Towards a Quantitative Evaluation Framework for Trustworthy AI in Facial Analysis

Annika Schreiner
Friedrich-Alexander-Universität Erlangen-Nürnberg
annika.schreiner@fau.de

Nils Kemmerzell
Friedrich-Alexander-Universität Erlangen-Nürnberg
nils.kemmerzell@fau.de

Abstract

As machine learning (ML) models are increasingly being used in real-life applications, ensuring their trustworthiness has become a rising concern. Previous research has extensively examined individual perspectives on trustworthiness, such as fairness, robustness, privacy, and explainability. Investigating their interrelations could be the next step in achieving an improved understanding of the trustworthiness of ML models.

By conducting experiments within the context of facial analysis, we explore the feasibility of quantifying multiple aspects of trustworthiness within a unified evaluation framework. Our results indicate the viability of such a framework, achieved through the aggregation of diverse metrics into holistic scores. This framework can serve as a practical tool to assess ML models in terms of multiple aspects of trustworthiness, specifically enabling the quantification of their interactions and the impact of training data. Finally, we discuss potential solutions to key technical challenges in developing the framework and the opportunities of its transfer to other use cases.

Keywords: Trustworthy AI, Evaluation, Facial Analysis

1. Introduction

As the use of machine learning (ML) systems in real-life applications increases, there is a growing demand for their trustworthiness. In response, several research fields, such as Explainable AI (XAI) or privacy-preserving machine learning, investigate certain perspectives of the trustworthiness of ML models. However, current literature also has started to analyze select combinations of aspects (Balagopalan et al., 2022; Schrouff et al., 2022; Shokri et al., 2021; C. Tran et al., 2021). In this context, Trustworthy AI represents an area of research that investigates how a safe, transparent, and responsible use of AI can be ensured, to increase user trust (X. Liu et al., 2021). It includes multiple perspectives from which the concept of the trustworthiness of ML models can be approached. The aspects of fairness, robustness, privacy, and explainability are addressed particularly frequently (Li et al., 2023). Operationalizing these perspectives could be the next step towards an improved understanding of the trustworthiness of ML models.

In this study, we leverage the context of facial analysis to assess the viability of quantifying diverse aspects of trustworthiness within a unified evaluation framework. We investigate the performance of ResNet models, trained on public facial image datasets, with regard to fairness, robustness, privacy, and explainability. Using this experimental basis, we explore the design of a framework that aggregates multiple quantitative metrics into comprehensive scores.

Overall, our findings indicate that this proposed framework can serve as an effective tool for gaining insights into the trustworthiness of ML models. We observe that a single score, which encompasses several dimensions of trustworthiness, is useful for a first overview, however, it may occlude trade-offs between different dimensions. Conversely, our results suggest the feasibility of formulating an aggregated score for each dimension, thereby facilitating performance comparisons across multiple models. Applying our framework to four different datasets indicates its potential utility in revealing the relationships between various perspectives of trustworthiness and in examining the impact of dataset characteristics on model

URI: https://hdl.handle.net/10125/107326
978-0-9981331-7-1

trustworthiness. We further discuss the key technical challenges associated with this framework, including the selection of relevant metrics, their normalization and aggregation, and evaluate the adaptability of our framework to other use cases.

2. Fundamentals

Our approach towards quantifying trustworthiness within an evaluation framework is based on four aspects of trustworthiness, namely fairness, robustness, privacy, and explainability.

2.1. Fairness

The concept of fairness in ML aims to ensure that decisions made by algorithms are not biased toward specific individuals or population groups. Unfair behavior can arise from potential biases in the training data or certain design decisions in the algorithm itself (Mehrabi et al., 2021). Various definitions of fairness exist, which can be categorized into group-based and individual-based approaches. Group fairness is based on the concept that certain groups, defined by sensitive attributes, should be equally likely to be classified into a particular category, whereas individual fairness relies on the idea that similar individuals should receive similar model outputs (Caton & Haas, 2020).

2.2. Robustness

Robustness ensures the performance of ML models in a variety of circumstances. When used in practice, ML models may encounter data containing various types of perturbations misleading the model, which can be naturally occurring or intentionally crafted (X. Liu et al., 2021). Robustness to distribution shift refers to the ability of an ML model of performing well under changes between the distribution of the data a model was trained on and the distribution of a test set (D. Tran et al., 2022). Robustness to adversarial attacks refers to the ability of a model to resist being misled by nearly imperceptible perturbations of the input data. White box evasion attacks, where an attacker can access the model and carefully craft fraudulent test samples, are often used to evaluate model robustness in this context (Brendel et al., 2019).

2.3. Privacy

Privacy in ML entails protecting personal data used by models (de Cristofaro, 2020). A breach occurs when training data or model parameters can be deduced from the model via various attack methods (de Cristofaro, 2020). Many of these assume the availability of a query interface allowing data input (de Cristofaro, 2020). However, usually, third parties only receive model outputs like confidence scores or predicted labels. Such outputs can be exploited through different attacks, such as membership inference, model inversion, or model extraction (Al-Rubaie & Chang, 2019).

