cdot \sum_{i=1}^{n} (GT_i - P_i)^2} \tag{69}$$

where: E_{max} = maximum possible error

3.18 Structural similarity (SSIM) [112, 113]

Calculates the structural similarity using the mean, variance, and covariance. (range: $[-1, 1]$)

+ Provides more accurate results by considering structural characteristics. [112]
− Increased calculation complexity.

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{70}$$

where: μ_x, μ_y = mean
σ_x, σ_y = variance
σ_{xy} = covariance
c_1, c_2 = division stabilizers, e.g., $(0.01 \cdot 2^8 - 1)^2$ and $(0.03 \cdot 2^8 - 1)^2$ (8 bits per value)

3.19 Structural dissimilarity (DSSIM) [112, 113]

Calculates the structural dissimilarity using the mean, variance, and covariance. (range: $[0, 1]$)

+ Provides more accurate results by considering structural characteristics. [112]
− Increased calculation complexity.

$$DSSIM = \frac{1 - SSIM}{2} \tag{71}$$

3.20 Intersection over union (IoU) [68, 69, 114, 115]

Intersection over union (IoU) [68, 69, 114, 115] is a fundamental metric used primarily to evaluate the accuracy of object detection algorithms in tasks such as image segmentation and computer vision. IoU assesses the overlap between predicted and ground truth objects by calculating the ratio of the area of overlap to the area of union between the predicted and the actual annotations. It provides a clear, quantitative measure of how closely the contours of detected objects match the true object boundaries, encompassing both the correctness and precision of the detection in a single score. A higher IoU score indicates a greater degree of overlap and, consequently, an improved model performance. A perfect score of 1.0 represents an exact match between the predicted and the ground truth area. Because of its ability to accurately measure the effectiveness of detection models in capturing the true object space, IoU is extensively utilized in evaluating models for tasks such as autonomous driving, aerial image analysis [133, 134], and medical imaging, where precise localization is critical. Its widespread adoption stems from its simplicity and effectiveness in providing a direct indicator of spatial accuracy.

Calculates the similarity of two sets of values via the intersection over the union of both sets. In machine learning also known as the Jaccard index (JI). (range: $[0, 1]$)

+ Captures overlap between predicted and true regions.
+ Commonly used in image segmentation and object detection tasks.
− Sensitive to small region misalignment.
− Does not distinguish between types of errors.

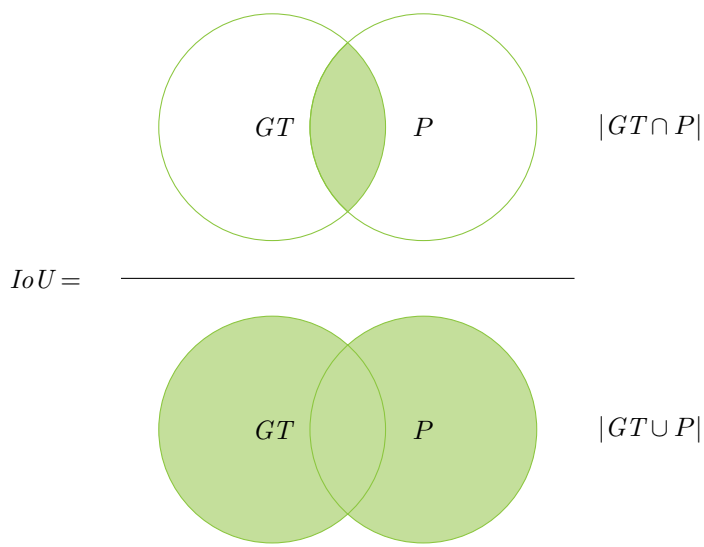

Figure 6: Intersection over union (IoU).

$$IoU = \frac{|GT \cap P|}{|GT \cup P|} \tag{72}$$

3.21 Dice coefficient (DC) [116, 117]

Calculates the similarity of two sets of values via twice the intersection over the sum of both sets. In machine learning also known as the F1-score. (range: $[0, 1]$)

+ Can be well used for image segmentation and object detection.
− No differentiation depending on the size of both sets is made.

