



AI+ Foundation™

Certification



AI + Foundation

Module 4

Hands – On 1:

Title: Getting Started with Aequitas – Fair by Design

Problem Statement:

Machine learning models often exhibit biases due to underrepresentation or skewed data, leading to unfair outcomes for certain demographic groups. For instance, medical imaging datasets may disproportionately represent lighter skin tones, causing models to perform poorly for darker skin tones. This guide addresses the need to audit and mitigate bias in AI systems using Aequitas, ensuring fairness in predictions.

Objectives:

1. Audit fairness in machine learning models using Aequitas.
2. Identify and mitigate bias at both the dataset and algorithmic levels.
3. Quantify fairness using metrics like Disparate Impact (DI) and Statistical Parity Difference (SPD).
4. Generate a de-biased dataset through synthetic data augmentation.
5. Evaluate the robustness of the model under edge cases.

Steps in Precise Manner:

1. Access the Aequitas Experimentation Environment.
2. Choose the dataset type (Image or Tabular).
3. Select or upload an image dataset.
4. Confirm dataset details and identify sensitive/target features.
5. Define proxy variables for sensitive features.
6. Select fairness metrics (e.g., DI, SPD).
7. Choose a data mitigation technique (e.g., Stable Diffusion-based Data Augmentation).

8. Configure augmentation settings (Batch Size, Epochs).
9. Run data mitigation to generate synthetic data.
10. View mitigation summary to confirm bias reduction.
11. Proceed to model mitigation or stress testing.

Tools Used:

- Aequitas Platform: Open-access tool for auditing fairness in AI models.
- Fairness Metrics: Disparate Impact (DI), Statistical Parity Difference (SPD).
- Data Augmentation: Stable Diffusion-based synthetic image generation.
- Dataset: Skin Disease Dataset (or custom dataset).

Steps in Detailed Manner:

Step 1: Access the Aequitas Experimentation Environment

1. Open your web browser (e.g., Chrome, Firefox).
2. Navigate to the Aequitas documentation page:

<https://aequitas-home.readthedocs.io/en/latest/fair-by-design.html#experimentation-environment>

3. Wait for the homepage to load.

AEQUITAS
0.1.1

- Context and background
- Glossary
- Framework Components
- Innovative Techniques for AI Fairness
- Use Cases
- Pills & Tutorials
- START EXPERIMENTING
- Fair-by-Design Methodology
- Experimentation Environment

START EXPERIMENTING

Fair-by-Design Methodology

PHASE	SCOPING	RISK ANALYSIS	DEVELOPMENT	EVALUATION	DEPLOYMENT & MONITORING	RE-EVALUATION
Meta-Methodology	Multi-Stakeholder Approach to AI Fairness-by-Design (MAP)					
Sub-Methodologies	Trustworthy AI Readiness Assessment (TAIRA)			Repeat TAIRA		
		Prohibited Social Scoring Assessment (PSSA)		Repeat PSSA		
		Fundamental Rights Impact Assessment for Fairness (FRIA-F)		Repeat FRIA-F		
		Fair Data Collection, Governance and Management (FDCGM)		Repeat FDCGM		
		Fair Model Methodology (FMM)		Repeat FMM		

- This is the main landing page of the Aequitas platform.
- Key sections:
 - **START EXPERIMENTING:** Entry point for new fairness audits.
 - **Fair-by-Design Methodology:** Ethical AI framework.
 - **Experimentation Environment:** Area for bias analysis.

- Context and background
- Glossary
- Framework Components
- Innovative Techniques for AI Fairness
- Use Cases
- Pills & Tutorials
- START EXPERIMENTING
- Fair-by-Design Methodology
- Experimentation Environment

Fair-by-Design Methodology

PHASE	SCOPING	RISK ANALYSIS	DEVELOPMENT	EVALUATION	DEPLOYMENT & MONITORING	RE-EVALUATION
Meta-Methodology	Multi-Stakeholder Approach to AI Fairness-by-Design (MAP)					
Sub-Methodologies	Trustworthy AI Readiness Assessment (TAIRA)			Repeat TAIRA		
		Prohibited Social Scoring Assessment (PSSA)		Repeat PSSA		
		Fundamental Rights Impact Assessment for Fairness (FRIA-F)		Repeat FRIA-F		
		Fair Data Collection, Governance and Management (FDCGM)		Repeat FDCGM		
		Fair Model Methodology (FMM)		Repeat FMM		