2.4. Explainability

Explainability in ML refers to an ML model's capacity to explain its decision-making process (Adadi & Berrada, 2018). Models are often classified as white-box or black-box based on their interpretability (Carvalho et al., 2019). White-box models, such as linear models or decision trees, provide explanations through their inherent functionality. Conversely, black-box models like deep neural networks or support vector machines necessitate post-hoc methods for explanations (Adadi & Berrada, 2018).

3. Quantifying Trustworthiness

Based on the context of facial analysis, we test the feasibility of quantifying multiple aspects of trustworthiness, namely fairness, robustness, privacy, and explainability within a single evaluation framework. According to Buolamwini and Gebru (2018), the field of facial analysis addresses a range of tasks connected to the ML-based perception of faces, such as face detection, face recognition, and gender classification. Given the pervasive use of facial analysis algorithms, ranging from smartphone access control to law enforcement, potential discriminatory behaviour based on gender or ethnicity are concerns. Their responsible use demands an extensive evaluation of the underlying ML models' trustworthiness.

To assess the trustworthiness of ML models, we select the use case of gender classification. Based on facial images we train a classification model that can predict if the person in an image is *male* or *female*. We employ a ResNet-50 model, pre-trained on ImageNet, which is a popular setup for image classification tasks. After training, model predictions on a test partition are evaluated to assess its trustworthiness.

3.1. Datasets and Training Procedure

We evaluate our gender classifier using four public facial image datasets, setting *gender* as target and *race* or *age* as sensitive attributes, based on dataset annotations. All attributes are binarized for simplification, grouping *race* into *white* and *non-white* if required, as in prior ML fairness studies (Buolamwini

& Gebru, 2018), although they are not inherently binary. The datasets' sizes and demographic distributions vary, as depicted in Figure 1. LFWA+ (Z. Liu et al., 2015) and UTKFace (Z. Zhang et al., 2017) have fewer samples than FairFace (Karkkainen & Joo, 2021) and CelebA (Z. Liu et al., 2015). LFWA+ and CelebA show notable demographic imbalances, FairFace is imbalanced only for *race*, while UTKFace is nearly balanced.

We use an 80-10-10 train-validation-test-split for each dataset, with the testset adjusted for distribution of target and sensitive attributes to ensure unbiased model evaluation. The model is trained with cross-entropy loss function and Adam, common choices for image classification tasks, with a batch size of 64. Training is halted if validation loss stagnates for 10 consecutive epochs to avoid overfitting.

3.2. Selection of Evaluation Metrics

Our approach focuses on four key dimensions: fairness, robustness, privacy and explainability. We started with a baseline of metrics that are commonly used in each respective research area, as well as those implemented in popular toolboxes. We tested their applicability and validity in our setting, resulting in a selection of 21 metrics forming a solid starting point for measuring trustworthiness.

Fairness. The dimension of fairness evaluates, if the performance of our gender classification model is influenced by peoples' characteristics of *race* or *age*. In current literature, mostly group-fairness metrics are considered to assess biased behaviour, comparing a model's performance on subgroups defined by a sensitive attribute (Caton & Haas, 2020). We select typical classification metrics, such as **Accuracy**, **Precision**, **Recall/True Positive Rate (TPR)** and **False Positive Rate (FPR)**, and calculate their value difference between subgroups. Additionally, we compute the two widely used fairness metrics of **Demographic Parity (DemP) difference** and **Equalized Odds (EOd) difference**. DemP requires that the probability of being classified with the positive label is equal across groups (Berk et al., 2021). EOd defines a model as fair if TPR and False Negative Rates are equal across groups (Hardt et al., 2016). All analyses are conducted using *scikit-learn* and Microsoft's *fairlearn* package.

Robustness. By including robustness in our evaluation, we aim to measure to which extent our classification model stays reliable under circumstances where image samples deviate from the training data. To cover different situations, we collect a set of seven metrics evaluating the robustness to distribution shift and adversarial attacks. The former can be measured by comparing a model's performance on an original and a shifted test set (Taori et al., 2020). We apply the augmentation technique Augmix (Hendrycks et al., 2019) to create perturbed images, which mixes randomly generated augmentations, e.g. rotations and color swapping, thereby simulating a shift in the data distribution. Inspired by D. Tran et al. (2022), we measure the model's **Accuracy**, the **ROC-AUC-Score**, and the **Brier Score Loss** on the original test set, as well as on a shifted test set by using *scikit-learn*, and calculate the value differences between both sets.

Another way to measure robustness is against adversarial attacks. These attacks craft adversarial samples intended to cause erroneous model predictions during inference. They are created by introducing small perturbations restricted by L_p-Norm to existing samples (Goodfellow et al., 2014). The model's robustness can be measured by evaluating its accuracy on the adversarial samples (Brendel et al., 2019). Due to the high computational effort of such attacks, we reduce ourselves to two exemplary and commonly known L_∞-Norm adversarial attacks: The **Fast Gradient Sign Method (FGSM)** (Goodfellow et al., 2014) and **Projected Gradient Descent (PGD)** (Madry et al., 2017). We compare the original accuracy to the model's accuracy under these attacks by calculating the respective differences. To additionally include attack-independent metrics, we select the **CLEVER-Score** (Weng et al., 2018), measuring the minimum amount of distortion required to fool a model, and **Loss Sensitivity**, measuring the effect of each data sample on the average loss (Arpit et al., 2017). For all named metrics, we use the implementations of the *Adversarial Robustness Toolbox (ART)*.

