Fair Machine Learning

An Open Educational Resource

Authors: Fiona Fischer, Linda Fernsel
in project "Fair Enough?" (16DHB4002)
of the research group Informatics and Society
at the University of Applied Sciences for Technology
and Economics (HTW) Berlin

Version: November 2023

Table of Contents

1. About this OER

Contents, learning objectives, prerequisites, and accompanying materials of this OER.

This OER provides structure, content, and task ideas for a teaching unit on the topic of *Fair Machine Learning*. The unit addresses both social and technical aspects of fair machine learning and aims to enable individuals who may engage with machine learning models in the future to critically examine these systems.

The focus of this OER is the application of machine learning in facial recognition technologies. All methods are demonstrated using the invented facial recognition tool *IdentiTOP*.

This OER pursues the following learning objectives:

- Participants understand use cases for facial recognition technology.
- Participants can identify common risks associated with the application of machine learning models (social perspective).
- Participants understand quality metrics for machine learning models (technological perspective).
- Participants can select appropriate quality metrics based on identified risks (technological perspective).

- Participants can calculate quality metrics based on the confusion matrix (technological perspective).
- Participants can interpret the significance of quality measurement results for the risks of machine learning models (technological perspective + social perspective).

This OER can supplement the topic of "Machine Learning," which is taught, for example, in upper-level computer science classes or at universities. It assumes that participants are already familiar with some basics of "Machine Learning." The use of the accompanying Jupyter Notebooks also presupposes a basic understanding of programming.

Python with Jupyter Notebook can be used to learn about the programmatic execution of audits. The Jupyter Notebooks are available in the repository at https://gitlab.com/iug-research/oer-faires-machine-learning (or shortened at https://bit.ly/identitop).

- Using the notebook fairness-audit-identitop.ipynb, the audit for *IdentiTOP* can be programmatically reproduced.
- The notebook fairness-audit-kreditwuerdigkeit-aufgabe.ipynb contains an advanced task, where an independent audit should be conducted under guidance.
- The solution is included in the notebook fairness-audit-kreditwuerdigkeit-loesung.ipynb.

Google Colab offers the possibility to use Jupyter Notebooks with only a browser.

Learners can also independently study this learning unit. For this, there is a handout available at https://iug.htw-berlin.de/wp-content/uploads/2023/11/Handout-Faires-Machine-Learning.pdf, which is intended to be used in combination with the Jupyter Notebook for *IdentiTOP*.

2. Introduction

This section introduces the topic and terminology and motivates engagement with the subject.

With the help of machine learning, we can develop models that classify data based on characteristics or make predictions. However, it has become evident that these models are not always fair.

An example of this is the project Gender Shades [4]. This project investigated the fairness of automated facial analysis services concerning gender and skin tone. The analysis of facial analysis algorithms revealed that women and darker-skinned individuals were more frequently misclassified compared to men and lighter-skinned individuals. 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 predominantly included lighter-skinned men [4].

Therefore, the topic of "Fair Machine Learning" aims to critically examine machine learning from various perspectives. This involves exploring why machine learning systems are unfair (Step 1), identifying the causes of unfairness (Step 2), and understanding how unfairness can be mitigated (Step 3).

The fairness of a model can be examined from various perspectives:

1. The social perspective concerns the assessment of social consequences and risks of a model when humans interact with and use it.
2. The technical perspective pertains to the technical measurement of model quality based on mathematical definitions.

3. Use Cases of Facial Recognition Technology

This section provides an overview of possible use cases for facial recognition technology. It also introduces the use case "IdentiTOP," which will serve as an example in the further course of the OER.

Facial recognition technology uses machine learning models for three different levels of tasks related to a person's face [3]:

1. **Face detection:** Identifying whether there is a face in the image.
2. **Feature Recognition:** Identifying what kind of face it is.
3. **Identity Recognition**: Identifying whose face it is.