Context and background
Glossary
Framework Components
Innovative Techniques for AI Fairness
Use Cases
Pills & Tutorials
START EXPERIMENTING
Fair-by-Design Methodology
Experimentation Environment

```

# No user data
ethicalads!
import dev
region: global
type: image

```

Monetize your audience: Fund an OSS project or website with EthicalAds.

a privacy-first ad network

Ad by EthicalAds

Sub-Methodologies

Prohibited Social Scoring Assessment (PSSA)
Fundamental Rights Impact Assessment for Fairness (FRIA-F)
Fair Data Collection, Governance and Management (FDCGM)
Fair Model Methodology (FMM)
Fair Outcome Interpretation Methodology (FOIM)

Repeat PSSA
Repeat FRIA-F
Evaluate against FDCGM
Evaluate against FMM
Evaluate against FOIM

Experimentation Environment

Note

Access to the Experimentation Environment

See a Demo

Previous

latest

- Ads from EthicalAds indicate this is a public, open-access tool.

Context and background
Glossary
Framework Components
Innovative Techniques for AI Fairness
Use Cases
Pills & Tutorials
START EXPERIMENTING
Fair-by-Design Methodology
Experimentation Environment

```

# No user data
ethicalads!
import dev
region: global
type: image

```

Monetize your audience: Fund an OSS project or website with EthicalAds.

Sub-Methodologies

Fundamental Rights Impact Assessment for Fairness (FRIA-F)
Fair Data Collection, Governance and Management (FDCGM)
Fair Model Methodology (FMM)
Fair Outcome Interpretation Methodology (FOIM)

Repeat PSSA
Evaluate against FDCGM
Evaluate against FMM
Evaluate against FOIM

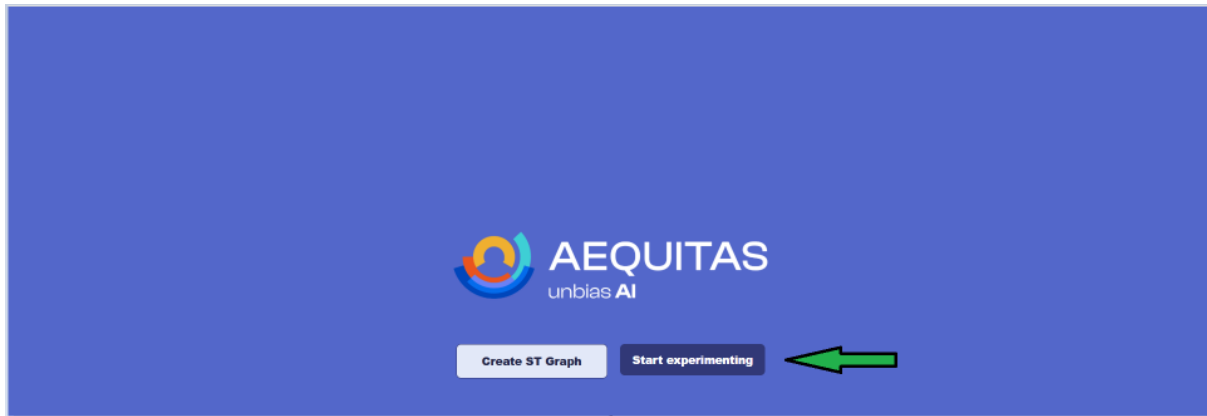
Experimentation Environment

Note

Access to the Experimentation Environment

See a Demo

Previous



✅ Step 2: Choose Dataset Type

1. Click on "**Experimentation Environment**".
2. On the next screen, select:

"Which type of dataset are you going to use? "

- ☐ Image Dataset
- ☐ Tabular Dataset

The image shows the "Dataset Type Selection" screen. It has a blue header with the text "Choose the dataset type you want to use in next steps." and a "Feedback" button. The main content area is white and contains the question "Which type of dataset are you going to use?" followed by two radio button options: "Image Dataset" and "Tabular Dataset". At the bottom right, there is a "Continue" button. The footer is dark blue and contains the AEQUITAS unbias AI logo, the text "Backend URL: http://aequitas.apice.unibo.it:4005", and two buttons: "Reset all" and "Create Report".

- This step sets the data modality for analysis.
- The system adjusts its tools depending on whether you're using images or tables.

Dataset Type Selection

Choose the dataset type you want to use in next steps.

Which type of dataset are you going to use?

