Uncertainty-Aware Pre-Trained Foundation Models for Patient Risk Prediction via Gaussian Process

Jiaying Lu*
Emory University
jiaying.lu@emory.edu

Shifan Zhao
Emory University
shifan.zhao@emory.edu

Wenjing Ma
University of Michigan
wenjinma@umich.edu

Hui Shao
Emory University
hui.shao@emory.edu

Xiao Hu
Emory University
xiao.hu@emory.edu

Yuanzhe Xi
Emory University
yuanzhe.xi@emory.edu

Carl Yang†
Emory University
j.carlyang@emory.edu

ABSTRACT

Patient risk prediction models are crucial as they enable healthcare providers to proactively identify and address potential health risks. Large pre-trained foundation models offer remarkable performance in risk prediction tasks by analyzing multimodal patient data. However, a notable limitation of pre-trained foundation models lies in their deterministic predictions (*i.e.*, lacking the ability to acknowledge uncertainty). We propose Gaussian Process-based foundation models to enable the generation of accurate predictions with instance-level uncertainty quantification, thus allowing healthcare professionals to make more informed and cautious decisions. Our proposed approach is principled and architecture-agnostic. Experimental results show that our proposed approach achieves competitive performance on classical classification metrics. Moreover, we observe that the accuracy of certain predictions is much higher than that of the uncertain ones, which validates the uncertainty awareness of our proposed method. Therefore, healthcare providers can trust low-uncertainty predictions and conduct more comprehensive investigations on high-uncertainty predictions, ultimately enhancing patient outcomes with less expert intervention.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Uncertainty Quantification; Clinical Foundation Models; Patient Risk Prediction; Gaussian Process Classification

*Authors' complete affiliations: J. Lu, Department of Computer Science & Nell Hodgson Woodruff School of Nursing, Emory University. S. Zhao, Y. Xi Department of Mathematics, Emory University. W. Ma, Department of Biostatistics, University of Michigan. H. Shao, Hubert Department of Global Health & Department of Family and Preventive Medicine, Emory University. X. Hu, Nell Hodgson Woodruff School of Nursing, Emory University; The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology & Emory University. C .Yang, Department of Computer Science, Emory University.

†Corresponding author: Carl Yang, j.carlyang@emory.edu.

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore.

ACM ISBN 979-8-4007-0172-6/24/05
https://doi.org/10.1145/3589335.3651456

1 INTRODUCTION

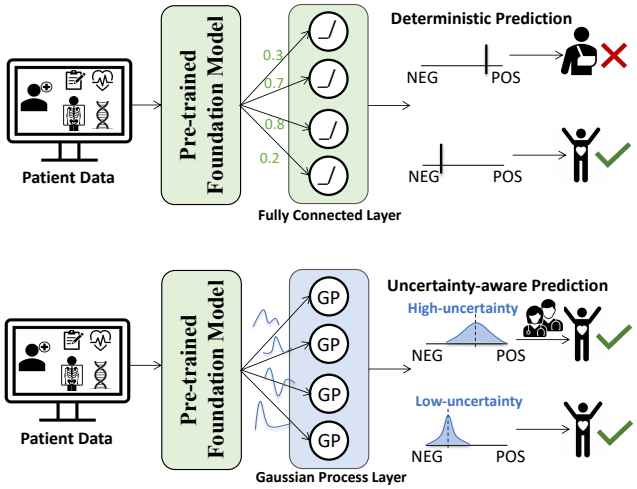


Figure 1: Pre-trained foundation models fine-tuned for patient risk prediction tasks. (Top) The conventional fully connected layer only supports deterministic prediction. (Bottom) Our Gaussian Process layer enables uncertainty-aware prediction. Therefore, healthcare professionals can investigate uncertain predictions and make cautious decisions.

Risk prediction models assist care providers in identifying patients at higher risks of future health-related events [24]. These events could include diseases, medical conditions, complications, or adverse health outcomes. Accurate risk prediction enables the implementation of preventative or early intervention measures to mitigate such risks. With the recent advancement of deep learning techniques, domain-specialized pre-trained foundation models (*PFMs*) [3, 25, 26] offer remarkable performance in risk prediction tasks by effectively utilizing scarce real-world clinical supervision signals. Moreover, PFMs can handle a variety of patient data modalities (*e.g.* textual clinical notes, visual histopathological images).