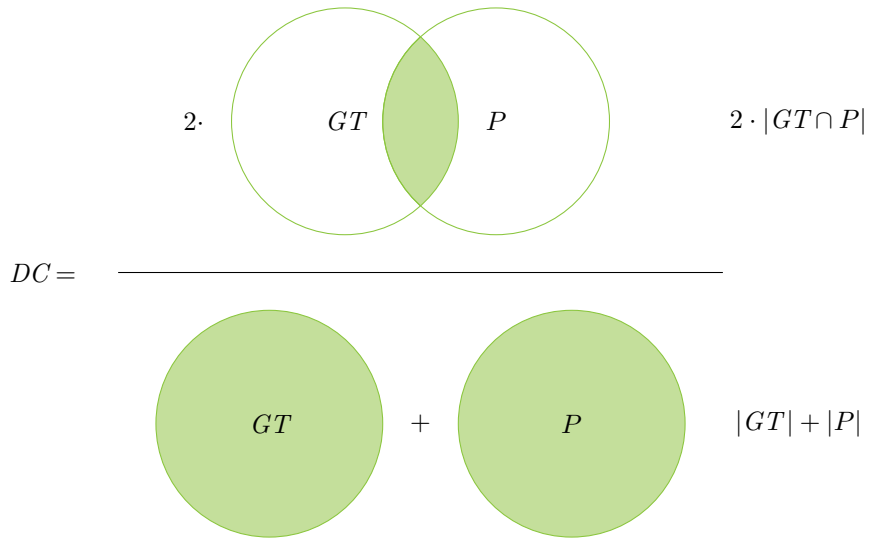

Figure 7: Dice coefficient (DC).

$$DC = \frac{2 \cdot |GT \cap P|}{|GT| + |P|} \tag{73}$$

3.22 Overlap coefficient (OC) [118–121]

Calculates the similarity of two sets of values via the intersection over the smaller set of both sets. (range: $[0, 1]$)

+ A differentiation depending on the size of both sets is made.

$$OC = \frac{|GT \cap P|}{\min(|GT|, |P|)} \tag{74}$$

3.23 Functions

The following table gives an overview of computer vision metrics commonly used with the Python programming language machine learning libraries scikit-learn[6], TensorFlow[7] (and Keras[8]), and PyTorch[9,10].

Equation	scikit-learn	TensorFlow	PyTorch
Mean absolute error (MAE) (59)	mean_absolute_error	keras.losses.MeanAbsoluteError	MeanAbsoluteError
Mean absolute percentage error (MAPE) (60)	mean_absolute_percentage_error	keras.losses.MeanAbsolutePercentageError	MeanAbsolutePercentageError
Mean square error (MSE) (65)	mean_squared_error	keras.losses.MeanSquaredError	MeanSquaredError
Root mean square error (RMSE) (66)	/	keras.metrics.RootMeanSquaredError	/
Peak signal-to-noise ratio (PSNR) (69)	/	image.psnr	/
Structural similarity (SSIM) (70)	/	image.ssim	/
Intersection over union (IoU) (72)	/	keras.metrics.IoU	detection.iou.IntersectionOverUnion
Dice coefficient (DC) (73)	/	/	Dice

Table 4: Selection of function calls for the available metrics in scikit-learn, TensorFlow, and PyTorch. The call for the respective metrics follows the corresponding scheme: scikit-learn – sklearn.metrics.<metric>, TensorFlow – tf.<metric>, and PyTorch – torchmetrics.<metric>.

[6] scikit-learn project page, https://scikit-learn.org/stable/
[7] TensorFlow project page, https://www.tensorflow.org/
[8] Keras project page, https://keras.io/
[9] PyTorch project page, https://pytorch.org/
[10] TorchMetrics project page, https://torchmetrics.readthedocs.io/en/stable/

Author contributions

Tobias Schlosser, Michael Friedrich, and Trixy Meyer conducted this contribution's conceptualization and writing process with the help of Danny Kowerko in extending this manuscript with examples.