Privacy. Our privacy evaluation assesses the model's vulnerability to malicious attacks that seek to deduce private information from the ML model, focusing on two attack types, following the implementations of Y. Liu et al. (2022). We utilize a **Membership Inference Attack (MembInf)** to assess the model's vulnerability to reveal if a specific sample was part of its training data (Shokri et al., 2017). We explore two scenarios: one in which an adversary trains the attack model using a shadow model, and another where the attacker has access to some original training data. The success of these attacks is quantified using the **ROC-AUC-Score** of the attack model on the attack test set.

We also conduct a **Model Extraction Attack (ModExt)** to examine the risk of unauthorized access to intellectual property or sensitive knowledge within

the model (Chandrasekaran et al., 2020). Here, the adversary tries to clone the original model's behavior. Two scenarios are assessed, based on the access to the original training data as outlined by Tramèr et al. (2016): one with a similar-distribution shadow dataset and the other with some actual training data. The effectiveness of these attacks is quantified by the attack model's **Accuracy**.

Explainability. Given the potential implications of facial analysis model decisions, it's crucial to provide interpretable explanations. We apply a post-hoc method to our black-box model, the ResNet-50, to derive explanations for its outputs. To evaluate the model's explainability, we follow a two-stage process. Firstly, saliency maps are generated via a method by Simonyan et al. (2013) that highlights each pixel's significance for the model's prediction. Secondly, we compute quantitative metrics to examine the explanations' characteristics, focusing on four key metrics. **Robustness** measures the sensitivity of an explanation to slight input data modifications (Yeh et al., 2019). **Faithfulness** refers to the accuracy of importance scores generated by the saliency function, reflecting each input's significance for the model's prediction (Bhatt et al., 2020). For **Randomization**, we compare explanations for model inputs to those generated on a random logit, as per Sixt et al. (2019); high similarity suggests greater explanation randomness. The explanation's **Complexity** denotes the number of input features required to clarify the model's output; effective explanations are concise and avoid irrelevant features (Chalasani et al., 2020). We implement this using the *quantus* framework, limiting the computations to a random sample of 512 test set images due to high computational demands.

3.3. Designing a Score-Based Framework

Having selected a base of metrics covering the four dimensions of fairness, robustness, privacy, and explainability, we aim to express their values in comprehensive scores to facilitate the estimation of a ML model's trustworthiness. We choose the following procedure to design our framework:

1. For each metric, its value is scaled between 0 to 10 resulting in a "Metric Score", where a score of 10 represents the optimal value for the corresponding metric. For this purpose, the respective value ranges of the metrics are projected onto the named scale by defining their limit values.

2. For each dimension, the metric scores are summarized into one "Dimension Score" (DS) by calculating their mean. Each DS displays how well the model performs in the corresponding dimension.

3. The DS for robustness, fairness, privacy, and explainability are used to calculate the average of all DS resulting in the overall "Trustworthiness Score" (TS). The TS represents the overall trustworthiness of the model as an aggregation across all dimensions.

4. Since a high prediction accuracy is a prerequisite for ML models, we compare our DS and TS to an "Accuracy Score" (AS) calculated by normalizing the model accuracy on the test set to the usual scale of 0 to 10.

3.4. Experimental Results

The results of applying our proposed evaluation framework to four different datasets, shown in Table 1, are described in the following, structured along the evaluated dimensions. Each experiment was conducted five times and we report the mean values. Due to the computational complexity of the experiments, especially for large datasets, we restricted ourselves to five runs. However, we found the variation between runs to be small for most of the examined metrics. The metric correlation matrices are depicted in the supplemental materials.[1]

Fairness Score. The fairness score displays substantial variations across datasets. The models trained on the more balanced datasets of FairFace and UTKFace achieve the highest fairness scores of almost 10. In comparison, models trained on the clearly imbalanced CelebA and especially LFWA+ show significantly lower results. The biases of the training data seem lead to more inaccurate predictions for the *not young* or *non-white* group, often being incorrectly classified as *male*.

For FairFace, the metrics of TPR and precision, as well as FPR difference are negatively correlated. Regarding the nearly equally distributed UTKFace dataset, the scores are at a slightly lower level than FairFace except for DemP difference, but all fairness metrics show positive correlations among each other. In contrast, a strong trade-off between TPR difference and all other fairness metrics is observed for CelebA. For LFWA+, additional to negative correlations between TPR and precision, as well as FPR difference, a trade-off relationship between DemP and accuracy difference can be noticed. The results also show a strong disparity between EOd difference and DemP difference.

[1] https://github.com/anniSc/TAI-Evaluation

Robustness Score. The robustness score shows significant variations across datasets, driven by a high degree of variability in metrics measuring robustness to adversarial samples. CelebA and LFWA+ models show the strongest overall performance, with a significantly high CLEVER-Score for CelebA, and accuracy under attacks for LFWA+. In contrast, the FairFace and UTKFace models exhibit a decrease in accuracy under distribution shift, and particularly low scores for adversarial attacks and CLEVER.

