VERIFAI - A Step Towards Evaluating the Responsibility of AI-Systems

Sabrina Göllner [1], Marina Tropmann-Frick [2]

Abstract: This work represents the first step towards a unified framework for evaluating an AI system's responsibility by building a prototype application. The python based web-application uses several libraries for testing the fairness, robustness, privacy, and explainability of a machine learning model as well as the dataset which was used for training the model. The workflow of the prototype is tested and described using images of a healthcare dataset since healthcare represents an area where automatic decisions affect human lives, and building responsible AI in this area is therefore indispensable.

Keywords: Artificial Intelligence; Responsible AI; Privacy-preserving AI; Explainable AI; Ethical AI; Trustworthy AI

1 Introduction

This paper is based on the Structured Literature Review 'Aspects and Views on Responsible AI' presented at the *LOD conference 2022* [LO22] and on the follow-up paper which is currently in the writing process. The aforementioned papers conclude from the current state of the art that *Responsible AI* encompasses the aspects of 'security, privacy, ethics, explainability, human-centeredness, and trust'. The trust aspects are also up to the user's perception and human-centeredness requires a human-in-the-loop setting and will be part of our future work. Our first goal is to verify the security, privacy, ethics, and explainability within an AI system through different metrics in a single framework. Therefore we have created 'VERIFAI' (e**V**aluating th**E** **R**esponsib**I**lity o**F** **AI**-systems), which is a first step towards putting this concept into practice. To the best of our current knowledge, there is no other framework that checks and evaluates multiple responsibility factors, so this is the novelty of the present work.

2 Implementation

This section is divided into two parts: the first part explains the selection of the data, and model architecture as well as the selection of toolkits for the evaluations and the second

[1] Hamburg University of Applied Sciences, Department of Computer Science, Berliner Tor 7, 20099 Hamburg, Germany sabrina.goellner@haw-hamburg.de
[2] marina.tropmann-frick@haw-hamburg.de

doi:10.18420/BTW2023-63

part consists of the presentation of the resulting web application based on an example walkthrough.

2.1 Dataset and model architecture

For the prototype implementation the healthcare dataset *HAM10000* [Ts18] was chosen because it satisfies two criteria: 1) it consists of dermatoscopic images from different populations including a representative collection of all important diagnostic categories in the realm of pigmented lesions and 2) because it consists not only of image data but also of metadata for the analysis.
The chosen model architecture for testing is *Xception* [Ch17], which is a network with a linear stack of depthwise separable convolution layers with residual connections. It achieved the best results on the dataset compared to other architectures.

2.1.1 Selection of toolkits

Since the project's goal was to verify the ethics, security, privacy, and explainability of both the data and model, the first step was to research state-of-the-art toolkits for testing. From the result of the libraries found, each of them could be classified into one of our four categories:

1. Evaluation of Explainability: Quantus [He22], IBM AI Explainability 360: [Ar19]
2. Evaluation of Ethics: Tensorflow Fairness Indicators [Te22], IBM AI Fairness 360 [Be18], Fairlearn [Bi20], Aequitas [Sa18], REVISE [Wa22], VISSL [Go21],
3. Evaluation of Security: IBM Adversarial Robustness 360 Toolkit [Ni18], Foolbox: [Ra20], Advbox [Go20], UnMask [Fr20]
4. Evaluation of Privacy: Privacy Meter [Sh22], IBM: differential privacy toolkit [Ho19], Tensorflow Privacy [ACP22]

