




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

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

Junior Professorship of Media Computing,  
Chemnitz University of Technology,  
09107 Chemnitz, Germany,  
`{firstname.lastname}@cs.tu-chemnitz.de`

September 30, 2023

## Abstract

Motivated by the recent developments in machine learning (ML) and deep learning (DL) as well as a general need for big data analysis in object detection and classification as well as image generation by utilizing learning-based approaches and models such as deep neural networks (DNN), this manuscript aims at providing a consolidated overview of evaluation and performance metrics for machine learning and computer vision (CV). For this purpose, well-established evaluation metrics are presented, for which their (dis-)advantages as well as their origins are emphasized. Following, all to this manuscript related data, including our  $\text{\LaTeX}$  source code, will be made publicly available and can be found under [https://github.com/TSchlosser13/Evaluation\\_Metrics](https://github.com/TSchlosser13/Evaluation_Metrics). As this manuscript is meant as a continuously consolidated overview, more evaluation metrics will be added over time.

# Table of Contents

<b>1 Machine Learning</b>	<b>4</b>
1.1 Precision / positive predictive value (PPV)	5
1.1.1 Macro average precision (APmacro)	5
1.1.2 Micro average precision (APmicro)	5
1.1.3 Weighted average precision (APweighted)	6
1.2 Negative predictive value (NPV)	6
1.3 Recall / true positive rate (TPR) / sensitivity / hit rate	6
1.3.1 Macro average recall (ARmacro)	7
1.3.2 Micro average recall (ARmicro)	7
1.3.3 Weighted average recall (ARweighted)	7
1.4 True negative rate (TNR) / specificity / selectivity	8
1.5 Prevalence	8
1.6 Accuracy (A)	8
1.7 Balanced accuracy (BA)	8
1.8 Balanced accuracy weighted (BAW)	9
1.9 Average accuracy (AA)	9
1.10 Average class accuracy (ACA)	9
1.11 Error rate (ER)	10
1.12 Average error rate (AER)	10
1.13 F-score / $F\beta$ -score	10
1.13.1 Macro average F-score	10
1.13.2 Micro average F-score	11
1.13.3 Weighted average F-score	11
1.14 F0-score	11
1.15 F0.5-score	12
1.16 F1-score	12
1.17 F2-score	12
1.18 False discovery rate (FDR)	12
1.18.1 Macro average FDR (FDRmacro)	13
1.18.2 Micro average FDR (FDRmicro)	13
1.18.3 Weighted average FDR (FDRweighted)	13
1.19 False omission rate (FOR)	13
1.20 False positive rate (FPR)	13
1.21 False negative rate (FNR) / miss rate	14
1.21.1 Macro average FNR (FNRmacro)	14
1.21.2 Micro average FNR (FNRmicro)	14
1.21.3 Weighted average FNR (FNRweighted)	14
1.22 Positive likelihood ratio (LR+)	14
1.23 Negative likelihood ratio (LR-)	15
1.24 Diagnostic odds ratio (DOR)	15
1.25 Fowlkes–Mallows index (FM)	15
1.26 Informedness / bookmaker informedness (BM) / Youden's J statistic / Youden's index	16
1.27 Markedness (MK)	16
1.28 Matthews correlation coefficient (MCC) / phi coefficient / Yule phi coefficient	16
1.29 Jaccard index (JI) / threat score (TS) / critical success index (CSI)	16

1.30	Receiver operating characteristic curve (ROC curve) and area under the curve (AUC)	17
1.31	Average precision (AP)	18
1.32	Mean average precision (mAP)	18
1.33	Cohen's kappa for binary classification	18
1.34	Gini impurity	19
1.35	P4 metric	19
1.36	Skill score (SS)	19
1.37	Relative improvement factor	20
<b>2</b>	<b>Computer Vision</b>	<b>24</b>
2.1	Error (E)	25
2.2	Absolute error / sum of absolute errors (AE)	25
2.3	Relative absolute error (RAE)	25
2.4	Mean error (ME)	25
2.5	Mean percentage error (MPE)	26
2.6	Mean absolute error (MAE)	26
2.7	Mean absolute percentage error (MAPE)	26
2.8	Mean absolute scaled error (MASE)	26
2.9	Mean normalized bias (MNB)	27
2.10	Normalized mean bias (NMB)	27
2.11	Squared error / sum of squared errors (SE)	27
2.12	Mean square error (MSE)	28
2.13	Root mean square error (RMSE)	28
2.14	Normalized root mean square error (NRMSE)	28
2.15	Root mean squared logarithmic error (RMSLE)	29
2.16	Peak signal-to-noise ratio (PSNR)	29
2.17	Structural similarity (SSIM)	29
2.18	Structural dissimilarity (DSSIM)	30
2.19	Intersection over union (IoU)	30
2.20	Dice coefficient (DC)	31
2.21	Overlap coefficient (OC)	31

# 1 Machine Learning

The following evaluation metrics within the context of machine learning are motivated by the contributions of [1, 2].

