Journal of Advanced College of Engineering and Management

Hybrid GAN-Transformer For Synthetic Medical Text Generation To Address Data Scarcity In Healthcare

[1*]Rajesh Raskoti, [2]Kobid Karkee, [3]Shreekrishna Timilsina
[1]MSc. in informatics and intelligent Systems Engineering, Thapathali Campus, Institute of Engineering, Tribhuvan University Kathmandu, Nepal
[2]Asst. Professor, Department of Electronics and Computer Engineering, Thapathali Campus, Institute of Engineering, Tribhuvan University Kathmandu, Nepal
[3]Lecturer, Department of Electronics and Computer Engineering, Himalaya College Of Engineering, Tribhuvan University Lalitpur, Nepal
Email: [2]karkeekobid09@tcioe.edu.np , [3]sktimilsina2010@gmail.com
Corresponding email: [1]rajeshraskoti100@gmail.com*

DOI: 10.3126/jacem.v12i01.93910

Abstract
Medical research often struggles to access patient data because privacy laws like HIPAA and GDPR protect sensitive health information. This study proposes a system that generates realistic synthetic medical notes without risking patient privacy. We developed a hybrid model using two neural networks: DistilGPT-2 as the text generator and BioBERT as the evaluator. BioBERT was trained on clinical notes from the MIMIC-III database. Training was completed in three phases. First, the generator learned basic medical writing and achieved a best validation perplexity of 8.87 (from an initial 9.34), with training perplexity reducing from 19.55 to 6.90. Second, we intentionally weakened the evaluator to maintain balance between the models. Third, both networks were trained at full strength with added controls. However, the evaluator became too accurate, which disrupted the training process. The generator's validation perplexity increased to 47.05, and text diversity (Distinct-1) decreased from 0.271 (Phase 2 start) to a minimum of 0.110 (Phase 3, epoch 8).

Keywords*—Generative Adversarial Network, Synthetic data, Transformer*

1. INTRODUCTION

The fast growth of AI in healthcare has created new opportunities in clinical decisions, disease prediction, and medical text analysis. Much of this progress depends on Natural Language Processing which requires large and well prepared datasets to work effectively. However healthcare data is highly sensitive and protected by strict privacy laws such as HIPAA and GDPR. Because of these rules patient information is difficult to access in the amounts needed to train strong NLP models. This gap between the need for data and its limited availability has become a major barrier to medical AI research especially in places where clinical datasets are limited or not available.

Synthetic medical text generation has become a promising way to solve the shortage of healthcare data. It creates artificial but realistic patient records and clinical notes that can expand small datasets without risking patient privacy. Although there has been progress in this field existing methods still have clear limits. Models based only on Generative Adversarial Networks often produce varied text but may lack clear meaning and logical flow. In contrast Transformer based models understand context well but may generate less diverse text when training data is limited. When used alone neither approach fully meets the complex language and clinical needs of medical text generation.

This research addresses this limitation by proposing a hybrid GAN Transformer framework for synthetic medical text generation. The idea is to combine the strengths of both models. GANs help create diverse text while Transformers improve context and meaning. By joining these strengths the system aims to generate high quality and privacy safe clinical text.

2. LITERATURE REVIEW

Libo Ren et al. [1] apply Bio-ClinicalBERT with varied masking strategies for synthetic clinical letter generation, where encoder-only models achieved a BERTScore of 0.85, outperforming encoder-decoder architectures. Despite only 204 clinical letters, the model effectively generates diverse, de-identified synthetic records for expanding annotated datasets.

Chao Yan et al. [2] present a tutorial on synthetic EHR generation using the EMR-WGAN model on MIMIC-IV, evaluated across five datasets on utility, privacy, and fairness, showing strong correlation and prediction accuracy but lower medical concept richness. The framework is limited to static EHR data and inadequately addresses rare clinical events and computational overhead.

Xiang Yue et al. [3] fine-tune GPT2-Large using DP-SGD with epsilon of four, achieving classification performance comparable to non-DP text across three feedback attributes with minimal utility loss. Limitations include shorter generated texts, evaluation on only two datasets, struggles with rare distributions, and unaddressed computational overhead.

Suranga N. Kasthurirathne et al. [4] use enhanced SeqGAN to generate 1,092 positive and 5,001 negative synthetic Salmonella lab messages, with BLEU/GLEU scores and 70–80% feature overlap confirming resemblance to real data, and Random Forest classifiers performing comparably on synthetic records. Generalizability is limited by seven uniform report types and only 69.6% feature coverage.

