Synthetic Data Generation for Skin Cancer Detection

This project builds a proof-of-concept AI pipeline that combines real dermatoscopic datasets with diffusion-based synthetic images to support skin-cancer detection. A Kedro backend and Streamlit UI enable data generation and model training to study privacy, fairness across skin types, and diagnostic performance.

Created by: Viktor_Pavlov
Created on: December 1, 2025 12:51 AM
Changed on: December 1, 2025 1:15 AM

Context of use: Education
Level of education: Master

Impact on society

What impact is expected from your technology?

This category is only partial filled.

What is exactly the problem? Is it really a problem? Are you sure?

Medical AI for skin-cancer detection needs many high-quality labelled images, but collecting and annotating dermatology data is slow, expensive and tightly constrained by privacy regulations. This technology explores whether synthetic images and semi-supervised learning can ease that bottleneck and still reach clinically useful performance, so diagnostic AI becomes more accessible and cost-efficient.

Are you sure that this technology is solving the RIGHT problem?

This question has not been answered yet.

How is this technology going to solve the problem?

This question has not been answered yet.

What negative effects do you expect from this technology?

This question has not been answered yet.

In what way is this technology contributing to a world you want to live in?

This question has not been answered yet.

Now that you have thought hard about the impact of this technology on society (by filling out the questions above), what improvements would you like to make to the technology? List them below.

This question has not been answered yet.

Hateful and criminal actors

What can bad actors do with your technology?

This category is only partial filled.

In which way can the technology be used to break the law or avoid the consequences of breaking the law?

Yes, under certain circumstances it could be misused. Synthetic or curated images and performance metrics could be falsely presented as if they came from fully validated clinical studies, helping bad actors market unapproved diagnostic tools or bypass proper regulatory and ethical review in other medical domains.

Can fakers, thieves or scammers abuse the technology?

This question has not been answered yet.

Can the technology be used against certain (ethnic) groups or (social) classes?

This question has not been answered yet.

In which way can bad actors use this technology to pit certain groups against each other? These groups can be, but are not constrained to, ethnic, social, political or religious groups.

This question has not been answered yet.

How could bad actors use this technology to subvert or attack the truth?

This question has not been answered yet.

Now that you have thought hard about how bad actors can impact this technology, what improvements would you like to make? List them below.

This question has not been answered yet.

Privacy

Are you considering the privacy & personal data of the users of your technology?

This category is only partial filled.

Does the technology register personal data? If yes, what personal data?

The prototype uses existing dermatology image datasets that are already pseudonymised by their providers and then generates synthetic images from them. It does not collect new identifiers such as names or contact details. However, because the source images depict real lesions, I treat them conceptually as health-related personal data and align the design with GDPR-style principles.

Do you think the technology invades the privacy of the stakeholders? If yes, in what way?

This question has not been answered yet.

Is the technology is compliant with prevailing privacy and data protection law? Can you indicate why?

This question has not been answered yet.

Does the technology mitigate privacy and data protection risks/ concerns (privacy by design)? Please indicate how.

This question has not been answered yet.

In which way can you imagine a future impact of the collection of personal data?

This question has not been answered yet.

Now that you have thought hard about privacy and data protection, what improvements would you like to make? List them below.

This question has not been answered yet.

Human values

How does the technology affect your human values?

This category is only partial filled.

How is the identity of the (intended) users affected by the technology?

Intended users are researchers and clinicians exploring synthetic data and semi-supervised learning, not patients. The tool can strengthen their sense of competence by making complex pipelines more approachable and experimental. At the same time it can encourage viewing patient images as technical resources, so I want that ethical tension to stay visible in documentation and discussions.

How does the technology influence the users' autonomy?

This question has not been answered yet.

What is the effect of the technology on the health and/or well-being of users?

This question has not been answered yet.

Now that you have thought hard about the impact of your technology on human values, what improvements would you like to make to the technology? List them below.

This question has not been answered yet.

Stakeholders

Have you considered all stakeholders?

This category is only partial filled.

Who are the main users/targetgroups/stakeholders for this technology? Think about the intended context by answering these questions.

Name of the stakeholder
Qin Zhao - Fontys Project Owner

How is this stakeholder affected?
-

Did you consult the stakeholder?
No

Are you going to take this stakeholder into account?
No

Name of the stakeholder
Ralf Raumanns - Dermatology / AI domain expert

How is this stakeholder affected?
-

Did you consult the stakeholder?
No

Are you going to take this stakeholder into account?
No

Name of the stakeholder
Study Research Group

How is this stakeholder affected?
-

Did you consult the stakeholder?
No

Are you going to take this stakeholder into account?
No

Name of the stakeholder
Dermatological AI researchers and students using the tool

How is this stakeholder affected?
-

Did you consult the stakeholder?
No

Are you going to take this stakeholder into account?
No

Did you consider all stakeholders, even the ones that might not be a user or target group, but still might be of interest?
-

Now that you have thought hard about all stakeholders, what improvements would you like to make? List them below.
This question has not been answered yet.