---

## General

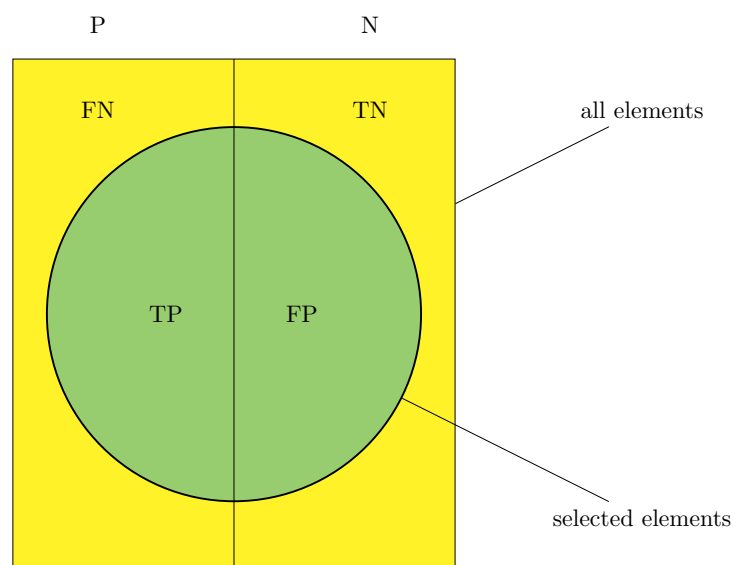


Figure 1: General definitions machine learning (reprinted from [3, Figure 4.1]).

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.

---

## 1.1 Precision / positive predictive value (PPV) [4, 5]

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

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

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

---

### 1.1.1 Macro average precision ( $AP_{macro}$ ) [6, 7, 8, 9]

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, thus 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.

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

---

### 1.1.2 Micro average precision ( $AP_{micro}$ ) [6, 7]

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.

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

---

### 1.1.3 Weighted average precision ( $AP_{\text{weighted}}$ ) [10]

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.

$$AP_{\text{weighted}} = \frac{1}{\sum_{i=1}^n w_i} \cdot \sum_{i=1}^n w_i \cdot \text{Precision}_i = \frac{1}{\sum_{i=1}^n w_i} \cdot \sum_{i=1}^n w_i \cdot \frac{TP_i}{TP_i + FP_i} \quad (4)$$

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

---

## 1.2 Negative predictive value (NPV) [4, 5]

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.

$$NPV = \frac{TN}{TN + FN} \quad (5)$$

---

## 1.3 Recall / true positive rate (TPR) / sensitivity / hit rate [11, 12]

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

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

$$\text{Recall} = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (6)$$

---

### 1.3.1 Macro average recall ( $AR_{macro}$ ) [6, 7, 13, 14]

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, thus 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.

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

---

### 1.3.2 Micro average recall ( $AR_{micro}$ ) [6, 7]

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.

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

---

### 1.3.3 Weighted average recall ( $AR_{weighted}$ ) [15, 10]

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.

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

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

---

## 1.4 True negative rate (TNR) / specificity / selectivity [11, 12]

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.

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} \quad (10)$$

---

## 1.5 Prevalence [16, 17]

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

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

$$Prevalence = \frac{P}{T} = \frac{P}{P + N} \quad (11)$$

---

## 1.6 Accuracy (A) [1, 18]

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).

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

---

## 1.7 Balanced accuracy (BA) [19, 20]

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

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

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



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


---

## 1.8 Balanced accuracy weighted (BAW) [21, 22]

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.

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

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

---

## 1.9 Average accuracy (AA) [19, 23]

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.

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


---

## 1.10 Average class accuracy (ACA) [24, 25]

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.

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

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

---

### 1.11 Error rate (ER) [26, 27]

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.

