AI in the Pharmaceutical Industry: The Case of Chugai Pharmaceutical

December 5, 2025

Chugai Pharmaceutical Co., Ltd.

Digital Transformation Unit

Isao Sano, Ph.D.

Agenda

1. AI Applications in the Pharmaceutical Industry

2. Generation and Classification of Chest X-ray Images Using GANs with Limited Training Data

[2-1] Introduction
[2-2] Materials and Methods
[2-3] Results and Discussion

3. Predicting Patient Enrollment Outcomes in Clinical Trials Using Large Language Models

[3-1] Introduction
[3-2] Materials and Methods
[3-3] Results and Discussion

4. Information Gathering at ACL, a Top Conference in Natural Language Processing

5. Conclusion

Agenda

1. AI Applications in the Pharmaceutical Industry

2. Generation and Classification of Chest X-ray Images Using GANs with Limited Training Data

[2-1] Introduction
[2-2] Materials and Methods
[2-3] Results and Discussion

3. Predicting Patient Enrollment Outcomes in Clinical Trials Using Large Language Models

[3-1] Introduction
[3-2] Materials and Methods
[3-3] Results and Discussion

4. Information Gathering at ACL, a Top Conference in Natural Language Processing

5. Conclusion

Isao Sano

Career

- Completed Graduate School of Life Sciences, Tohoku University.

Ph.D. in Life Sciences.

- Joined Chugai Pharmaceutical Co., Ltd. as a new graduate in 2021.

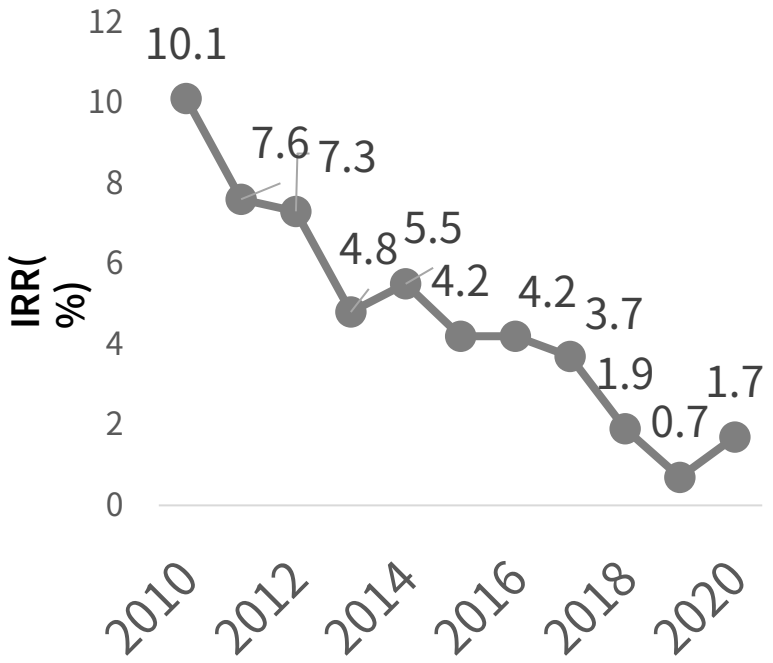
Main Responsibilities

- Image analysis and marketing analytics using machine learning
- Natural language processing and knowledge graph applications using clinical trial data
- Planning data science events and promoting knowledge sharing and collaboration among data scientists

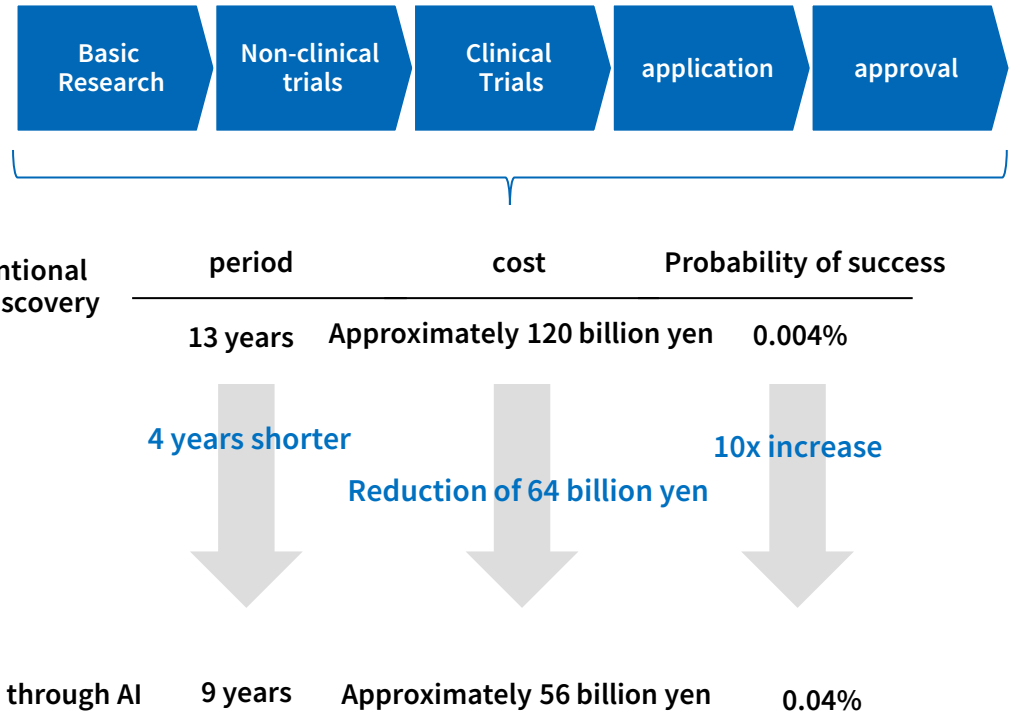
Why digital now in the pharmaceutical industry?

The R&D productivity of pharmaceuticals has been declining year by year, and there is an increasing need to leverage “digital” technologies in drug discovery.

Mega Pharma's R&D Productivity (IRR)*1



The potential of AI in drug discovery*2

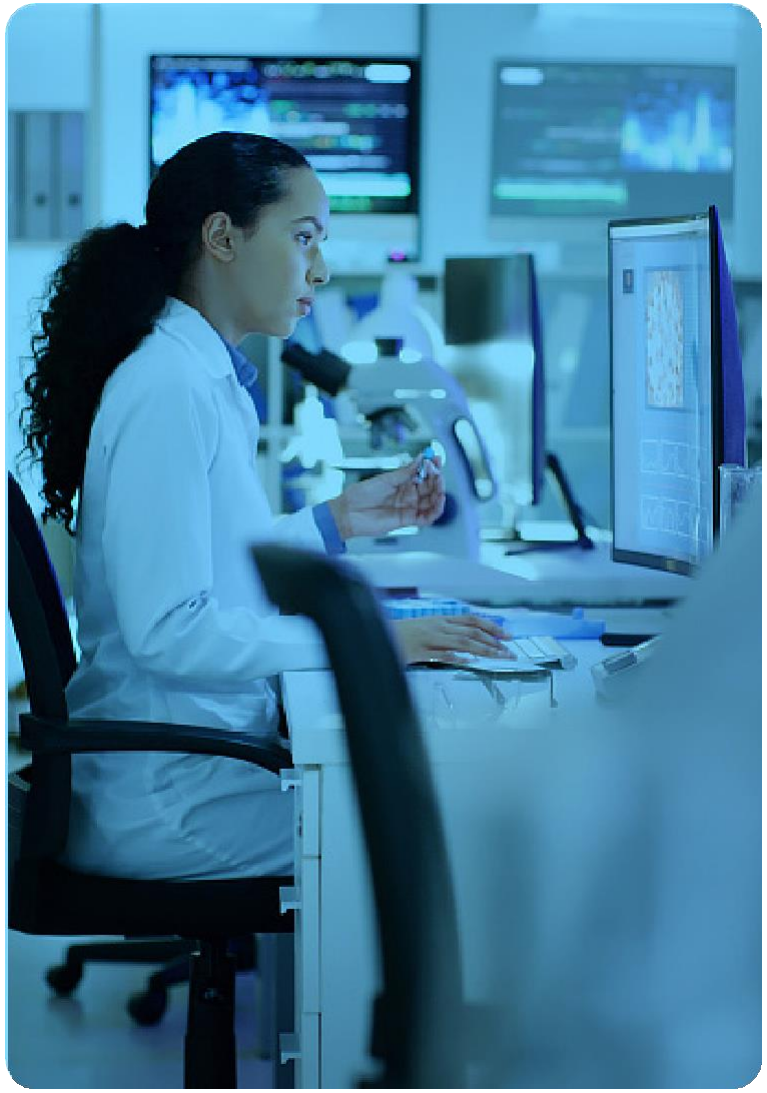


*1: Deloitte "Measuring the return from pharmaceutical innovation 2020" (covering 12 major global companies)

**2: Prepared by processing and creating "The 3rd Roundtable Meeting on the Promotion of AI Utilization in the Health and Medical Field (March 29) Submitted by Okuno Members" (Ministry of Health, Labour and Welfare)

About Chugai Pharmaceutical

A research-driven pharmaceutical company with unique technologies and science as its strengths.



One of Japan's leading prescription pharmaceutical companies

- Revenue: approx. ¥1.17 trillion
- Operating profit: approx. ¥556 billion
- Operating profit margin: 47.5%

Unique Business Model

- Entered into a strategic alliance with Roche in 2002
- Delivering innovative in-house products to patients worldwide
- Number of countries with global approval for Chugai-originated products: over 110

Unique Technologies and Science

- World-class antibody engineering technologies
- Drug discovery capabilities based on diverse modalities, including antibodies, small molecules, and mid-sized molecules
- 9 Breakthrough Therapy*1 designations

Leading the industry through DX (Digital Transformation)

- Selected as a "DX Platinum Company 2023–2025"

*1 The system was introduced by the U.S. Food and Drug Administration (FDA) in July 2012 with the aim of promoting the development and review of drugs to treat serious or fatal diseases and conditions.

DX – “CHUGAI DIGITAL VISION 2030”

We promote company-wide DX under the slogan “CHUGAI DIGITAL VISION 2030.”
Through DX, we aim to continuously deliver innovative medicines while improving productivity across the entire value chain.

CHUGAI DIGITAL VISION 2030

Value Creation

Utilizing digital
Increasing the sophistication of the early phase of research and development
Utilizing digital technologies such as AI to improve the probability of success in drug development, and significantly reduce the time and cost of drug discovery

Dramatic productivity gains across the entire value chain
Utilizing digital technologies such as AI and robotics to significantly improve the efficiency of business processes

Productivity innovation

Building a company-wide foundation to support innovation creation
Promote the establishment of a global-level IT infrastructure and the development of digital human resources