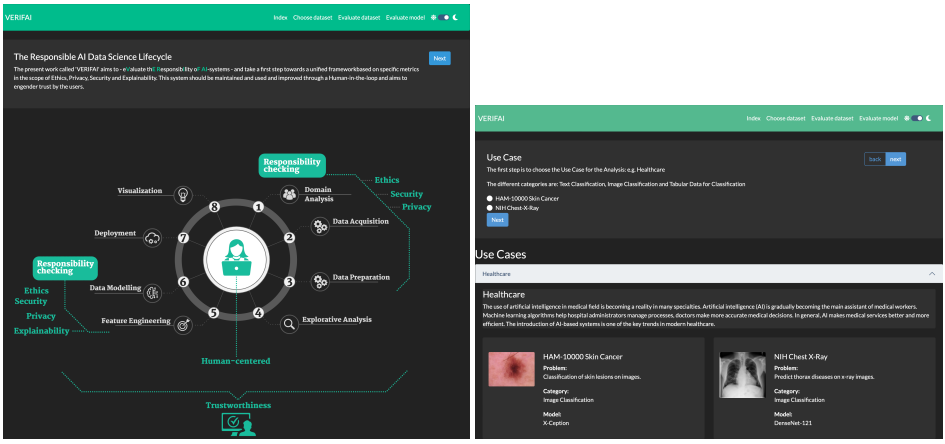
Based on the features, metrics, quality, and usage limitations of the analyzed toolkits and libraries we came up with the following decisions for the prototype:
For the ethics/fairness evaluation of the model, there was, unfortunately, no suitable library that could handle medical image data properly, so the results were calculated without the usage of a toolkit, but with self-written python functions. The well-documented *Tensorflow privacy* was chosen for the privacy verification. For the security verification, the robustness test was performed using *Foolbox* because the calculation is reliable and fast, and the set of metrics can be extended in the future with others from the same library. In terms of explainability toolboxes, the choice fell on the *Quantus* toolbox because it supports evaluations of all kinds of neural networks and provides many different metrics to test

against that can be used for comparison in the future. To counteract confusion, the terms 'robustness' will be used instead of security and 'fairness' instead of ethics hereafter, as these are also referred to as such in the evaluations in this context.

2.2 Exemplary Walkthrough

This section shows the exemplary program flow based on an example with the presented data set using screenshots of the results and explanations.

2.2.1 The responsible data science lifecycle and the selection of the use case



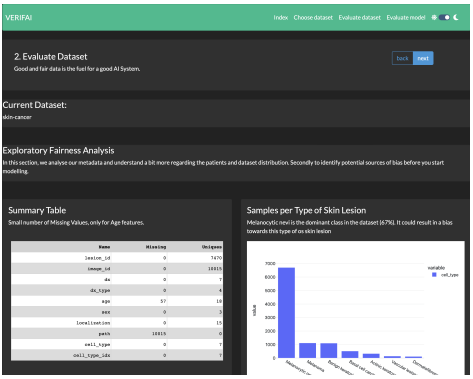
(a) Screenshot: index / responsible data science lifecycle (b) Screenshot: use cases

Fig. 1: Dataset

The index page, shown in figure 1a, is intended to introduce the systems' workflow to the user. It also offers a figure of the *Responsible Data Science Lifecycle*, which is the data science lifecycle extended through responsibility checks. At the top of the web page, there is an explanation of the current step, this continues throughout the application. Figure 1b is a screenshot of the step, where the user can choose from different use cases and corresponding data sets to run the evaluation on. In this case, we choose the HAM10000 image dataset with pigmented skin lesions which belongs to the healthcare use cases. By choosing the dataset we can go to the next step.

2.2.2 Exploratory fairness analysis of the dataset

In this step, we can analyze our dataset based on an exploratory fairness analysis.



(a) Screenshot: Evaluate Dataset (part 1)

(b) Screenshot: Evaluate Dataset (part 2)

Fig. 2: Screenshot: exploratory fairness analysis of the dataset

Figure 2a and 2b display the data exploration for detecting potential biases to the user. For example, we can see that class *melanocytic nevi* is the dominant one in the dataset (67%). Using the data for modeling could result in a bias towards this type of skin lesion. This analysis is to help users detect such biases and prevent bad modeling.

2.2.3 Evaluate Model

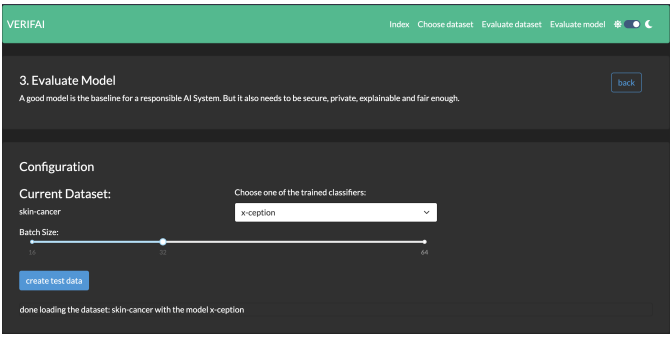


Fig. 3: Screenshot: load model and set configurations

In this step (see figure 3), the user can select the model and configuration for evaluation. The model in this example is an already trained *Xception* model. Choosing the batch size for the test dataset is also possible. The evaluations for robustness, fairness, privacy, and explainability are explained next.