☒ Image Dataset

☐ Tabular Dataset

Feedback

Continue

3. Select Image Dataset and click Continue.

Dataset Type Selection

Image Dataset Selection

Choose or provide a dataset which will be subject to the fairness process.

Choose a dataset or load your own.

Available datasets

☐ Skin Disease Dataset

☐ Custom

Feedback

Continue

Backend URL: <http://aequitas.apice.unibo.it:4005> [Reset all](#) [Create Report](#)

4. Now, select the “**Skin Disease dataset**” or if you wish to upload your own dataset than select “**Custom**” option.

✅ Step 3: Select an Image Dataset

1. After choosing "Image Dataset", click: “**Available datasets**”

Dataset Type Selection

Choose or provide a dataset which will be subject to the fairness process. Feedback

Image Dataset Selection

Choose a dataset or load your own.
Available datasets

☐ Skin Disease Dataset

☐ Custom

Continue

2. Choose:

- ☒ **Skin Disease Dataset**
- OR upload your own via **Custom**

☒ **Step 4: Confirm Dataset and Select Features**

Dataset Type Selection

Choose or provide a dataset which will be subject to the fairness process. Feedback

Image Dataset Selection

Choose a dataset or load your own.
Available datasets

☒ **Skin Disease Dataset**

☐ Custom

Skin Disease Dataset

A dataset used for skin disease classification, containing images of various skin conditions to assist in diagnosis.

Id	skin-disease
Size	773
Rows	49269
Columns	3
Created at	12 Nov 2024

Continue

Backend URL: <http://aequitas.apice.unibo.it:4005> Reset all Create Report

- In the above fig, we can see the dataset details.

Dataset Type Selection

Choose or provide a dataset which will be subject to the fairness process.
Feedback

Image Dataset Selection

Choose a dataset or load your own.
Available datasets

☒ Skin Disease Dataset

☐ Custom

Skin Disease Dataset

A dataset used for skin disease classification, containing images of various skin conditions to assist in diagnosis.

Id

skin-disease

Size

773

Rows

49269

Columns

3

Created at

12 Nov 2024

Continue

1. Review the dataset preview:

- Rows: ~49,268
- Columns: 3
- Created: 12 Nov 2024

2. Click Continue to confirm successful dataset loading.

3. Dataset metadata includes:




- URIs: Links to image files.
- skin_color: Intermediate, Brown, Dark, etc.
- disease: Chickenpox, Urticaria, etc.

Dataset Type Selection

Do you want to proceed with the selected dataset?
Feedback

Image Dataset Selection

Dataset Confirmation

URIs	skin_color	disease
	tan	Iatrogenic Drug Induced Exanth...
	brown	Iatrogenic Drug Induced Exanth...
	tan	Iatrogenic Drug Induced Exanth...

Continue

- This screen confirms the dataset has been loaded successfully.

- The Skin Disease Dataset contains dermatological images labeled by condition and patient skin tone.

✅ Step 5: Identify Sensitive and Target Features

1. Identify features:

- ✅ **skin_color**: Mark as Sensitive Feature.
- ✅ **disease**: Mark as Target Feature (to predict).

2. Click Continue.



Which are the sensitive and output features? Select features on the left. Columns display all statistical information about the features. On the right, choose the sensitive features and a target feature to predict. Features potentially sensitive based on their names are preselected automatically but you can deselect them.

	feature	missing_values	unique	top	freq	values	distribution	sensitive	target
<input checked="" type="checkbox"/>	URIs	0	49268	https://dvcs.apice.unibo.it/ae...	1	https://dvcs.apice.unibo.it/ae...		<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	skin_color	0	6	Intermediate	15654	brown, dark, intermediate, lig...		<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	disease	0	9	Urticaria	12916	Chickenpox, Iatrogenic Drug In...		<input type="checkbox"/>	<input type="checkbox"/>

Total features selected: 3 Sensitive features selected: 0 Target feature: None

Continue

- Metadata includes:
 - **URIs**: Links to image files
 - **skin_color**: Intermediate, brown, dark, etc.
 - **disease**: Chickenpox, Urticaria, etc.
- In this example, it's focused on medical image data (skin disease classification).
- This dataset may have bias risks due to underrepresentation of darker skin tones in medical imaging.