References

[1] T. J. Sejnowski, *The deep learning revolution.* MIT press, 2018.
(1 citation on 1 page: 4)

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
(1 citation on 1 page: 4)

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
(1 citation on 1 page: 4)

[4] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.
(1 citation on 1 page: 4)

[5] Y. Liu, P. Sun, N. Wergeles, and Y. Shang, "A survey and performance evaluation of deep learning methods for small object detection," *Expert Systems with Applications*, vol. 172, p. 114602, 2021.
(1 citation on 1 page: 4)

[6] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *Digital Signal Processing*, vol. 126, p. 103514, 2022.
(1 citation on 1 page: 4)

[7] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.
(1 citation on 1 page: 4)

[8] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," *Electronics*, vol. 10, no. 5, p. 593, 2021.
(1 citation on 1 page: 4)

[9] A. F. Cooper, E. Moss, B. Laufer, and H. Nissenbaum, "Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 864–876.
(1 citation on 1 page: 4)

[10] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern recognition letters*, vol. 30, no. 1, pp. 27–38, 2009.
(1 citation on 1 page: 4)

[11] D. G. Altman and J. M. Bland, "Statistics notes: Diagnostic tests 2: predictive values," *Bmj*, vol. 309, no. 6947, p. 102, 1994.
(5 citations on 3 pages: 5, 7, and 8)

[12] G. S. Fletcher, *Clinical epidemiology: the essentials.* Lippincott Williams & Wilkins, 2019.
(5 citations on 3 pages: 5, 7, and 8)

[13] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in medicine*, vol. 23, no. 1, pp. 89–109, 2001.
(1 citation on 1 page: 5)

[14] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: A comprehensive review," in *Healthcare*, vol. 10, no. 3. MDPI, 2022, p. 541.
(1 citation on 1 page: 5)

[15] M. N. Ashtiani and B. Raahemi, "Intelligent fraud detection in financial statements using machine learning and data mining: a systematic literature review," *Ieee Access*, vol. 10, pp. 72 504–72 525, 2021.
(1 citation on 1 page: 5)

[16] A. Ali, S. Abd Razak, S. H. Othman, T. A. E. Eisa, A. Al-Dhaqm, M. Nasser, T. Elhassan, H. Elshafie, and A. Saif, "Financial fraud detection based on machine learning: a systematic literature review," *Applied Sciences*, vol. 12, no. 19, p. 9637, 2022.
(1 citation on 1 page: 5)

[17] Y. Yang, "An evaluation of statistical approaches to text categorization," *Information retrieval*, vol. 1, no. 1-2, pp. 69–90, 1999.
(8 citations on 5 pages: 5, 7, 8, 9, and 10)

[18] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.
(8 citations on 5 pages: 5, 7, 8, 9, and 10)

[19] M. Zhu, "Recall, precision and average precision," *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, vol. 2, no. 30, p. 6, 2004.
(2 citations on 2 pages: 5 and 7)

[20] K. He, Y. Lu, and S. Sclaroff, "Local descriptors optimized for average precision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 596–605.
(2 citations on 2 pages: 5 and 7)

[21] L. Han, S. Luo, J. Yu, L. Pan, and S. Chen, "Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes," *IEEE journal of biomedical and health informatics*, vol. 19, no. 2, pp. 728–734, 2014.
(4 citations on 3 pages: 5, 8, and 10)

[22] J. Yerushalmy, "Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques," *Public Health Reports (1896-1970)*, pp. 1432–1449, 1947.
(5 citations on 3 pages: 5, 9, and 10)

[23] D. G. Altman and J. M. Bland, "Diagnostic tests. 1: Sensitivity and specificity." *BMJ: British Medical Journal*, vol. 308, no. 6943, p. 1552, 1994.
(5 citations on 3 pages: 5, 9, and 10)

[24] A. Rosenberg, "Classifying skewed data: Importance weighting to optimize average recall," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.
(2 citations on 2 pages: 5 and 9)

[25] S. Yang, Z. Gong, K. Ye, Y. Wei, Z. Huang, and Z. Huang, "Edgernn: a compact speech recognition network with spatio-temporal features for edge computing," *IEEE Access*, vol. 8, pp. 81 468–81 478, 2020.
(2 citations on 2 pages: 5 and 9)

[26] R. A. Gordon, R. M. Rozelle, and J. C. Baxter, "The effect of applicant age, job level, and accountability on the evaluation of job applicants," *Organizational Behavior and Human Decision Processes*, vol. 41, no. 1, pp. 20–33, 1988.
(2 citations on 2 pages: 5 and 10)

