Pancreatic Cancer Prediction Using LLM based Embeddings

Center for Artificial Intelligence and Genome Informatics
Made K. Prasadha, BS, Bingyu Mao, PhD, Michael Ghebranious, Degui Zhi, PhD, Laila Rasmy, PhD
McWilliams School of Biomedical Informatics at UTHealth Houston

For questions via email: made.k.prasadha@uth.tmc.edu

INTRODUCTION

Background: Pancreatic cancer (PaCa) is estimated to be the second leading cause of cancer death by 2030.[1]

Question: Do we still need specialized clinical foundation models or can we simply leverage the advancements of large language models (LLMs) to boost the accuracy of specialized clinical models trained on structured EHR data, even with the sparsity of the samples?

Objective: To explore the feasibility of using LLM based embeddings for developing accurate PaCa predictive models.

METHODS

- For reproducibility we used the EHRSHOT[2] Pancreatic Cancer (PaCa) cohort, including 3810 patients and 217 positive PaCa cases. Unlike the original work our prediction unit is the patient and not visits, so we mainly utilized the last eligible patient visit and further cleaned the cohort.
- We encoded the patient trajectory into dense embeddings with frozen LLMs.
- We formatted patient histories in different forms: summary text, XML including visits data in reverse order, and Markdown file with a summary at the top of the file followed by detailed visit level events.[3]
- We generated embeddings from eight different LLMs with and without instructions using the first 4096 tokens from the most recent visits.
- We evaluated binary classifier performance using the area under receiver operating curve (AUROC) and the area under precision recall curve (AUPRC).

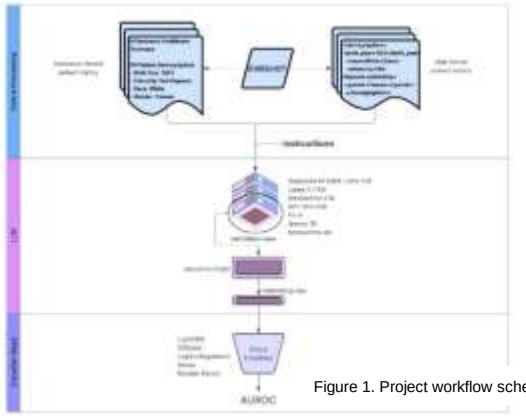

Figure 1. Project workflow schematic

RESULTS

- We found that the Markdown input format was associated with the best results.
- Scikit-learn's Logistic Regression (LR) or MLP showed the best performance compared to other tree or vector machine-based classifiers.
- We also found that adding instructions even for embedding models such as Qwen3 Embedding 8B was associated with better results.
- DeepSeek R1 Distill shows the highest performance (AUROC 89.9% / 59.9% AUPRC) followed by Llama 70B (88.5% / 59.9%) and GPT OSS 20B (88.4% / 66.1%).

Table 1. AUROC and AUPRC values for various model generated embeddings

Model	LR		MLP	
	AUROC	AUPRC	AUROC (STD)	AUPRC (STD)
DeepSeek R1 Distill Llama 70B	**89.9**	*59.9*	87.1 (0.5)	53.9 (3.4)
GPT OSS 20B	88.4	**66.1**	86.6 (0.4)	54 (3.5)
Llama 3.1 70B	88.5	59.9	86.6 (0.7)	51.7 (3.60
MedGemma 27B	86.9	59.4	87.2 (0.60	57.9 (1.8)
MedGemma 4B	85.8	54.9	86 (0.2)	48.9 (2.0)
Phi 4	84.6	46.3	85.1 (0.5)	46 (2.1)
Qwen2 7B	84.9	53.4	86 (1.3)	53.6 (3.7)

Table 2. Comparing Llama-3.1-70B-Instruct results across various workflows

Llama 3.1 70B	AUROC		AUPRC	
Text Format	Fine-tuning	No Finetuning	Fine-tuning	No Finetuning
	84.7	**87.0**	46.6	**50.6**
Additional File Format - with No Fine-tuning	XML	Markdown	XML	Markdown
	83.4	**88.5**	37.0	**59.9**