Examples of Use Cases:

Facial Recognition	**Feature Recognition**	**Identity Recognition**
Photo filters in apps like Snapchat [1]	Detecting gender and age to suggest targeted advertisements [2]	Identifying individuals sought by police in video surveillance [5]
Monitoring whether a person is alone at a computer during an online exam [8]	Measuring interest in class as feedback for teachers [7]	Identifying participants during an online exam [6]

Use Case: IdentiTOP - A Hypothetical System for Identifying Participants in Online Exams

Online exams pose the risk that students may not complete the exam themselves but instead hire someone else to do it for them. The goal of IdentiTOP is to prevent this type of cheating during online exams.

When a person registers with an educational institution (e.g., a university), they must submit a copy of their ID card. The individuals' ID images are stored in a database. IdentiTOP has access to the database of ID images of all registered students and also knows which course they are enrolled in. During the registration process at the educational institution and at the start of the online exam, examinees are informed about the use of IdentiTOP and the consequences of cheating.

During the exam, IdentiTOP takes random snapshots of the person via the webcam while they are writing the exam. The system then retrieves the ID image of the registered person from the database. This ID image is compared to the webcam photo using a machine learning model, which predicts whether the individuals in the photos are the same person.

If it is the same person, they can continue working undisturbed, and their identity will be checked again by IdentiTOP in a few minutes.

However, if the model determines that it is not the same person, the exam attempt is terminated by the system, and the cheating individual receives a "failed" grade for the exam attempt.

The developers of IdentiTOP emphasize that invigilators no longer need to constantly monitor possibly hundreds of examinees via webcam during the exam, as the system eliminates the need for their personal intervention in cases of cheating. This allows cheating to be consistently identified and addressed. Cheating is further reduced as students are informed in advance about the system and its consequences.

Teaching Idea: Ask participants if they have ever come into contact with facial recognition technology and for what purpose facial recognition was used in that case. Collect the use cases on paper or on the board. Categorize the use cases into the three stages of facial recognition technology.

4. Critically Examining Machine Learning Models

This section explains the steps for critically examining machine learning models.

The critical examination of machine learning models at Stage 1 is conducted in four steps: (1) Identification of model risks, (2) Selection of quality metrics,

(3) calculation of quality metrics, and (4) interpretation of quality metrics. Finally, an additional step can be added, but this already corresponds to Stages 2 and 3 of the critical examination: (5) Developing improvement proposals.

4.1. Identifying the Risks of a Model

With the help of the social perspective, the application-specific social consequences and risks of a machine learning model, when used by humans, can be identified.
The following questions can help to identify risks:

- What are the potential disadvantages and negative consequences of a system?
- What errors can occur?
- How do these problems affect people?
- Could these issues occur structurally and lead to discrimination?

Examples forof Discussions on Use Cases of Facial Recognition

Photo Filters in Apps like Snapchat [1]

- Technology:

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 may occur, leading to unwanted filter effects. Additionally, the use of beautifying filters may distort one's self-image and that of other users.

Recognizing Gender and Age to Suggest Targeted Advertising [2]

Recognizing gender and age enables targeted advertising that may be more relevant to the intended audience. Incorrect recognition of gender or age could lead to irrelevant advertisements. This may also have psychological effects, for example, on trans individuals. Even with accurate recognition, people could feel misrepresented or stereotyped. It is crucial to ensure that no stereotypes are reinforced and that individuals have the ability to manually indicate their preferences.

Identifying Individuals Wanted by Police Through Video Surveillance [5]

The identification of individuals wanted by police could enhance public safety and make law enforcement more efficient. However, false identifications could bring innocent individuals under criminal suspicion. This can lead to psychological and physical harm [10], as well as a loss of trust in law enforcement agencies. This technology should be used cautiously and in combination with human oversight.

Monitoring Whether a Person Is Alone at a Computer During an Online Exam [8]

Monitoring whether a person is alone at a computer during an online exam ensures that cheating during exams is prevented. However, the technology could interpret other objects as a second person by mistake, or fail to recognize people as such. Test takers could be falsely accused of cheating, which might have serious academic consequences. Moreover, the feeling of being watched could put students under pressure, causing them to behave differently than they normally would.