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

---

### 1.12 Average error rate (AER) [28, 29]

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.

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

---

### 1.13 F-score / F $\beta$ -score [30, 31]

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

- + The recall can be weighted  $\beta$  times as important as the precision.
- Finding the correct  $\beta$  can be different from application to application.

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

---

#### 1.13.1 Macro average F-score [32, 33]

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.

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

where:  $F_i$  = F-score of class  $i$  with a chosen  $\beta$

---

### 1.13.2 Micro average F-score [34, 33]

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.

$$\text{Micro average F-score} = 2 \cdot \frac{AP_{micro} \cdot AR_{micro}}{AP_{micro} + AR_{micro}} \quad (22)$$

---

### 1.13.3 Weighted average F-score [35, 36]

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

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

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

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  $i$

---

## 1.14 F0-score [30, 31]

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.

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

---

### 1.15 F0.5-score [30, 31]

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.

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

---

### 1.16 F1-score [30, 31]

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

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

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

---

### 1.17 F2-score [30, 31]

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.

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

---

### 1.18 False discovery rate (FDR) [37, 38]

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

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

$$FDR = \frac{FP}{FP + TP} \quad (28)$$

---

### 1.18.1 Macro average FDR (FDRmacro)

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

---

### 1.18.2 Micro average FDR (FDRmicro)

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

---

### 1.18.3 Weighted average FDR (FDRweighted)

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

---

## 1.19 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.

$$FOR = \frac{FN}{FN + TN} \quad (32)$$

---

## 1.20 False positive rate (FPR) [40]

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.

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \quad (33)$$

---

## 1.21 False negative rate (FNR) / miss rate [40]

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

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

$$FNR = \frac{FN}{P} = \frac{FN}{FN + TP} \quad (34)$$

---

### 1.21.1 Macro average FNR (FNRmacro)

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

---

### 1.21.2 Micro average FNR (FNRmicro)

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

---

### 1.21.3 Weighted average FNR (FNRweighted)

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

---

## 1.22 Positive likelihood ratio (LR+) [41, 42]

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).

$$LR+ = \frac{Sensitivity}{1 - Specificity} = \frac{TPR}{FPR} \quad (38)$$

---

### 1.23 Negative likelihood ratio (LR−) [41, 42]

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).

$$LR- = \frac{1 - Sensitivity}{Specificity} = \frac{FNR}{TNR} \quad (39)$$

---

### 1.24 Diagnostic odds ratio (DOR) [43, 44]

In medicine, DOR measures the classification effectiveness of a diagnostic test. [43] (range:  $[0, \infty)$ )

- + Emphasizes the accuracy of the LR+ in relation to the LR−.
- + Independent of the prevalence.
- − Increased calculation and interpretation complexity.

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

---

### 1.25 Fowlkes–Mallows index (FM) [45, 46]

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.
- − F1-score geometric mean equivalent (no weighting factor  $\beta$  provided).
- − Less well known and used.

$$FM = \sqrt{PPV \cdot TPR} = \sqrt{Precision \cdot Recall} \quad (41)$$

---

### 1.26 Informedness / bookmaker informedness (BM) / Youden's J statistic / Youden's index [47, 48]

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

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

$$\text{Informedness} = \text{Sensitivity} + \text{Specificity} - 1 = \text{TPR} + \text{TNR} - 1 \quad (42)$$

---

### 1.27 Markedness (MK) [49]

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  $1$ .

$$\text{MK} = \text{PPV} + \text{NPV} - 1 \quad (43)$$

---

### 1.28 Matthews correlation coefficient (MCC) / phi coefficient / Yule phi coefficient [50, 51, 52]

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

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

$$\text{MCC} = \sqrt{\text{TPR} \cdot \text{TNR} \cdot \text{PPV} \cdot \text{NPV}} \quad (44)$$

---

### 1.29 Jaccard index (JI) / threat score (TS) / critical success index (CSI) [53, 54]

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.



$$JI = \frac{TP}{TP + FN + FP} \quad (45)$$


---

### 1.30 Receiver operating characteristic curve (ROC curve) and area under the curve (AUC) [55, 56, 2]

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

- + Multiple results allow for a more meaningful evaluation.
- Dependent on TPR and FPR.
- Increased calculation complexity.