Overall, the distribution shift metrics exhibit only minor variations across datasets while being positively correlated to each other. Regarding the attacks, the robustness to PGD is generally very low compared to FGSM and a trade-off between both attacks exists, except for LFWA+. Regarding the attack-independent metrics, the CLEVER-Score appears to be highly variable and dependent on the dataset, while generally being positively correlated to the attack-based metrics. The loss sensitivity shows the opposite behaviour to this.

Privacy Score. The privacy score results demonstrate minor discrepancies across datasets. FairFace-trained models achieve the highest privacy score, balancing resistance to all attacks. Models trained on LFWA+ withstand model extraction attacks better, yet are more vulnerable to membership inference attacks. The UTKFace and CelebA-trained models display opposite characteristics.

These scores reveal a trade-off between resisting both membership inference and model extraction attacks, consistent with prior studies (Y. Liu et al., 2022). Our findings also support the idea that a model's vulnerability to privacy attacks is tied to its generalization capability (Yeom et al., 2018). Models trained on larger datasets, CelebA and Fairface, are less susceptible to membership inference attacks compared to the LFWA+ trained model, the smallest dataset. This suggests stronger overfitting to training samples, compromising generalization and increasing vulnerability to membership inference. However, this dynamic is inverted for model extraction attacks.

Explainability Score. Our selection of four explainability metrics aims to assess the quality of explanations generated via saliency maps. Overall, we find the explainability scores to be the lowest out of all trustworthiness dimensions. Especially, the faithfulness scores are consistently low and show a comparatively strong deviation across different runs. The other examined metrics vary strongly between datasets. Balanced datasets yield lower robustness and complexity scores, implying a broader array of input features is needed for significant explanations, which are also more perturbation-sensitive. This aligns with the less robust nature of models examined in the robustness section. Explanations for more imbalanced datasets, however, appear more random yet more robust and less complex.

Trustworthiness Score and Accuracy Score. The findings from our experiments indicate that the TS across different datasets are fairly similar, attributed largely to the compensating effects of differing DS. In our study, AS were found to correlate positively with the overall TS. For instance, models trained on the CelebA dataset delivered the highest TS and AS. Despite achieving the same TS, FairFace and UTKFace differed in terms of robustness and explainability scores, with UTKFace having a higher AS compared to FairFace. Models trained on the LFWA+ dataset displayed the lowest TS and a notably low AS.

Relationship between Dimensions. Our findings highlight a trade-off involving privacy and various other dimensions. Models that show less vulnerability to privacy breaches often exhibit higher levels of noise, reduced overall accuracy, potential inaccuracies for specific groups, decreased explainability, and an increased susceptibility to being misled by changing situations. Moreover, our experimentation uncovers a trade-off between fairness and robustness. This might arise from the conflicting objectives: one seeks to bolster the impact of sensitive attributes to attain equality among different groups, while the other aims to diminish the influence of certain attributes and samples to enhance resilience. We also note a negative correlation between fairness and explainability. On the contrary, there exists a positive relationship between robustness and explainability, possibly due to the fact that heightened model robustness also results in more resilient and simpler explanations.

Influence of Dataset Characteristics. Our experiments reveal noticeable differences in fairness and robustness scores across different datasets compared to privacy and explainability, suggesting that dataset attributes significantly affect these dimensions. Specifically, an imbalanced distribution of target and sensitive attribute classes tends to lower fairness but enhance robustness and explainability. The increased robustness in the CelebA dataset, despite its imbalance, might be due to its larger size, leading to better generalization. Similarly, LFWA+'s lower overall accuracy and stronger randomness in predictions might minimize performance drops due to distribution shift and adversarial samples. However, these observations may be correlative rather than causative, requiring further research. Notably, training on larger datasets can heighten the vulnerability to model extraction, while tending to impede membership inference attacks.

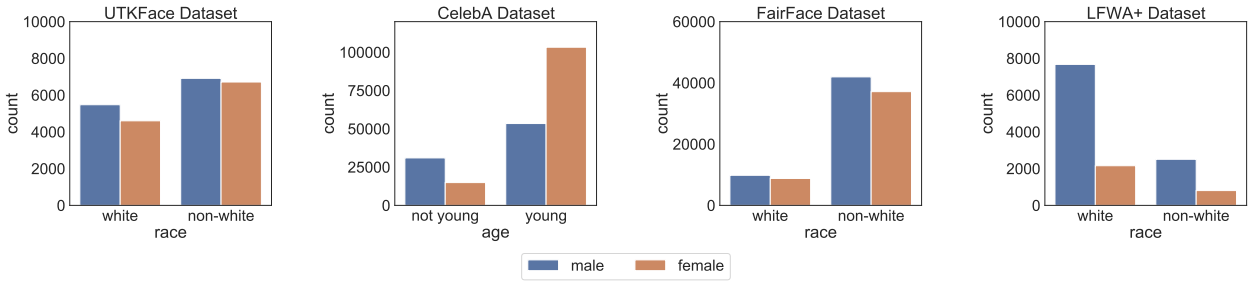

Figure 1. Demographic distributions of the selected facial image datasets

Table 1. Metric Scores, Dimension Scores, Trustworthiness Scores, and Accuracy Scores on four selected datasets with the respective standard deviations in parantheses