Climate Reform

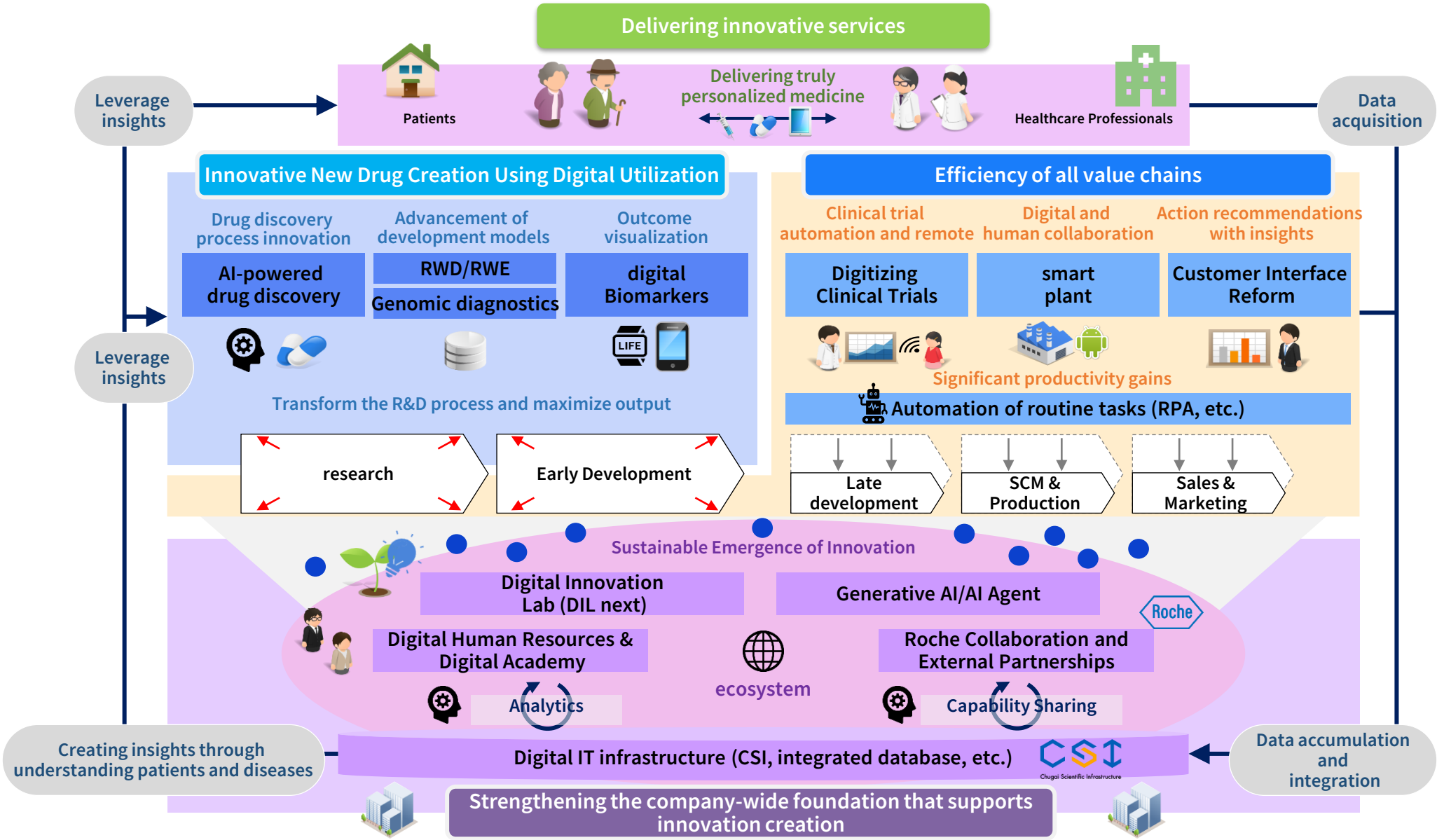
DXプラチナ企業 2023-2025 Digital Transformation

In "DX Stocks"

Selected as a "DX Platinum Company 2023-2025"

Chugai Pharmaceutical is the only one selected as a DX stock in the pharmaceutical industry for the fifth consecutive year

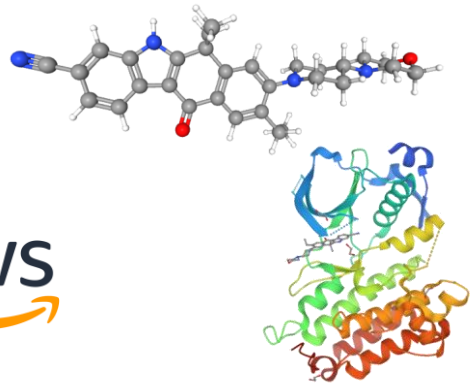
CHUGAI DIGITAL VISION 2030

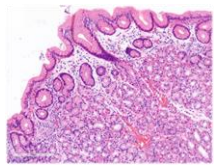


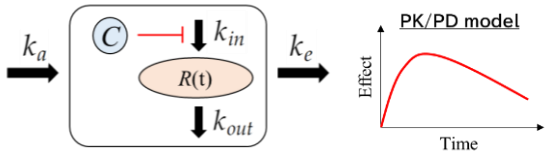
Our Data Science Activities

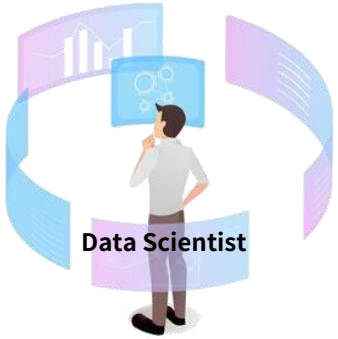
10 ~ 15 years

Idea Ideas	Drug discovery research		Development Research		Clinical Trials			After the application is approved, To the market
	Target molecule identification	Search and discovery of new drugs	Selection of product candidates for development	Preclinical Studies	Phase I study	Phase II study	Phase III Study	
	Evaluation system construction Targeted Assessment	screening (in vitro) (in vivo)	Pharmacology Pharmacokinetics toxicity	Pharmacology Pharmacokinetics GLP Security	Pharmacokinetics safety	validity Method of administration dose	validity safety	









aws

Google Cloud

ChatGPT

Gemini

We apply data science techniques to solve problems across a wide range of domains—from chemical structures and medical images to omics data, sensor data, and real-world data.

Agenda

1. AI Applications in the Pharmaceutical Industry

2. Generation and Classification of Chest X-ray Images Using GANs with Limited Training Data

[2-1] Introduction

[2-2] Materials and Methods

[2-3] Results and Discussion

3. Predicting Patient Enrollment Outcomes in Clinical Trials Using Large Language Models

[3-1] Introduction

[3-2] Materials and Methods

[3-3] Results and Discussion

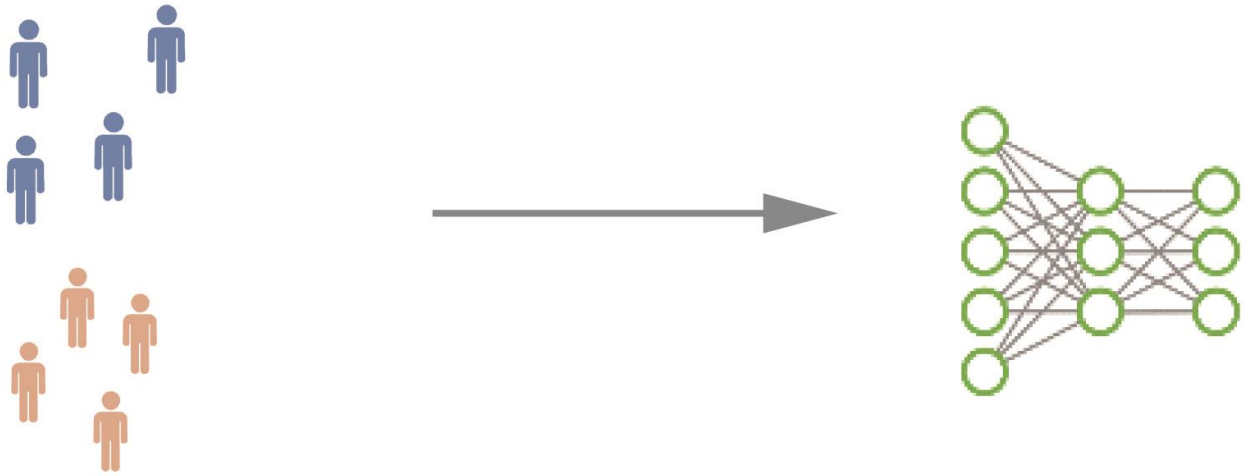
4. Information Gathering at ACL, a Top Conference in Natural Language Processing

5. Conclusion

Challenge: Limited Data in Pharmaceuticals

In the pharmaceutical field, the use of data science has often been hindered by limited data availability.

When data are scarce, model performance tends to be low.



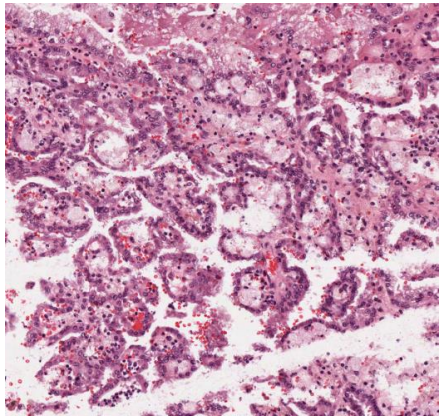
Here, using image analysis as an example, we introduce a case where model performance was improved even with a small dataset.

Background: Image-Based Diagnosis in Medicine

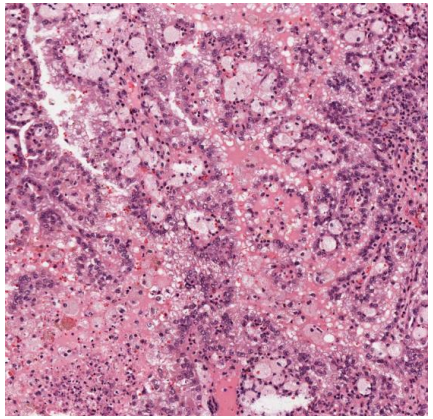
Developing accurate image-based diagnostic technologies is one of the key challenges in healthcare [Niazi 19, Topol 19].

Pathologic images

Immune cells A

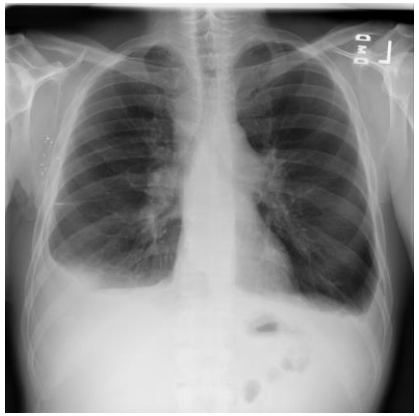


Immune cells B



X-ray images

Diseases A



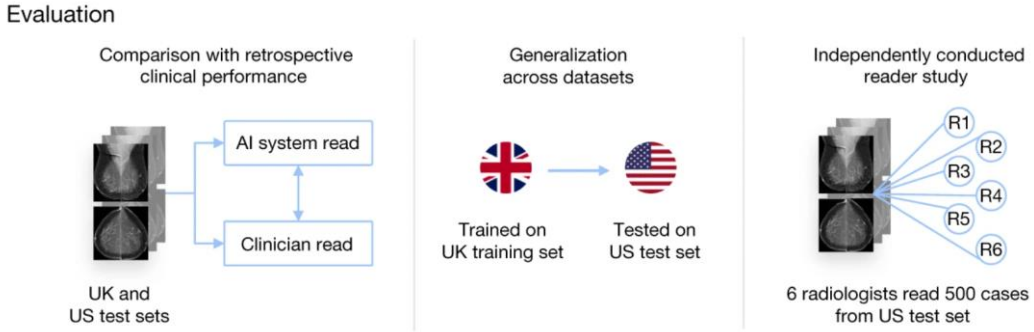
Diseases B

Visual diagnosis by human experts can be subjective, and it is known that different observers may arrive at different diagnoses even when looking at the same image [Lawton 14, Ayan 19].

https://portal.gdc.cancer.gov/files/0300fdaa-a8c4-4d71-a6b5-28af757285d9
https://nihcc.app.box.com/v/ChestXray-NIHCC

Image Classification Models and Their Limitations

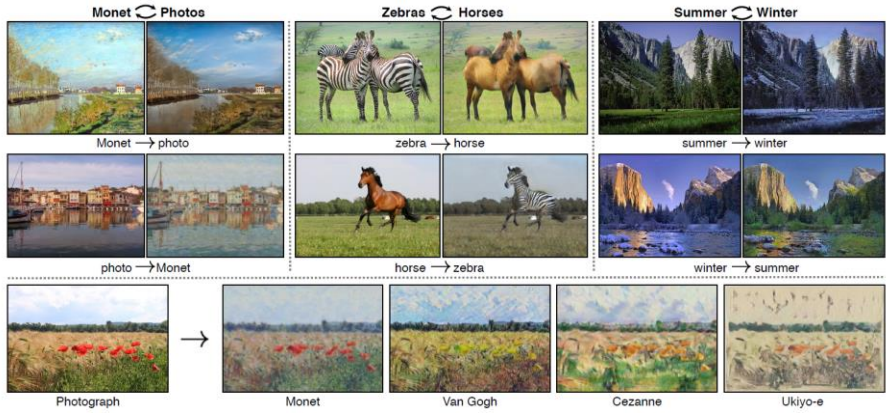
Image classification models have in some cases achieved performance surpassing that of human experts [Rajpurkar 17, McKinney 20].



McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ... & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89-94.

Traditionally, in developing image-based diagnostic technologies using machine learning, it has been difficult to build highly accurate image classification models because the amount of data available for training was limited [Greenspan 16].

GAN（Generative Adversarial Networks）



Generative adversarial networks (GANs) have been shown to be capable of generating high-fidelity images [Brock 18, Karras 20].

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223-2232).

Data Challenges in the Medical Domain

In the medical field, it is often difficult to collect large volumes of image data linked to diagnostic labels.

The use of light-weight GANs suggests that GAN-based approaches can be applied even in the medical domain.

Liu, B., Zhu, Y., Song, K., & Elgammal, A. (2020). Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In International Conference on Learning Representations.

Agenda

1. AI Applications in the Pharmaceutical Industry

2. Generation and Classification of Chest X-ray Images Using GANs with Limited Training Data

[2-1] Introduction

[2-2] Materials and Methods

[2-3] Results and Discussion

3. Predicting Patient Enrollment Outcomes in Clinical Trials Using Large Language Models

[3-1] Introduction

[3-2] Materials and Methods

[3-3] Results and Discussion

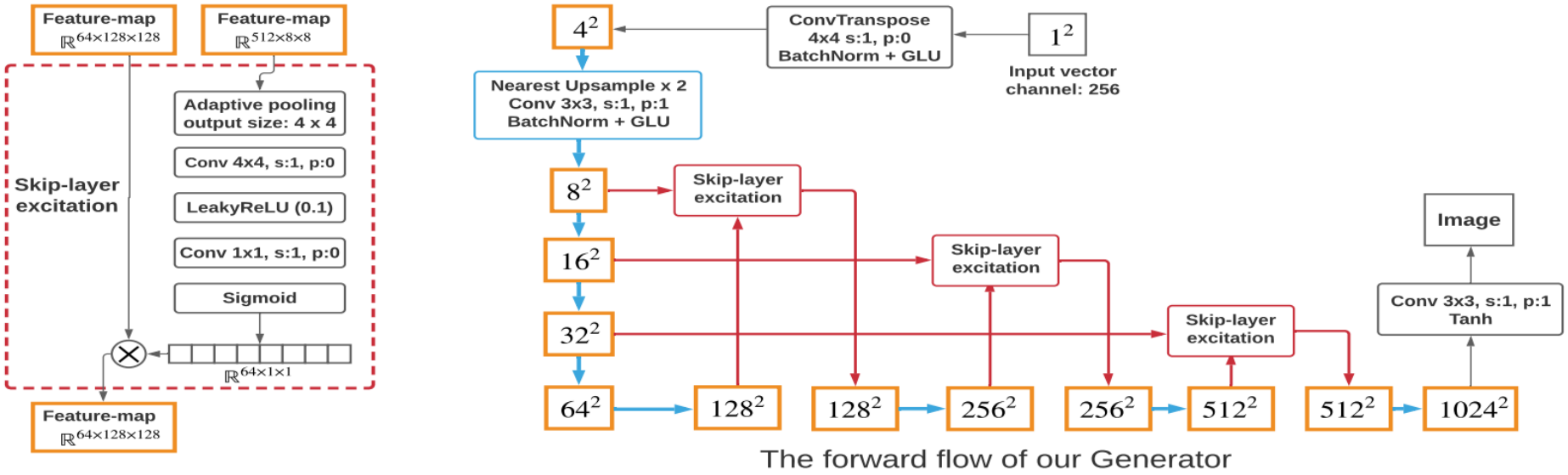
4. Information Gathering at ACL, a Top Conference in Natural Language Processing

5. Conclusion

Key Feature 1 of Light-weight GAN

Addressing the need for deeper networks to generate high-quality images

By introducing the SLE Module as a skip connection in the Generator, we reduced network depth and computational cost while improving training speed [Liu 20].

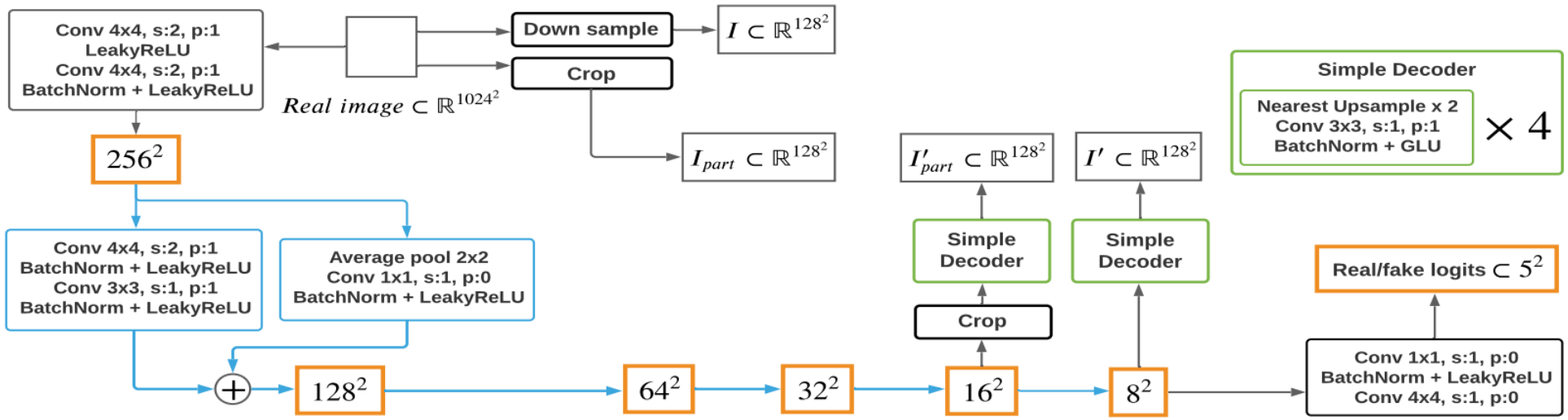


Liu, B., Zhu, Y., Song, K., & Elgammal, A. (2020). Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In International Conference on Learning Representations.

Key Feature 2 of Light-weight GAN

Addressing the need for large mini-batches to improve discriminator accuracy

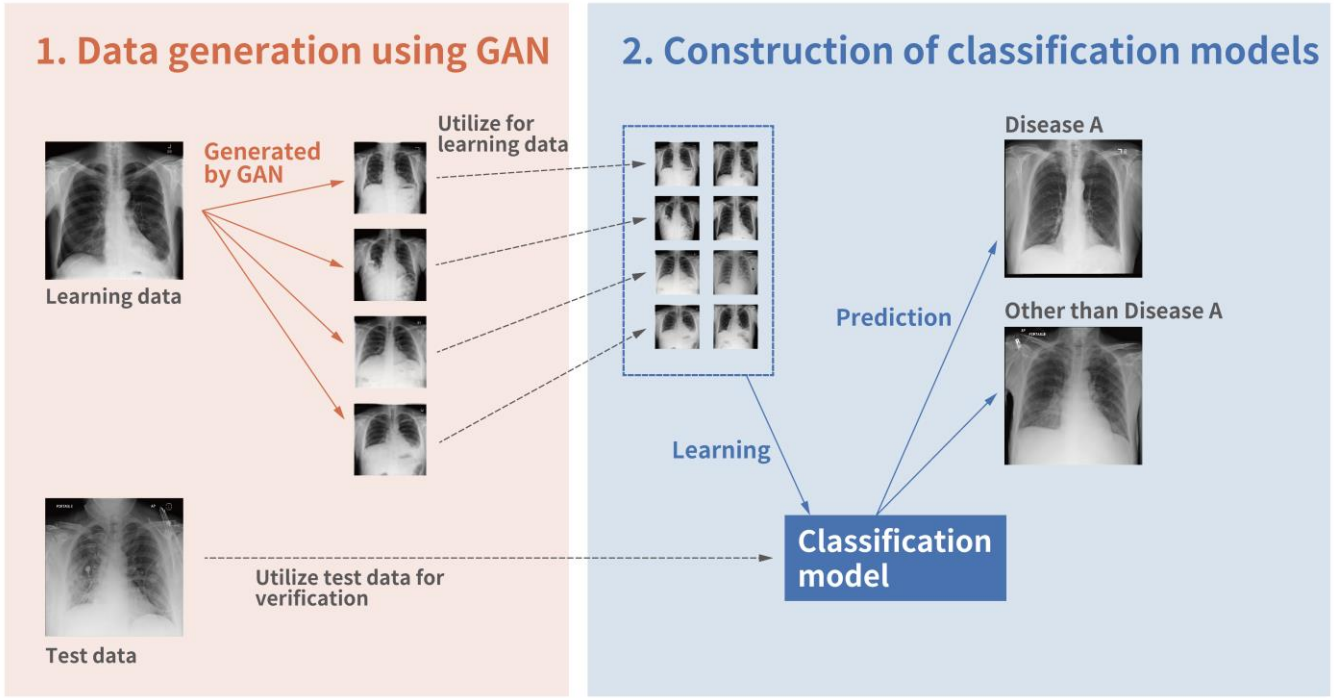
By introducing self-supervised learning into the Discriminator, we stabilized training and improved accuracy even when using small batch sizes [Liu 20].



Liu, B., Zhu, Y., Song, K., & Elgammal, A. (2020). Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In International Conference on Learning Representations.

Objective

We examined whether high-performance image classification models can be built even with limited data.

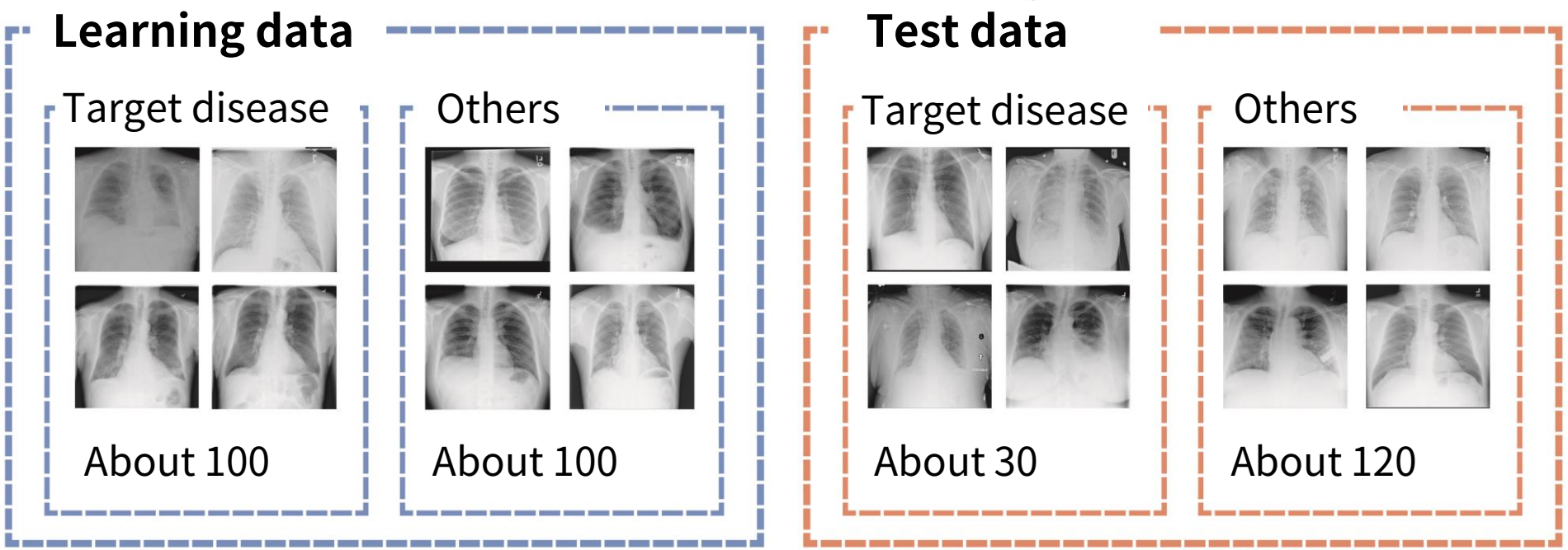


- We used chest X-ray images to generate additional data with a light-weight GAN and constructed image classification models.
- We then evaluated how this approach affected the performance of the models.

https://nihcc.app.box.com/v/ChestXray-NIHCC

Dataset

We used chest X-ray images released by the U.S. National Institutes of Health (NIH).