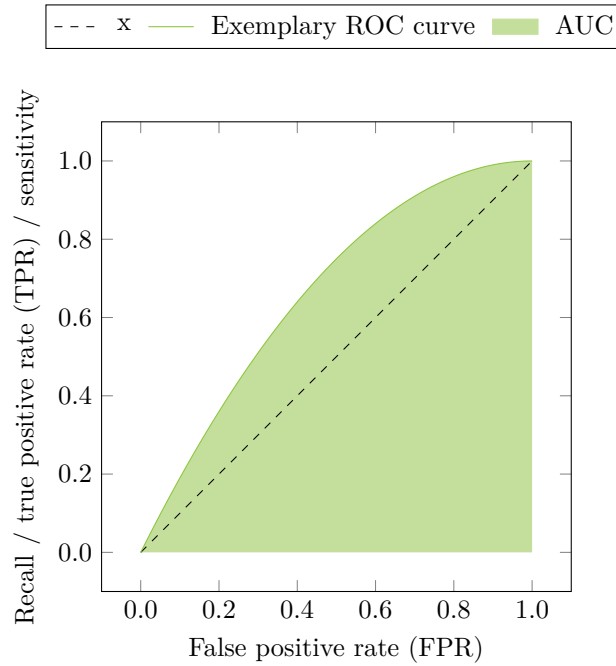


Figure 2: Receiver operating characteristic curve (ROC curve) and area under the curve (AUC).

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


---

### 1.31 Average precision (AP) [57, 58]

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

- + Multiple results allow for a more meaningful evaluation.
- Dependent on precision and recall.
- Increased calculation complexity.

$$AP = \int_{x=0}^1 \text{Precision}(\text{Recall}^{-1}(x)) dx = \frac{1}{n} \cdot \sum_{i=1}^n \text{Precision}(\text{Recall}_i) \quad (47)$$

---

### 1.32 Mean average precision (mAP) [57, 58]

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.

$$mAP = \frac{1}{n} \cdot \sum_{i=1}^n AP_i \quad (48)$$

---

### 1.33 Cohen's kappa for binary classification [59, 60, 61]

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

- + Robust against class imbalance.
- More difficult to interpret than other metrics.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (49)$$
$$P_e = \left( \frac{TP + FP}{T} \right) \cdot \left( \frac{TP + FN}{T} \right) + \left( \frac{TN + FN}{T} \right) \cdot \left( \frac{TN + FP}{T} \right)$$

where:  $\kappa$  = Cohen's kappa

$P_e$  = expected accuracy

$P_o$  = model accuracy

---

### 1.34 Gini impurity [62, 63, 64]

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.

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

where:  $D$  = data set

$p$  = probability of samples

---

### 1.35 $P_4$ metric [65]

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.

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

---

### 1.36 Skill score (SS) [66]

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.

$$SS = 1 - \frac{Metric_{GT}}{Metric_P} \quad (52)$$

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

---

### 1.37 Relative improvement factor [67]

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.

$$\text{Relative improvement factor} = \frac{1 - \text{Metric}_{GT}}{1 - \text{Metric}_P} \quad (53)$$

where:  $\text{Metric}_{GT}$  = best possible expectation for a model based on a given metric  
 $\text{Metric}_P$  = actual prediction of a model based on a given metric

---

Figure 3 shows an example result for a classification problem with 3 imbalanced classes, while Fig. 4 illustrates an example result for a classification problem with 3 balanced classes created from the Iris flower data set of R. A. Fisher [68]. 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 while 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 dataset while this is not the case for the imbalanced dataset. 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).

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

1: *setosa*

2: *versicolor*

3: *virginica*

Ground truth class	Predicted class			$\Sigma$	$Recall_i$	$FNR_i$	Eq. 6	$\frac{TP}{TP + FN}$	Eq. 34	$\frac{FN}{FN + TP}$	
	1: <i>setosa</i>	50	0	0	50	100%	0%	$Recall_1$	$\frac{50}{50 + 0 + 0}$	$FNR_1$	$\frac{0 + 0}{0 + 0 + 50}$
	2: <i>versicolor</i>	0	47	3	50	94%	6%	$Recall_2$	$\frac{47}{47 + 0 + 3}$	$FNR_2$	$\frac{0 + 3}{0 + 3 + 47}$
	3: <i>virginica</i>	0	2	48	50	96%	4%	$Recall_3$	$\frac{48}{48 + 0 + 2}$	$FNR_3$	$\frac{0 + 2}{0 + 2 + 48}$
$\Sigma$	50	49	51	150							
$Precision_i$	100%	95.9%	94.1%								
$FDR_i$	0%	4.1%	5.9%								