	UTKFace	CelebA	FairFace	LFWA+
Fairness Score	**9.69 (0.14)**	**9.40 (0.18)**	**9.74 (0.14)**	**7.20 (0.79)**
Accuracy difference	9.65 (0.16)	9.71 (0.16)	9.81 (0.14)	7.42 (1.11)
Precision difference	9.68 (0.12)	9.20 (0.29)	9.76 (0.21)	7.06 (0.92)
TPR difference	9.62 (0.24)	9.70 (0.12)	9.76 (0.29)	7.22 (1.79)
FPR difference	9.68 (0.12)	9.17 (0.30)	9.74 (0.22)	6.83 (1.23)
DemP difference	9.92 (0.04)	9.44 (0.13)	9.79 (0.14)	8.61 (1.23)
EOd difference	9.57 (0.19)	9.15 (0.28)	9.59 (0.26)	6.03 (1.23)
Robustness Score	**5.66 (0.17)**	**7.19 (0.41)**	**6.19 (0.27)**	**7.04 (0.48)**
Accuracy drop	8.75 (0.47)	9.39 (0.14)	8.94 (0.51)	9.85 (0.92)
ROC-AUC-Score drop	9.71 (0.12)	9.86 (0.01)	9.44 (0.14)	9.41 (0.43)
Brier Score Loss increase	9.01 (0.26)	9.55 (0.08)	9.13 (0.45)	9.44 (0.39)
Accuracy drop under FGSM	1.03 (0.61)	4.47 (0.44)	4.63 (1.07)	6.01 (0.33)
Accuracy drop under PGD	0.48 (0.05)	0.46 (0.03)	1.92 (0.06)	2.20 (0.24)
CLEVER-Score	0.72 (0.31)	9.24 (3.16)	0.28 (0.11)	3.95 (3.77)
Loss Sensitivity	9.93 (0.04)	9.90 (0.06)	8.97 (0.27)	9.81 (0.02)
Privacy Score	**6.26 (0.26)**	**5.12 (0.05)**	**6.61 (0.27)**	**6.34 (0.45)**
ROC-AUC-Score MembInf 1	9.74 (0.05)	10.0 (0.00)	9.80 (0.05)	8.75 (0.33)
ROC-AUC-Score MembInf 2	9.84 (0.11)	9.62 (0.05)	9.85 (0.14)	8.62 (0.27)
Accuracy ModExt 1	3.73 (1.01)	0.62 (0.20)	4.50 (0.50)	5.16 (1.30)
Accuracy ModExt 2	1.75 (0.29)	0.24 (0.03)	2.29 (0.78)	2.82 (0.33)
Explainability Score	**4.89 (0.10)**	**5.45 (0.31)**	**3.92 (0.18)**	**5.33 (0.21)**
Robustness	8.31 (0.37)	9.12 (0.11)	4.51 (0.77)	9.32 (0.12)
Faithfulness	0.17 (0.22)	2.32 (1.38)	0.19 (0.36)	1.90 (0.84)
Randomization	5.65 (0.21)	3.20 (0.23)	5.67 (0.08)	2.58 (0.11)
Complexity	5.35 (0.11)	7.15 (0.11)	5.33 (0.05)	7.52 (0.08)
Trustworthiness Score	**6.62 (0.07)**	**6.79 (0.17)**	**6.62 (0.11)**	**6.48 (0.30)**
Accuracy Score	**9.76 (0.02)**	**9.77 (0.01)**	**9.04 (0.03)**	**8.90 (0.11)**

4. Insights and Challenges of the Framework

Our findings demonstrate the inherent difficulty in achieving a singular, comprehensive metric for measuring trustworthiness. Although the TS serves as a valuable initial indicator, relying solely on it may obscure potential trade-offs between different dimensions, as the DS can counterbalance each other, leading to similar TS values. Furthermore, we highlight the necessity of considering prediction accuracy as a measure of general utility for a comprehensive model evaluation. Nonetheless, introducing an aggregated score for each dimension of trustworthiness considerably simplifies the assessment of these trustworthiness requirements. The DS condense a complex array of metrics into interpretable scores, enabling a rapid evaluation of an ML model's vulnerabilities. When comparing multiple models, e.g. trained on different datasets, these scores promptly

highlight variations in performance. Additionally, patterns can be observed across datasets, suggesting the impact of specific dataset characteristics, such as distribution and size, on particular dimensions. Furthermore, our framework allows for easy investigation of the interrelationships between dimensions, overall trustworthiness, and accuracy, enabling the identification of similarities and trade-offs among these aspects. The framework significantly reduces complexity while providing valuable insights into trustworthiness. However, during the development of the framework, several technical challenges emerged, which are discussed in the subsequent sections.

4.1. Metric Selection

One of the challenges in developing our framework is the selection of appropriate metrics for each dimension. Currently, our framework is based on 21 metrics that cover the four dimensions of fairness, robustness, privacy, and explainability. In the following, we provide insights into the interactions between the metrics and their perceived suitability for the framework.