[27] C. E. Metz, "Basic principles of roc analysis," in *Seminars in nuclear medicine*, vol. 8, no. 4. Elsevier, 1978, pp. 283–298.
(3 citations on 3 pages: 5, 6, and 11)

[28] J. Taylor, *Introduction to error analysis, the study of uncertainties in physical measurements*. University Science Books, 1997.
(2 citations on 2 pages: 5 and 11)

[29] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 3121–3124.
(3 citations on 3 pages: 5, 11, and 12)

[30] J. D. Kelleher, B. Mac Namee, and A. D'arcy, *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press, 2020.
(2 citations on 2 pages: 5 and 11)

[31] A. Salman, U. Sharaha, E. Rodriguez-Diaz, E. Shufan, K. Riesenberg, I. J. Bigio, and M. Huleihel, "Detection of antibiotic resistant escherichia coli bacteria using infrared microscopy and advanced multivariate analysis," *Analyst*, vol. 142, no. 12, pp. 2136–2144, 2017.
(2 citations on 2 pages: 5 and 11)

[32] P. Infante, G. Jacinto, A. Afonso, L. Rego, P. Nogueira, M. Silva, V. Nogueira, J. Saias, P. Quaresma, D. Santos *et al.*, "Factors that influence the type of road traffic accidents: A case study in a district of portugal," *Sustainability*, vol. 15, no. 3, p. 2352, 2023.
(2 citations on 2 pages: 5 and 11)

[33] K. J. Rothman, *Epidemiology: an introduction*. Oxford university press, 2012.
(2 citations on 2 pages: 5 and 10)

[34] N. Bruce, D. Pope, and D. Stanistreet, *Quantitative methods for health research: a practical interactive guide to epidemiology and statistics*. John Wiley & Sons, 2018.
(2 citations on 2 pages: 5 and 10)

[35] C. J. Van Rijsbergen, *The geometry of information retrieval*. Cambridge University Press, 2004.
(9 citations on 4 pages: 5, 13, 14, and 15)

[36] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC medical imaging*, vol. 15, no. 1, pp. 1–28, 2015.
(9 citations on 4 pages: 5, 13, 14, and 15)

[37] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.
(3 citations on 2 pages: 5 and 15)

[38] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of statistics*, pp. 1165–1188, 2001.
(3 citations on 2 pages: 5 and 15)

[39] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1171–1180.
(2 citations on 2 pages: 5 and 16)

[40] A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. Bossuyt, "The diagnostic odds ratio: a single indicator of test performance," *Journal of clinical epidemiology*, vol. 56, no. 11, pp. 1129–1135, 2003.
(3 citations on 2 pages: 5 and 19)

[41] J. A. Doust, P. P. Glasziou, E. Pietrzak, and A. J. Dobson, "A systematic review of the diagnostic accuracy of natriuretic peptides for heart failure," *Archives of internal medicine*, vol. 164, no. 18, pp. 1978–1984, 2004.
(2 citations on 2 pages: 5 and 19)

[42] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *Journal of the American statistical association*, vol. 78, no. 383, pp. 553–569, 1983.
(2 citations on 2 pages: 5 and 19)

[43] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of intelligent information systems*, vol. 17, pp. 107–145, 2001.
(2 citations on 2 pages: 5 and 19)

[44] J. A. Swets, "The relative operating characteristic in psychology: A technique for isolating effects of response bias finds wide use in the study of perception and cognition." *Science*, vol. 182, no. 4116, pp. 990–1000, 1973.
(4 citations on 2 pages: 5 and 18)

[45] J. J. Deeks and D. G. Altman, "Diagnostic tests 4: likelihood ratios," *Bmj*, vol. 329, no. 7458, pp. 168–169, 2004.
(4 citations on 2 pages: 5 and 18)

[46] C. S. Peirce, "The numerical measure of the success of predictions," *Science*, no. 93, pp. 453–454, 1884.
(2 citations on 2 pages: 5 and 19)

[47] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.
(2 citations on 2 pages: 5 and 19)

[48] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.
(3 citations on 2 pages: 6 and 21)

[49] J. Huang, B. Chen, B. Yao, and W. He, "Ecg arrhythmia classification using stft-based spectrogram and convolutional neural network," *IEEE access*, vol. 7, pp. 92 871–92 880, 2019.
(1 citation on 1 page: 12)