When GAN was used, the target disease of the learning data and the other data were doubled.

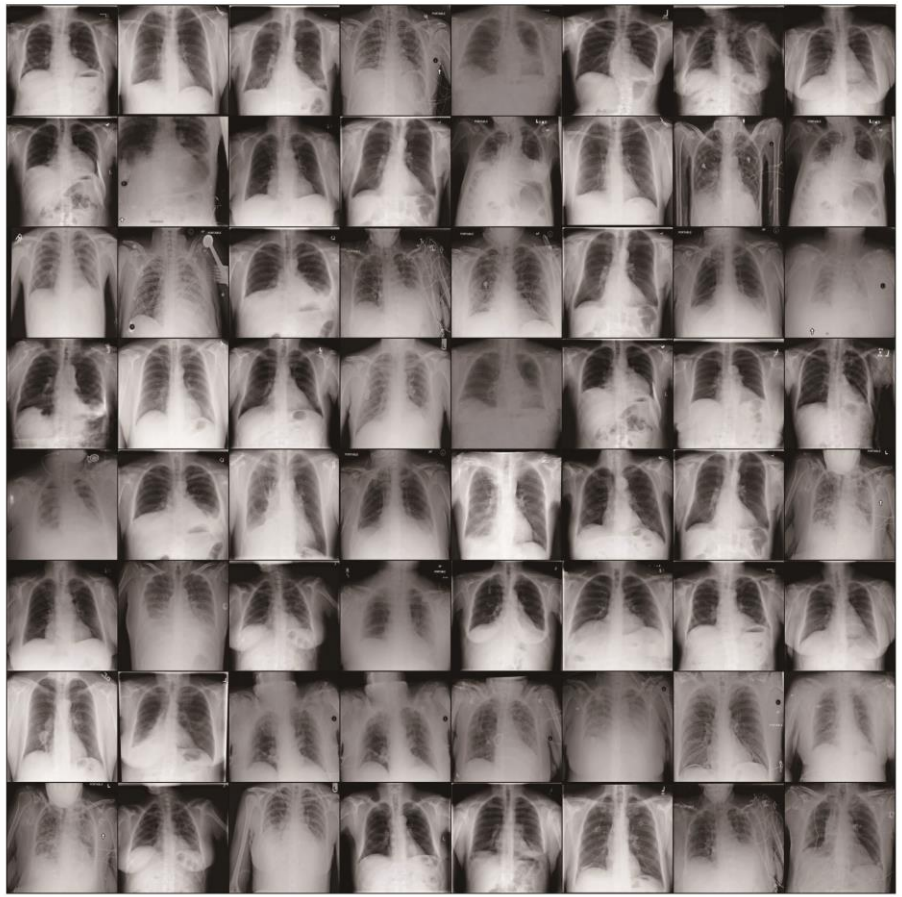
To avoid overlap between the training and test datasets, we split the images into training and test sets before starting the analysis.

https://nihcc.app.box.com/v/ChestXray-NIHCC

Data Generation with GAN

Using a light-weight GAN, we generated new images that possess the characteristics of each disease (Atelectasis and Cardiomegaly) from their respective training datasets.

Atelectasis



Cardiomegaly

Agenda

1. AI Applications in the Pharmaceutical Industry

2. Generation and Classification of Chest X-ray Images Using GANs with Limited Training Data

[2-1] Introduction
[2-2] Materials and Methods
[2-3] Results and Discussion

3. Construction of an Enrollment Prediction Model Using Summaries of Clinical Trial Criteria with LLMs

[3-1] Introduction
[3-2] Materials and Methods
[3-3] Results and Discussion

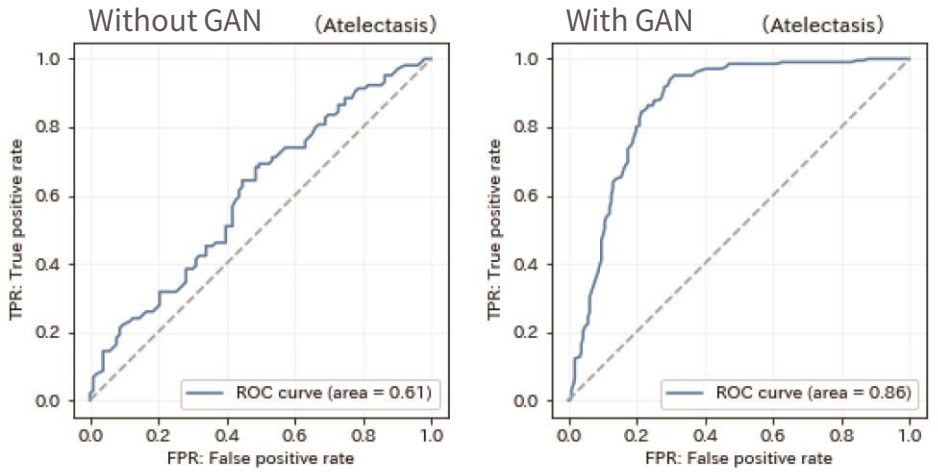
4. Information Gathering at ACL, a Top Conference in Natural Language Processing

5. Conclusion

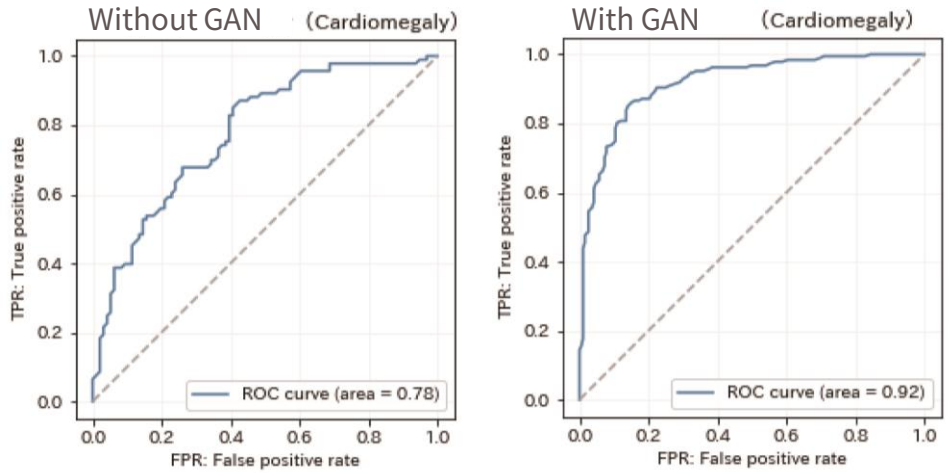
Building Image Classification Models with ResNet (1)

We built image classification models using ResNet and calculated ROC curves for each setting. As a result, the AUC values on the validation set were higher when using data generated by the light-weight GAN.

1) Atelectasis



2) Cardiomegaly

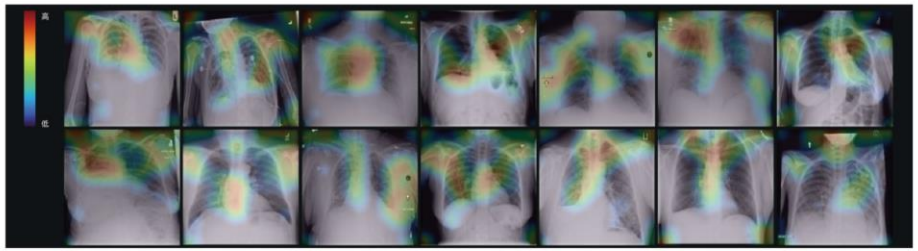


Building Image Classification Models with ResNet (2)

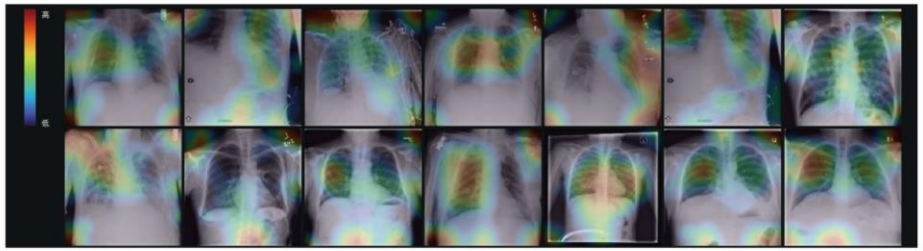
We computed activation maps for each model. The results showed that, in all cases, the constructed image classification models primarily relied on features in the chest region for classification.

1) Atelectasis

Without GAN（Atelectasis）

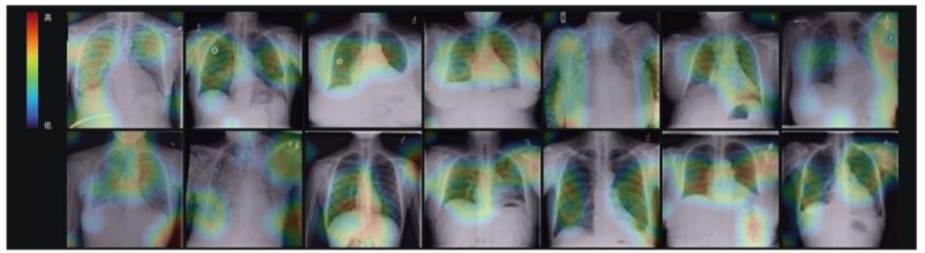


With GAN（Atelectasis）

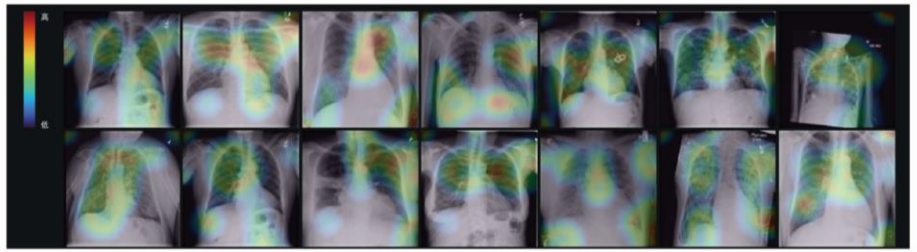


2) Cardiomegaly

Without GAN （Cardiomegaly）



With GAN（Cardiomegaly）



Comparison of Model Performance (Accuracy and AUC)

We calculated Accuracy and AUC in each setting. For both Atelectasis and Cardiomegaly, the models trained with data generated by the light-weight GAN achieved higher Accuracy and AUC than those trained without GAN-generated data.

	Accuracy	AUC
Without GAN （Atelectasis）	0.62	0.54
With GAN（Atelectasis）	0.76	0.60
Without GAN （Cardiomegaly）	0.58	0.73
With GAN（Cardiomegaly）	0.79	0.80

Comparison of Model Performance (Precision, Recall, F1-score)

We calculated precision, recall, and F1-score in each setting. The results showed that, for both the target diseases and the “Others” category, precision and F1-score were higher when using data generated by the light-weight GAN.

