

Fair Machine Learning

Handout

Authors: Fiona Fischer, Linda Fernsel
in project 'Fair Enough?' (16DHB4002) of the
research group Computer Science and Society, at
the University of Applied Sciences (HTW) Berlin.



Version: November 2023

License: [CC-BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/)

Content

This handout addresses the topic of 'Fair Machine Learning' using the example of facial recognition technology. It begins with an introduction to facial recognition technology and demonstrates the value of critically examining machine learning through the 'Gender Shades' study. Then, it explains how machine learning models can be critically viewed from both a social and a technological perspective.

IdentiTOP Use Case & Jupyter Notebook

The contents of the handout are demonstrated using the fictional use case *IdentiTOP* in a Python [Jupyter Notebook](#). [Google Colab](#) offers the possibility to use Jupyter Notebooks with just a browser.

The notebook is available at bit.ly/identitop.

Learning Objectives

- Participants know use cases for facial recognition technology.
- Participants can identify common risks when applying machine learning models (Social Perspective).
- Participants know quality metrics for machine learning models (Technological Perspective).
- Participants can select appropriate quality metrics for specific risks (Technological Perspective).
- Participants can calculate quality metrics based on the confusion matrix (Technological Perspective).
- Participants can interpret the significance of quality metric results for the risks of machine learning models (Technological Perspective + Social Perspective).

1. Facial Recognition Technology

With the help of machine learning, we can develop models that classify data based on features or make predictions about it. Such models are used, for example, in facial recognition technology. Facial recognition technology has three stages [3]:

Facial Recognition

Identifying whether there is a face in the image.

1

Feature Recognition

Identifying what kind of face it is.







2

Identity Recognition

Identifying whose face it is.

3

Examples of use cases for these technologies are:

1 Facial Recognition	2 Feature Recognition	3 Identity Recognition
 Photo filters in apps like SnapChat [1]	 Identifying gender and age to suggest targeted advertisements [2]	 Identification of wanted individuals in video surveillance [5]
 Monitoring during an online exam to ensure a person is alone in front of the computer [8]	 Measuring interest during a lesson as feedback for teachers [7]	 Identification of participants during an online exam [6]

Check the Jupyter Notebook for the description of *IdentiTOP*.

2. Gender Shades

"Gender Shades" is one of many studies that critically examine machine learning. In this study, automated facial recognition services were investigated for their fairness regarding gender and skin color [4]. The analysis of facial analysis algorithms found that women, compared to men, and darker-skinned individuals, compared to lighter-skinned individuals, were more frequently misclassified [4]. Dark-skinned women were the most frequently misclassified [4]. Researchers found that this was primarily due to the training data of the algorithms: existing datasets mainly included lighter-skinned men [4].






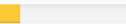





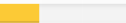





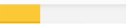
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% 	79.2% 	100% 	98.3% 	20.8% 
 FACE++	99.3% 	65.5% 	99.2% 	94.0% 	33.8% 
 IBM	88.0% 	65.3% 	99.7% 	92.9% 	34.4% 

Figure 1: Results of the "Gender Shades" study. Source: gendershades.org

3. Critically Examining Machine Learning

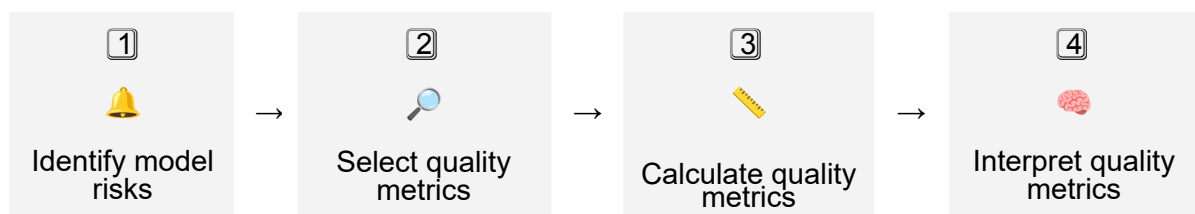
The critical examination can be carried out in three sequential stages, of which primarily Stage 1 is addressed here:

How fair is the machine learning system? 1

What are the causes of unfairness? 2

How can unfairness be addressed? 3

The critical examination at Stage 1 follows four steps:



In conclusion, an additional step can be added, which, however, already touches on Stages **2** and:



All steps are explained in more detail below:

Step **1**: Identify model risks

The following questions can help identify the risks of a model:

- What are potential **disadvantages** and **negative consequences** of a system?
- What **errors** can occur?
- What are the **impacts** of these problems on people?
- Can these problems occur **structurally** and cause **discrimination**?

The following section discusses the previously presented use cases for facial recognition technology with regard to their risks:



Photo filters in apps like SnapChat [1]

The use of facial recognition technology in apps like SnapChat allows users to apply interactive, entertaining, or beautifying filters to their faces. However, errors in recognizing facial features can occur, resulting in unwanted filter effects. Additionally, the use of beautifying filters could disturb users' self-image and the perception of others.