[50] U. Bhowan, M. Johnston, and M. Zhang, "Developing new fitness functions in genetic programming for classification with unbalanced data," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 406–421, 2011.
(1 citation on 1 page: 12)

[51] D. Devarriya, C. Gulati, V. Mansharamani, A. Sakalle, and A. Bhardwaj, "Unbalanced breast cancer data classification using novel fitness functions in genetic programming," *Expert Systems with Applications*, vol. 140, p. 112866, 2020.
(1 citation on 1 page: 12)

[52] D. J. Hand, "Recent advances in error rate estimation," *Pattern Recognition Letters*, vol. 4, no. 5, pp. 335–346, 1986.
(1 citation on 1 page: 12)

[53] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016.
(1 citation on 1 page: 12)

[54] Y. Hamamoto, S. Uchimura, M. Watanabe, T. Yasuda, Y. Mitani, and S. Tomita, "A gabor filter-based method for recognizing handwritten numerals," *Pattern recognition*, vol. 31, no. 4, pp. 395–400, 1998.
(1 citation on 1 page: 12)

[55] H. Han, X. Guo, and H. Yu, "Variable selection using mean decrease accuracy and mean decrease gini based on random forest," in *2016 7th ieee international conference on software engineering and service science (icsess)*. IEEE, 2016, pp. 219–224.
(1 citation on 1 page: 12)

[56] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets," *arXiv preprint arXiv:1308.6242*, 2013.
(1 citation on 1 page: 13)

[57] K. Takahashi, K. Yamamoto, A. Kuchiba, and T. Koyama, "Confidence interval for micro-averaged f 1 and macro-averaged f 1 scores," *Applied Intelligence*, vol. 52, no. 5, pp. 4961–4972, 2022.
(2 citations on 2 pages: 13 and 14)

[58] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27*. Springer, 2005, pp. 345–359.
(1 citation on 1 page: 14)

[59] M. Al-Badrashiny and M. Diab, "Lili: A simple language independent approach for language identification," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 1211–1219.
(1 citation on 1 page: 14)

[60] N. Alswaidan and M. E. B. Menai, "Hybrid feature model for emotion recognition in arabic text," *IEEE Access*, vol. 8, pp. 37 843–37 854, 2020.
(1 citation on 1 page: 14)

[61] A. Banerjee, U. Chitnis, S. Jadhav, J. Bhawalkar, and S. Chaudhury, "Hypothesis testing, type i and type ii errors," *Industrial psychiatry journal*, vol. 18, no. 2, p. 127, 2009.
(3 citations on 1 page: 17)

[62] G. Sreenu and S. Durai, "Intelligent video surveillance: a review through deep learning techniques for crowd analysis," *Journal of Big Data*, vol. 6, no. 1, pp. 1–27, 2019.
(1 citation on 1 page: 17)

[63] M. Xue, C. Yuan, H. Wu, Y. Zhang, and W. Liu, "Machine learning security: Threats, countermeasures, and evaluations," *IEEE Access*, vol. 8, pp. 74 720–74 742, 2020.
(1 citation on 1 page: 17)

[64] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.
(1 citation on 1 page: 20)

[65] G. U. Yule, "On the methods of measuring association between two attributes," *Journal of the Royal Statistical Society*, vol. 75, no. 6, pp. 579–652, 1912.
(1 citation on 1 page: 20)

[66] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.
(1 citation on 1 page: 20)

[67] H. Cramér, *Mathematical methods of statistics*. Princeton university press, 1999, vol. 26.
(1 citation on 1 page: 20)

[68] P. Jaccard, "The distribution of the flora in the alpine zone. 1," *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.
(4 citations on 3 pages: 20, 29, and 36)

[69] A. H. Murphy, "The finley affair: A signal event in the history of forecast verification," *Weather and forecasting*, vol. 11, no. 1, pp. 3–20, 1996.
(4 citations on 3 pages: 20, 29, and 36)

[70] D. M. Green, J. A. Swets *et al.*, *Signal detection theory and psychophysics*. Wiley New York, 1966, vol. 1.
(2 citations on 1 page: 21)

[71] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine," *Clinical chemistry*, vol. 39, no. 4, pp. 561–577, 1993.
(2 citations on 1 page: 21)