Pre-trained foundation models refer to large Transformer-based models [1, 14] that have become the *de-facto standard* for predictive and generative tasks with textual, visual, and other modality data. The successful paradigm is to pre-train PFMs with massive unlabeled data in pretext tasks and then fine-tune PFMs on a smaller task-specific dataset [4]. Originally proposed by natural language processing researchers, the Transformer architecture can effectively represent input tokens into latent dense vectors (*embeddings*) via

the multi-head self-attention mechanism [22]. More modalities are further expanded by conducting various tokenization procedures in the input data (*e.g.* Vision Transformers [6] split each image into a sequence of patches) with minimum architecture modifications [15]. Therefore, large-scale PFMs with billions of parameters [12] can be trained with the support of the parallelization of their self-attention mechanism and powerful hardware accelerators.

Although achieving impressive accuracy in a wide range of patient risk prediction tasks recently, a notable limitation of PFMs lies in their deterministic predictions (*i.e.,* lacking the ability to express its own predictive uncertainty) [5, 19]. Providing uncertainty estimates alongside predictions helps healthcare professionals make more informed and cautious decisions. It allows them to gauge the reliability of the model's predictions and consider the uncertainty when deciding on a course of action for a patient. Conventional statistical models such as Bayesian models [7, 21] provide uncertainty-aware risk predictions, supported by robust theoretical bases. However, the drawbacks of these statistical models include hand-crafted feature engineering and ad-hoc technical designs, which impose significant challenges when adapting them to new applications with limited prior knowledge or training samples.

In this work, we propose uncertainty-aware PFMs to improve trust between predictive models and healthcare professionals. Our proposed approach is principled and architecture-agnostic. A key technical contribution is to utilize Gaussian Process Classification (*GPC*) [17, 27] as the prediction layer. The original prediction layer of PFMs is mainly the fully connected layer that assigns deterministic probabilities to candidate class labels, which can not provide uncertainty quantification. A GPC layer is a Bayesian approach that inherently tackles the quantification of uncertainty. We follow the dominant learning paradigm to freeze the parameters of PFMs and only fine-tune our GPC layer on downstream patient risk prediction tasks. The frozen PFMs serve as the automated feature extractors based on rich domain-specific prior knowledge. Then the GPC layer is responsible for learning the mapping from the input embedding space to the output class space, providing uncertainty quantification based on Bayesian inference. To evaluate the performance of our uncertainty-aware PFM models, we conduct experiments on (a) one clinical note-based clinical outcome prediction dataset [18] and (b) one histopathological image-based breast cancer classification dataset [20]. Each dataset contains thousands of training instances for fine-tuning the PFMs. Four popular foundation models spanning text Transformers [16, 23] and vision transformers [6, 13] are tested. Empirical results demonstrate our uncertainty-aware PFMs achieve competitive performance to deterministic PFMs on classical classification metrics. Moreover, our uncertainty-aware PFMs are capable of generating stochastic predictions for each test instance, which naturally reflect the uncertainty levels of predictions. Empirically, we find the accuracy of certain predictions to be much higher than that of the uncertain ones, thus validating the uncertainty-aware property and real-world utility of our proposed approach.

2 PROPOSED APPROACH

2.1 Preliminaries

We focus on the classification setting for patient risk prediction.

Definition 2.1 (Patient Risk Prediction (Classification)). Given a set of N training instances $\mathbf{X} = \{x_1, \ldots, x_N\}$ and their corresponding labels $\mathbf{Y} = \{y_1, \ldots, y_N\}$, a classifier produces a predicted label $\hat{y}_* = \mathbf{F}(x_*)$ as the function of any new test instance x_*.

Empirically, pre-trained foundation models PFM_Θ contain:

(1) A Transformer ϕ_θ that represents each input instance x in latent dense vector by $\mathbf{h} = \phi_\theta(x)$. Here, $\mathbf{h} \in \mathbb{R}^{1\times d}$, where d denotes the dimension of Transformer embedding space.
(2) A fully-connected layer f_ω that maps h into the unnormalized classification output (*i.e.* logits) by $\mathbf{z} = f_\omega(\mathbf{h})$. Here, $\mathbf{z} \in \mathbb{R}^{1\times C}$, where C denotes number of target classes. A common implementation for the fully-connected layer is $f_\omega(\mathbf{h}) = \mathbf{hW} + \mathbf{b}$, where $W \in \mathbb{R}^{d\times C}$ and $b \in \mathbb{R}^{1\times C}$ denote learnable parameters ω of f_ω. Moreover, f_ω can be extended to more complicated forms when non-linearity is necessary.