Which are the sensitive and output features? Select features on the left. Columns display all statistical information about the features. On the right, choose the sensitive features and a target feature to predict. Features potentially sensitive based on their names are preselected automatically but you can deselect them.

	feature	missing_values	unique	top	freq	values	distribution	sensitive	target
<input checked="" type="checkbox"/>	URIs	0	49268	https://dvcx.apice.unibo.it/ae...	1	https://dvcx.apice.unibo.it/ae...		<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	skin_color	0	6	intermediate	15654	brown, dark, intermediate, lig...		<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	disease	0	9	Urticaria	12916	Chickenpox, Iatrogenic Drug In...		<input type="checkbox"/>	<input checked="" type="checkbox"/>

Total features selected: 3 Sensitive features selected: 1 Target feature: disease

Continue

1. Select:

- ☒ skin_color → mark as Sensitive Feature
- ☒ disease → mark as Target Feature (to predict)

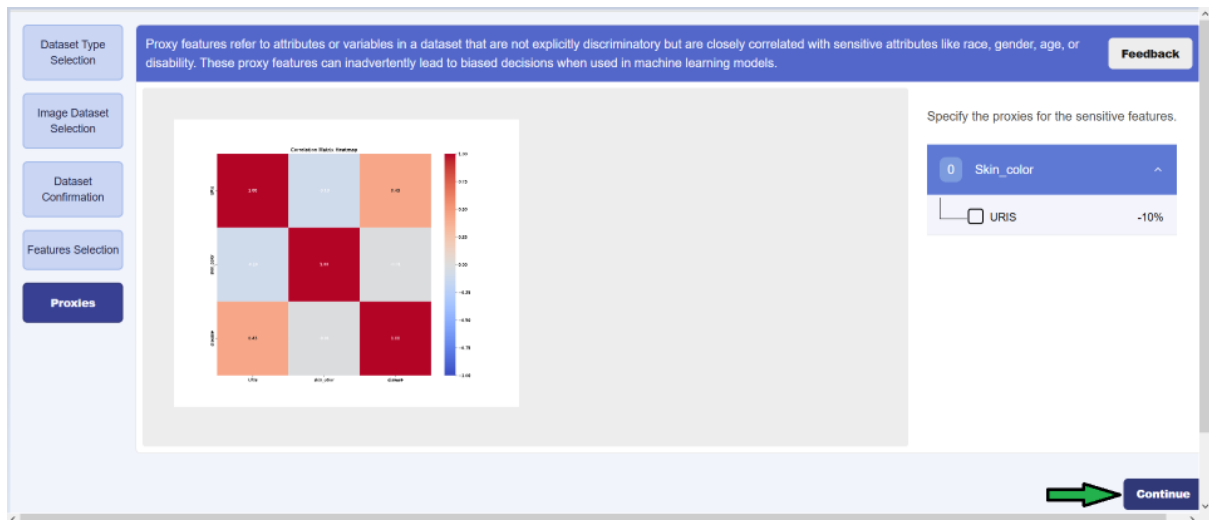
2. Click on “**Continue**” button.

- You can verify:
 - Number of samples
 - Source (via URIs)
 - Structure
- Ensures transparency before proceeding.
- Sensitive attributes are those that could lead to discrimination (e.g., race, gender, age).
- Here, skin_color is flagged because it correlates with racial demographics in healthcare.
- disease is the prediction target — the model should diagnose skin conditions fairly across skin tones.
- The system auto-detects potential sensitive features based on naming patterns.

☒ **Step 6: Define Proxy Variables**

1. Specify proxies for sensitive features:

- For skin_color, ensure SKIN_COLOR is listed as a proxy.



2. Click on “**Continue**” button.

- A proxy variable is a non-sensitive feature that strongly correlates with a sensitive one.
- Example: ZIP code might act as a proxy for race.
- In images, color histograms or brightness levels might indirectly reveal skin color.
- By identifying proxies, Aeiquitas helps prevent covert discrimination even if direct identifiers are removed.

✅ Step 7: Select Fairness Metrics

1. Choose fairness metrics:

- ✅ **Disparate Impact (DI)**
- ✅ **Statistical Parity Difference (SPD)**

Fairness metrics are quantitative tools used to assess the fairness of AI systems, ensuring all individuals or groups are treated equitably. These metrics help pinpoint biases that may arise from training data, algorithm design, or unintended model outcomes. The framework recommends the most suitable metrics based on the context analysis. For a detailed understanding of each metric's meaning and formulation, please see <https://aequitas-home.readthedocs.io/en/latest/detection.html#definitions-based-on-predicted-outcome>.