[72] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.
(1 citation on 1 page: 21)

[73] J. N. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment," *Journal of Thoracic Oncology*, vol. 5, no. 9, pp. 1315–1316, 2010.
(1 citation on 1 page: 21)

[74] A. Jiménez-Valverde, "Insights into the area under the receiver operating characteristic curve (auc) as a discrimination measure in species distribution modelling," *Global Ecology and Biogeography*, vol. 21, no. 4, pp. 498–507, 2012.
(1 citation on 1 page: 21)

[75] C. D. Manning, *An introduction to information retrieval*. Cambridge university press, 2009.
(2 citations on 2 pages: 22 and 23)

[76] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.
(2 citations on 2 pages: 22 and 23)

[77] B. Ozenne, F. Subtil, and D. Maucort-Boulch, "The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases," *Journal of clinical epidemiology*, vol. 68, no. 8, pp. 855–859, 2015.
(1 citation on 1 page: 22)

[78] H. R. Sofaer, J. A. Hoeting, and C. S. Jarnevich, "The area under the precision-recall curve as a performance metric for rare binary events," *Methods in Ecology and Evolution*, vol. 10, no. 4, pp. 565–577, 2019.
(1 citation on 1 page: 22)

[79] J. Cook and V. Ramadas, "When to consult precision-recall curves," *The Stata Journal*, vol. 20, no. 1, pp. 131–148, 2020.
(1 citation on 1 page: 22)

[80] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.
(1 citation on 1 page: 23)

[81] P. Ranganathan, C. Pramesh, and R. Aggarwal, "Common pitfalls in statistical analysis: Measures of agreement," *Perspectives in clinical research*, vol. 8, no. 4, p. 187, 2017.
(1 citation on 1 page: 23)

[82] D. Chicco, M. J. Warrens, and G. Jurman, "The matthews correlation coefficient (mcc) is more informative than cohen's kappa and brier score in binary classification assessment," *IEEE Access*, vol. 9, pp. 78 368–78 381, 2021.
(1 citation on 1 page: 23)

[83] R. G. Pontius Jr and M. Millones, "Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment," *International journal of remote sensing*, vol. 32, no. 15, pp. 4407–4429, 2011.
(1 citation on 1 page: 23)

[84] P. Olofsson, G. M. Foody, M. Herold, S. V. Stehman, C. E. Woodcock, and M. A. Wulder, "Good practices for estimating area and assessing accuracy of land change," *Remote sensing of Environment*, vol. 148, pp. 42–57, 2014.
(1 citation on 1 page: 23)

[85] G. M. Foody, "Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification," *Remote sensing of environment*, vol. 239, p. 111630, 2020.
(1 citation on 1 page: 23)

[86] C. Gini, *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.[Fasc. I.]*. Tipogr. di P. Cuppini, 1912.
(1 citation on 1 page: 24)

[87] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.
(1 citation on 1 page: 24)

[88] A. S. Manek, P. D. Shenoy, and M. C. Mohan, "Aspect term extraction for sentiment analysis in large movie reviews using gini index feature selection method and svm classifier," *World wide web*, vol. 20, pp. 135–154, 2017.
(1 citation on 1 page: 24)

[89] M. Sitarz, "Extending f1 metric, probabilistic approach. advances in artificial intelligence and machine learning. 2023; 3 (2): 61," 2023.
(1 citation on 1 page: 24)

[90] A. H. Murphy, "Skill scores based on the mean square error and their relationships to the correlation coefficient," *Monthly weather review*, vol. 116, no. 12, pp. 2417–2424, 1988.
(1 citation on 1 page: 24)

[91] T. Schlosser, M. Friedrich, F. Beuth, and D. Kowerko, "Improving automated visual fault inspection for semiconductor manufacturing using a hybrid multistage system of deep neural networks," *Journal of Intelligent Manufacturing*, vol. 33, no. 4, pp. 1099–1123, 2022.
(1 citation on 1 page: 25)

[92] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
(1 citation on 1 page: 26)

[93] I. E. Richardson, *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia*. John Wiley & Sons, 2004.
(2 citations on 2 pages: 29 and 30)