1) Atelectasis

Without GAN（**Atelectasis**）

	precision	recall	f1-score
Atelectasis	0.18	0.32	0.23
Others	0.82	0.69	0.75

With GAN （**Atelectasis**）

	precision	recall	f1-score
Atelectasis	0.30	0.24	0.27
Others	0.84	0.88	0.86

2) Cardiomegaly

Without GAN （**Cardiomegaly**）

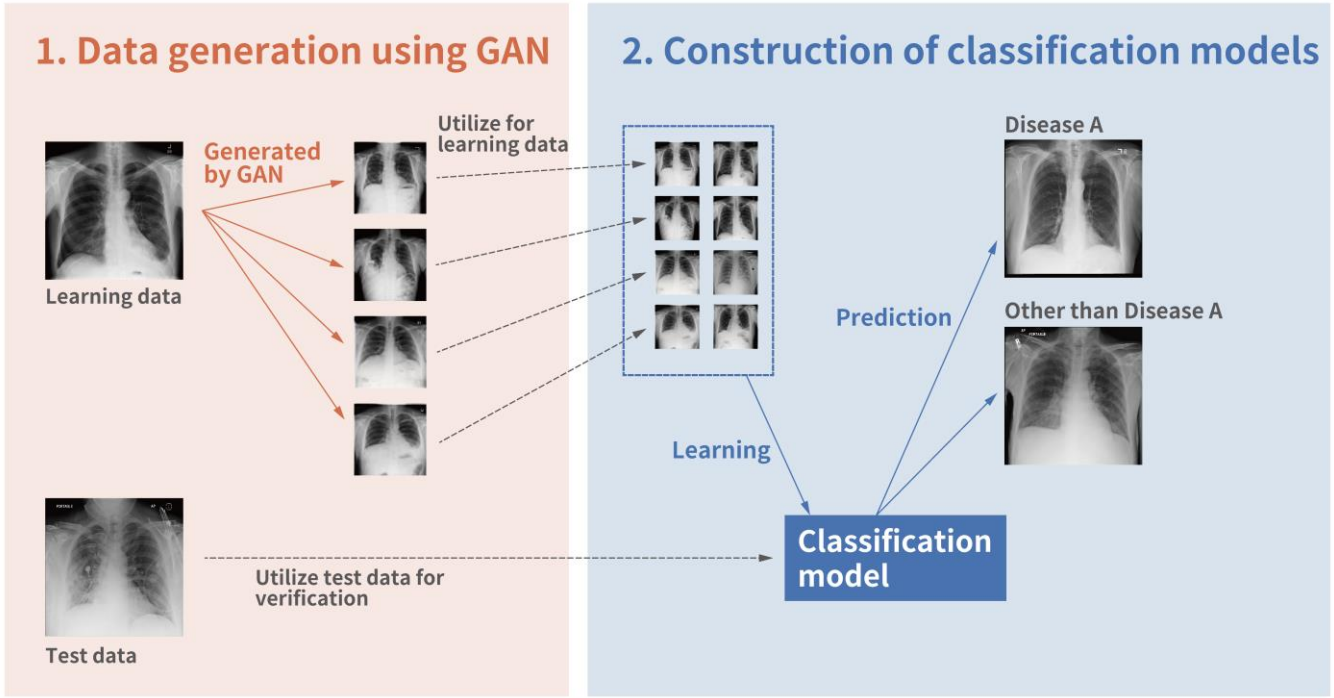
	precision	recall	f1-score
Cardiomegaly	0.25	0.75	0.38
Others	0.91	0.54	0.68

With GAN （**Cardiomegaly**）

	precision	recall	f1-score
Cardiomegaly	0.42	0.58	0.49
Others	0.91	0.84	0.87

Summary of Results

When we increased the training data using light-weight GANs for chest X-ray images, we were able to construct high-performance image classification models.



In the medical field, the amount of data available for training is often limited [Greenspan 16]. Therefore, data generation using light-weight GANs may serve as a useful approach for building high-accuracy classification models based on small medical datasets.

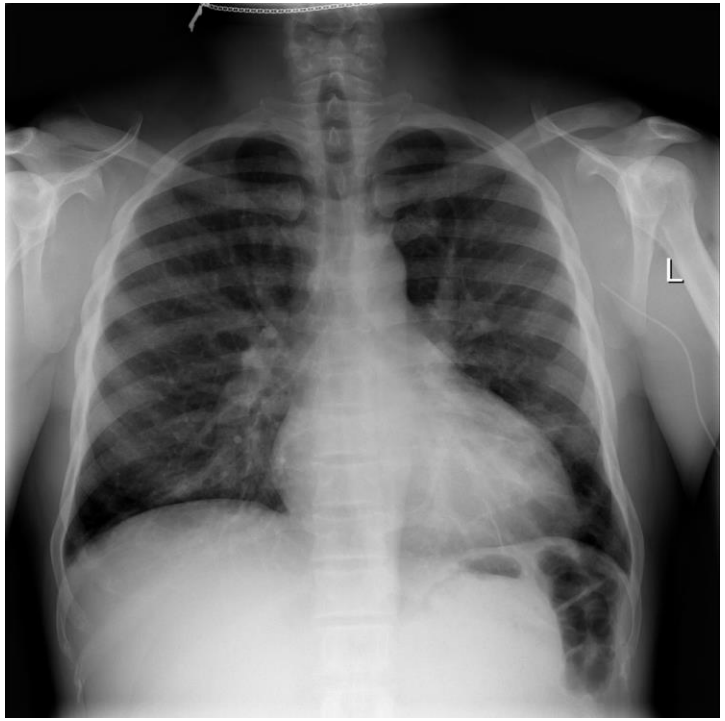
https://nihcc.app.box.com/v/ChestXray-NIHCC

Differences Between Diseases

We found that, compared to Atelectasis, Cardiomegaly showed a larger performance improvement when GAN-generated data were used.

Atelectasis

Cardiomegaly



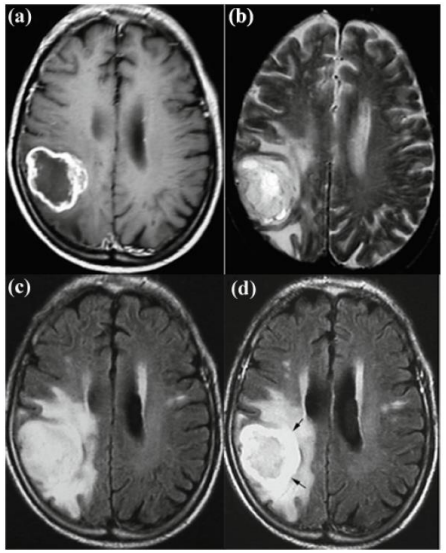
One possible explanation is that the features of Cardiomegaly in X-ray images are more pronounced, so generating additional training data with GANs had a greater impact on the performance of the image classification models.

https://nihcc.app.box.com/v/ChestXray-NIHCC

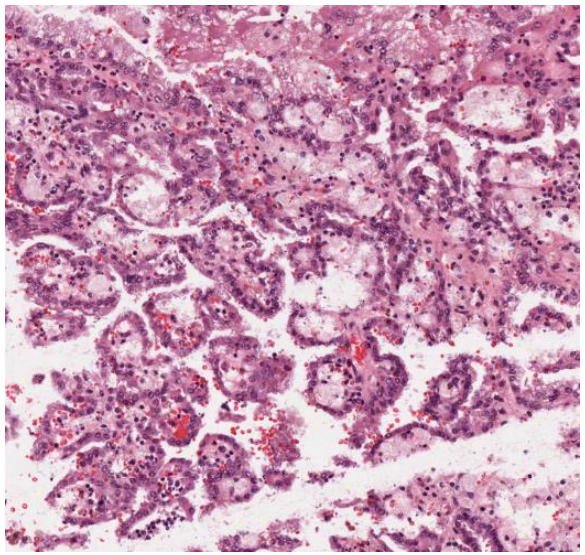
Delivering Better Medicines to Patients as Early as Possible

This study suggests the usefulness of GANs for analyzing specific diseases in chest X-ray images.

MRI image



Pathologic image



As future work, we plan to:
- Extend our analysis to other types of images and diseases beyond Atelectasis and Cardiomegaly
- Further refine the algorithms

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6961028
https://portal.gdc.cancer.gov/files/0300fdaa-a8c4-4d71-a6b5-28af757285d9

Agenda

1. AI Applications in the Pharmaceutical Industry

2. Generation and Classification of Chest X-ray Images Using GANs with Limited Training Data
[2-1] Introduction
[2-2] Materials and Methods
[2-3] Results and Discussion

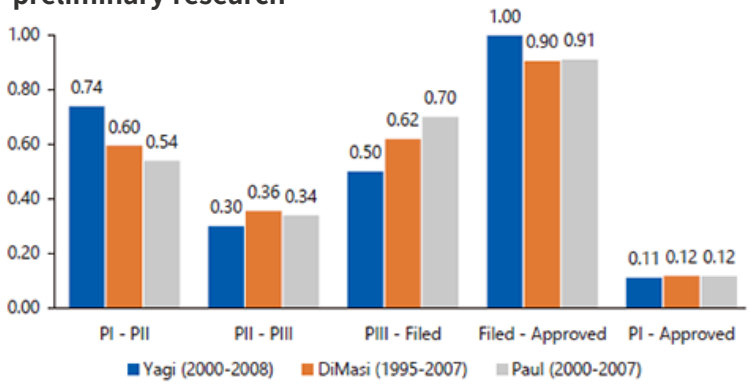
3. Predicting Patient Enrollment Outcomes in Clinical Trials Using Large Language Models
[3-1] Introduction
[3-2] Materials and Methods
[3-3] Results and Discussion

4. Information Gathering at ACL, a Top Conference in Natural Language Processing

5. Conclusion

Predicting clinical trial success is crucial

Fig. Success rate of drug discovery according to preliminary research



- Clinical trials are crucial processes for verifying the efficacy and safety of new drugs and treatments.
- Obtaining approval for a new drug requires a long period and substantial costs.

However

The probability of new drug approval is low, and development failure leads to significant losses for companies.

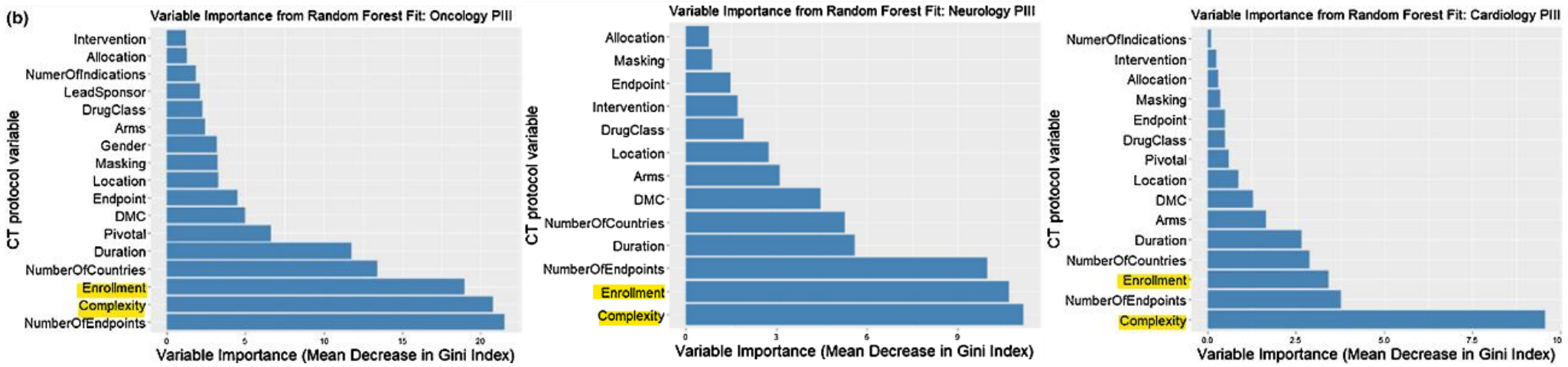
■ Benefits of predicting probability of success

- If low success probability trials can be identified early, risks can be avoided by determining the causes.
- This would be helpful for strategic decision-making, such as prioritizing product pipelines.
- etc.

https://www.jpma.or.jp/opir/news/066/07.html

Enrollment is crucial for success of clinical trials

- Patient enrollment is a crucial factor for the success or failure of clinical trials
  - In previous research, it has been shown that patient enrollment is a parameter that highly contributes to the prediction of clinical trial success in each disease field.
  - If we can predict patient enrollment, it could lead to predicting clinical trial success.



https://www.sciencedirect.com/science/article/pii/S1359644620300052

Quantifying eligibility criteria is crucial

- The success of enrollment is deeply influenced by the setting of eligibility criteria
  - Eligibility criteria is the set of conditions participants must meet to join a clinical trial.
  - In previous research, the complexity of eligibility criteria was extracted and used for predicting clinical trial success.
- Eligibility criteria include various pieces of information other than complexity.