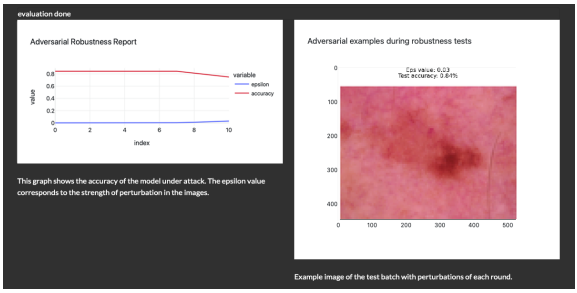


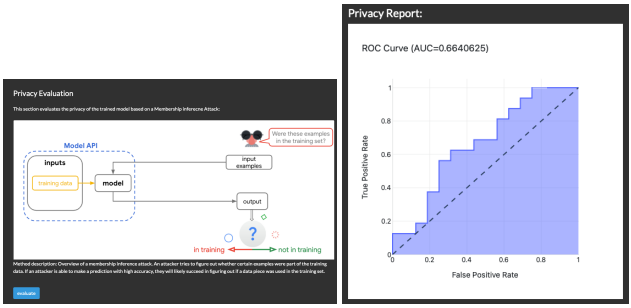
Fig. 4: Screenshot: evaluate model robustness (result)

Robustness Evaluation Figure 4 shows the results of testing if the model is robust against adversarial attacks with perturbated images, a metric which is called *adversarial robustness*. This is tested using the Projected Gradient Descent (PGD) attack. It attempts to find the perturbation that maximizes the loss of a model on a particular input while keeping the size of the perturbation smaller than a specified amount referred to as epsilon (eps) while max. eps = 0.3, as this is used for benchmarking in the *RobustBench* [Cr20]). Each round the *eps* is increasing and the robustness score is measured as shown in the plots: starting from 82%, the model has still an accuracy of *72%* in the last round, using a perturbation of *eps=0.3*, which is still a good accuracy. The image on the right is an example image from the test batch with the maximum perturbations (eps=0.3) added.

Fig. 5: Screenshot: evaluate fairness (result)

Fairness Evaluation Figure 5 shows the correctly and incorrectly classified images through a confusion matrix. We can see, that the tested model is very biased in the direction of the class *melanocytic nevi*. The reason for this result is probably because we already had an imbalanced dataset before. The second plot is the F1-score for the different classes, which is suitable for imbalanced data. Because we have a bias in the model the end results are not good enough for a good fairness score.

Privacy Evaluation Figure 6 shows the check of the privacy leakage through a *membership inference attack*, which tries to find out if specific examples were in the training set (see fig. 6a). The results show that the membership inference attack was successful but had only an



(a) Screenshot: membership inference explanation

(b) Screenshot: ROC curve of the membership inference attack

Fig. 6: Screenshots: Evaluate model privacy (results)

AUC of 66% (see fig. 6b). This means, that the privacy leakage was only satisfactory for the attacker, which is better for the model's privacy score, we can therefore determine that the leak of information was only moderate in this case.

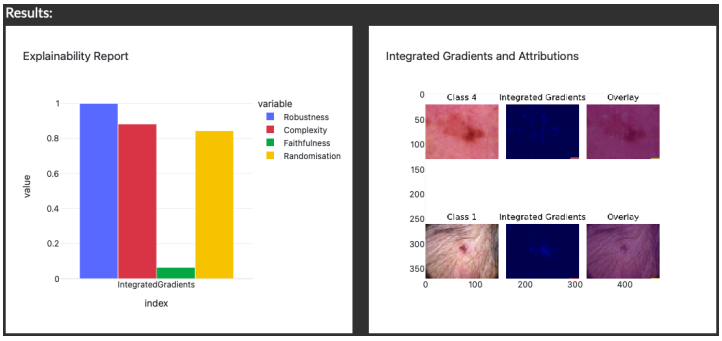


Fig. 7: Screenshot: Evaluate Model Explainability

Explainability Evaluation Figure 7 displays the quantitative evaluation of the model's explanation in combination with the chosen explainability method. We chose the suitable XAI-Method called *Integrated Gradients*. The used metrics are: *Robustness*, which measures the probability that the inputs with the same explanation have the same prediction label [Ye19], *Complexity* that measures if only highly attributed features are truly predictive of the model output [Ch18], *Faithfulness* which iteratively replaces a random subset of given attributions with a baseline value and then measures the correlation between the sum of this attribution subset and the difference in function output [BWM20], and *Randomisation*, which computes for the distance between the original explanation and the explanation for a random other class [SGL20]. In the barplot on the left, we can see that Robustness, Complexity, and Randomisation scored well with relatively high percentages, but Faithfulness did not. All four scores contribute with equal weighting to the final result and their scores were therefore

averaged. In the figure next to the bar chart, we can see also examples of images from the test batch, and the corresponding explanation. This is so that the user can also get an idea of whether the explanation is good enough or not.