Fairness. Our fairness assessment uses six metrics that capture different aspects of group fairness. As previous studies have explored (Kleinberg et al., 2016; Wick et al., 2019), these metrics, and overall accuracy, involve theoretical trade-offs. DemP aims to equalize positive classification probabilities regardless of the actual label, while EOd equalizes proportions based on the ground truths. Precision difference focuses on equalizing false positives, and TPR difference focuses on equalizing false negatives. The distribution of the training and test set significantly affects the extent of these trade-offs. In perfectly balanced settings, excellent results can be achieved for fairness metrics and overall accuracy, as most likely fulfilled by UTKFace models, while imbalanced training sets normally amplify trade-offs. Our results reflect this assumption by showing greater value variations and trade-offs for more imbalanced training sets.

Our current selection of several group fairness metrics provides an initial understanding of general fairness behavior and the impact of the dataset distribution. Since the values and relations of these metrics highly depend on the datasets, it is not possible to provide a universal recommendation for metric selection. Given the subjective nature of fairness, it is crucial to establish a primary definition of fairness that aligns with the specific use case and select metrics accordingly. By focusing on metrics that specifically reflect the fairness goals, unnecessary trade-offs can be avoided, leading to a more representative aggregated fairness score. Moreover, apart from group fairness metrics, approaches of individual fairness (Dwork et al., 2011), might be considered in future work.

Robustness. Our robustness assessment encompasses seven metrics that address distribution shift and adversarial attacks, including both attack-based and attack-independent metrics. In our experimental results, the metrics for distribution shift do only exhibit minor variations across datasets and thus have a minor impact on the variation of the robustness score compared to the adversarial metrics. Measuring not only the accuracy drop, but also ROC-AUC, and Brier Score difference under distribution shift illuminates different performance aspects, which was especially valuable for assessing UTKFace and FairFace models. A more diverse approach is also recommended in terms of the shift simulation. Exploring distribution shift beyond our used augmentations, e.g. by using other augmentation techniques or another testset, could provide a more fine-grained estimation of model robustness to real-world perturbations.

In terms of adversarial robustness, our findings reveal notable variations in scores. It is important to note that being robust against one type of adversarial attack does not guarantee robustness against other attacks. Therefore, we recommend conducting multiple attacks to gain a comprehensive understanding of the model's vulnerabilities, although they are highly computationally demanding. Apart from our selected FGSM and PGD, various other attack types exist, which could be considered for evaluating adversarial robustness (Brendel et al., 2019). Regarding loss sensitivity, we observed that a low impact of each data sample on the average loss is not necessarily indicative of the model's robustness against a specific attack in our case. Conversely, the CLEVER-Score shows a positive correlation with the attack metrics, although its results vary significantly across datasets, and can highly deviate within multiple runs. These variations may arise because the score is evaluated only on a small subset due to its high computational demands. Further expanding this subset could allow for more reliable conclusions to be drawn.

Privacy. In creating our evaluation framework, we use four primary metrics to gauge the performance of black-box membership inference and model extraction attacks under two distinct data scenarios. Black-box membership inference attacks are preferred for their computational efficiency in comparison to white-box attacks and their robustness in assessing privacy leaks. The two data scenarios reflect realistic threats: one where the attacker trains a model using a shadow model,

and another where the attacker accesses parts of the training dataset, like in a data breach. We minimize overfitting and overestimation of attack performance, as in common benchmarks (Y. Liu et al., 2022), by using early stopping in training. Model extraction attacks, evaluated under the same conditions, supplement our privacy assessment.

However, our chosen metrics might not cover all potential privacy vulnerabilities. During our research, we also attempted to implement model inversion attacks (Fredrikson et al., 2015). However, we were unable to successfully invert the models, leading to their exclusion from our framework. This situation underscores the complexities inherent in the current state of model inversion attacks, and points to the need for further research in this area. Furthermore, researchers could consider exploring additional assumptions regarding membership inference attacks (Ye et al., 2022), evaluating differential privacy measures (Dwork, 2008) or assessing the effectiveness of data anonymization techniques such as face de-identification (Gross et al., 2006).

Explainability. Our two-stage approach to quantifying explainability begins with generating saliency maps using the simple gradient method (Simonyan et al., 2013), due to its lesser computational demand compared to more intricate methods like GradCAM (Selvaraju et al., 2017) or Integrated Gradients (Sundararajan et al., 2017).

We adopt four metrics from the *quantus* framework, namely Complexity, Randomization, Faithfulness, and Robustness. This selection aims to cover diverse aspects of explanation while keeping computational efficiency in mind. Despite the Faithfulness metric demonstrating notable variance across runs, the other metrics show stability, indicating their potential reliability in capturing specific dimensions of explainability.

Overall, we find that explainability is the toughest dimension to quantify, due to its computational demands and the variation induced through the additional step of generating explanations. Future work could also consider alternative methods like counterfactual explanations (Sauer & Geiger, 2021) or surrogate models like LIME (Ribeiro et al., 2016).

4.2. Normalization

We normalize the metric values on a scale of 0 to 10 for meaningful comparisons. To carry out this normalization, a consensus on the minimum and maximum values for each metric is necessary. In theory, most metrics possess well-defined limits, often ranging from 0 to 1, as is the case with accuracy. However, empirical metric values often exhibit much narrower ranges. For example, in our study, the distribution shift metrics fall within the range of 0 to 0.1, while accuracy drops under attacks can span the whole range in between 0 and 1. By maintaining the theoretical limits for all metrics, those with higher variations in values may overshadow those with smaller variations in the aggregated scores. Consequently, we recommend adjusting the minimum and maximum values in accordance with the observable metric ranges in practical scenarios. However, such adjustments necessitate a thorough understanding of the behavior of each metric.