- **If we can extract clinically meaningful information other than complexity from eligibility criteria, it may be useful for predicting clinical trial success.**

Purpose

We target to quantify the eligibility criteria of clinical trials, by constructing a success/failure model for Enrollment.

Agenda

1. AI Applications in the Pharmaceutical Industry

2. Generation and Classification of Chest X-ray Images Using GANs with Limited Training Data
[2-1] Introduction
[2-2] Materials and Methods
[2-3] Results and Discussion

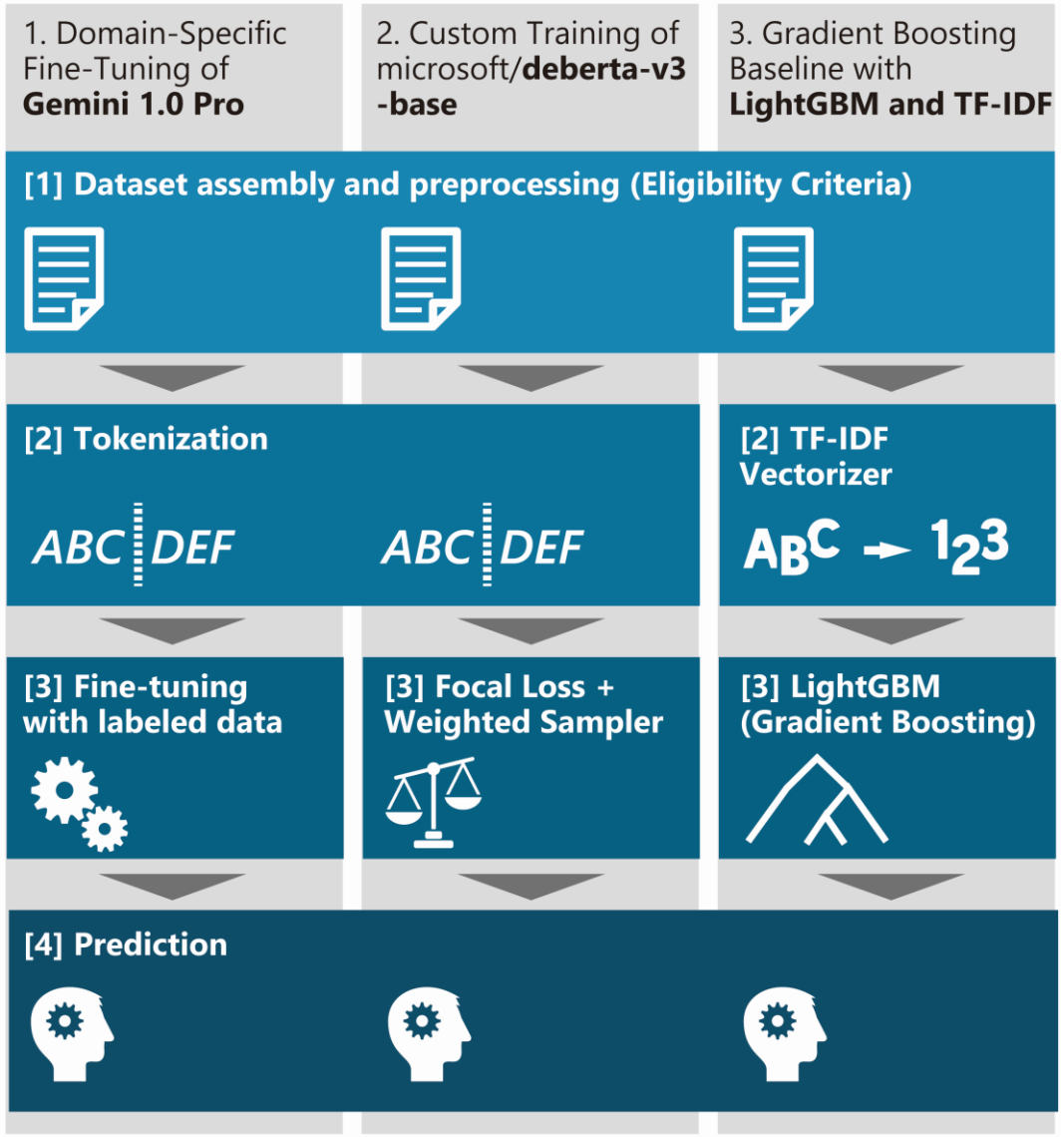
3. Predicting Patient Enrollment Outcomes in Clinical Trials Using Large Language Models
[3-1] Introduction
[3-2] Materials and Methods
[3-3] Results and Discussion

4. Information Gathering at ACL, a Top Conference in Natural Language Processing

5. Conclusion

Task and dataset

■**Data source**:
Aggregate Analysis of ClinicalTrials.gov (AACT)
■**Input**:
free-text eligibility criteria for each trial
■**Outcome**:
binary enrollment success vs. failure
- “COMPLETED” = success
- Trials with plausible accrual attempts but not completed = failure
- Trials that never opened or had “NO PATIENTS ENROLLED” for non-enrollment reasons were excluded
■**Final dataset**:
290,449 trials, labeled by
- Enrollment outcome
- Oncology vs. non-oncology (MeSH neoplasm terms)
- Trial phase (1–4)
For LLM fine-tuning, 50,000 trials were sampled (stratified) for training/validation; remaining trials used as test set



Three modeling approaches

#	models	approches
1	Gemini 1.0 Pro (fine-tuned LLM)	[1] Conversational JSON format: user = eligibility text, model = enrollment label [2] Supervised fine-tuning on Vertex AI with tuned hyperparameters (learning rate, batch size, sequence length) [3] Evaluated using accuracy, precision, recall, F1; trials without valid outputs removed from evaluation
2	DeBERTa-v3-base (open-source transformer)	[1] Hugging Face implementation with max sequence length = 512, padding & truncation [2] 80/20 train–validation split, stratified; separate held-out test set [3] Class imbalance handled via weighted random sampler + Focal Loss [4] Custom Trainer with fp16, gradient accumulation, 10 epochs, learning rate = 5e-6 [5] Threshold tuned on validation set to maximize F1; metrics reported on test set
3	LightGBM + TF-IDF (classical baseline)	[1] Eligibility text → TF-IDF unigram–bigram features [2] 80/20 train–validation split; separate test set with identical preprocessing [3] LightGBM (gbdt): 31 leaves, learning rate = 0.05, feature/bagging fraction = 0.8, up to 1000 iterations [4] Performance evaluated via accuracy, precision, recall, F1, and confusion matrix

Agenda

1. AI Applications in the Pharmaceutical Industry

2. Generation and Classification of Chest X-ray Images Using GANs with Limited Training Data
[2-1] Introduction
[2-2] Materials and Methods
[2-3] Results and Discussion

3. Predicting Patient Enrollment Outcomes in Clinical Trials Using Large Language Models
[3-1] Introduction
[3-2] Materials and Methods
[3-3] Results and Discussion

4. Information Gathering at ACL, a Top Conference in Natural Language Processing

5. Conclusion

Overall performance of three models

Task: predict enrollment success vs. failure from textual eligibility criteria in 290,449 trials

[1] Gemini 1.0 Pro (fine-tuned)
Accuracy: 0.9252, F1: 0.9610
Slightly lower than LightGBM but still strong and robust

[2] DeBERTa-v3-base (customized)
Accuracy: 0.9426, F1: 0.9704
However, these scores are misleading because the model predicts almost all cases as "Enrollment Success"

[3] LightGBM + TF-IDF
Accuracy: 0.9421, F1: 0.9702 (best overall balance)

		All	Oncology	Non-oncology
1) Domain-Specific Fine-Tuning of Gemini 1.0 Pro	Accuracy	0.9252	0.8426	0.9166
	Precision	0.9446	0.8933	0.9387
	Recall	0.9781	0.9344	0.9747
	F1 score	0.9610	0.9134	0.9563
	Confusion matrix	[[759, 12851], [4913, 219102]]	[[108, 833], [490, 6977]]	[[705, 10891], [4331, 166632]]
2) Custom Training of microsoft/deberta-v3-base	Accuracy	0.9426	0.8878	0.9363
	Precision	0.9426	0.8878	0.9363
	Recall	1.0000	1.0000	1.0000
	F1 score	0.9704	0.9406	0.9671
	ROC-AUC	0.6150	0.4852	0.5599
	Confusion matrix	[[0, 13814], [0, 226643]]	[[0, 953], [0, 7543]]	[[0, 11767], [0, 172924]]
3) Gradient Boosting Baseline with LightGBM and TF-IDF	Accuracy	0.9421	0.8825	0.9353
	Precision	0.9430	0.8894	0.9369
	Recall	0.9990	0.9909	0.9982
	F1 score	0.9702	0.9374	0.9666
	ROC-AUC	0.6816	0.6472	0.6707
	Confusion matrix	[[139, 13675], [237, 226406]]	[[24, 929], [69, 7474]]	[[138, 11629], [317, 172607]]

Subgroup analyses and error patterns

Oncology vs. non-oncology, Phases 1–4

[1] Gemini: modest but stable performance in both oncology and non-oncology
[2] DeBERTa: competitive accuracy in some partitions, but driven by predicting almost all trials as success
[3] LightGBM: consistently high accuracy and F1 across most subgroups

		All				Oncology				Non-oncology			
		Phase 1	Phase 2	Phase 3	Phase 4	Phase 1	Phase 2	Phase 3	Phase 4	Phase 1	Phase 2	Phase 3	Phase 4
1) Domain-Specific Fine-Tuning of Gemini 1.0 Pro	Accuracy	0.9509	0.8659	0.9343	0.8965	0.8287	0.7584	0.8672	0.8316	0.9340	0.8736	0.9325	0.8974
	Precision	0.9597	0.9029	0.9462	0.9227	0.8850	0.8575	0.9149	0.8956	0.9444	0.9021	0.9435	0.9208
	Recall	0.9904	0.9532	0.9866	0.9687	0.9275	0.8591	0.9416	0.9209	0.9881	0.9637	0.9876	0.9715
	F1 score	0.9748	0.9273	0.9660	0.9452	0.9058	0.8583	0.9281	0.9081	0.9658	0.9319	0.9650	0.9455
	Confusion matrix	[[8, 234], [54, 5573]]	[[66, 615], [281, 5718]]	[[5, 264], [63, 4645]]	[[20, 308], [119, 3679]]	[[6, 123], [74, 947]]	[[68, 307], [303, 1847]]	[[8, 60], [40, 645]]	[[0, 19], [14, 163]]	[[12, 210], [43, 3570]]	[[56, 591], [205, 5445]]	[[3, 254], [53, 4238]]	[[29, 267], [91, 3103]]
2) Custom Training of microsoft/deberta-v3-base	Accuracy	0.9586	0.8980	0.9455	0.9198	0.8874	0.8512	0.9091	0.8995	0.9425	0.8959	0.9437	0.9149
	Precision	0.9586	0.8980	0.9455	0.9198	0.8874	0.8512	0.9091	0.8995	0.9425	0.8959	0.9437	0.9152
	Recall	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997
	F1 score	0.9789	0.9462	0.9720	0.9582	0.9403	0.9196	0.9524	0.9471	0.9704	0.9451	0.9710	0.9556
	ROC-AUC	0.5713	0.5347	0.4223	0.4797	0.5562	0.4795	0.4545	0.4232	0.4377	0.4668	0.4439	0.4383
	Confusion matrix	[[0, 250], [0, 5793]]	[[0, 695], [0, 6116]]	[[0, 275], [0, 4772]]	[[0, 336], [0, 3851]]	[[0, 132], [0, 1040]]	[[0, 382], [0, 2185]]	[[0, 69], [0, 690]]	[[0, 20], [0, 179]]	[[0, 227], [0, 3720]]	[[0, 668], [0, 5748]]	[[0, 260], [0, 4358]]	[[0, 300], [1, 3236]]
3) Gradient Boosting Baseline with LightGBM and TF-IDF	Accuracy	0.9565	0.8955	0.9459	0.9178	0.8865	0.8535	0.9078	0.8945	0.9407	0.8950	0.9433	0.9124
	Precision	0.9593	0.9008	0.9464	0.9208	0.8886	0.8591	0.9090	0.9031	0.9435	0.8990	0.9446	0.9154
	Recall	0.9969	0.9930	0.9994	0.9964	0.9971	0.9904	0.9986	0.9888	0.9968	0.9944	0.9984	0.9963
	F1 score	0.9777	0.9446	0.9722	0.9571	0.9397	0.9201	0.9517	0.9440	0.9694	0.9443	0.9708	0.9541
	ROC-AUC	0.7680	0.6773	0.6587	0.6420	0.6569	0.6468	0.5847	0.6028	0.7066	0.6672	0.6580	0.6279
	Confusion matrix	[[5, 245], [18, 5775]]	[[26, 669], [43, 6073]]	[[5, 270], [3, 4769]]	[[6, 330], [14, 3837]]	[[2, 130], [3, 1037]]	[[27, 355], [21, 2164]]	[[0, 69], [1, 689]]	[[1, 19], [2, 177]]	[[5, 222], [12, 3708]]	[[26, 642], [32, 5716]]	[[5, 255], [7, 4351]]	[[2, 298], [12, 3225]]