The final classification prediction can be expressed as

$$\hat{y} = \arg\max_c (\Pr(y = c|x)) = \arg\max_c \left(\frac{e^{\mathbf{z}_{(c)}}}{\sum_{c'=1}^{C} e^{\mathbf{z}_{(c')}}}\right). \quad (1)$$

$\mathbf{z}_{(c)}$ denotes the c^{th} element of the high-dimensional vector $\mathbf{z}$, which refers to the logits of predicting x's label as c^{th} class. The cross-entropy loss is employed to fine-tune PFM_Θ:

$$\mathcal{L}(\Theta) = -\sum_{n=1}^{N} \mathbb{1}(y_n) \cdot \log(Pr(\hat{y}_n|x_n)), \quad (2)$$

where $\mathbb{1}(\cdot)$ is an indicator function that returns the one-hot encoding for the true class label y_n. Practically, only the parameters ω of the fully-connected layer g_ω are updated, while the parameters θ of the Transformer ϕ_θ are frozen. Fine-tuning only the last layer allows the PFMs to adapt specifically to the downstream task at hand without disrupting the general knowledge encoded in the lower layers. It is also parameter efficient considering the PFMs often have a large number of parameters. Moreover, it can prevent overfitting when the scale of the training dataset is limited.

2.2 Gaussian Process Classification Layer to Enable Uncertainty-Aware Prediction

The fully-connected layer can only generate deterministic point estimation. Our insight is to employ a Gaussian Process-based [17] uncertainty-aware layer (GPC layer) $\mathbf{g}_\beta$ to augment PFMs. For GPC layer capable of generating stochastic distribution estimation, it still implies Eq.(1). However, $\mathbf{z}_{(c)}$ now is a random variable, thus capable of generating stochastic predictions that naturally reflect instance-level uncertainty through the variance of $\mathbf{z}_c$. Following the noise case Gaussian Process property,

$$\mathbf{z}_{(c)} \sim GP_c(m_c(\mathbf{H}), \mathcal{K}_c(\mathbf{H}, \mathbf{H}) + \delta_c^2\mathbf{I}), \quad (3)$$

where $\mathbf{H} = \phi_\theta(\mathbf{X}) \in \mathbb{R}^{N\times d}$ denotes embeddings of all training instances, $m_c(\cdot)\colon \mathbb{R}^{N\times d} \mapsto \mathbb{R}^{N\times 1}$ and $\mathcal{K}_c(\cdot,\cdot)\colon \mathbb{R}^{N\times d} \times \mathbb{R}^{N'\times d} \mapsto \mathbb{R}^{N\times N'}$ are the mean function and kernel function of one GP that corresponds to the logits for c^{th} class, δ^2 denotes the observation noise variance, $\mathbf{I} \in \mathbb{R}^{N\times N}$ denotes an identity matrix. According to Bayes' theorem, the Gaussian posterior of $\mathbf{z}_{(\mathbf{c}),*}$ for one new test instance x_* can be expressed as

$$Pr(\mathbf{z}_{(\mathbf{c}),*}|\mathbf{H}, \mathbf{Y}, \mathbf{h}_*) \sim \mathcal{N}(\mu_{c,*}, \sigma_{c,*}), \quad (4)$$

where $\mathcal{N}(\mu_{c,*}, \sigma_{c,*})$ denotes a Gaussian distribution with mean prediction $\mu_{c,*}$ and variance $\sigma_{c,*}$. Thanks to the great mathematical properties of Gaussian Process, we can get closed-form expression

$$\mu_{c,*} = m_c(\mathbf{h}_*) + \mathcal{K}_c^{\intercal}(\mathbf{H}, \mathbf{h}_*)\mathcal{K}_c^{-1}(\mathbf{H}, \mathbf{h}_*)(\mathbf{Y} - m_c(\mathbf{h}_*)); \quad (5)$$

$$\sigma_{c,*} = \mathcal{K}_c(\mathbf{h}_*, \mathbf{h}_*) - \mathcal{K}_c^{\intercal}(\mathbf{H}, \mathbf{h}_*)(\mathcal{K}_c(\mathbf{H}, \mathbf{H}) + \delta_c^2\mathbf{I})^{-1}\mathcal{K}_c(\mathbf{H}, \mathbf{h}_*). \quad (6)$$