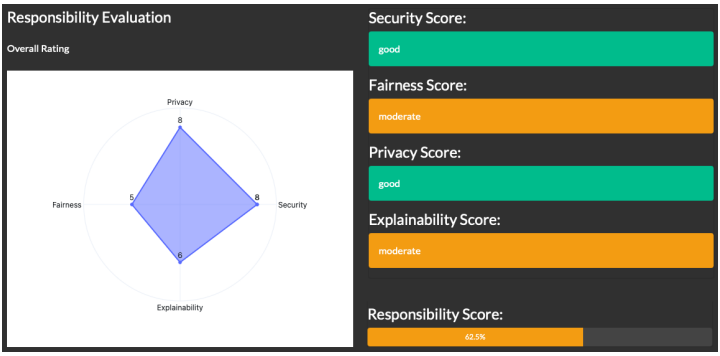


Fig. 8: Screenshot: Responsibility Evaluation

Responsibility Evaluation Finally, the *Responsibility Evaluation* in figure 8 summarizes the calculated scores using the proposed metrics and highlights them with different colors according to their scores. The rating was calculated as follows: A 'perfect' model would score full points in every aspect, which equals 10 points. In our test case, the security evaluation was tested with a good result (8/10) as well as privacy (6/10), while the other metrics fairness and explainability (6/10) show still some weaknesses and achieved therefore moderate scores. The worst result was the fairness score of the model (5/10) because of the bias. Thus, our model achieved a final score of 62.5% (25/40) with the metrics currently implemented.

3 Open challenges and future work

In this work, we created a prototype implementation of VERIFAI, an application for evaluating an AI system's responsibility based on several aspects. In the present prototype, we used a healthcare dataset. The tool can evaluate the dataset for fairness and a trained machine learning model for fairness, privacy leakage, adversarial robustness, and explainability using a variety of state-of-the-art metrics. Even though this work only covers a limited number of metrics so far, it is a good basis for future work. The following extensions are planned for future work: We will add more data sets belonging to different suitable scenarios, different machine learning models for each scenario, extend the set of metrics for each category, choose between selectable or auto-selection of the right metrics for the given problem, selectable target user, selectable focus for which aspect is most important for the target user, the tolerance level for each aspect, suggestions for mitigations, evaluation of trustworthiness and human-in-the-loop aspects. We are also working on making VERIFAI as transparent as possible for the users for helping to create more responsible AI systems.

Bibliography

[ACP22] Andrew, Galen; Chien, Steve; Papernot, Nicolas: , Tensor Flow Privacy. https://github.com/tensorflow/privacy, 2022.

[Ar19] Arya, Vijay; Bellamy, Rachel KE; Chen, Pin-Yu; Dhurandhar, Amit; Hind, Michael; Hoffman, Samuel C; Houde, Stephanie; Liao, Q Vera; Luss, Ronny; Mojsilović, Aleksandra et al.: One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv preprint arXiv:1909.03012, 2019.

[Be18] Bellamy, Rachel K. E.; Dey, Kuntal; Hind, Michael; Hoffman, Samuel C.; Houde, Stephanie; Kannan, Kalapriya; Lohia, Pranay; Martino, Jacquelyn; Mehta, Sameep; Mojsilovic, Aleksandra; Nagar, Seema; Ramamurthy, Karthikeyan Natesan; Richards, John; Saha, Diptikalyan; Sattigeri, Prasanna; Singh, Moninder; Varshney, Kush R.; Zhang, Yunfeng: AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. October 2018.

[Bi20] Bird, Sarah; Dudík, Miro; Edgar, Richard; Horn, Brandon; Lutz, Roman; Milan, Vanessa; Sameki, Mehrnoosh; Wallach, Hanna; Walker, Kathleen: Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May 2020.