The confusion matrix highlighted in yellow shows the following: the upper-left cell represents cases predicted as Negative that were actually Negative; the upper-right cell represents cases predicted as Positive that were actually Negative; the lower-left cell represents cases predicted as Negative that were actually Positive; and the lower-right cell represents cases predicted as Positive that were actually Positive.

limitations, and future direction

■Limitations

[1] Inputs limited to textual eligibility criteria; no structured data such as site characteristics or timelines
[2] Gemini 1.0 Pro version used did not expose probabilities, preventing full threshold-based analysis
[3] Results for DeBERTa highlight sensitivity to hyperparameters and sampling strategies

■Future work

[1] Integrate structured metadata (site info, historical enrollment, demographics) with text
[2] Explore newer LLMs (e.g., Gemini 2.0, Med-LM, GPT-o1) and use probability outputs for better calibration
[3] Move toward multitask modeling, predicting not only success/failure but also reasons for failed enrollment or withdrawal

Agenda

1. AI Applications in the Pharmaceutical Industry

2. Generation and Classification of Chest X-ray Images Using GANs with Limited Training Data
[2-1] Introduction
[2-2] Materials and Methods
[2-3] Results and Discussion

3. Predicting Patient Enrollment Outcomes in Clinical Trials Using Large Language Models
[3-1] Introduction
[3-2] Materials and Methods
[3-3] Results and Discussion

4. Information Gathering at ACL, a Top Conference in Natural Language Processing

5. Conclusion

ACL 2025 Travel Report

ACL (Association for Computational Linguistics) is a top conference in natural language processing (NLP) and computational linguistics.

Conference locations typically rotate across the U.S., Asia, and Europe.
- This year: Vienna
- Next year: San Diego

As the use of NLP has been increasing in analytics projects for the Clinical Development and Sales Divisions, I applied to attend the conference.

https://2025.aclweb.org/

Conference Schedule and Networking

- **6 days in total, with sessions from 9:00–18:00 and networking over lunch and dinner**
- **Schedule**

Day 1: Tutorials, Welcome reception

Day 2: Keynote, poster & oral sessions

Day 3: Keynote, poster & oral sessions, social event

Day 4: Poster & oral sessions, closing session

Day 5: Workshops

Day 6: Workshops

The number of posters was very large: about 450 posters per session (in the venue shown above ×2).

Use of LLMs in Healthcare (1/2)

EC-RAFT: Automated Generation of Clinical Trial Eligibility Criteria through Retrieval-Augmented Fine-Tuning

Challenge

[1] Eligibility criteria (EC) for clinical trials are a critical factor in determining whether patients can participate.

[2] However, traditional EC creation has several issues:

- Highly dependent on expert knowledge and labor-intensive
- Lack of consistency
- Sometimes overly restrictive, which can hinder patient recruitment

Objective

Using large-scale clinical trial data (approx. 210,000 trials from ClinicalTrials.gov), the study developed EC-RAFT (Retrieval-Augmented Fine-Tuning) to automatically generate clinically plausible and high-quality eligibility criteria.

Results

[1] EC-RAFT significantly outperformed base models (e.g., LLaMA 3.1, Gemini 1.5) and conventional fine-tuning methods.

[2] Using LLM-as-a-Judge, the model's outputs showed high agreement with human annotators (correlation coefficient ≥ 0.90).

[3] Even relatively small models achieved performance comparable to or better than large models by introducing intermediate reasoning steps (Reasoning Paths).

Discussion

[1] Combining Retrieval-Augmented Fine-Tuning with intermediate reasoning generation improved both the accuracy and clinical validity of EC generation.

[2] A method for converting free-text evaluations by LLMs into a standardized format enabled quantitative evaluation while preserving clinical nuance.

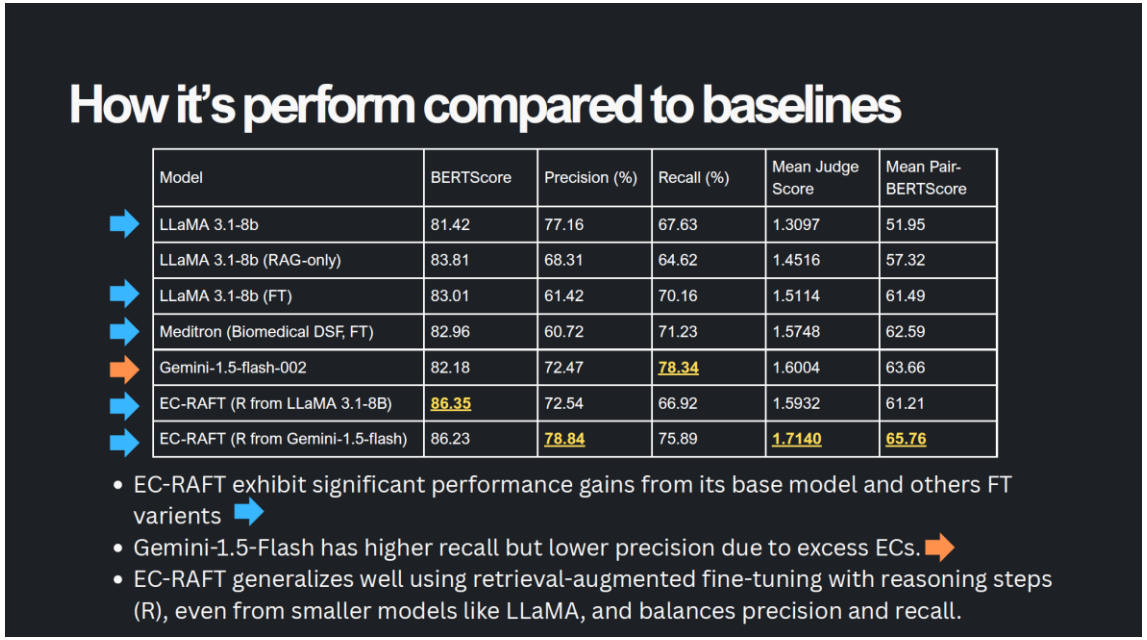
Potential Applications in Chugai

[1] Clinical Trial Design Support

Automatically generate candidate eligibility criteria during trial planning for rapid review.

[2] Optimization of Patient Recruitment

Automatically check whether ECs are overly restrictive, helping improve recruitment success rates.



Lekuthai, N., Pewngam, N., Sokrai, S., & Achakulvisut, T. (2025, July). EC-RAFT: Automated Generation of Clinical Trial Eligibility Criteria through Retrieval-Augmented Fine-Tuning. In Findings of the Association for Computational Linguistics: ACL 2025 (pp. 9432-9444).

Use of LLMs in Healthcare (2/2)

Aligning AI Research with the Needs of Clinical Coding Workflows Eight Recommendations Based on US Data Analysis and Critical Review

- **Challenge**

Clinical coding (assigning ICD codes) in real-world healthcare is highly burdensome because:

- Clinical notes are long and complex
- Descriptions can be ambiguous
- Coding guidelines are frequently updated
- ICD is a large and detailed code system (over 100,000 codes)

- **Objective**

Based on analyses using U.S. MIMIC data and a review of prior research, the study clarifies how current evaluation methods and automation approaches should be improved to align better with real-world needs.

- **Results**

The presentation proposed several recommendations to bridge the gap between research and practice:

[1] Report threshold-independent metrics

[2] Compare performance directly with human coders

[3] Explore partial or subset-based automation

[4] Build diverse datasets

- **Discussion**

[1] Current research is biased toward specific evaluation setups and does not fully meet real-world requirements for comprehensiveness, accuracy, and usability.

[2] Full automation is not realistic; hybrid approaches that assume collaboration with human coders are more practical.

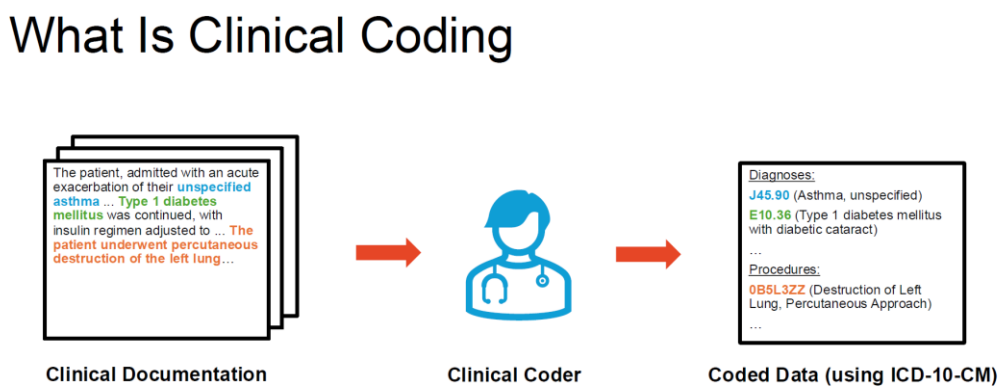
- **Potential Applications in Chugai**

[1] Improving Clinical Trial Data Quality

Automating standardized ICD coding for trial data can streamline cross-trial comparisons and regulatory submissions.

[2] Internal AI R&D

Applying LLM-based auto-coding to Chugai's clinical data as "human-in-the-loop AI tools" could build a practical AI foundation aligned with on-site needs.



Gan, Y., Rybinski, M., Hachey, B., & Kummerfeld, J. K. (2025, July). Aligning AI research with the needs of clinical coding workflows: Eight recommendations based on US data analysis and critical review. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 909-922).

Leveraging Knowledge Graphs × LLMs (1/2)

Knowledge Graph Retrieval-Augmented Generation for LLM-based Recommendation

Challenges

[1] Hallucination: LLM-based recommender systems may recommend items that do not exist

Example: recommending a non-existent movie “Godmother” to a user who watched “The Godfather.”

[2] Lack of Knowledge: Insufficient domain knowledge (e.g., movies, pharma, healthcare).

[3] Data Limitations: Training corpora for recommendation are limited during LLM pretraining, leading to poor performance.

Objectives

[1] Propose K-RagRec, a Retrieval-Augmented Generation (RAG) method leveraging Knowledge Graphs (KG), to improve the accuracy and reliability of LLM-based recommendation.

[2] Retrieve top-K similar subgraphs from the KG and use them in recommendation generation, to reduce hallucinations and better incorporate up-to-date and domain-specific knowledge.