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

$FDR_1$

$FDR_2$

$FDR_3$

$\frac{FP}{FP + TP}$

$\frac{0 + 0}{0 + 0 + 50}$

$\frac{0 + 2}{0 + 2 + 47}$

$\frac{0 + 3}{0 + 3 + 48}$

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

Class	Precision [%]	Recall [%]	F <sub>1</sub> [%]	$\Sigma$
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}

## Functions

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

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 <sub>macro</sub> ) (2)	average_precision_score	/	AveragePrecision
Micro average precision (AP <sub>micro</sub> ) (3)	average_precision_score	/	AveragePrecision
Weighted average precision (AP <sub>weighted</sub> ) (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>.

<sup>1</sup>scikit-learn project page, <https://scikit-learn.org/stable/>

<sup>2</sup>TensorFlow project page, <https://www.tensorflow.org/>

<sup>3</sup>Keras project page, <https://keras.io/>

<sup>4</sup>PyTorch project page, <https://pytorch.org/>

<sup>5</sup>TorchMetrics project page, <https://torchmetrics.readthedocs.io/en/stable/>

## 2 Computer Vision

The following evaluation metrics within the context of machine learning are motivated by the contributions of [69, 70].

---

### General

Abbreviation	Meaning
$GT$	Ground truth
$P$	Prediction
$n$	Number of values

Table 3: General definitions computer vision.

---



## 2.1 Error (E)

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

- + Intuitive and straightforward to apply.

$$E = GT - P \quad (54)$$

---

## 2.2 Absolute error / sum of absolute errors (AE) [71]

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

- + Straightforward to calculate.
- No differentiation depending on the number of compared values is made.
- Large individual differences equal to many small ones (distribution problem).

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

---

## 2.3 Relative absolute error (RAE) [72, 73, 74]

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.

$$RAE = \frac{AE}{\sum_{i=1}^n |GT_i - \overline{GT}|} = \frac{\sum_{i=1}^n |GT_i - P_i|}{\sum_{i=1}^n |GT_i - \overline{GT}|} \quad (56)$$
$$\overline{GT} = \frac{1}{n} \cdot \sum_{i=1}^n GT_i$$

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

---

## 2.4 Mean error (ME) [75, 76]

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

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

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


---

## 2.5 Mean percentage error (MPE) [77, 78]

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.

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


---

## 2.6 Mean absolute error (MAE) [79, 80]

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.

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


---

## 2.7 Mean absolute percentage error (MAPE) [72, 80]

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.

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


---

## 2.8 Mean absolute scaled error (MASE) [80, 81]

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.

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


---

## 2.9 Mean normalized bias (MNB) [82, 83]

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.

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


---

## 2.10 Normalized mean bias (NMB) [84, 82]

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.

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


---

## 2.11 Squared error / sum of squared errors (SE) [85]

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.
- Large individual differences equal to many small ones (distribution problem).

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


---

## 2.12 Mean square error (MSE) [86]

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.

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

---

## 2.13 Root mean square error (RMSE) [79, 80, 87]

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.

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

---

## 2.14 Normalized root mean square error (NRMSE) [88, 89]

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

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

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

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

---

## 2.15 Root mean squared logarithmic error (RMSLE) [90]

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.

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

---

## 2.16 Peak signal-to-noise ratio (PSNR) [91, 92]

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.

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

where:  $E_{max}$  = maximum possible error

---

## 2.17 Structural similarity (SSIM) [93, 94]

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

- + Provides more accurate results by considering structural characteristics. [93]
- Increased calculation complexity.