Select the fairness metrics and the features to check. The selected metrics will be mitigated in the next steps.

Disparate Impact
1 Disparate Impact measures fairness by comparing favorable outcomes between groups. A ratio below 0.8 may indicate discrimination.

☒ SKIN_COLOR

Statistical Parity Difference
0 Statistical Parity Difference measures fairness by comparing the difference in favorable outcomes between groups. A value close to 0 indicates minimal bias.

Disparate Impact - Disease/Skin_color

Chickenpox

Iatrogenic Drug Induced Exanthema

Continue

2. Apply them to:

- Group: skin_color
- Outcome: disease

Fairness metrics are quantitative tools used to assess the fairness of AI systems, ensuring all individuals or groups are treated equitably. These metrics help pinpoint biases that may arise from training data, algorithm design, or unintended model outcomes. The framework recommends the most suitable metrics based on the context analysis. For a detailed understanding of each metric's meaning and formulation, please see <https://aequitas-home.readthedocs.io/en/latest/detection.html#definitions-based-on-predicted-outcome>.

Select the fairness metrics and the features to check. The selected metrics will be mitigated in the next steps.

Disparate Impact
0 Disparate Impact measures fairness by comparing favorable outcomes between groups. A ratio below 0.8 may indicate discrimination.

☐ SKIN_COLOR

Statistical Parity Difference
1 Statistical Parity Difference measures fairness by comparing the difference in favorable outcomes between groups. A value close to 0 indicates minimal bias.

☒ SKIN_COLOR

Statistical Parity Difference - Disease/Skin_color

Chickenpox

Iatrogenic Drug Induced Exanthema

Continue

3. Click on the “Continue” button.

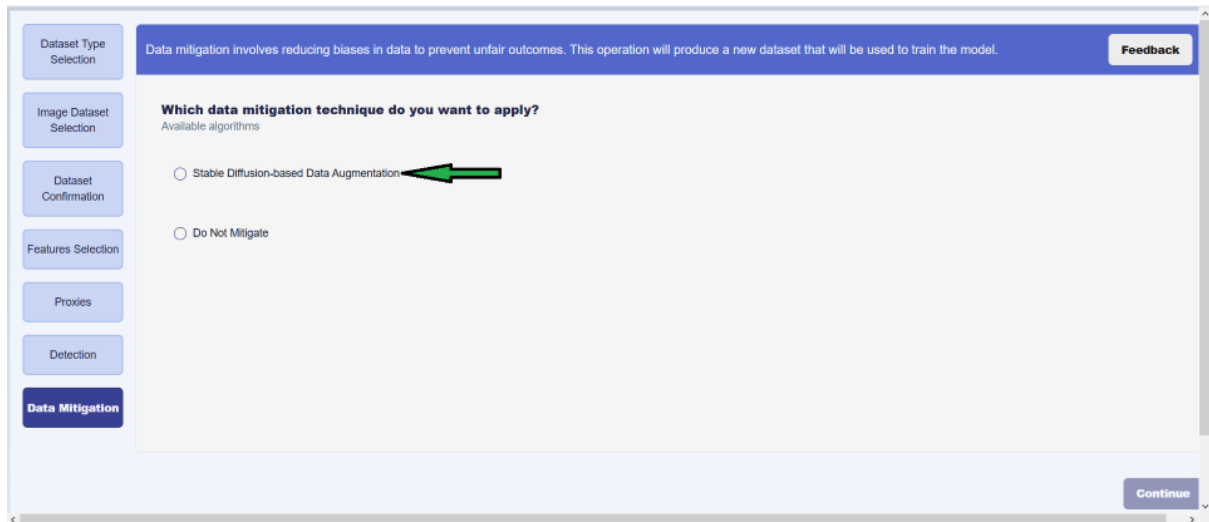
What These Metrics Mean:

METRIC	DEFINITION	FAIRNESS THRESHOLD
Disparate Impact (DI)	Ratio of positive prediction rates between groups	≥ 0.8 ("80% rule")
Statistical Parity Difference (SPD)	Difference in positive prediction rates	Close to 0

- If models predict diseases less often for darker skin tones, $DI < 0.8 \rightarrow$ **unfair**.

- These metrics help quantify group-level bias.