Mohammad Loni et al. [5] review 52 studies on generative AI for synthesizing medical text, time series, and longitudinal healthcare data, emphasizing evaluation in practical healthcare environments. However, the review inadequately addresses ethical dimensions and potential biases in generated outputs, risking fairness in downstream clinical analyses.

Y Kim et al. [6] evaluate distilGPT2, BioGPT, CerebroGPT, and ChatGPT for synthetic cerebrovascular medical reports, where distilGPT2 achieved the highest TF-IDF of 0.4549

and ROUGE-1/2/L of 0.4217/0.2618/0.4146, while ChatGPT excelled in grammatical coherence. The narrow disease focus limits broader generalizability.

3. METHODOLOGY

A. Theoretical Formulations

Following PHI removal, a DistilGPT-2-based generator leverages self-attention to capture long-range linguistic dependencies, while a BioBERT discriminator drives adversarial training to improve clinical realism. The Transformer architecture processes tokenized medical input through multi-head self-attention, feed-forward networks, and residual connections, generating contextually coherent synthetic text via masked self-attention and encoder-decoder attention. The system produces a labeled synthetic medical corpus annotated for clinical entities, supporting diagnostic model training and clinical simulation, representing a scalable approach to responsible synthetic data generation in healthcare.

B. Mathematical Modeling

The given approach fuses two major paradigms in generative modeling, Generative Adversarial Networks (GANs) and Transformers. A GAN consists of a generator G and a discriminator D playing a minimax game. The objective function is:

$$\frac{min}{G}\frac{max}{D}E_{x\sim pdata(x)}[logD(x)] + E_{z\sim p_z(z)}[log(1 - D(G(z)))] \quad (1)$$

In our instance, we utilise a Transformer model (DistilGPT-2) and a pre-trained BioBERT model that has been fine-tuned for the discrimination of medical text. The Transformer's attention mechanism makes things more coherent in context, and the GAN framework promotes variety.

The Generator G_θ takes a noise vector z sampled from a Gaussian distribution and outputs a synthetic text sequence $\hat{x} = G_\theta(z)$. The Discriminator D_φ receives either real data x or synthetic data $\hat{x}$, and outputs a probability of the data being real. The discriminator is optimized using cross-entropy loss:

$$L_D = -E_{x\sim pdata}\left[logD_\emptyset(x)\right] - E_{z\sim p_z}\left[log\left(1 - D_\emptyset\left(G_\theta(z)\right)\right)\right] \quad (2)$$

For the generator, the objective is to maximize the discriminator's mistake:

$$L_G = -E_{z\sim p_z}\left[log\left(1 - D_\emptyset\left(G_\theta(z)\right)\right)\right] \quad (3)$$

An attention function takes a query and a set of key-value pairs, all represented as vectors, and produces an output vector. This output is calculated as a weighted sum of the values, with the weights derived from a compatibility function that evaluates the relationship between the query and each key [7].

C. System Block Diagram

The system block diagram for the whole system is shown here.