[94] J. S. Armstrong and F. Collopy, "Error measures for generalizing about forecasting methods: Empirical comparisons," *International journal of forecasting*, vol. 8, no. 1, pp. 69–80, 1992.
(4 citations on 3 pages: 29, 31, and 32)

[95] ——, "Another error measure for selection of the best forecasting method: The unbiased absolute percentage error," *International Journal of Forecasting*, vol. 8, no. 2, pp. 69–80, 2000.
(2 citations on 2 pages: 29 and 31)

[96] É. O. Rodrigues, V. Pinheiro, P. Liatsis, and A. Conci, "Machine learning in the prediction of cardiac epicardial and mediastinal fat volumes," *Computers in biology and medicine*, vol. 89, pp. 520–529, 2017.
(2 citations on 2 pages: 29 and 31)

[97] R. A. Fisher *et al.*, "A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error." *Monthly Notices of the Royal Astronomical Society*, vol. 80, pp. 758–770, 1920.
(2 citations on 2 pages: 29 and 31)

[98] T. Anjali, K. Chandini, K. Anoop, and V. Lajish, "Temperature prediction using machine learning approaches," in *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, vol. 1. IEEE, 2019, pp. 1264–1268.
(2 citations on 2 pages: 29 and 31)

[99] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.
(4 citations on 3 pages: 29, 32, and 34)

[100] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International journal of forecasting*, vol. 22, no. 4, pp. 679–688, 2006.
(8 citations on 3 pages: 29, 32, and 34)

[101] K. Pearson, "X. contributions to the mathematical theory of evolution.—ii. skew variation in homogeneous material," *Philosophical Transactions of the Royal Society of London.(A.)*, no. 186, pp. 343–414, 1895.
(2 citations on 2 pages: 29 and 31)

[102] Y. Jiang, "Prediction of monthly mean daily diffuse solar radiation using artificial neural networks and comparison with other empirical models," *Energy policy*, vol. 36, no. 10, pp. 3833–3837, 2008.
(2 citations on 2 pages: 29 and 31)

[103] A. T. Mohan and D. V. Gaitonde, "A deep learning based approach to reduced order modeling for turbulent flow control using lstm neural networks," *arXiv preprint arXiv:1804.09269*, 2018.
(2 citations on 2 pages: 29 and 32)

[104] N. R. Draper and H. Smith, *Applied regression analysis*. John Wiley & Sons, 1998, vol. 326.
(2 citations on 2 pages: 29 and 33)

[105] P. J. Bickel and K. A. Doksum, *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. CRC Press, 2015.
(2 citations on 2 pages: 29 and 33)

[106] R. G. Pontius, O. Thontteh, and H. Chen, "Components of information for multiple resolution comparison between maps that share a real variable," *Environmental and ecological statistics*, vol. 15, pp. 111–142, 2008.
(2 citations on 2 pages: 29 and 34)

[107] J. C. Chang and S. R. Hanna, "Air quality model performance evaluation," *Meteorology and Atmospheric Physics*, vol. 87, no. 1-3, pp. 167–196, 2004.
(2 citations on 2 pages: 29 and 34)

[108] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for dna microarray gene expression data: local least squares imputation," *Bioinformatics*, vol. 21, no. 2, pp. 187–198, 2005.
(2 citations on 2 pages: 29 and 34)

[109] A. Nafees, M. F. Javed, S. Khan, K. Nazir, F. Farooq, F. Aslam, M. A. Musarat, and N. I. Vatin, "Predictive modeling of mechanical properties of silica fume-based green concrete using artificial intelligence approaches: Mlpnn, anfis, and gep," *Materials*, vol. 14, no. 24, p. 7531, 2021.
(2 citations on 2 pages: 29 and 34)

[110] D. Salomon, *Data compression: the complete reference*. Springer Science & Business Media, 2004.
(2 citations on 2 pages: 29 and 35)

[111] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.
(2 citations on 2 pages: 29 and 35)

[112] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
(6 citations on 2 pages: 29 and 35)

[113] V. Ghodrati, J. Shao, M. Bydder, Z. Zhou, W. Yin, K.-L. Nguyen, Y. Yang, and P. Hu, "Mr image reconstruction using deep learning: evaluation of network structure and loss functions," *Quantitative imaging in medicine and surgery*, vol. 9, no. 9, p. 1516, 2019.
(4 citations on 2 pages: 29 and 35)