Results

[1] Overall Performance: K-RagRec outperformed conventional LLM-based recommender systems in both recommendation accuracy and consistency.

[2] Efficiency: Combining retrieval and generation enabled efficient recommendation.

[3] Generalization: The method showed potential applicability across different domains.

Discussion

[1] RAG techniques are effective for challenges that are difficult for LLMs alone, such as hallucination, knowledge gaps, and cold start problems.

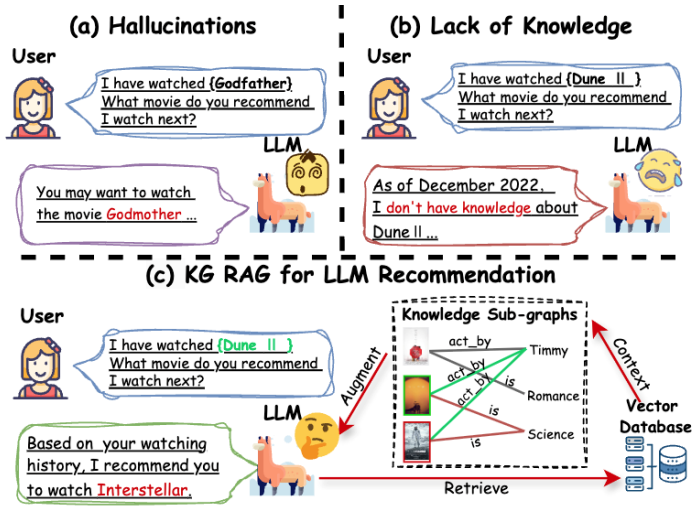
[2] Using KGs greatly improves the reliability and correctness of recommendations.

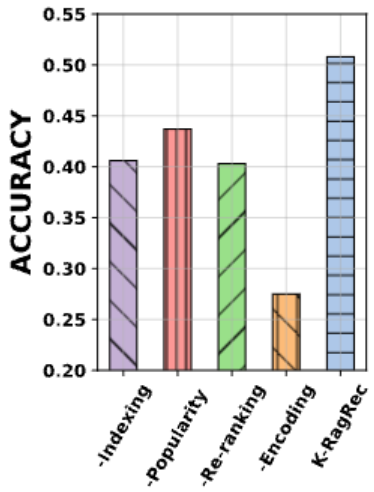
[3] There is future potential for more efficient and scalable RAG methods.

Potential Applications in Chugai

[1] Clinical Trial Design Support: Extract patient populations and treatment combinations from KGs and generate candidate trial designs.

[2] Knowledge Management: Structuring internal and external knowledge (papers, patents, regulatory information) into KGs allows researchers and MRs to instantly access needed information.





Wang, S., Fan, W., Feng, Y., Shanru, L., Ma, X., Wang, S., & Yin, D. (2025, July). Knowledge graph retrieval-augmented generation for llm-based recommendation. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 27152-27168).

Leveraging Knowledge Graphs × LLMs (2/2)

GNN-RAG Graph Neural Retrieval for Efficient Large Language Model Reasoning on Knowledge Graphs

- **Challenge**

[1] In Knowledge Graph Question Answering (KGQA), we must extract relevant parts from a large KG and use language models to generate answers.

[2] Current challenges include:

- Graph Neural Networks (GNNs): Strong at handling large graphs but weak at understanding natural language.

- Large Language Models (LLMs): Excellent at natural language understanding but computationally expensive and inefficient for reasoning directly over large graphs.

- **Objective**

Reuse GNNs as Retrievers and combine them with LLMs to achieve efficient and accurate KGQA. Especially aim to improve performance on multi-hop reasoning and questions involving multiple entities.

- **Results**

Proposed GNN-RAG (Graph Neural Retrieval-Augmented Generation).

- GNNs efficiently extract relevant subgraphs.

- LLMs are used only for final reasoning based on the extracted subgraphs.

- **Discussion**

[1] By using GNNs as Retrievers, we can clearly separate roles:

- GNNs for understanding graph structure

- LLMs for natural language understanding and generation

[2] Compared to using LLMs alone, this approach maintains or improves accuracy while reducing computational cost.

[3] Future directions include hybrid approaches with LLM-based retrievers and optimizing routing for long-context scenarios.

- **Potential Applications in Chugai**

[1] Knowledge Exploration for Drug Discovery and Clinical Trial Design

- Efficiently extract relevant subgraphs from large biological KGs (gene–disease–drug relationships) using GNNs.

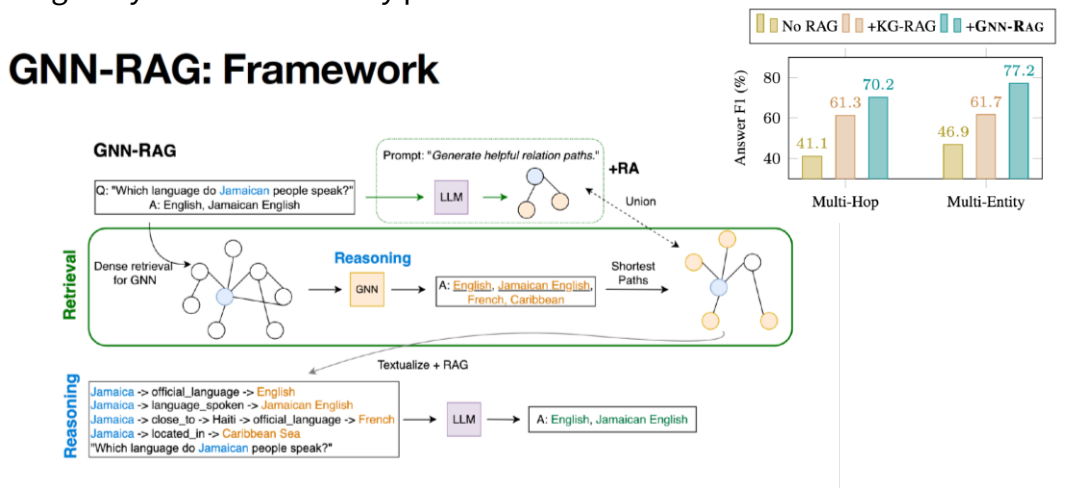
- LLMs then interpret these subgraphs based on researchers' natural language queries.

- Helps accelerate candidate drug discovery and side effect prediction.

[2] Supporting Interpretation of RWD

- Automatically extract important patterns from KGs built from medical data (patient characteristics, treatment courses, adverse events).

- Researchers can ask natural language questions and quickly obtain insights on eligibility criteria or efficacy prediction.



Mavromatis, C., & Karypis, G. (2025, July). Gnn-rag: Graph neural retrieval for efficient large language model reasoning on knowledge graphs. In Findings of the Association for Computational Linguistics: ACL 2025 (pp. 16682-16699).

Agenda

1. AI Applications in the Pharmaceutical Industry

2. Generation and Classification of Chest X-ray Images Using GANs with Limited Training Data
[2-1] Introduction
[2-2] Materials and Methods
[2-3] Results and Discussion

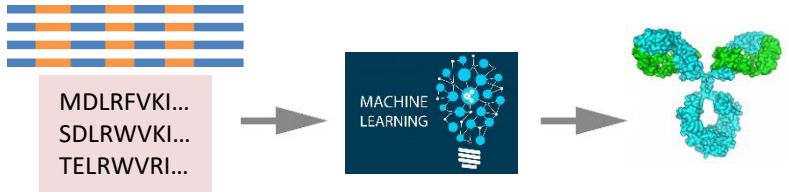
3. Predicting Patient Enrollment Outcomes in Clinical Trials Using Large Language Models
[3-1] Introduction
[3-2] Materials and Methods
[3-3] Results and Discussion

4. Information Gathering at ACL, a Top Conference in Natural Language Processing

5. Conclusion

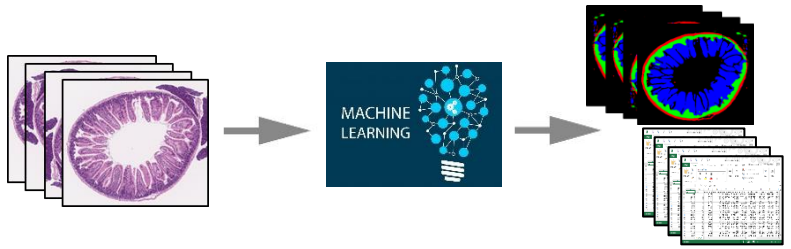
Data Science at Chugai

At Chugai, we tackle various challenges through big data analytics and AI model development, not limited to any single field.



■ Drug Discovery

- Developing AI technologies to improve R&D efficiency
- Developing NLP models and analyzing real-world data



■ Pathology / Pharmacology

- Developing novel technologies through image analysis

■ Digital Biomarkers

- Developing visualization technologies for patient conditions to improve clinical trial efficiency
- Utilizing wearable device data and more

■ Marketing

- Analyzing marketing data to improve ROI for B2B activities targeting physicians and medical institutions
- Demand forecasting, sales talk analysis, and customer data analytics

Our goal is to make data science a central driver for creating breakthrough medicines that could not have been discovered with conventional technologies alone.

References

[Ayan 19] Ayan, E., and Ünver, H. M.: Diagnosis of pneumonia from chest X-ray images using deep learning, *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pp. 1－5 (2019)

[Brock 18] Brock, A., Donahue, J., and Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis, *arXiv preprint arXiv:1809.11096* (2018)

[Gan 25] Gan, Y., Rybinski, M., Hachey, B., & Kummerfeld, J. K. (2025, July). Aligning AI research with the needs of clinical coding workflows: Eight recommendations based on US data analysis and critical review. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 909-922).

[Greenspan 16] Greenspan, H., van Ginneken, B., and Summers, R. M.: Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique, *IEEE transactions on medical imaging*, Vol. 35, No. 5, pp. 1153－1159 (2016)

[Karras 20] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T.: Analyzing and improving the image quality of stylegan. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110－8119 (2020)

[Lawton 14] Lawton, T. J., Acs, G., Argani, P., Farshid, G., Gilcrease, M., Goldstein, N., Koerner, F., Rowe, J. J., Sanders, M., Shah, S. S., and Reynolds, C.: Interobserver variability by pathologists in the distinction between cellular fibroadenomas and phyllodes tumors, *International journal of surgical pathology*, Vol. 22, No. 8, pp. 695－698 (2014)

[Lekuthai 25] Lekuthai, N., Pewngam, N., Sokrai, S., & Achakulvisut, T. (2025, July). EC-RAFT: Automated Generation of Clinical Trial Eligibility Criteria through Retrieval-Augmented Fine-Tuning. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 9432-9444).

[Liu 20] Liu, B., Zhu, Y., Song, K., and Elgammal, A.: Towards faster and stabilized gan training for high-fidelity few-shot image synthesis, *International Conference on Learning Representations* (2020)

[Mavromatis 25] Mavromatis, C., & Karypis, G. (2025, July). Gnn-rag: Graph neural retrieval for efficient large language model reasoning on knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 16682-16699).

[McKinney 20] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., Fauw, J. D., and Shetty, S.: International evaluation of an AI system for breast cancer screening, *Nature*, Vol. 577, No. 7788, pp. 89－94 (2020)

[Niazi 19] Niazi, M. K. K., Parwani, A. V., and Gurcan, M. N.: Digital pathology and artificial intelligence, *The lancet oncology*, Vol. 20, No. 5, pp. e253－e261 (2019)

[Rajpurkar 17] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., and Ng, A. Y.: CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning, *arXiv preprint arXiv:1711.05225* (2017)

[Topol 19] Topol, E. J.: High-performance medicine: the convergence of human and artificial intelligence, *Nature medicine*, Vol. 25, No. 1, pp. 44－56 (2019)

[Wang 17] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M.: ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097－2106 (2017)

[Wang 25] Wang, S., Fan, W., Feng, Y., Shanru, L., Ma, X., Wang, S., & Yin, D. (2025, July). Knowledge graph retrieval-augmented generation for llm-based recommendation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 27152-27168).

創造で、想像を超える。

CHUGAI 中外製薬

Roche ロシュ グループ