In the general multi-class classification setting, our proposed GPC layer $\mathbf{g}_\beta$ utilizes C independent GPs to transform the instance embedding into the class logits, denoted as $\mathbf{g}_\beta = [GP_1, \ldots, GP_C]$. Therefore, we can use the expectation to express predicted probability for a test instance x_* (as compared to Eq.(1)):

$$\Pr(y_* = c|x_*) = \int \frac{e^{\mathbf{z}_{(c),*}}}{\sum_{c'=1}^{C} e^{\mathbf{z}_{(c'),*}}} Pr(\mathbf{z}_{(c),*}|\mathbf{H}, \mathbf{Y}, \mathbf{h}_*)d\mathbf{z}_*. \quad (7)$$

Unfortunately, we can't get closed-form estimates of the probabilities in Eq.(7). An approximation using J samples can be used:

$$\Pr(y_* = c|x_*) \approx \frac{1}{J}\sum_{j=1}^{J} \frac{e^{\mathbf{z}_{(c),*,j}}}{\sum_{c'=1}^{C} e^{\mathbf{z}_{(c'),*,j}}}, \quad (8)$$

where $\mathbf{z}_{(c'),*,j}$ denotes the j-th sampling of $\mathbf{z}_{(c'),*}$ following Eq.(4). As can be seen from Eq.(8), our GPC layer is capable of generating stochastic classification predictions with regard to each test instance, which inherently reflects the uncertainty quantification through the variance over the predicted samples. Instead of manually setting a hard threshold to determine whether a prediction is uncertain or not, we propose utilizing the t-test for that purpose to maximize the easy utility of our uncertainty-aware PFMs (please refer to Sec.3.3 for a detailed discussion).

In our GPC implementation, we further add a linear projector to the Transformer output embedding $\mathbf{h}$ before applying Gaussian Process, because Gaussian Process is more effective in input vectors with small dimensionality and modern Transformer embedding dimension d is typically 768 or 1024. The linear projector can be denoted as $\mathbf{h}' = \tanh(\mathbf{h}\mathbf{W}' + \mathbf{b}')$, thus $\mathbf{h}' \in \mathbb{R}^{d'}$, where $d' \ll d$. Moreover, We select zero mean function for $m(\cdot)$, and radial basis function kernel with learnable lengthscale for $\mathcal{K}(\cdot,\cdot)$, for efficient implementation. We follow the Dirchlet-based Gaussian Process [17] that transforms the classification targets into regression ones using an approximate Dirichlet classification likelihood [9].

3 EXPERIMENTS

3.1 Experimental Settings

Risk Prediction Datasets. We use the following datasets to evaluate our proposed method, with statistics presented in Tab.1.

(1) MedNLI [18]: a natural language inference dataset, which focuses on identifying potential clinical outcome (*hypotheses*) based on past medical history (the *premise*), extracted from MIMIC-III clinical notes annotated by care providers.
(2) BreakHis [20]: an image classification dataset, which focuses on distinguishing benign tumors and malignant tumors in breast cancer histopathological images, collected from multiple patients using various magnifying factors.

Backbone Pre-Trained Foundation Models. For *text-modality* backbone foundation models, we test on: (a.1) the bi-directional BERT-based architecture ClinicalBERT [23]; (a.2) the uni-directional auto-regressive GPT-based architecture BioGPT [16]. Both PFMs have been extensively pre-trained on the healthcare corpus. For *image-modality* backbone foundation models, we test on: (b.1) vision transformer-based architecture ViT [6]; (b.2) hierarchical vision transformer-based architecture SwinV2 [13]. Both PFMs have been extensively pre-trained on general-domain high-resolution images.

Table 1: Statistics of the used datasets.

Dataset	Modality	#Category	#Train	#Test
MedNLI	Text	3	11,232	1,422
BreakHis	Image	2	5,005	2,904

Compared Approaches. The following approaches are compared: (1) Fully-connected layer-based PFM which is the classical deterministic classification approach. (2) *Monte Carlo Dropout*-based PFM [8] which is a stochastic classification approach casting the dropout as approximate Bayesian inference.