✅ Step 8: Choose Data Mitigation Technique



The screenshot shows a web interface for data mitigation. On the left is a vertical sidebar with buttons: 'Dataset Type Selection', 'Image Dataset Selection', 'Dataset Confirmation', 'Features Selection', 'Proxies', 'Detection', and 'Data Mitigation' (which is highlighted in dark blue). The main content area has a blue header with the text 'Data mitigation involves reducing biases in data to prevent unfair outcomes. This operation will produce a new dataset that will be used to train the model.' and a 'Feedback' button. Below the header, the question 'Which data mitigation technique do you want to apply?' is followed by 'Available algorithms'. Two radio button options are listed: 'Stable Diffusion-based Data Augmentation' (which is selected, indicated by a green checkmark and a green arrow pointing to it) and 'Do Not Mitigate'. A 'Continue' button is located at the bottom right of the main area.

1. Under Data Mitigation, choose:
 - ☒ Stable Diffusion-based Data Augmentation
 - ☐ Do Not Mitigate
2. Click **Launch Stable Diffusion-based Data Augmentation**

✅ Step 9: Configure Augmentation Settings

1. Set parameters:
 - **Augmentation Criterion:** Balanced
 - **Batch Size:** 256
 - **Epochs:** 100
2. Click Run Data Mitigation

Data mitigation involves reducing biases in data to prevent unfair outcomes. This operation will produce a new dataset that will be used to train the model. Feedback

Which data mitigation technique do you want to apply?
Available algorithms

☒ **Stable Diffusion-based Data Augmentation**

☐ Do Not Mitigate

Launch Stable Diffusion-based Data Augmentation
An augmentation based on Stable Diffusion training via DreamBooth followed by a classification step via a Swin Transformer.

Augmentation Criterion
none

Augmentation Criterion for the bias remover

Batch Size: 256
Batch size for the data augmentation

Epochs: 100
Number of epochs for the data augmentation

Continue

- This is where bias reduction begins.
- **Problem:** Some skin tones may be underrepresented → model performs poorly on them.
- Solution: Use AI-generated synthetic images to balance the dataset.

✅ Step 10: Run Data Mitigation

Which data mitigation technique do you want to apply?
Available algorithms

☒ **Stable Diffusion-based Data Augmentation**

☐ Do Not Mitigate

Launch Stable Diffusion-based Data Augmentation
An augmentation based on Stable Diffusion training via DreamBooth followed by a classification step via a Swin Transformer.

Augmentation Criterion
balanced

Augmentation Criterion for the bias remover

Batch Size: 256
Batch size for the data augmentation

Epochs: 100
Number of epochs for the data augmentation

Run

Continue

1. Wait for the process to complete.
2. Look for confirmation: “✅ Data Mitigation – Completed”.
3. System actions:
 - a. Generated synthetic images for underrepresented skin tones.

- b. Balanced class distribution.
- c. Created a de-biased version of the dataset.

Dataset Type Selection

Image Dataset Selection

Dataset Confirmation

Features Selection

Proxies

Detection

Data Mitigation

Which data mitigation technique do you want to apply?
Available algorithms

- ☒ **Stable Diffusion-based Data Augmentation**
- ☐ Do Not Mitigate

Launch Stable Diffusion-based Data Augmentation
An augmentation based on Stable Diffusion training via DreamBooth followed by a classification step via a Swin Transformer.

