Appendix C: Use case HC1 detailed findings

C.1 Fair Model Methodology

C.1.1 Use Case Summary

The AEQUITAS use case of AI Assisted Identification of Dermatological Disease for Diversity and Inclusion in Dermatology in Pediatric Patients was evaluated for this workshop. The AI system aims to detect and identify dermatological diseases in children, whilst reducing bias in the model and supporting fair and accurate diagnosis. The AI system is based on a transformer model which has been developed using ~300 images of pediatric dermatological diseases between 2010 and 2020.

C.1.2 Attendees

- Technical Expert from UNIBO (University of Bologna)
- Social Sciences Expert from ThinkTankPeriod
- Social Sciences Expert from ADECCO
- Sociotechnical Expert from Phillips
- Socio-technical Expert from UMU (Umeå University)
- Technical Expert from UMU (Umeå University)
- Technical Expert from UNIBO (University of Bologna)
- Moderator from ALLAI
- Moderator from ALLAI

C.1.3 Methodology

The Fair Model Methodology (FMM) is grounded in the EU AI Act's high-risk requirements. The FMM assists development decision-makers (DMM) during the development stage of the AI system to select and/or design a fair algorithm and model. The primary purpose is to ensure that technical, ethical, legal, and social constraints of fairness are adequately addressed through the development process. As such, for a model or an algorithm to be deemed fair it must successfully complete the following assessments:

- Code Audits: assessment of the design steps, algorithmic training, and risk management tools adopted to mitigate the risk of unfairness.
- Process Audits: evaluating the extent fairness principles and values are included in the process shaping the design, development, and use of the AI system. This audit entails:
- Risk evaluation of Fundamental Rights
- Risk evaluation of social/societal impacts

Funded by the European Union

C.1.4 Workshop Summary

Prior to the workshop, participants received access to the FMM workshop template which included details about the necessary sections and questions to complete the audit. Participants also received information about the AI use case, including a brief explanation of the AI with key details of the system's description, intended functionality, its specific aims and description of the dataset involved.

The workshop was held via WEBEX and began with a brief introduction. During the session the moderator explained the goals and objectives of the workshop and provided a summary of the methodology at hand. The workshop took a plenary format with the moderator facilitating an open discussion whilst guiding the attendees through the workshop. After the introduction, the moderator presented the AI use case and proceeded to conduct the Code Audits assessment of the model's requirements. Each question was posed and then discussed by participants to evaluate whether the AI system's model met the necessary criteria to be considered fair. During the workshop attendees provided feedback on the quality of the methodology to ensure its validity. Since this workshop was the first for the use-case, the methodological process is still ongoing. Indeed, this workshop partially completed the Code Audits assessment, which will be continued in future sessions. The filled-out workshop template can be seen in the Annex of this report.

C.1.5 Results

The methodology outlined in the FFM template was used to evaluate whether the AI model is fair in its development stage. Due time constraints, the template was partially completed during the workshop. The results cover a part of Code Audits' section 'Model Requirements' up until sub-section 5 'Validation Details'. The rest of the template still needs to be validated and reviewed by technical experts and will be updated accordingly.

C.1.5.1 Part 1 - Section 1: Model Requirements

Model Description

Developed by UNIBO for AEQUITAS project funded by Horizon Europe Project AEQUITAS. Model type & Language is a Swin Transformer developed with Python using dermatological images of skin disease. Documentation language was English. AI model Code is open and was sourced from Huggingface.com.

Model Characteristics

Intended purpose & context

The intended purpose is skin disease predication in pediatric patients; however, the model itself will not be deployed but rather illustrates real-world bias in a model trained on light skin. With final application, some hospitals might have these models deployed, and with proper safeguards, specifications, and documentations in place the model is appropriate enough to be used. In terms of accuracy, the model presents errors, however it also displayed higher level of accuracy than human doctors. AI model errors can differ

Funded by
the European Union

due to other aspects of the model architecture; this emphasizes the criticality of correct model choice.

Model assumptions

Splitting an image into small pieces is a valid way of analyzing the case to correlate different small parts with each other. The effects of damaged skin and assumptions related to user information need to be considered. In particular, it is important to be clear on what kind of information doctors may need to understand or predict the progression of diseased skin.

Model architecture & parameters

Swin transformer with 26 million parameters and 100 hidden layers. UMU and UNIBO technical experts noted here that listing all parameters and coefficients would be too specific. To maintain relevance, it was instead proposed that such a question should prompt technical experts to explain the global choices they have made and why they made them. In particular, an UMU technical expert remarked that it would be valuable to conduct a risk analysis at this stage and identify what parameters and choices are important understanding and conveying the model risks.

Inputs & outputs

The inputs in this model were non-dermoscopic images in the medical domain taken with a phone camera. A technical expert from UMU noted that the question should also prompt technical experts to explain assumptions regarding the inputs, what requirements the model places on them and the relevant constraints. In this use case, the constraints around the input were that the lighting and patient position in the images was inconsistent.

The outputs were identified to be the classifications of various skin diseases. Again, a technical expert from UMU noted that it would be valuable to prompt developers to explain why such an output was chosen and motivate the design decisions for presenting the output in such a manner. It was also noted that technical limitations that may have led designers to choose such an output should also be listed. Furthermore, a new sub question was also proposed, requiring designers to answer how they considered the user when designing the output and how do they intend the output to be interpreted and used.

Role & Context

A predictive AI model mainly trained on lighter skin, potential use of generated synthetic data of darker skin tones. Integration of workflows to be considered as the model questions might give different answers.

