

Deep Learning Data Generation for Medical Prediction Systems

Creating a System to Generate Complex Patient Data



AI-generated images of surgery equipment¹

It is becoming more and more clear that **AI systems** can improve healthcare in numerous ways. For example, a Machine Learning model was created that can outperform doctors in recognizing breast cancer.² Researchers have also created a system that performs better than 72% of general practitioners when diagnosing test cases of illnesses.³ Moreover, multiple studies with Convolutional Neural Networks in medical care have shown promising results.⁴

It is safe to say that the digital revolution provides many opportunities. For example, with large amounts of vital data, we might be able to create a Machine Learning tool which would provide a doctor with information to make better decisions. We can potentially **save years of life** with these kinds of systems. So, what is holding us back?

We have still a **big hurdle** to overcome: **the availability of enough high-quality data**. This is where you come in: we want you to help solve the data shortage by **generating it yourself**. You could think of training **GANs (Generative Adversarial Networks)**⁵ that generate new data based on big openly available datasets. To test your project, you can generate data from real surgery datasets provided by our partners at the renowned Charité hospital. You will be able to **consult with surgeons** that have a technical background and are excited to bring **new technologies into practice**.

You are offered the chance to generate data from various datasets and develop ML-models to evaluate the generated data. Your solution can be used for, for example, **hospitals with little resources**. Using some of their data with your system, they could generate custom data to get **accurate predictions** in a short period of time. Your data could also be used in research to improve clinical practices and **improve quality of life**.

¹ Generated using Dall-E Mini (<https://github.com/borisdayma/dalle-mini>) with the query: "surgery equipment", see: Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021, July). Zero-shot text-to-image generation. In *International Conference on Machine Learning* (pp. 8821-8831). PMLR.

² McKinney, S.M., Sieniek, M., Godbole, V. et al. International evaluation of an AI system for breast cancer screening. *Nature* 577, 89–94 (2020).

³ Richens, J.G., Lee, C.M. & Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat Commun* 11, 3923 (2020).

⁴ Nagendran M., Chen Y, Lovejoy C A, Gordon A C, Komorowski M, Harvey H et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies *BMJ* 2020;

⁵ See: <https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/> for an introduction

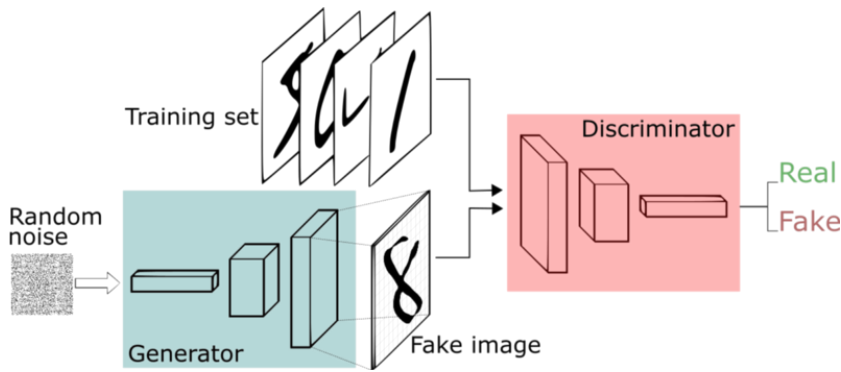


Figure 2: Schematic showing the basics of a GAN used to generate images

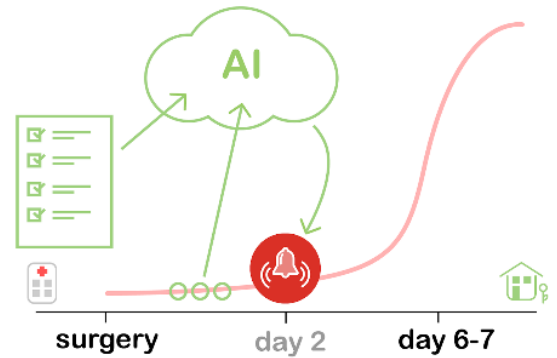


Figure 3: A system that alerts doctors of complications before they are likely to happen

Goal

We are interested in exciting **new, practical, innovative ideas for improving AI in health**. You will be building different types of **Machine Learning models** and **data generation algorithms** and implement them **with scalable software systems**. You will have access to a powerful cluster to do this. The goal is to create an **accurate tool**, which can be used to create data **for a variety of use cases**. We want to use the results of your project in **research and in clinical practice**.

Requirements and Expectations

This is a project focused on **computer science** applied to **health care event prediction**. Ideally, you have a **data science** or **it-engineering** background and are motivated and **excited to learn** about putting this to use to **make a real, state-of-the-art solution**. The following topics will be relevant for this project, and we expect prior knowledge **of one or more** of the following:

- **Creating a reliable, extensible, and explainable software system**
 - ➔ Experience with software engineering and modelling
- **Processing static, and time-series data**
 - ➔ The use of different techniques to analyse and enhance data
- **Machine learning and deep learning**
 - ➔ Experience with frameworks like Scikit-Learn, TensorFlow, and Keras or equivalents
 - ➔ *Bonus: experience with data generation frameworks techniques*
- **Knowledge of statistical methods**
 - ➔ Some fundamental experience with basic mathematical methods
- **Combining science and engineering**
 - ➔ Advancing the status quo by publishing your results and software

Contact

If you have any questions about the project, want to see what is possible, or you are curious about the skills involved, please contact us. We are always happy with the input of motivated students.

Robin van de Water

Room: G-2.1.11

Phone: +49 331 5509-3436

E-Mail: robin.vandewater@hpi.de

Bjarne Pfitzner

Room: G-2.1.12

Phone: +49 331 5509-1374

E-Mail: bjarne.pfitzner@hpi.de

Bert Arrrich

Room: G-2.1.14

Phone: +49 331 5509-4850

E-Mail: bert.arrrich@hpi.de

⁶ Source: <https://www.freecodecamp.org/news/an-intuitive-introduction-to-generative-adversarial-networks-gans-7a2264a81394>