Measuring Interest in Class as Feedback for Teachers [7]

If students' interest in class is measured, teachers could receive valuable feedback to adjust and improve their teaching. However, not all facial expressions necessarily reflect actual interest or understanding. This might affect individuals from different cultural backgrounds with varying facial expressions or people who express their emotions differently due to inherent traits. Students could be unfairly classified as disinterested or distracted, leading to unjustified educational measures. This technology should not be used as the sole means for evaluating interest.

Identifying Participants in an Online Exam [6]

The identification of participants in an online exam ensures that the correct person is taking the test. However, false identification could result in a person being unfairly excluded from an exam, leading to academic and emotional distress.

Teaching Idea: Discuss the previously gathered use cases with students to practice identifying risks.

Teaching Note: The topic of "facial recognition technology" lends itself well to discussions about the history and social implications of surveillance, as well as further discussions on data privacy.

Risks of IdentiTOP:

- (R1) If IdentiTOP falsely assumes that the person cannot be identified as themselves (i.e., the system fails to recognize the correct individual), this could cause significant emotional distress for the innocent individual. Additionally, it could result in organizational burdens (e.g., appealing the judgment by IdentiTOP to obtain a new exam attempt) or academic disadvantages (if a new attempt is not granted).
- (R2) If IdentiTOP falsely assumes that the individual is the correct person, even though they only look similar, cheating could go undetected, and the actual performance of the individual might not be properly assessed. This could result in a person being hired for a job they are not qualified for, which could have negative effects on all involved parties in the specific situation.
- If IdentiTOP works better for lighter-skinned men than for other groups (see “Gender Shades” [4]), some groups may face greater disadvantages than others, reinforcing existing structural discrimination.

- (R4) If IdentiTOP works better for individuals with higher-quality webcams, financially disadvantaged groups may be discriminated against.
- (R5) Test-takers may feel their privacy is violated, as IdentiTOP takes unexpected photos of them. This could lead to emotional distress and consequently poorer exam performance.

4.2. Selecting Quality Metrics

In examining the social perspective, we found that every model brings its own individual risks. Therefore, it is necessary to select metrics that help assess specific risks. For each risk, an 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 the actual outcomes. The results of a machine learning model are classified as either "positive" or "negative."

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

The confusion matrix is structured as follows [11]:

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

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

Confusion Matrix for IdentiTOP

The result "both images show the same person" is considered a positive outcome, and the result "both images show different persons" is considered a negative outcome.

IdentiTOP was deliberately tested. 160 out of 300 test participants "cheated" to determine how well IdentiTOP functions.

This is the confusion matrix for IdentiTOP:

	Predicted as Positive	Predicted as Negative
Actually Positive	113	27
Actually Negative	10	150

The confusion matrix indicates the following:

- 113 cases were correctly predicted as "both images show the same person."
- 10 cases were falsely predicted as "both images show the same person," when they did not.
- 27 cases were falsely predicted as "both images show different persons," when they were the same.
- 150 cases were correctly predicted as "both images show different persons."

This means the model made 37 errors out of 300 cases and was correct in the remaining 263 cases.

Teaching Suggestion: Ask students to analyze the above statements from the confusion matrix with questions such as:

"How many predictions were correct that the two images showed the same person?"

What Quality Metrics Exist?

Common metrics for evaluating model quality include accuracy, precision, and recall [9]. Accuracy refers to all predictions (TP + TN / TP + TN + FP + FN), precision only to the cases with positive predictions (TP / TP + FP), and recall to the truly positive cases (TP / TP + FN).

These metrics are calculated as follows based on the confusion matrix [11]:

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)

Calculation of Common Metrics for IdentiTOP:

- Accuracy = 263 / 300 ≈ 0,88
- Precision = 113 / 123 ≈ 0,92
- Recall = 113 / 140 ≈ 0,81

Teaching Idea: Ask students to calculate the metrics based on the confusion matrix and the metric definitions.