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

where:  $\mu_x, \mu_y$  = mean

$\sigma_x, \sigma_y$  = variance

$\sigma_{xy}$  = 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)

---

## 2.18 Structural dissimilarity (DSSIM) [93, 94]

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

- + Provides more accurate results by considering structural characteristics. [93]
- Increased calculation complexity.

$$DSSIM = \frac{1 - SSIM}{2} \quad (71)$$

---

## 2.19 Intersection over union (IoU) [53, 54, 95, 96]

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:  $[0, 1]$ )

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

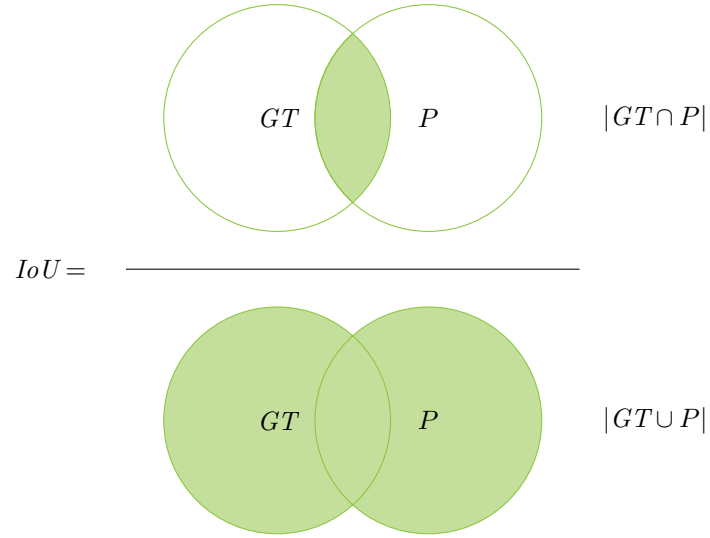


Figure 5: Intersection over union (IoU).

$$IoU = \frac{|GT \cap P|}{|GT \cup P|} \quad (72)$$

---

## 2.20 Dice coefficient (DC) [97, 98]

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.

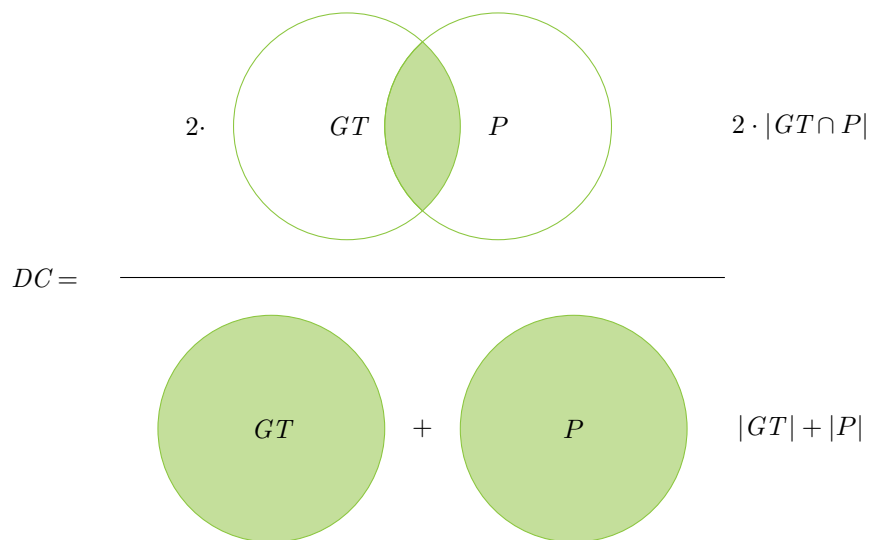


Figure 6: Dice coefficient (DC).

$$DC = \frac{2 \cdot |GT \cap P|}{|GT| + |P|} \quad (73)$$

## 2.21 Overlap coefficient (OC) [99, 100, 101, 102]

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

- + Can be well used for image segmentation and object detection.
- + A differentiation depending on the size of both sets is made.
- A differentiation based on set size can be detrimental.

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

## Functions

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

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

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>`.