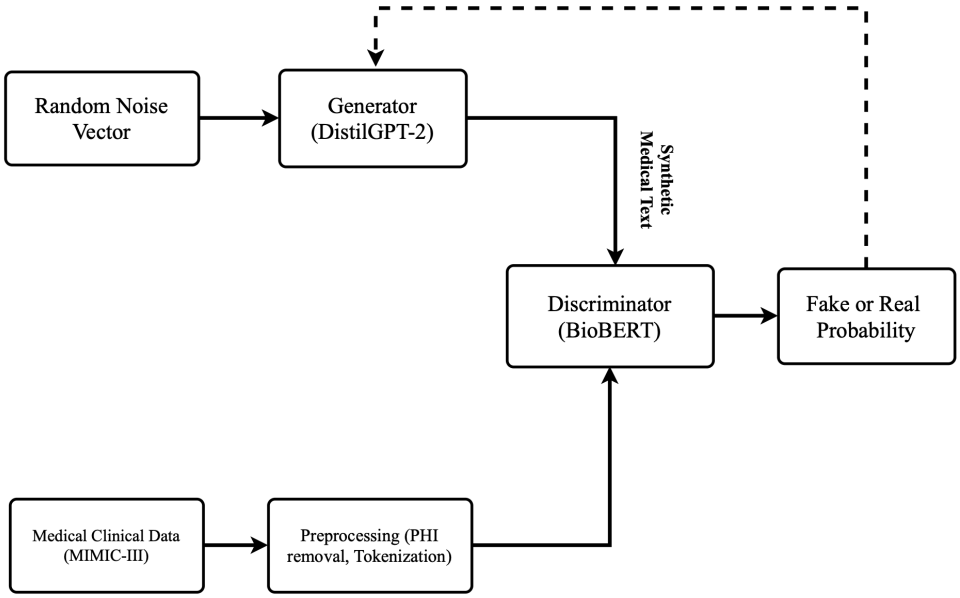

Figure 3.1: System Block Diagram

This diagram illustrates the hybrid GAN-Transformer architecture for generating synthetic medical text. The real clinical data from the MIMIC-III dataset is preprocessed (removal of PHI and tokenization) and input into the discriminator based on BioBERT, and a random noise vector is input into the generator based on DistilGPT-2 to generate synthetic medical text. Both the real and synthetic medical texts are input into the discriminator, which predicts a probability of being real or fake, and this result is back-propagated to update the generator to generate more realistic clinical text. The solid arrow represents forward data flow and dashed arrow represents gradient back-propagation.

D. Transformer Architecture

a. Bidirectional Encoder Representations from Transformers for Biomedical Text Mining

We used BioBERT described by Lee et. al [8] as a discriminator. The classifier uses a Sequential block with two dropout layers (0.2), two linear layers reducing dimensions from hidden size to 256 and then to 64, each followed by LayerNorm and ReLU activation. A final dropout (0.1) and linear layer map to a single output for binary classification. The top 6 layers were frozen during phase 2 training

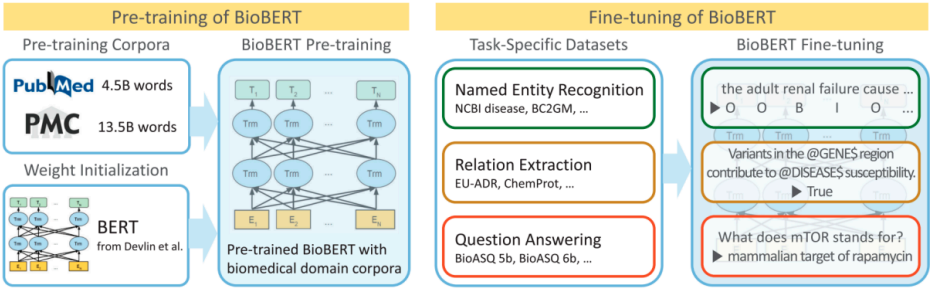

Figure 3.2: Overview of pre-training and fine tuning of BioBERT[8]

b. DistilGPT-2

The Generator uses DistilGPT-2 (loaded via AutoModel For CausalLM) with a Sequential noise layer consisting of a Linear layer, LayerNorm, GELU activation, and a second Linear layer. A 128-dimensional random noise input is used to improve generation diversity, and an additional linear layer is added to refine the final output.

E. Dataset Explanation

a. Contents of datasets.

The primary challenge involves adapting the MIMIC-III dataset de-identified ICU patient data for privacy-preserving synthetic medical text generation, as no dedicated dataset exists for this purpose. The project utilizes structured clinical notes representing diverse medical conditions and treatments, ensuring broad applicability for training a sentence-level synthetic text generation model in healthcare.

F. Evaluation Metrics

For the evaluation of the synthetic medical text generated by the hybrid-GAN model, we use

$$BLEU = BP.exp\left\{\sum_{n=1}^{N} w_n logP_n\right\} \quad (4)$$

$$CVS = \frac{1}{M}\sum_{1}^{M} F\left(\left(\hat{x}_j\right) BioBERT_{med} = valid\right) \quad (5)$$

$$Distinct - 1 = \frac{number\ of\ unique\ unigram}{total\ number\ of\ unigrams} \quad (6)$$

$$Distinct - 2 = \frac{number\ of\ unique\ bigram}{total\ number\ of\ bigrams} \quad (7)$$

Distinct-N metrics were computed over a held-out generation set at fixed evaluation checkpoints, 50 samples every 500 training steps in Phase 2, and 100 samples per epoch in Phase 3. Higher Distinct-N values indicate lower inter-sample repetition,

reflecting broader vocabulary coverage. Diversity tracking also serves as an indirect privacy indicator: highly repetitive outputs suggest the model may be memorizing individual training records from MIMIC-III, whereas diverse outputs confirm generalization beyond specific patient notes, a key transparency requirement for privacy-preserving synthetic text generation.