Augmentation Criterion
balanced

Augmentation Criterion for the bias remover

Batch Size: 256
Batch size for the data augmentation

Epochs: 100
Number of epochs for the data augmentation

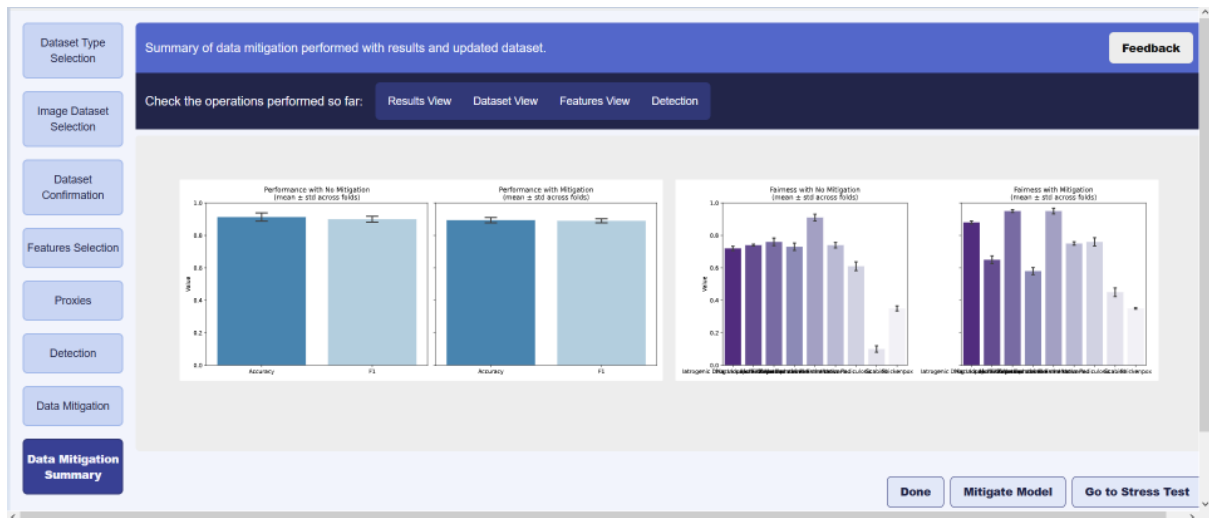
Completed

Continue

- The system has now:
 - Generated synthetic images for underrepresented skin tones.
 - Balanced class distribution.
 - Created a **de-biased version** of the dataset.
- This new dataset can now be used to train a fairer AI model.

✓ Step 11: View Mitigation Summary

1. Click View Results or Continue.
2. Explore:
 - Updated dataset stats.
 - Feature distributions.
 - Detection results.
3. Confirms:
 - Increased sample count for minority groups.
 - More uniform feature distribution.
 - Bias reduction at the data level.

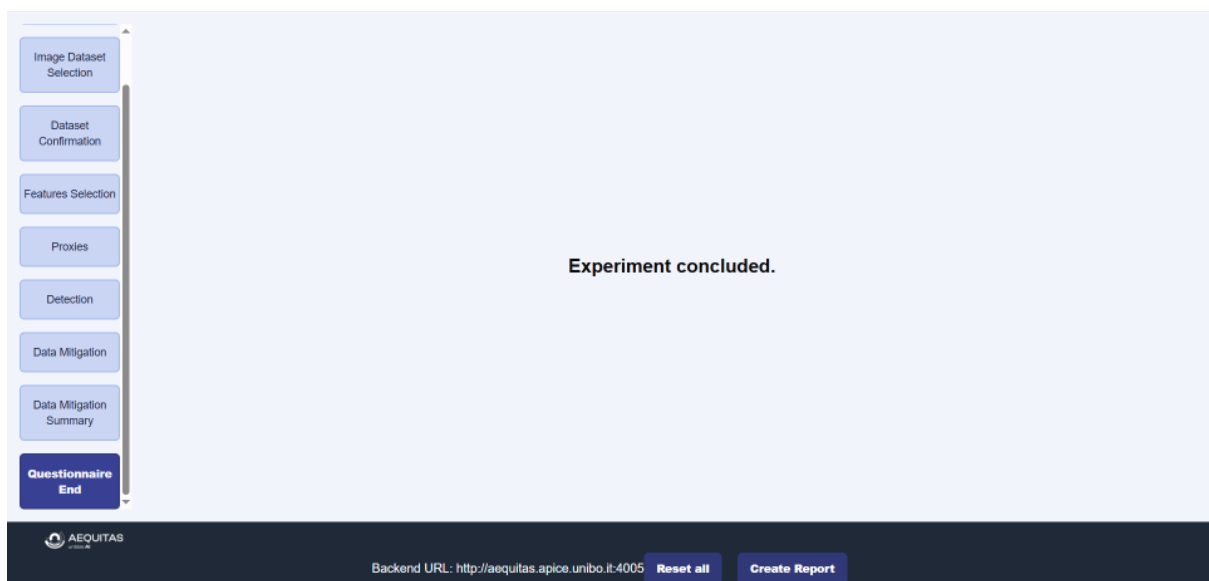


- Shows impact of mitigation:
 - Increased sample count for minority groups
 - More uniform feature distribution
- Confirms that **bias has been reduced at the data level**.

✅ Step 12: Proceed to Model Mitigation or Stress Test

1. Click:

- **Mitigate Model** → Apply algorithmic fairness techniques
- **Go to Stress Test** → Evaluate robustness under edge cases



- This concludes the data-centric fairness phase.



aicerts.ai

Contact

252 West 37th St., Suite 1200W
New York, NY 10018