<sup>6</sup>scikit-learn project page, <https://scikit-learn.org/stable/>

<sup>7</sup>TensorFlow project page, <https://www.tensorflow.org/>

<sup>8</sup>Keras project page, <https://keras.io/>

<sup>9</sup>PyTorch project page, <https://pytorch.org/>

<sup>10</sup>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] C. E. Metz, “Basic principles of roc analysis,” in *Seminars in nuclear medicine*, vol. 8, no. 4. Elsevier, 1978, pp. 283–298.  
(2 citations on 2 pages: 4 and 8)
- [2] T. Fawcett, “An introduction to roc analysis,” *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.  
(2 citations on 2 pages: 4 and 17)
- [3] T. Schlosser, “Biologically Inspired Hexagonal Deep Learning for Hexagonal Image Processing,” Ph.D. dissertation, Chemnitz University of Technology, 2023.  
(1 citation on 1 page: 4)
- [4] D. G. Altman and J. M. Bland, “Statistics notes: Diagnostic tests 2: predictive values,” *Bmj*, vol. 309, no. 6947, p. 102, 1994.  
(2 citations on 2 pages: 5 and 6)
- [5] G. S. Fletcher, *Clinical epidemiology: the essentials*. Lippincott Williams & Wilkins, 2019.  
(2 citations on 2 pages: 5 and 6)
- [6] Y. Yang, “An evaluation of statistical approaches to text categorization,” *Information retrieval*, vol. 1, no. 1-2, pp. 69–90, 1999.  
(4 citations on 2 pages: 5 and 7)
- [7] F. Sebastiani, “Machine learning in automated text categorization,” *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.  
(4 citations on 2 pages: 5 and 7)
- [8] M. Zhu, “Recall, precision and average precision,” *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, vol. 2, no. 30, p. 6, 2004.  
(1 citation on 1 page: 5)
- [9] 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.  
(1 citation on 1 page: 5)
- [10] 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.  
(2 citations on 2 pages: 6 and 7)
- [11] 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.  
(2 citations on 2 pages: 6 and 8)
- [12] D. G. Altman and J. M. Bland, “Diagnostic tests. 1: Sensitivity and specificity,” *BMJ: British Medical Journal*, vol. 308, no. 6943, p. 1552, 1994.  
(2 citations on 2 pages: 6 and 8)
- [13] A. Rosenberg, “Classifying skewed data: Importance weighting to optimize average recall,” in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.  
(1 citation on 1 page: 7)

- [14] 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.  
(1 citation on 1 page: 7)
- [15] 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.  
(1 citation on 1 page: 7)
- [16] K. J. Rothman, *Epidemiology: an introduction*. Oxford university press, 2012.  
(1 citation on 1 page: 8)
- [17] N. Bruce, D. Pope, and D. Stanistreet, *Quantitative methods for health research: a practical interactive guide to epidemiology and statistics*. John Wiley & Sons, 2018.  
(1 citation on 1 page: 8)
- [18] J. Taylor, *Introduction to error analysis, the study of uncertainties in physical measurements*. University Science Books, 1997.  
(1 citation on 1 page: 8)
- [19] 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.  
(2 citations on 2 pages: 8 and 9)
- [20] 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.  
(1 citation on 1 page: 8)
- [21] A. Salman, U. Sharaha, E. Rodriguez-Diaz, E. Shufan, K. Riesenber, 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.  
(1 citation on 1 page: 9)
- [22] 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.  
(1 citation on 1 page: 9)
- [23] 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: 9)
- [24] 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: 9)
- [25] 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: 9)

- [26] 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: [10](#))
- [27] 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: [10](#))
- [28] 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: [10](#))
- [29] 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: [10](#))
- [30] C. J. Van Rijsbergen, *The geometry of information retrieval*. Cambridge University Press, 2004.  
(5 citations on 3 pages: [10](#), [11](#), and [12](#))
- [31] 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.  
(5 citations on 3 pages: [10](#), [11](#), and [12](#))
- [32] 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: [10](#))
- [33] 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: [10](#) and [11](#))
- [34] 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: [11](#))
- [35] 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: [11](#))
- [36] 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: [11](#))
- [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.  
(1 citation on 1 page: [12](#))

- [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.  
(1 citation on 1 page: 12)
- [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.  
(1 citation on 1 page: 13)
- [40] 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.  
(2 citations on 2 pages: 13 and 14)
- [41] 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.  
(2 citations on 2 pages: 14 and 15)
- [42] J. J. Deeks and D. G. Altman, “Diagnostic tests 4: likelihood ratios,” *Bmj*, vol. 329, no. 7458, pp. 168–169, 2004.  
(2 citations on 2 pages: 14 and 15)
- [43] 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.  
(2 citations on 1 page: 15)
- [44] 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.  
(1 citation on 1 page: 15)
- [45] 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.  
(1 citation on 1 page: 15)
- [46] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” *Journal of intelligent information systems*, vol. 17, pp. 107–145, 2001.  
(1 citation on 1 page: 15)
- [47] C. S. Peirce, “The numerical measure of the success of predictions,” *Science*, no. 93, pp. 453–454, 1884.  
(1 citation on 1 page: 16)
- [48] W. J. Youden, “Index for rating diagnostic tests,” *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.  
(1 citation on 1 page: 16)
- [49] 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: 16)