Additionally, there are less commonly used metrics, such as the **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 as follows [11]:

- Specificity = TN / (TN + FP)
- FNR = FN / (FN + TP)
- FPR = FP / (FP + TN)
- FDR = FP / (FP + TP)
- FOR = FN / (FN + TN)

Fairness

Fairness can be measured using mathematical fairness definitions. Fairness definitions state that a model works **equally well** for **different groups**.
Groups refer to the division of individuals based on various attributes, such as "gender." One group in this case would be "women." Groups can also be "intersectional," meaning individuals are divided based on multiple attributes, such as "gender" and "skin color." In this case, "dark-skinned women" would represent one group.
Equal means that the quality of the model should be similar for all groups. Measures of similarity between two numbers include the difference (e.g., FPR for Group 1 minus FPR for Group 2) or the ratio (e.g., FPR for Group 1 divided by FPR for Group 2).
Good refers to the measurable quality of the model, such as FPR. The quality metric is selected based on the risk.

Some named fairness definitions are [13]:

- **Predictive Parity:** This corresponds to precision parity. This means the precision for each group must be equally high.
- **Equal Opportunity**: This corresponds to False Negative Rate Equality (FNR Equality). This means the FNR for each group must be equally low.
- Predictive Equality**:** This corresponds to False Positive Rate Equality (FPR Equality). This means the FPR for each group must be equally low.

Selection of Quality Metrics for IdentiTOP:

- (R1) *IdentiTOP* falsely assuming that the individual is not the same person is acceptable if truly positive cases result in few false negatives (FN). This can be measured with the FNR. A lower FNR makes it less likely for R1 to occur.
- (R2) *IdentiTOP* falsely assuming that the individual is the same person is acceptable if truly negative cases result in few false positives (FP). This can be measured with the FPR. A lower FPR makes it less likely for R2 to occur.
- (R3) *IdentiTOP* works better for lighter-skinned men than for other groups. This consists of two parts:

  - Risk a) is acceptable if the quality criterion for (R1) is equally valid for people of all genders and skin tones. The corresponding fairness definition is FNR Equality or Equal Opportunity.
  - Risk b) is acceptable if the quality criterion for (R2) is equally valid for people of all genders and skin tones. The corresponding fairness definition is FPR Equality or Predictive Equality.

- (R4) corresponds to (R3), but the groups are further subdivided based on webcam quality. The treatment of this risk is similar to (R3) and does not need to be shown separately. In reality, however, it is still very important to examine R4.
- (R5) is acceptable if most test-takers report that their privacy is not significantly disrupted by *IdentiTOP*. To verify this, a survey of *IdentiTOP* users would need to be conducted and evaluated. This example omits R5 for simplicity. In reality, however, it is crucial to examine R5 as well.

Teaching Idea: Encourage students to think about how risks can be verified using the already known quality metrics. They can collaboratively work toward a solution.

4.3. Calculating Quality Metrics

The selected metrics are now calculated based on the model's results (predictions and actual outcomes).

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

To compare the quality metrics of two groups, the difference or the ratio of the results is examined. To compare the quality across more than two groups, for example, the value of the group with the worst performance can be compared with the group with the best performance. Alternatively, a comparison with the average quality across all groups can be a useful option.

Not every metric can be calculated directly from the model's results. Additional data may need to be collected or requested. To verify fairness definitions, demographic data, for instance, may need to be available.

Demographic Data for Fairness Assessment of *IdentiTOP*:
Für R3 muss das Geschlecht und die Hautfarbe der Personen im Testdatenset bekannt sein. Um ein valides Ergebnis zu erhalten, wurde beim gezielten Test von IdentiTOP eine balancierte Auswahl an Testpersonen getroffen und ihre demographischen Daten als binärer Wert erhoben.

The following confusion matrices were derived for the different demographic groups:

Group	TP	FN	FP	TN	∑
Male	56	9	5	80	150
Female	57	18	5	70	150
Light-skinned	64	3	4	77	148
Dark-skinned	49	24	6	73	152
Male and light-skinned	30	2	3	39	74
Male and dark-skinned	26	7	2	41	76
Female and light-skinned	34	1	3	38	74
Female and dark-skinned	23	17	4	32	76

Teaching Idea: To practice understanding the confusion matrix, ask students questions about the data, such as: "How many light-skinned men who attempted to cheat were caught by IdentiTOP?"