G. Multi-phase adversarial training strategy

The training comprised three phases, each using a structured discriminator control strategy to prevent mode collapse. (A) Generator-to-Discriminator Update Ratio: the generator was updated k times per discriminator update, with k=3 in Phase 2 and k=2 in Phase 3. (B) Layer Freezing (Phase 2 only): the top 6 of 12 BioBERT encoder layers were frozen during Phase 2 to restrict discriminator capacity, then all layers were unfrozen in Phase 3. (C) Adaptive Learning Rate Control: discriminator accuracy was monitored every 500 training steps (Phase 2) or every 200 seconds (Phase 3), if accuracy exceeded 80%, the generator learning rate was increased and the discriminator learning rate decreased, with opposite adjustments if accuracy fell below 60%, using adjustment factors of 1.5 in Phase 2 and 1.3 in Phase 3.

H. Privacy Mechanism

This study implements a multi-layer privacy protection strategy. At the preprocessing stage, PHI is removed from all MIMIC-III source texts using regular-expression-based de-identification. During training, gradient clipping is applied to both the generator (C=1.0) and discriminator (C=0.5) at every step, bounding each model's sensitivity to individual samples. Additionally, the R1 gradient penalty (gamma=1.0) applied in Phase 3 penalizes large gradients on real data, reducing the risk of the discriminator memorizing individual training records.

I. Label smoothing was applied to discriminator target

Real samples were assigned a label of 0.85 (rather than 1.0) and fake samples a label of 0.15 (rather than 0.0), reducing discriminator overconfidence and improving gradient signal to the generator.

4. RESULTS AND DISCUSSION

This section presents the results from using the hybrid-GAN architecture, which was specifically designed to generate synthetic medical text.

A. Explicit failure analysis and adversarial dynamics interpretation

a. Generator loss and discriminator loss over training phases

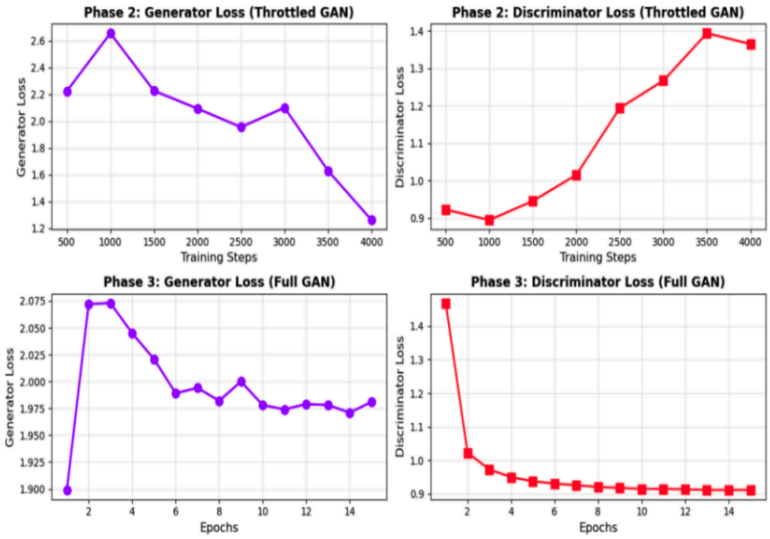

Figure 4.1: Generator loss and discriminator loss

Figure 4.1 shows GAN loss curves across two phases. In Phase 2, generator loss falls from 2.6 to 1.3, while discriminator loss rises from 0.9 to 1.4, suggesting better balance. In Phase 3, generator loss spikes then stabilizes near 1.97–1.98, and discriminator loss drops from 1.5 to 0.9 before flattening. Overall, the generator improves, making discrimination harder.