Data

Is data in your technology properly used?

This category is only partial filled.

Are you familiar with the fundamental shortcomings and pitfalls of data and do you take this sufficiently into account in the technology?

Yes. I assume datasets are incomplete, subjective and biased by clinic, device and selection effects. Synthetic images are always evaluated on real, held-out data, and I check for overfitting to artefacts or obvious domain shift. Limitations and remaining risks are documented together with the results, rather than treated as solved.

How does the technology organize continuous improvement when it comes to the use of data?

This question has not been answered yet.

How will the technology keep the insights that it identifies with data sustainable over time?

This question has not been answered yet.

In what way do you consider the fact that data is collected from the users?

This question has not been answered yet.

Now that you have thought hard about the impact of data on this technology, what improvements would you like to make? List them below.

This question has not been answered yet.

Inclusivity

Is your technology fair for everyone?

This category is only partial filled.

Will everyone have access to the technology?

This question has not been answered yet.

Does this technology have a built-in bias?

Yes. Using popular benchmark datasets bakes in their class balance, clinical setting and demographic mix. Synthetic generation and model choices inevitably reflect those starting points and my own assumptions as a single researcher. Part of the project is to surface these biases through exploratory analysis and transparent reporting, not to pretend they are fully removed.

Does this technology make automatic decisions and how do you account for them?

This question has not been answered yet.

Is everyone benefitting from the technology or only a a small group? Do you see this as a problem? Why/why not?

This question has not been answered yet.

Does the team that creates the technology represent the diversity of our society?

This question has not been answered yet.

Now that you have thought hard about the inclusivity of the technology, what improvements would you like to make? List them below.

This question has not been answered yet.

Transparency

Are you transparent about how your technology works?

This category is only partial filled.

Is it explained to the users/stakeholders how the technology works and how the business model works?

The idea, main workflows and data sources in the project documents and in the Streamlit POC interface. It is explicit that this is a personal research prototype, not a certified medical device, and that there is no commercial business model behind it. Known limitations, risks and open questions are described in documentations so stakeholders can interpret the results cautiously.

If the technology makes an (algorithmic) decision, is it explained to the users/stakeholders how the decision was reached?

This question has not been answered yet.

Is it possible to file a complaint or ask questions/get answers about this technology?

This question has not been answered yet.

Is the technology (company) clear about possible negative consequences or shortcomings of the technology?

This question has not been answered yet.

Now that you have thought hard about the transparency of this technology, what improvements would you like to make? List them below.

This question has not been answered yet.

Sustainability

Is your technology environmentally sustainable?

This category is only partial filled.

In what way is the direct and indirect energy use of this technology taken into account?

Training and running generative and classification models consumes GPU time and therefore energy. I try to limit this by re-using pre-trained models where possible, keeping experiment grids small and using shared research infrastructure instead of separate long-running servers. Future work could add simple tracking of compute per run to guide greener choices.

Do you think alternative materials could have been considered in the technology?

This question has not been answered yet.

Do you think the lifespan of the technology is realistic?

This question has not been answered yet.

What is the hidden impact of the technology in the whole chain?

This question has not been answered yet.

Now that you have thought hard about the sustainability of this technology, what improvements would you like to make? List them below.

This question has not been answered yet.

Future

Did you consider future impact?

This category is only partial filled.

What could possibly happen with this technology in the future?
If approaches like this mature, synthetic dermatology data and semi-supervised learning could lower annotation costs and let more teams prototype diagnostic tools with less real patient data. At the same time, weak governance around synthetic data and validation could hide important performance gaps, leading to over-confident deployment of fragile models.

Sketch a or some future scenario (s) (20-50 years up front) regarding the technology with the help of storytelling. Start with at least one utopian scenario.
This question has not been answered yet.

Sketch a or some future scenario (s) (20-50 years up front) regarding the technology with the help of storytelling. Start with at least one dystopian scenario.
This question has not been answered yet.

Would you like to live in one of this scenario's? Why? Why not?
This question has not been answered yet.

What happens if the technology (which you have thought of as ethically well-considered) is bought or taken over by another party?
This question has not been answered yet.

Impact Improvement: Now that you have thought hard about the future impact of the technology, what improvements would you like to make? List them below.
This question has not been answered yet.