Quality Metric Calculation for IdentiTOP:

- (R1) FNR ≈ 0,19
- (R2) FPR ≈ 0,06
- (R3 a)
  - Gender
    - FNR(male) ≈ 0,14
    - FNR(female) ≈ 0,24
    - FNR(f) / FNR(m) ≈ 1,73 > 1,2
      FNR(f) - FNR(m) ≈ 0,10
  - Skin Color
    - FNR(light-skinned) ≈ 0,04
    - FNR(dark-skinned) ≈ 0,33
    - FNR(d) / FNR(l) ≈ 7,34 > 1,2
      FNR(d) - FNR(l) ≈ 0,28 > 0,2
  - Intersectional Groups
    - FNR (male and light-skinned) ≈ 0.06
    - FNR (male and dark-skinned) ≈ 0.21
    - FNR (female and light-skinned) ≈ 0.03
    - FNR (female and dark-skinned) ≈ 0.43

        - The best FNR value is ~0.03 (for female and light-skinned). The average FNR value is ~0.18. Using these FNR values, each remaining group is now compared:
            - FNR(ml) / FNR(min) ≈ 2,19 > 1,2
              FNR(ml) - FNR(min) ≈ 0,03
              FNR(ml) / FNR(avg) ≈ 0,34 < 0,8
              FNR(ml) - FNR(avg) ≈ -0,12
            - FNR(md) / FNR(min) ≈ 7,42 > 1,2
              FNR(md) - FNR(min) ≈ 0,18
              FNR(md) / FNR(avg) ≈ 1,17
              FNR(md) - FNR(avg) ≈ 0,03
            - FNR(fl) / FNR(avg) ≈ 0,16 < 0,8
              FNR(fl) - FNR(avg) ≈ -0,15
            - FNR(fd) / FNR(min) ≈ 14,88 > 1,2
              FNR(fd) - FNR(min) ≈ 0,40 > 0,2
              FNR(fd) / FNR(avg) ≈ 2,33 > 1,2
              FNR(fd) - FNR(avg) ≈ 0,24 > 0,2

- (R3 b)
    - Gender
        - FPR(male) ≈ 0,06
        - FPR(female) ≈ 0,07
        - FPR(f) / FPR(m) ≈ 1,13
          FPR(f) - FPR(m) ≈ 0,01
    - Skin Color
        - FPR(light-skinned) ≈ 0,05
        - FPR(dark-skinned) ≈ 0,08
        - FPR(d) / FPR(l) ≈ 1,54 > 1,2
          FPR(d) - FPR(l) ≈ 0,03
    - Intersectional Groups
        - FPR (male and light-skinned) ≈ 0.07
        - FPR (male and dark-skinned) ≈ 0.05
        - FPR (female and light-skinned) ≈ 0.03
        - FPR (female and dark-skinned) ≈ 0.11
        - The best FPR value is ~0.03 (for females and light-skinned). The average FPR value is ~0.06. Using these FPR values, each remaining group is now compared:
            - FPR(ml) / FPR(min) ≈ 2,79 > 1,2
              FPR(ml) - FPR(min) ≈ 0,05
              FPR(ml) / FPR(avg) ≈ 1,12
              FPR(ml) - FPR(avg) ≈ 0,01
            - FPR(md) / FPR(min) ≈ 1,81 > 1,2
              FPR(md) - FPR(min) ≈ 0,02
              FPR(md) / FPR(avg) ≈ 0,73 < 0,8

FPR(md) - FPR(avg) ≈ -0,02

- FPR(fl) / FPR(avg) ≈ 0,40 < 0,8
  FPR(wh) - FPR(avg) ≈ -0,04
- FPR(fl) / FPR(min) ≈ 4,33 > 1,2
  FPR(fd) - FPR(min) ≈ 0,09 < 0,2
  FPR(fd) / FPR(avg) ≈ 1,75 > 1,2
  FPR(wd) - FPR(avg) ≈ 0,05

Teaching Idea: Calculate FOR and TPR together and divide the calculation of the fairness definitions among the students. The calculations can be done on paper, in Excel, or using any programming language—for example, in Python with the accompanying Jupyter Notebook.