3.2 Classification Results

Table 2: Results of various uncertainty-aware PFMs.

(a) Textual patient risk prediction results.

Method	NLL	Brier	Acc
ClinicalBERT			
Fully Connected	0.85	0.51	60.34%
MC Droput	0.87	0.52	59.42%
GPC (**Ours**)	**0.81**	**0.47**	**65.26**%
BioGPT			
Fully Connected	**0.80**	0.50	63.78%
MC Droput	0.81	**0.47**	63.29%
GPC (**Ours**)	0.86	0.51	**67.65**%

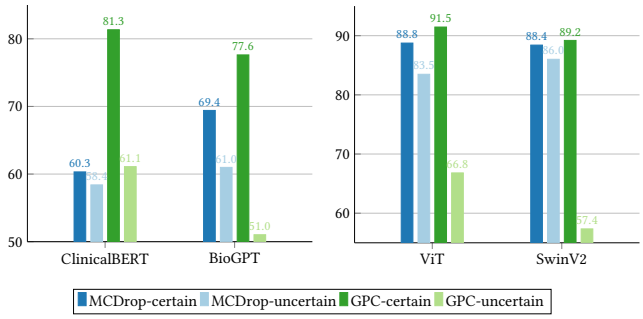
(b) Visual patient risk prediction results.

Method	NLL	Brier	Acc
ViT			
Fully Connected	0.37	0.23	84.47%
MC Droput	0.37	0.23	84.37%
GPC (**Ours**)	**0.34**	**0.20**	**86.47**%
SwinV2			
Fully Connected	**0.31**	**0.19**	87.29%
MC Droput	0.33	0.20	86.26%
GPC (**Ours**)	0.63	0.24	**87.74**%

We first examine the performance of all methods using conventional classification metrics, including negative log-loss (NLL), Brier Score [2], and accuracy, which are reasonable and widely used in previous studies. Among these metrics, both NLL and Brier Score are the less the better, and accuracy is the more the better. As can be seen from Tab.2, our GPC-based PFMs achieve competitive performance in NLL and Brier Score across all four foundation models over two datasets. Moreover, our method consistently outperforms baselines in accuracy for all two risk prediction tasks.

3.3 Uncertainty Quantification

Other than the conventional classification metrics, we are interested in the uncertainty quantification of the proposed method. Existing uncertainty quantification works [10, 17] have been focusing on *confidence calibration* (*i.e.*, the predicted probabilities



(a) Textual risk prediction. **(b) Visual risk prediction.**

Figure 2: Uncertainty quantification results in accuracy (%).

by a classification model should be aligned with the actual likelihood of the corresponding outcomes), which mainly provides a group-level uncertainty. In this work, we adopt an instance-level uncertainty quantification metric to reflect how sure the model is about each of its predictions. The uncertainty quantification is conducted on the stochastic outputs from non-deterministic classification models. Specifically, we apply paired two-sample t-test on model predictions [11]: we obtain the most and second-most predicted probabilities for each instance, and test whether the difference is statistically significant. The rejection status ($\alpha = 0.01$) reflects whether the predictive model is certain about its prediction for a specific test instance. As shown in Fig.2, we obtain the average accuracy over predictions with low-uncertainty (denoted as "[*method*]-certain") v.s. ones with high-uncertainty (denoted as "[*method*]-uncertain"). The accuracy of certain predictions made by our GPC-based PFMs is always the highest among all backbone foundation models and all datasets. Moreover, clear accuracy gaps can be consistently observed between GPC-certain predictions and GPC-uncertain predictions, while such gaps are vague for the baseline method MC Dropout. This outcome validates the uncertainty-awareness property of our proposed method. Therefore, by trusting low-uncertainty predictions and requesting care providers' further investigation on high-uncertainty predictions, the proposed GPC-based PFMs can achieve trustworthy patient risk prediction with minimum manual intervention.