b. Discriminator dominance despite handicapping in Phase 2

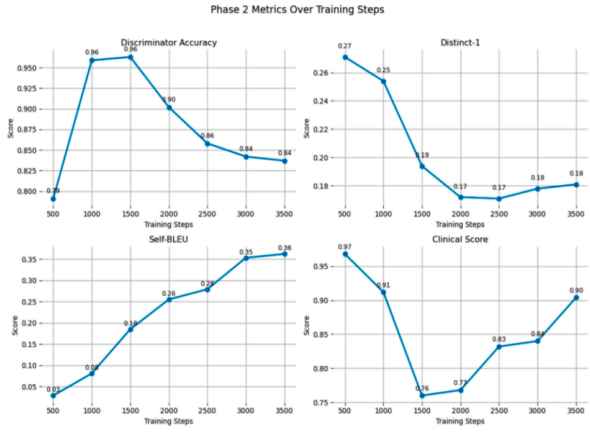

Figure 4.2: Training Metrics in phase 2

Although layer freezing and generator update ratio (k=3) partially reduced discriminator dominance by step 3500 (accuracy recovering to 0.84), the discriminator reached 96% accuracy at step 1500, confirming that Phase 2 controls were insufficient to fully prevent discriminator overpowering. Distinct-1 drops from 0.27 and stabilizes near 0.18, showing lower but steady diversity. Self-BLEU

increases from 0.03 to 0.36, meaning outputs become more similar. The clinical score falls from 0.97 to 0.76, then recovers to 0.90 by step 3500.

c. Generator gradient collapse in phase 3

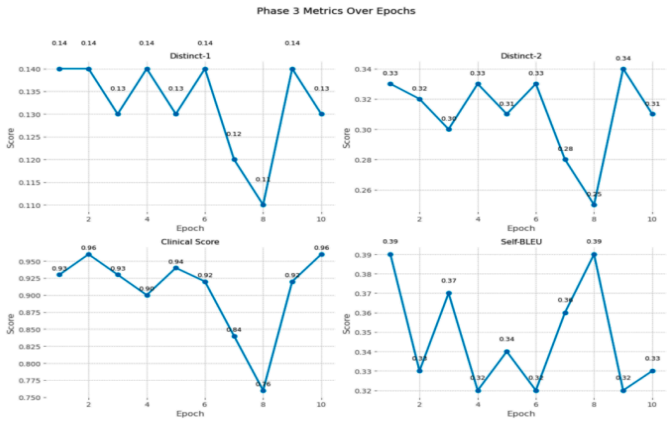

Figure 4.3: Training Metrics in phase 3

During Phase 3 (10 epochs), the metrics fluctuated, with a notable drop at epoch 8. Distinct-1 stayed nearly constant (0.11–0.14), hitting its minimum at epoch 8. Distinct-2 ranged from 0.25 to 0.34, lowest at epoch 8 and highest at 0.34. The Clinical Score varied most, from 0.96 (epochs 2 and 10) down to 0.76 (epoch 8). Self-BLEU oscillated between 0.32 and 0.39, peaking at epochs 1 and 9.

d. Severe mode collapse problem

The model was trained as a basic language model for three epochs in Phase 1. The training perplexity went down from 19.55 to 6.90, and the validation perplexity went up from 9.34 to a best value of 8.87. The samples that were made after training showed a lot of variety with a Self-BLEU score of only 0.002. They also had a good meaning score of 0.923 and a clinical score of 0.894, which made them look very clinical. During phase 2, the GAN training went on for 3500 steps. The discriminator became very strong over time, with accuracy rising to around 96% at some points. Diversity got worse as training went on. Phase 2 demonstrated that even with layer freezing and a 3:1 generator update ratio, the discriminator consistently dominated, driving diversity down and output repetition up throughout training. In Phase 3, the full GAN was trained for up to 10 epochs, but it stopped early after epoch 10 because performance didn't improve anymore. Diversity stayed very low the whole time (Distinct between 0.11 and 0.14, Distinct-2 between 0.25 and 0.34). All metrics dropped noticeably at epoch 8 and only partly recovered later. At epoch 2, the best-performing checkpoint recorded a Clinical Validity Score of 0.96, Distinct-1 of 0.130, and Distinct-2 of 0.300 (at epoch 2), representing the most balanced combination of clinical quality and output diversity across all training epochs. Final samples showed Self-BLEU of 0.299, Embedding Score of 0.933, and Clinical Score of 0.848.