4.4. Interpreting Quality Metrics

As 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 deemed similar enough to be fair must be determined based on context. A common guideline is the "Disparate Impact" rule [12], which states that the ratio of two quality metric results should fall between 0.8 and 1.2, or that the absolute difference should lie between -0.2 and 0.2. If the quality metric is very small and close to 0, the difference ratio is often more appropriate as a measure of equality.

It is often debated which metrics to compare and what tolerance ranges should apply. Therefore, all calculations should always be disclosed and transparently interpreted.

Interpretation of the Results of the Quality Calculation for *IdentiTOP*:

- (R1) Of all positive cases (where the person did not cheat), 20% were falsely classified as cheating. This means one in five cheating alerts disrupts an innocent person during their exam attempt. This risk hypothesis has been confirmed.
- (R2) 6% of cheating test-takers were not detected by the model. This indicates that the risk of undetected cheating during exams is relatively low and therefore acceptable.
- (R3 a) Among non-cheating individuals, dark-skinned people, particularly dark-skinned women, are more likely to be falsely classified as cheating. Dark-skinned women are misclassified twice as often as dark-skinned men and more often than light-skinned individuals. Light-skinned women are the least likely to receive a false alert, while light-skinned men have only slightly more alerts. Therefore, there is no general discrimination based on gender,

but rather based on skin color, with additional discrimination against dark-skinned women.

- (R3 b) The FPR is relatively low for all groups, but it becomes evident that the model is slightly less effective at detecting dark-skinned cheaters compared to light-skinned cheaters. Light-skinned women are most often detected as cheaters, followed by dark-skinned men, then light-skinned men, and finally dark-skinned women, whose cheating goes undetected in 11% of cases.
- Conclusion: It can be concluded that *IdentiTOP* generally recognizes cheating attempts as such; however, dark-skinned cheaters are less likely to be detected compared to light-skinned cheaters. This difference is particularly noticeable among women. Furthermore, a non-negligible portion of innocent test-takers is disrupted during their exam attempt, leading to undue stress for the students. Dark-skinned individuals, especially dark-skinned women, experience significantly more false alerts than light-skinned individuals. *IdentiTOP* also disproportionately disadvantages dark-skinned individuals, particularly dark-skinned women.

Teaching Idea: Ask Learners to interpret results.

4.5. Developing Improvement Suggestions

Individuals with prior knowledge of machine learning can also reflect on the reasons behind the results (see also [14]) and provide suggestions on how the system could be improved.

Recommendation:
The high burden on innocent test-takers could be avoided through a "human-in-the-loop" strategy, where a human reviews the identity of the person flagged by the cheating alert based on the photos and only approves the exam termination if cheating is confirmed by the human reviewer.
However, to stay true to *IdentiTOP's* original goals of reducing the need for invigilators, ensuring fairness, and improving model quality, the system should be further optimized. The reason for the high FNR and the FNR—and, to some extent, FPR—inequalities might lie in the training data of IdentiTOP. *IdentiTOP* should be trained on a dataset containing more dark-skinned individuals, especially dark-skinned women, and then tested again accordingly.

Teaching Idea: Ask students to consider where the unfairness in the model arises and what could be done to minimize the risks of the model.

5. Summary & Conclusion

Machine learning enables the development of models that classify and predict data. The topic of "Fair Machine Learning" examines the quality of machine learning models not only from a technical perspective but also from a social perspective, identifying causes of unfairness and seeking solutions. Ensuring fairness in AI systems, particularly in sensitive areas such as facial recognition, is crucial. Continuous evaluation and improvement are necessary to avoid discrimination and bias.

Teaching Idea: Ask students to write three sentences summarizing their insights about the fairness of machine learning and the significance of facial recognition technology from this lesson.

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.basicthinking.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 in 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). Universitäten spähen 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 Guttag, 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