- [50] 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: 16)
- [51] 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: 16)
- [52] H. Cramér, *Mathematical methods of statistics*. Princeton university press, 1999, vol. 26.  
(1 citation on 1 page: 16)
- [53] P. Jaccard, “The distribution of the flora in the alpine zone. 1,” *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.  
(2 citations on 2 pages: 16 and 30)
- [54] 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.  
(2 citations on 2 pages: 16 and 30)
- [55] D. M. Green, J. A. Swets *et al.*, *Signal detection theory and psychophysics*. Wiley New York, 1966, vol. 1.  
(1 citation on 1 page: 17)
- [56] 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.  
(1 citation on 1 page: 17)
- [57] C. D. Manning, *An introduction to information retrieval*. Cambridge university press, 2009.  
(2 citations on 1 page: 18)
- [58] 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 1 page: 18)
- [59] 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: 18)
- [60] 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: 18)
- [61] 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: 18)
- [62] 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: 19)

- [63] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.  
(1 citation on 1 page: 19)
- [64] 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: 19)
- [65] M. Sitarz, “Extending f1 metric, probabilistic approach. advances in artificial intelligence and machine learning. 2023; 3 (2): 61,” 2023.  
(1 citation on 1 page: 19)
- [66] 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: 19)
- [67] 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: 20)
- [68] 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: 21)
- [69] 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: 24)
- [70] 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: 24)
- [71] I. E. Richardson, *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia*. John Wiley & Sons, 2004.  
(1 citation on 1 page: 25)
- [72] 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.  
(2 citations on 2 pages: 25 and 26)
- [73] —, “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.  
(1 citation on 1 page: 25)
- [74] 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.  
(1 citation on 1 page: 25)

- [75] 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.  
(1 citation on 1 page: 25)
- [76] 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 (ICICT)*, vol. 1. IEEE, 2019, pp. 1264–1268.  
(1 citation on 1 page: 25)
- [77] 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.  
(1 citation on 1 page: 26)
- [78] 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.  
(1 citation on 1 page: 26)
- [79] 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.  
(2 citations on 2 pages: 26 and 28)
- [80] 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.  
(4 citations on 2 pages: 26 and 28)
- [81] 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.  
(1 citation on 1 page: 26)
- [82] 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 1 page: 27)
- [83] 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: 27)
- [84] 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: 27)
- [85] N. R. Draper and H. Smith, *Applied regression analysis*. John Wiley & Sons, 1998, vol. 326.  
(1 citation on 1 page: 27)



- [86] P. J. Bickel and K. A. Doksum, *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. CRC Press, 2015.  
(1 citation on 1 page: 28)
- [87] 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.  
(1 citation on 1 page: 28)
- [88] 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.  
(1 citation on 1 page: 28)
- [89] 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.  
(1 citation on 1 page: 28)
- [90] 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.  
(1 citation on 1 page: 29)
- [91] D. Salomon, *Data compression: the complete reference*. Springer Science & Business Media, 2004.  
(1 citation on 1 page: 29)
- [92] 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.  
(1 citation on 1 page: 29)
- [93] 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.  
(4 citations on 2 pages: 29 and 30)
- [94] 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.  
(2 citations on 2 pages: 29 and 30)
- [95] 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.  
(1 citation on 1 page: 30)
- [96] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, “Object detection in 20 years: A survey,” *Proceedings of the IEEE*, 2023.  
(1 citation on 1 page: 30)
- [97] L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.  
(1 citation on 1 page: 31)

- [98] 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.  
(1 citation on 1 page: 31)
- [99] D. Szymkiewicz, “Une contribution statistique à la géographie floristique,” *Acta Societatis Botanicorum Poloniae*, vol. 11, no. 3, pp. 249–265, 1934.  
(1 citation on 1 page: 31)
- [100] 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.  
(1 citation on 1 page: 31)
- [101] C. Bell, “Mutual information and maximal correlation as measures of dependence,” *The Annals of Mathematical Statistics*, pp. 587–595, 1962.  
(1 citation on 1 page: 31)
- [102] D. W. Goodall, “Sample similarity and species correlation,” in *Ordination of plant communities*. Springer, 1978, pp. 99–149.  
(1 citation on 1 page: 31)