[BWM20] Bhatt, Umang; Weller, Adrian; Moura, José MF: Evaluating and aggregating feature-based model explanations. arXiv preprint arXiv:2005.00631, 2020.

[Ch17] Chollet, Francois: Xception: Deep Learning With Depthwise Separable Convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). July 2017.

[Ch18] Chalasani, P; Chen, J; Chowdhury, AR; Jha, S; Wu, X: Concise explanations of neural networks using adversarial training. arXiv arXiv–1810. arXiv preprint arXiv:1810.06583, 2018.

[Cr20] Croce, Francesco; Andriushchenko, Maksym; Sehwag, Vikash; Debenedetti, Edoardo; Flammarion, Nicolas; Chiang, Mung; Mittal, Prateek; Hein, Matthias: Robustbench: a standardized adversarial robustness benchmark. arXiv preprint arXiv:2010.09670, 2020.

[Fr20] Freitas, Scott; Chen, Shang-Tse; Wang, Zijie J; Chau, Duen Horng: Unmask: Adversarial detection and defense through robust feature alignment. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE, pp. 1081–1088, 2020.

[Go20] Goodman, Dou; Xin, Hao; Yang, Wang; Yuesheng, Wu; Junfeng, Xiong; Huan, Zhang: Advbox: a toolbox to generate adversarial examples that fool neural networks. 2020.

[Go21] Goyal, Priya; Duval, Quentin; Reizenstein, Jeremy; Leavitt, Matthew; Xu, Min; Lefaudeux, Benjamin; Singh, Mannat; Reis, Vinicius; Caron, Mathilde; Bojanowski, Piotr; Joulin, Armand; Misra, Ishan: , VISSL. https://github.com/facebookresearch/vissl, 2021.

[He22] Hedström, Anna; Weber, Leander; Bareeva, Dilyara; Motzkus, Franz; Samek, Wojciech; Lapuschkin, Sebastian; Höhne, Marina M.-C.: Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations. 2022.

[Ho19] Holohan, Naoise; Braghin, Stefano; Mac Aonghusa, Pól; Levacher, Killian: Diffprivlib: the IBM differential privacy library. ArXiv e-prints, 1907.02444 [cs.CR], July 2019.

[LO22] LOD Conference: . https://lod2022.icas.cc/program/, 2022.

[Ni18] Nicolae, Maria-Irina; Sinn, Mathieu; Tran, Minh Ngoc; Buesser, Beat; Rawat, Ambrish; Wistuba, Martin; Zantedeschi, Valentina; Baracaldo, Nathalie; Chen, Bryant; Ludwig, Heiko et al.: Adversarial Robustness Toolbox v1. 0.0. arXiv preprint arXiv:1807.01069, 2018.

[Ra20] Rauber, Jonas; Zimmermann, Roland; Bethge, Matthias; Brendel, Wieland: Foolbox Native: Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX. Journal of Open Source Software, 5(53):2607, 2020.

[Sa18] Saleiro, Pedro; Kuester, Benedict; Stevens, Abby; Anisfeld, Ari; Hinkson, Loren; London, Jesse; Ghani, Rayid: Aequitas: A Bias and Fairness Audit Toolkit. arXiv preprint arXiv:1811.05577, 2018.

[SGL20] Sixt, Leon; Granz, Maximilian; Landgraf, Tim: When explanations lie: Why many modified bp attributions fail. In: International Conference on Machine Learning. PMLR, pp. 9046–9057, 2020.

[Sh22] Shokri, Reza: ML Privacy Meter: A Tool to Quantify Information Leakage through Machine Learning Models. 2022.

[Te22] Tensorflow: , Tensor Flow Fairness Indicators. `https://github.com/tensorflow/fairness-indicators`, 2022.

[Ts18] Tschandl, Philipp: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. 2018.

[Wa22] Wang, Angelina; Liu, Alexander; Zhang, Ryan; Kleiman, Anat; Kim, Leslie; Zhao, Dora; Shirai, Iroha; Narayanan, Arvind; Russakovsky, Olga: REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. International Journal of Computer Vision (IJCV), 2022.

[Ye19] Yeh, Chih-Kuan; Hsieh, Cheng-Yu; Suggala, Arun; Inouye, David I; Ravikumar, Pradeep K: On the (in) fidelity and sensitivity of explanations. Advances in Neural Information Processing Systems, 32, 2019.