4.3. Aggregation

Regarding our scoring methodology, we employ equal weights when aggregating metric values to derive DS, and when combining DS to form the TS. This approach allows for an initial evaluation of an ML model's trustworthiness. However, it is important to acknowledge the trade-offs existing among metrics within each dimension, such as competing fairness definitions or privacy attacks, which could potentially outweigh one another. In order to highlight particularly significant risks in a specific use case, while keeping a broader metric selection, it may be advantageous to assign specific weights to selected metrics. Similarly, the TS can assign greater importance to certain dimensions if they are deemed especially critical for a particular ML application.

5. Limitations

Our proposed framework has the potential to enhance the understanding of the trustworthiness of ML models. However, certain limitations impede its applicability to other use cases. While our procedure of normalizing and aggregating metric values to comprehensive scores is universally applicable, the choice of metrics is closely tied to the specific use case and dependent on three key factors, namely the ML task, model architecture and data type.

Our framework can be applied to other image classification use cases without any modifications, provided that the underlying dataset has been labeled with binary target and sensitive attribute labels. In the event of multi-class settings, the calculation of fairness metrics needs to be adjusted. As our framework is founded on commonly used classification metrics, it may be necessary to adapt the metrics to specific tasks, such as object detection and image segmentation.

Regarding the model architecture, our framework is currently tailored to neural networks. When employing

other types of ML models, attack-based metrics for privacy and robustness may become irrelevant, and other explanation methods may be more suitable. Conversely, the fairness metrics are independent from the algorithm used.

Concerning the data type, group fairness metrics are generally applicable to all data types, with tabular data being particularly well-suited (Bird et al., 2020). Alternative ways to measure fairness might become relevant for NLP, e.g. gender-swapping (Bansal, 2022). Regarding robustness to distribution shifts, comparing performance on an original versus a shifted dataset can be applied to other fields, while the shift simulation might be adapted. Furthermore, the construction of adversarial examples is feasible for all data types and tasks, however text or tabular data might require specific pre-processing techniques due to their discrete features (W. E. Zhang et al., 2019). Similarly, privacy-targeting attacks can be applied to NLP after certain preprocessing, such as calculating word embeddings (Mahloujifar et al., 2021). Finally, the general methodology for evaluating explainability is suitable for all data types. However, different methods to generate the feature importance values might be required.

6. Conclusion

In this paper, we used the context of facial analysis to experimentally explore the design of a framework for evaluating the trustworthiness of ML models, that aggregates multiple quantitative metrics into comprehensive scores. Applying our framework to four different datasets indicated its utility in revealing the relationships between aspects of trustworthiness, and in examining the impact of dataset characteristics. We also discussed the primary technical challenges in developing the framework, which include the selection, normalization, and aggregation of metrics. Regarding the transferability of our framework, the procedure of calculating scores is universally applicable, while the effort of modifying the metric implementation mainly depends on the ML task, data type and model. Overall, our results serve as a foundation for future work focused on refining the framework, deepening the analysis of the behaviour of selected metrics and the influence of models and hyperparameters on trustworthiness. Finally, we aim to expand its scope to a broader range of applications.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Al-Rubaie, M., & Chang, J. M. (2019). Privacy-preserving machine learning: Threats and solutions. *IEEE Security & Privacy*, *17*(2), 49–58. https://doi.org/10.1109/MSEC.2018.2888775

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., & Lacoste-Julien, S. (2017). A closer look at memorization in deep networks. *Proceedings of the 34th International Conference on Machine Learning*, *70*, 233–242.

Balagopalan, A., Zhang, H., Hamidieh, K., Hartvigsen, T., Rudzicz, F., & Ghassemi, M. (2022). The road to explainability is paved with bias: Measuring the fairness of explanations. https://doi.org/10.1145/3531146.3533179

Bansal, R. (2022). A survey on bias and fairness in natural language processing. https://arxiv.org/pdf/2204.09591

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, *50*(1), 3–44. https://doi.org/10.1177/0049124118782533

Bhatt, U., Weller, A., & Moura, J. M. F. (2020). Evaluating and aggregating feature-based model explanations. https://arxiv.org/pdf/2005.00631

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in ai.

Brendel, W., Rauber, J., Kümmerer, M., Ustyuzhaninov, I., & Bethge, M. (2019). Accurate, reliable and fast robustness evaluation. *Advances in Neural Information Processing Systems*, *32*.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 77–91.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832. https://doi.org/10.3390/electronics8080832

Caton, S., & Haas, C. (2020). *Fairness in machine learning: A survey*. arXiv. https://doi.org/10.48550/arXiv.2010.04053

Chalasani, P., Chen, J., Chowdhury, A. R., Wu, X., & Jha, S. (2020). Concise explanations of neural networks using adversarial training. *Proceedings of the 37th International Conference on Machine Learning*, *119*, 1383–1391.