Identifying gender and age to suggest targeted advertisements [2]

Identifying gender and age enables targeted advertising that is potentially more relevant to the addressed individual. Incorrect identification of gender or age could lead to irrelevant ads. This could also have psychological effects, for example, for trans people. Even when recognized correctly, individuals might feel misunderstood or stereotyped. It is important to ensure that stereotypes are not reinforced and that people have the option to manually specify their preferences.



Identification of wanted individuals through video surveillance [5]

The identification of wanted individuals could enhance public safety and make police work more efficient. However, false identifications could target innocent citizens for law enforcement. Innocent people could be arrested or harassed, leading to mental and physical issues [10], as well as a loss of trust in law enforcement agencies. Therefore, this technology should be used with great caution and in combination with human review.



Monitoring the presence of a person during an online exam on a PC [8]

Monitoring whether a person is alone in front of the PC during an online exam is intended to prevent cheating during exams. However, the technology might falsely interpret other objects as a second person or fail to recognize other people who are present. Exam candidates could be unjustly accused of cheating, which could have serious academic consequences. Furthermore, the feeling of being monitored could put examinees under pressure, causing them to behave differently than usual.



Measuring interest in class as feedback for teachers [7]

If student interest in the classroom is measured, teachers could receive valuable feedback to adjust and improve their teaching. However, not all facial expressions reflect genuine interest or understanding. This is especially true for individuals from different cultural backgrounds who may have different facial expressions or who express their emotions differently due to inherent characteristics. Students may be unjustly labeled as disinterested or distracted, which could lead to unwarranted educational interventions. Therefore, this technology should not be used as the sole means for assessing interest.



Identification of participants in an online exam [6]

The identification of participants in an online exam ensures that the correct person is taking the exam. However, false identification could lead to someone being wrongly excluded from an exam, resulting in academic and emotional strain.

Check the Jupyter Notebook for the identification of potential risks of *IdentiTOP*.

Step 2: Select quality metrics

In the previous step, we determined that each model carries individual risks. It is therefore necessary to select metrics that help to assess specific risks. For each risk, the appropriate metric must be identified. Many metrics are based on the confusion matrix.

The Confusion Matrix

A confusion matrix provides an overview of the accuracy of a model's results by comparing the model's predictions with reality (i.e., the correct result). The results of a machine learning model can be divided into "positive" and "negative."

The confusion matrix includes the number of correct predictions: the number of true positive cases (abbreviated as **TP**) and the number of true negative cases (abbreviated as **TN**). It also includes the number of incorrect predictions: the number of false positive cases (abbreviated as **FP**) and the number of false negative cases (abbreviated as **FN**).

Specifically, the confusion matrix is structured as follows [11]:

Prediction → Reality ↓	Predicted as Positive	Predicted as Negative
Actually Positive	TP	FN
Actually Negative	FP	TN

A high model quality is achieved when TP and TN are high, and FP and FN are low.

Check the Jupyter Notebook for the confusion matrix of *IdentiTOP*.

What quality metrics are available?

Common metrics for evaluating model quality are accuracy, precision, and recall [9]. Accuracy refers to all predictions (TP + TN + FP + FN), precision to the cases with positive predictions (TP + FP), and recall to the actual positive cases (TP + FN). These metrics are calculated based on the confusion matrix as follows [11]:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

In addition, there are less common metrics, such as specificity, the false negative rate (FNR), the false positive rate (FPR), the false discovery rate (FDR), and the false omission rate (FOR), which are based on the confusion matrix [11]. These metrics are calculated based on the confusion matrix as follows:

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$$

$$\text{FNR} = \text{FN} / (\text{FN} + \text{TP})$$

$$\text{FPR} = \text{FP} / (\text{FP} + \text{TN})$$

$$\text{FDR} = \text{FP} / (\text{FP} + \text{TN})$$

$$\text{FOR} = \text{FN} / (\text{FN} + \text{TN})$$

Fairness

Fairness can be measured using mathematical fairness definitions. Fairness definitions indicate whether a model performs equally well for different groups.

What does "equally well for different groups" mean?

Groups	Equally	Well
The division of individuals based on different attributes , e.g., "gender." One group in this case would be "women." Groups can also be "intersectional." This means that people are divided based on multiple attributes, e.g., "gender" and "skin color." In this case, for example, "dark-skinned women" would be a group.	The quality of the model should be similar for all groups. Measures of similarity between two values are the difference (e.g., FPR of group 1 minus FPR of group 2) or the ratio (e.g., FPR of group 1 divided by FPR of group 2).	The measurable quality of the model, e.g., FPR. The quality metric is chosen based on the risk.

Some well-known, concrete fairness definitions include [13]:

- **Predictive Parity:** This corresponds to precision parity. That is, precision must be equally high for each group.
- **Equal Opportunity:** This corresponds to false negative rate (FNR) equality. That is, FNR must be equally low for each group.

- **Predictive Equality:** This corresponds to false positive rate (FPR) equality. That is, FPR must be equally low for each group.

Check the Jupyter Notebook for setting quality metrics for *IdentiTOP*.

Step 3: Calculate quality metrics

Based on the model results (predictions and reality), the selected metrics are now calculated.