Trustworthiness Considerations

Originally, the trustworthiness considerations were separated out by model role (e.g. predictive, prescriptive, descriptive), a technical expert from UMU noted that the listed trustworthiness concerns could potentially be present in all models, regardless of the

Funded by
the European Union

model's role. In this use case, two trustworthiness considerations were found to be pressing, namely:

- *Automation bias:* It was noted that while this is just a model, an automated system that is intended for skin disease prediction should be designed in such a way that it gives disclaimers, lists limitations and stresses that the outputs are only suggestions and that human oversight must be incorporated. Human oversight and input can also be incorporated during the design stage, it was noted that once the system has been trained, doctors should be asked to validate the outputs, turn feedback into rules, inject these rules into the generator to teach/constrain to generate according to some rules.
- *Integration into workflows*: This question required stakeholders to review at which point of the workflow the model will be integrated. They were prompted to consider whether the model will be used at an essential decision-making point, how it might affect workplace flows/relations. This model was likely to be embedded in a larger system that would be integrated into the workflow alongside humans. Participants noted here that it would be important to elucidate the workflows and the organizations flows.

C.1.5.2 Model Architecture

This sub-section was discussed and it was decided that because it shows more relevance at the system rather than model level, it was postponed for future assessment.

C.1.5.3 Training Details

Training Data used

Dataset arrangement: 60% training, 20% testing,20% validation.

A pre-processing step of anonymization, cropping at different patches, deletion of certain identifiers and removal of extremely blurry and poorly lit images was completed. Additionally, detection of diseases was configured via masks to identify white as disease and black as rest.

Mitigated risks in the pre-processing stage

Dataset was found to be unbalanced; thus, it was augmented to account for darker skin tones.

Training Hyperparameters & procedure

Supervised AI model with an on-going implementation of a loop coming back with additional knowledge provided by doctors.

Trustworthiness considerations

Once again, originally the trustworthiness considerations were separated out by model type (e.g. supervised, unsupervised, etc.). Here too, a technical expert from UMU noted

Funded by the European Union

that the listed trustworthiness concerns could potentially be present in all models. Thus, all trustworthiness considerations were evaluated for this use case, namely:

- *Construct gap:* Mask was created first to highlight part of a picture, then a doctor labels the disease that was highlighted. Negative examples included except for tattoos, birthmarks, blotches for anonymization purposes. It cannot be filtered out as it might lead to the model misidentifying non-diseased skin as diseased.
- *Labelling bias:* Expertise bias and clerical error can lead to incorrect labels; thus, it is assumed that labels are equally incorrect to doctor errors.
- *Ground truth error rate:* A final validation with doctors to identify bias. Bias is not always a result of discrimination towards minority, but circumstances such as geographic distribution of patients.
- *Transparency:* Currently not in place as explanation of results is needed.
- *Domain shifts:* Important to consider textural items on skin such as tattoos and birthmarks. The model demonstrated a higher accuracy rate for all skin colors when pigment generated images were added to the training dataset. Slightly lower accuracy rates for lighter skin tones. Finally, synthetic data shows promise, thus potentially removing issues related to privacy.

Internal benchmarks

The common quality benchmarks for accuracy were used. Regarding fairness, benchmarks were selected as the ones that are used in skin-disease AI applications, based on a literature, used the same domain. Here there was a larger discussion around benchmarking, with technical experts from UNIBO noting that it's better to keep the benchmarking question more open rather than narrowly specific, as benchmarking is usually domain dependent and the benchmarks chosen vary across use cases.

Biases identified during training

Towards darker skin tones, biases were found in the data not the process.

Model overfitting/underfitting & model shifts

Requires a technical angle and perspective, can be technique specific. Concept drift changes in training data overtime, input does not mirror training data.

C.1.6 Areas of Improvement

The methodology was a suitable medium to address and assess AI model fairness. Questions remained mostly relevant and interrelated cohesively with one another, which provided a clearer view of the EU AI Act's criteria. However, areas of improvement were highlighted, consisting of the clarity, relevance, and reformulation of certain methodology questions. Each component of improvement is interconnected but will be addressed separately to emphasize their unique impact.

First, certain questions were indicated to lack clarity in terms explanation and intention behind the question. Furthermore, the context was flagged as needing to be elucidated

Funded by the European Union

to enhance better understanding of the methodology. Additionally, clarity is needed for technical experts to better understand and answer the questions. For example, in the question regarding the language of the AI model, language can mean the coding language, the natural language used within the AI model or documentation language.

Second, some questions were indicated as lacking relevance to the broader purpose they are there to serve. Feedback from the consortium highlighted the importance of understanding why these questions are being answered, thus a justification needs to be incorporated. For example, the question regarding Model Parameters requested that model weights and coefficients be listed. However, in this use case there were 26 million parameters, rendering this information largely unreadable. It was instead suggested that the question be reformulated to specify the most 'relevant' parameters in terms of the model's purpose, that could give deployers a better indication of how the model works. Similar logic should be applied to the input and output question, requesting developers to reason and motivate why they were chosen.

Lastly, it was noted to space out specific questions to ensure different components are addressed. On par, trustworthiness considerations were communicated to be all relevant for all models, rather than specific consideration being allocated to a certain AI model (i.e., supervised, unsupervised & reinforcement).

In terms of general feedback, it was suggested to engage doctors in the workshops which would be domain experts and AI users.

Funded by
the European Union