4 CONCLUSION

Our innovative Gaussian Process-based approach enables pre-trained foundation models to output accurate predictions with instance-level uncertainty quantification. The uncertainty-aware PFMs acknowledge the inherent complexity of medical conditions and respect the uncertainties involved in predicting individual patient outcomes. In the future, we aim to (1) extend our uncertainty-aware PFMs into more modalities such as waveforms (human physiological data) and tabular (electronic health records); (2) explore the mini-batch stochastic gradient descent technique of Gaussian Process to reduce the computational burden.

REFERENCES

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[2] Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review* 78, 1 (1950), 1–3.

[3] Zhaoliang Chen, Cheng Ding, Nirbhay Modhe, Jiaying Lu, Carl Yang, and Xiao Hu. 2024. Adapting a Generative Pretrained Transformer Achieves SOTA Performance in Assessing Diverse Physiological Functions Using Only Photoplethysmography Signals: A GPT-PPG Approach. In *AAAI-Clinical Foundation Models*.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] James M Dolezal, Andrew Srisuwananukorn, Dmitry Karpeyev, Siddhi Ramesh, Sara Kochanny, Brittany Cody, Aaron S Mansfield, Sagar Rakshit, Radhika Bansal, Melanie C Bois, et al. 2022. Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology. *Nature communications* 13, 1 (2022), 6572.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

[7] Robert Dürichen, Marco AF Pimentel, Lei Clifton, Achim Schweikard, and David A Clifton. 2014. Multitask Gaussian processes for multivariate physiological time-series analysis. *IEEE Transactions on Biomedical Engineering* 62, 1 (2014), 314–322.

[8] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*. 1050–1059.

[9] Jacob R Gardner, Geoff Pleiss, David Bindel, Kilian Q Weinberger, and Andrew Gordon Wilson. 2018. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. In *NeurIPS*.

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *ICML*. 1321–1330.

[11] Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. 2022. Card: Classification and regression diffusion models. *NeurIPS* 35 (2022), 18100–18115.

[12] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[13] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*. 12009–12019.

[14] Jiaying Lu, Yongchen Qian, Shifan Zhao, Yuanzhe Xi, and Carl Yang. 2023. MuG: A Multimodal Classification Benchmark on Game Data with Tabular, Textual, and Visual Fields. In *Findings-EMNLP'23*.

[15] Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. 2024. Evaluation and Enhancement of Semantic Grounding in Large Vision-Language Models. In *AAAI-ReLM Workshop*.

[16] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics* 23, 6 (2022), bbac409.

[17] Dimitrios Milios, Raffaello Camoriano, Pietro Michiardi, Lorenzo Rosasco, and Maurizio Filippone. 2018. Dirichlet-based gaussian processes for large-scale calibrated classification. *NeurIPS* 31 (2018).

[18] Alexey Romanov and Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain. In *EMNLP*. 1586–1596.

[19] Jonas Schuett, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel. 2023. Towards best practices in AGI safety and governance: A survey of expert opinion. *arXiv preprint arXiv:2305.07153* (2023).

[20] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. 2015. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering* 63, 7 (2015), 1455–1462.

[21] Marcel AJ Van Gerven, Babs G Taal, and Peter JF Lucas. 2008. Dynamic Bayesian networks as prognostic models for clinical patient management. *Journal of biomedical informatics* 41, 4 (2008), 515–529.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS* 30 (2017).

[23] Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine* 29, 10 (2023), 2633–2642.

[24] Ran Xiao, Cheng Ding, Xiao Hu, Gari D Clifford, David W Wright, Amit J Shah, Salah Al-Zaiti, and Jessica K Zègre-Hemsey. 2023. Integrating multimodal information in machine learning for classifying acute myocardial infarction. *Physiological Measurement* 44, 4 (2023), 044002.

[25] Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. 2023. Large language models for healthcare data augmentation: An example on patient-trial matching. In *AMIA Annual Symposium Proceedings*. 1324.

[26] Yunkun Zhang, Jin Gao, Zheling Tan, Lingfeng Zhou, Kexin Ding, Mu Zhou, Shaoting Zhang, and Dequan Wang. 2024. Data-Centric Foundation Models in Computational Healthcare: A Survey. *arXiv preprint arXiv:2401.02458* (2024).

[27] Shifan Zhao, Tianshi Xu, Edmond Chow, and Yuanzhe Xi. 2023. An Adaptive Factorized Nystr\" om Preconditioner for Regularized Kernel Matrices. *arXiv preprint arXiv:2304.05460* (2023).