[114] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 658–666.
(3 citations on 2 pages: 29 and 36)

[115] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, 2023.
(3 citations on 2 pages: 29 and 36)

[116] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.
(2 citations on 2 pages: 29 and 37)

[117] T. Sorenson, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, and its application to analysis of vegetation on danish commons," *Kong Dan Vidensk Selsk Biol Skr*, vol. 5, pp. 1–5, 1948.
(2 citations on 2 pages: 29 and 37)

[118] D. Szymkiewicz, "Une conlribution statistique à la géographie floristique," *Acta Societatis Botanicorum Poloniae*, vol. 11, no. 3, pp. 249–265, 1934.
(2 citations on 2 pages: 29 and 37)

[119] G. G. Sempson, "Holarctic mammalian faunas and continental relationships during the cenozoic," *Geological Society of America Bulletin*, vol. 58, no. 7, pp. 613–688, 1947.
(2 citations on 2 pages: 29 and 37)

[120] C. Bell, "Mutual information and maximal correlation as measures of dependence," *The Annals of Mathematical Statistics*, pp. 587–595, 1962.
(2 citations on 2 pages: 29 and 37)

[121] D. W. Goodall, "Sample similarity and species correlation," in *Ordination of plant communities*. Springer, 1978, pp. 99–149.
(2 citations on 2 pages: 29 and 37)

[122] K. Suzuki, "Overview of deep learning in medical imaging," *Radiological physics and technology*, vol. 10, no. 3, pp. 257–273, 2017.
(1 citation on 1 page: 29)

[123] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision," *NPJ digital medicine*, vol. 4, no. 1, p. 5, 2021.
(1 citation on 1 page: 29)

[124] N. Buch, S. A. Velastin, and J. Orwell, "A review of computer vision techniques for the analysis of urban traffic," *IEEE Transactions on intelligent transportation systems*, vol. 12, no. 3, pp. 920–939, 2011.
(1 citation on 1 page: 29)

[125] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, 2018.
(1 citation on 1 page: 29)

[126] D. Feng, A. Harakeh, S. L. Waslander, and K. Dietmayer, "A review and comparative study on probabilistic object detection in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 9961–9980, 2021.
(1 citation on 1 page: 29)

[127] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," *International Journal of Computer Vision*, vol. 130, no. 10, pp. 2425–2452, 2022.
(1 citation on 1 page: 29)

[128] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.
(1 citation on 1 page: 30)

[129] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369.
(1 citation on 1 page: 30)

[130] S. Yu, B. Eder, R. Dennis, S.-H. Chu, and S. E. Schwartz, "New unbiased symmetric metrics for evaluation of air quality models," *Atmospheric Science Letters*, vol. 7, no. 1, pp. 26–34, 2006.
(2 citations on 2 pages: 32 and 33)

[131] K. Tsigaridis, N. Daskalakis, M. Kanakidou, P. Adams, P. Artaxo, R. Bahadur, Y. Balkanski, S. Bauer, N. Bellouin, A. Benedetti *et al.*, "The aerocom evaluation and intercomparison of organic aerosol in global models," *Atmospheric Chemistry and Physics*, vol. 14, no. 19, pp. 10 845–10 895, 2014.
(1 citation on 1 page: 32)

[132] M. R. Mebust, B. K. Eder, F. S. Binkowski, and S. J. Roselle, "Models-3 community multiscale air quality (cmaq) model aerosol component 2. model evaluation," *Journal of Geophysical Research: Atmospheres*, vol. 108, no. D6, 2003.
(1 citation on 1 page: 33)

[133] A. Al-Kaff, D. Martin, F. Garcia, A. de la Escalera, and J. M. Armingol, "Survey of computer vision algorithms and applications for unmanned aerial vehicles," *Expert Systems with Applications*, vol. 92, pp. 447–463, 2018.
(1 citation on 1 page: 36)

[134] Y. Akbari, N. Almaadeed, S. Al-Maadeed, and O. Elharrouss, "Applications, databases and open computer vision research from drone videos and images: a survey," *Artificial Intelligence Review*, vol. 54, pp. 3887–3938, 2021.
(1 citation on 1 page: 36)