For fairness-related risks, the appropriate quality metrics are calculated separately for each (intersectional) group. This separate calculation of quality metrics is also known as "slicing analysis."

For comparing quality metrics between two groups, the difference or the ratio of the results is calculated. To compare the quality between more than two groups, you can use, for example, the value of the best-performing group as the reference. Comparing with the average quality of all groups can also be a meaningful option.

⚠ Note: Not every metric can be directly calculated from the model results. It may be necessary to collect or request additional data. For assessing fairness definitions, for example, demographic data must be available.

Check the Jupyter Notebook for calculating the quality metrics of *IdentiTOP*.

Step 4: Interpret quality metrics

In the final step, the calculations are interpreted. Do the values indicate fulfillment of the quality criteria? Which risk hypotheses have been confirmed?

The threshold at which quality is considered similar enough to be fair must be determined depending on the context. Often, the "Disparate Impact" rule is used as a reference [12]. According to this rule, the ratio of two quality results should be between 0.8 and 1.2, or, under another interpretation, the difference should lie between -0.2 and 0.2. If the quality values are very small or would have to be divided by zero, it is advisable to use the difference instead of the ratio as a measure of equality.

The interpretation often varies depending on the comparison method and the tolerance limits chosen. Therefore, all calculations should always be disclosed and interpreted comprehensively.

Check the Jupyter Notebook for the interpretation of the results for *IdentiTOP*.

Step 5: Develop improvement suggestions

Based on the results of the quality assessment, one can consider the reasons for the results (see also [14]) and provide recommendations on how to attempt to improve the system.

Check the Jupyter Notebook for the recommendations for improving *IdentiTOP*.

Sources

- 1) BesteTipps (2023). Gesichtserkennung auf Snapchat: wie geht das? Lösung. Available at: <https://www.bestetipps.de/computer/handy/gesichtserkennung-auf-snapchat-wie-geht-das-loesung/> (Last accessed: 02.08.2023)
- 2) Vossen, R. (2013). Zukunft der Werbung: „Cara“ erkennt Alter und Geschlecht – und zeigt das passende Plakatmotiv. Available at: <https://www.basichthinking.de/blog/2013/05/22/zukunft-der-werbung-cara-erkennt-alter-und-geschlecht-und-zeigt-das-passende-plakatmotiv/> (Last accessed: 02.08.2023)
- 3) adaLearning (2019). Joy Boulamwini on Face Recognition Technology. Available at: <https://www.youtube.com/watch?v=rWMLcNaWfe0> (Last accessed: 07.08.2023)
- 4) Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of Machine Learning Research (Vol. 81, pp. 1-15). Conference on Fairness, Accountability, and Transparency. (Last accessed: 07.08.2023)
- 5) Leisegang, D. (2020). Biometrische Videoüberwachung: Die neue Hochrisikotechnologie. Blätter für deutsche und internationale Politik. Available at: <https://www.blaetter.de/ausgabe/2020/maerz/biometrische-videoueberwachung-die-neue-hochrisikotechnologie> (Last accessed: 07.08.2023)
- 6) Stiebert, J. (2022). Klage gegen Gesichtserkennung bei Uni-Prüfungen. Posteo News. Available at: <https://posteo.de/news/klage-gegen-gesichtserkennung-bei-uni-pr%C3%BCfungen> (Last accessed: 07.08.2023)
- 7) Wang, N. (2019). Totale Überwachung Chinas Schulen: Wenn Kameras jede Gesichtsregung auswerten. Der Tagesspiegel. Available at: <https://www.tagesspiegel.de/politik/wenn-kameras-jede-gesichtsregung-auswerten-4657504.html> (Last accessed: 07.08.2023)
- 8) Schiller, A. (2021). Universitätsräten Studenten mit Software aus. Frankfurter Allgemeine Zeitung. Available at: <https://www.faz.net/aktuell/karriere-hochschule/proctorio-und-wiseflow-hochschulen-spaehen-studenten-aus-17455837.html> (Last accessed: 07.08.2023)
- 9) Oppermann, A. (2021): Accuracy, Precision, Recall, F1-Score und Specificity. Available at: <https://artemoppermann.com/de/accuracy-precision-recall-f1-score-und-specificity/> (Last accessed: 08.08.2023)
- 10) Hill, K., (2023): Eight Months Pregnant and Arrested After False Facial Recognition Match. Available at: <https://www.nytimes.com/2023/08/06/business/facial-recognition-false-arrest.html> (Last accessed: 09.08.2023)
- 11) Wikipedia (2023): Confusion matrix. Available at: https://en.wikipedia.org/wiki/Confusion_matrix (Last accessed: 19.10.2023)
- 12) Wikipedia (2023): Disparate impact. Available at: https://en.wikipedia.org/wiki/Disparate_impact (Last accessed: 19.10.2023)
- 13) Verma, S. und Rubin, J. (2018). Fairness Definitions Explained. FairWare. Available at: <https://doi.org/10.1145/3194770.3194776>
- 14) Suresh, H. und Gutttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. EEAMO'21: Equity and Access in Algorithms, Mechanisms, and Optimization. Available at: <https://doi.org/10.1145/3465416.3483305>