Chandrasekaran, V., Chaudhuri, K., Giacomelli, I., Jha, S., & Yan, S. (2020). Exploring connections between active learning and model extraction. *29th USENIX Security Symposium (USENIX Security 20)*, 1309–1326.

de Cristofaro, E. (2020). An overview of privacy in machine learning. https://arxiv.org/pdf/2005.08679

Dwork, C. (2008). Differential privacy: A survey of results. *International conference on theory and applications of models of computation*, 1–19.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness through awareness. https://arxiv.org/pdf/1104.3913

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1322–1333. https://doi.org/10.1145/2810103.2813677

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. https://doi.org/10.48550/arXiv.1412.6572

Gross, R., Sweeney, L., de La Torre, F., & Baker, S. (2006). Model-based face de-identification. *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, 161. https://doi.org/10.1109/CVPRW.2006.125

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, *29*.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2019). Natural adversarial examples.

Karkkainen, K., & Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1548–1558.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. https://arxiv.org/pdf/1609.05807

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy ai: From principles to practices. *ACM Computing Surveys*, *55*(9). https://doi.org/10.1145/3555803

Liu, X., Xie, L., Wang, Y., Zou, J., Xiong, J., Ying, Z., & Vasilakos, A. V. (2021). Privacy and security issues in deep learning: A survey. *IEEE Access*, *9*, 4566–4593. https://doi.org/10.1109/ACCESS.2020.3045078

Liu, Y., Wen, R., He, X., Salem, A., Zhang, Z., Backes, M., de Cristofaro, E., Fritz, M., & Zhang, Y. (2022). Ml-doctor: Holistic risk assessment of inference attacks against machine learning models. *31st USENIX Security Symposium (USENIX Security 22)*, 4525–4542.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. *Proceedings of the IEEE international conference on computer vision*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. https://doi.org/10.48550/arXiv.1706.06083

Mahloujifar, S., Inan, H. A., Chase, M., Ghosh, E., & Hasegawa, M. (2021). Membership inference on word embedding and beyond. https://arxiv.org/pdf/2106.11384

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning.

Sauer, A., & Geiger, A. (2021). Counterfactual generative networks.

Schrouff, J., Harris, N., Koyejo, O., Alabdulmohsin, I., Schnider, E., Opsahl-Ong, K., Brown, A., Roy, S., Mincu, D., Chen, C., Dieng, A., Liu, Y., Natarajan, V., Karthikesalingam, A., Heller, K., Chiappa, S., & D'Amour, A. (2022). Maintaining fairness across distribution shift: Do we have viable solutions for real-world applications? https://doi.org/10.48550/arXiv.2202.01034

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Shokri, R., Strobel, M., & Zick, Y. (2021). On the privacy risks of model explanations. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 231–241. https://doi.org/10.1145/3461702.3462533

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. *2017 IEEE symposium on security and privacy (SP)*, 3–18.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. https://arxiv.org/pdf/1312.6034

Sixt, L., Granz, M., & Landgraf, T. (2019). When explanations lie: Why many modified bp attributions fail. https://arxiv.org/pdf/1912.09818

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*, *70*, 3319–3328.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., & Schmidt, L. (2020). Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, *33*, 18583–18599.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction apis. *25th USENIX security symposium (USENIX Security 16)*, 601–618.

Tran, C., Dinh, M., & Fioretto, F. (2021). Differentially private empirical risk minimization under the fairness lens. *Advances in Neural Information Processing Systems*, *34*, 27555–27565.

Tran, D., Liu, J., Dusenberry, M. W., Du Phan, Collier, M., Ren, J., Han, K., Wang, Z., Mariet, Z., Hu, H., Band, N., Rudner, T. G. J., Singhal, K., Nado, Z., van Amersfoort, J., Kirsch, A., Jenatton, R., Thain, N., Yuan, H., … Lakshminarayanan, B. (2022). Plex: Towards reliability using pretrained large model extensions. https://arxiv.org/pdf/2207.07411

Weng, T.-W., Zhang, H., Chen, P.-Y., Yi, J., Su, D., Gao, Y., Hsieh, C.-J., & Daniel, L. (2018). Evaluating the robustness of neural networks: An extreme value theory approach. https://doi.org/10.48550/arXiv.1801.10578

Wick, M., Panda, S., & Tristan, J.-B. (2019). Unlocking fairness: A trade-off revisited. *Advances in neural information processing systems*, *32*.

Ye, J., Maddi, A., Murakonda, S. K., Bindschaedler, V., & Shokri, R. (2022). Enhanced membership inference attacks against machine learning models. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 3093–3106. https://doi.org/10.1145/3548606.3560675

Yeh, C.-K., Hsieh, C.-Y., Suggala, A. S., Inouye, D. I., & Ravikumar, P. (2019). On the (in)fidelity and sensitivity for explanations. https://arxiv.org/pdf/1901.09392

Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018). Privacy risk in machine learning: Analyzing the connection to overfitting. *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, 268–282. https://doi.org/10.1109/CSF.2018.00027

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2019). Adversarial attacks on deep learning models in natural language processing: A survey. https://arxiv.org/pdf/1901.06796

Zhang, Z., Song, Y., & Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.