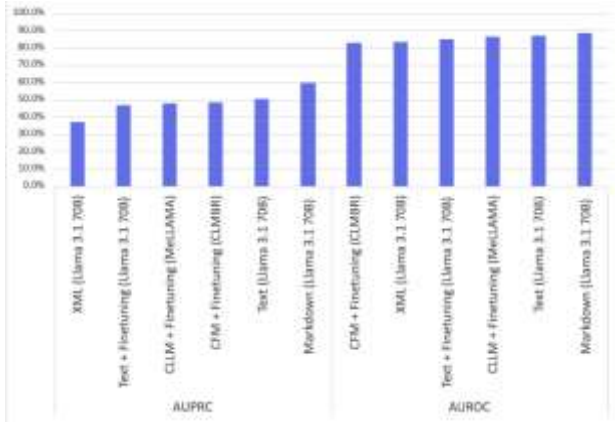

Figure 2. Performance of current models and methods (Markdown and XML) compared to previous work

Table 3. Comparing various data formats' Qwen3 8B, 30B model performance

Model	LR		MLP	
	AUROC	AUPRC	AUROC (STD)	AUPRC (STD)
Qwen3 Embedding + Text	**89.1**	**56.4**	**90.8 (0.44)**	**64.7 (1.98)**
Qwen3 Embedding + Markdown	85.1	55.8	88.1 (0.20)	52.2 (4.70)
Qwen3 Thinking + Text	88.2	50.2	86.0 (1.00)	34.4 (2.30)
Qwen3 Thinking + Markdown	86.5	53.8	86.4 (0.50)	46.6 (3.20)
Qwen3 Instruct + Text	88.0	47.9	86.6 (1.20)	38.1 (3.20)
Qwen3 Instruct + Markdown	86.5	54.0	86.4 (0.60)	47.6 (2.60)

DISCUSSION

- Our results show that generating patient level embeddings from high parameter based LLMs without the need of further finetuning of LLM weights can achieve state of the art performance.
- DeepSeek, our best performing embedding, took approximately nine hours to generate for our cohort while GPT OSS took around two hours. This indicates the promise of LLMs as foundations for clinical predictive models, with continuous efforts of efficiency improvements.
- Input format and the instructions had an impact on the generated LLMs embeddings quality and accordingly the classifier performance.
- In the majority of our results just a simple logistic regression was associated with the best predictive performance compared to other ML algorithms
- Testing with Qwen3 8B models, there is no difference between the Instruct or Thinking models and the embedding model was the one showing the best performance.
- Future work include testing LLMs' based embedding for time to event prediction tasks as well as

CONCLUSION

Pancreatic Cancer is one of the leading causes of cancer deaths due to difficulty in early detection and with the advancements of AI and LLMs we will be able to improve the performance and the generalizability of predictive models that could be easily adopted at the clinical side. Our findings showed promising results even without the need for supervised fine-tuning

ACKNOWLEDGEMENTS

This project is funded by NIH/NLM R01 grant: R01 LM014249

REFERENCES

1. He, J., Rasmy, L., Zhi, D., & Tao, C. (2025). *Advancing Pancreatic Cancer Prediction with a Next Visit Token Prediction Head on Top of Med-BERT. Cancers, 17*(3), 516. https://doi.org/10.3390/cancers17030516
2. Fleming, S. L., Lozano, A., Haberkorn, W. J., Jindal, J. A., Reis, E. P., Thapa, R., Brunskill, E. P. (2023). *MedAlign: A clinician-generated dataset for instruction following with electronic medical records* (arXiv:2308.14089). ArXiv. https://doi.org/10.48550/arXiv.2308.14089
3. Hegselmann, S., von Arnim, G., Rheude, T., Kronenberg, N., Sontag, D., Hindricks, G., Eils, R., & Wild, B. (2025) *Large language models are powerful electronic health record encoders* (arXiv:2502.17403v3). ArXiv. https://doi.org/10.48550/arXiv.2502.17403