B. Quantitative evaluation

a. Overall metric summary

Table 1: Overall metric summary during training phases

Metrics	Phase 1	Phase 2	Phase 3 (final)
Self-BLEU	0.002	0.419	0.299
Embedding Score	0.923	0.944	0.933
Clinical Score	0.894	0.784	0.848
Distinct - 1	-	0.181	0.130
Distinct - 2	-	-	0.310

† Distinct-2 was not tracked during Phase 1 and Phase 2, as inter-sample bi-gram diversity evaluation was introduced from Phase 3 onward.

b. Ablation study: No transformer, Transformer-only and hybrid architecture

Table 2: Metric summary during ablation study

Metrics	Clinical Validity Score	Self-BLEU
Generator only model	0.894	0.002
LSTM-GAN	0.650	0.550
GAN-Transformer hybrid	0.848	0.299

Notably, the GAN-Transformer hybrid does not surpass the generator- only model in clinical validity (0.848 vs. 0.894), indicating that adversarial training introduces a quality-diversity trade-off: the hybrid produces less repetitive outputs (Self-BLEU 0.299 vs. 0.002 in generator-only, though generator-only diversity is driven by unconstrained generation rather than adversarial pressure) at the cost of some clinical precision. This trade-off is an inherent challenge of GAN-based text generation.

5. CONCLUSION

This study developed a hybrid GAN-Transformer system for privacy-preserving synthetic medical text generation, combining a DistilGPT-2 generator with a BioBERT discriminator trained on MIMIC-III clinical notes across three phases. In Phase 1,

language model pre-training achieved a best validation perplexity of 8.87 (from an initial 9.34), establishing a strong clinical language foundation. Phase 2 adversarial training maintained discriminator accuracy between 0.84 and 0.96, with Distinct-1 stabilizing near 0.181, confirming that controlled discriminator updates preserved generation quality during early adversarial training. Phase 3 full adversarial training achieved its best-performing checkpoint at epoch 2, where the Clinical Validity Score reached 0.96, Distinct-1 = 0.130, and Distinct-2 = 0.300 (at epoch 2), representing the most balanced trade-off between clinical quality and output diversity. The final Phase 3 samples yielded a Clinical Validity Score of 0.848, Embedding Score of 0.933, and Self-BLEU of 0.299. Compared to the LSTM-GAN baseline (CVS = 0.650, Self-BLEU = 0.550), the hybrid model achieved superior clinical accuracy while producing more natural text variety. However, a gradient collapse at epoch 8 caused Distinct-1 to reach its minimum (0.110), revealing the fundamental challenge of maintaining adversarial balance in discrete text generation.

6. SUGGESTIONS AND RECOMMENDATIONS

Future research should prioritize implementation of larger generators such as GPT-Neo 1.3B or architecturally lighter discriminators, along with training on longer token sequences. Gradient flow challenges with discrete tokens can be addressed through reinforcement learning approaches like SeqGAN or REINFORCE, supplemented by stabilization techniques including spectral normalization and Wasserstein loss with gradient penalties. Immediate integration of differential privacy via DP-SGD is recommended to assess privacy-utility trade-offs, alongside validation studies involving medical professionals to verify that generated clinical text meets real-world quality and safety standards.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Department of Electronics and Computer Engineering, Thapathali Campus, Institute of Engineering and Himalaya College of Engineering for providing the necessary infrastructure and academic environment.

REFERENCES

[1] Libo Ren, Samuel Belkadi, Lifeng Han, Warren Del-Pinto, and Goran Nenadic. Synthetic4health: Generating annotated synthetic clinical letters. arXiv preprint arXiv:2409.09501, 2024.

[2] Chao Yan, Ziqi Zhang, Steve Nyemba, and Zhuohang Li. Generating synthetic electronic health record data using generative adversarial networks: Tutorial. JMIR AI, 3:e52615, 2024.

[3] Xiang Yue, Huseyin A Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. Synthetic text generation with differential privacy: A simple and practical recipe. arXiv preprint arXiv:2210.14348, 2022.

[4] Suranga N Kasthurirathne, Gregory Dexter, and Shaun J Grannis. Generative adversarial networks for creating synthetic free-text medical data: a proposal for collaborative research and re-use of machine learning models. AMIA Summits on Translational Science Proceedings, 2021:335, 2021.

[5] Mohammad Loni, Fatemeh Poursalim, Mehdi Asadi, and Arash Gharehbaghi. A review on generative AI models for synthetic medical text, time series, and longitudinal data. npj Digital Medicine, 8(1):1–10, 2025.

[6] Byoung-Doo Oh, Gi-Youn Kim, Chulho Kim, and Yu-Seop Kim. How to use language models for synthetic text generation in cerebrovascular disease-specific medical reports. In Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024), pages 10–17, 2024.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Sep. 2019.