Faculty of Engineering International Conference
FEIC 2026 | Biomedical Engineering | CMT-33

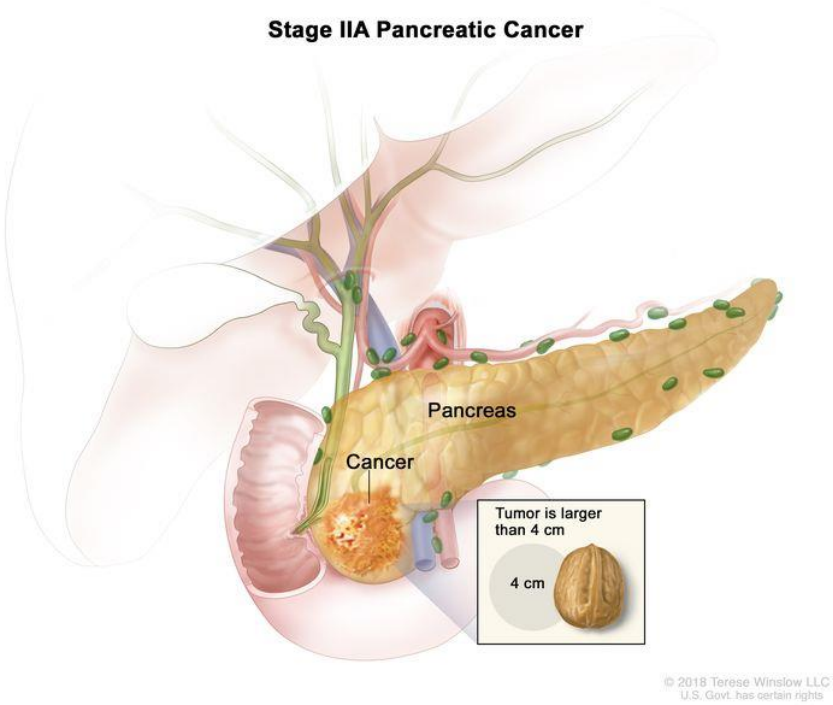
CAN ROUTINE DEMOGRAPHIC AND LIFESTYLE DATA PREDICT LATE-STAGE PANCREATIC CANCER?

A Machine Learning Evaluation for Low-Resource Settings

Sunny O. Agbai & **Solomon C. Nwaneri**
Department of Biomedical Engineering, University of Lagos

Background

< 12%	five-year survival
~ 1 in 5	can have surgery
Stage	decides the chemo route



PDAC is usually found late, after it has spread

Problem and Research Gap

1. Strong models need CA19-9, imaging or dense records
2. Such data are rarely captured in Nigerian hospitals
3. Demographic and lifestyle data alone: untested

Question: can routine data predict late-stage PDAC?

Aim and Objectives

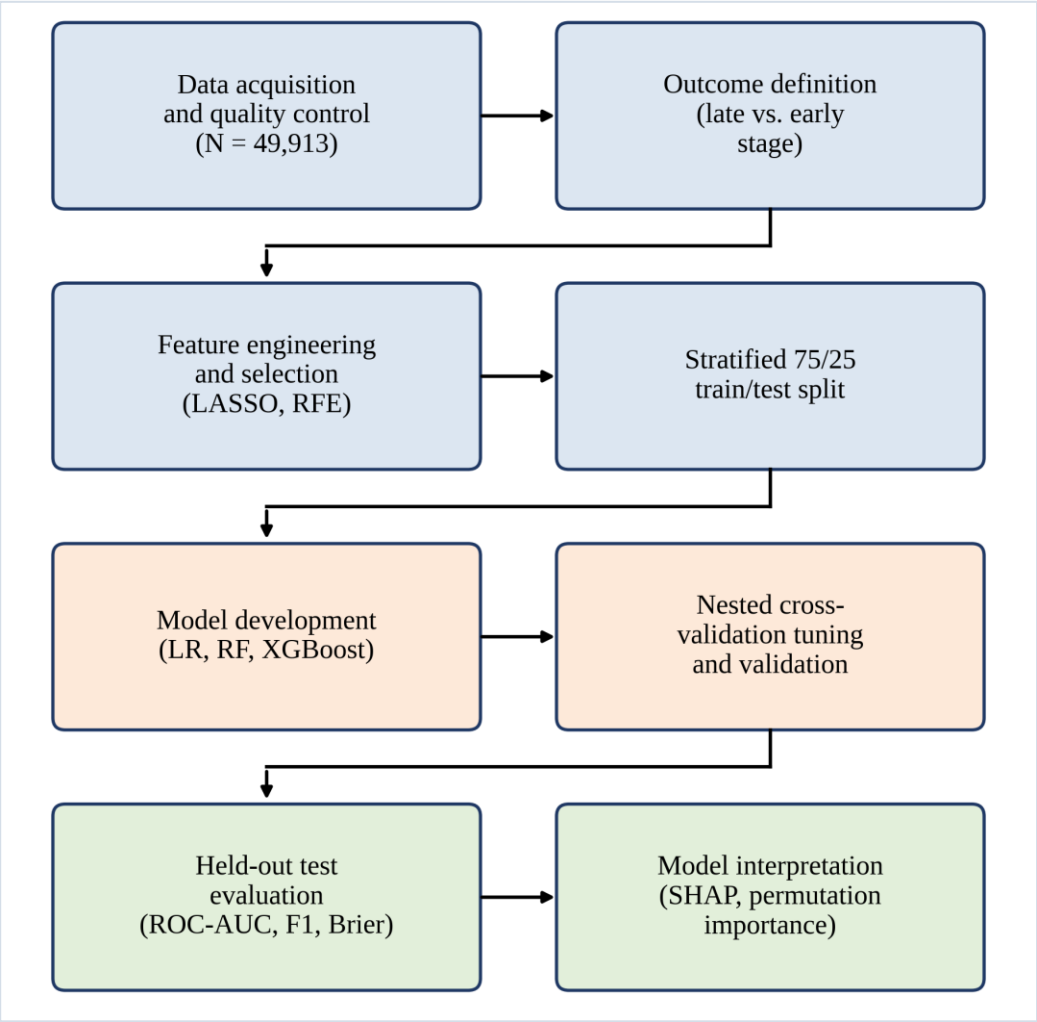
Aim: predict late-stage (III/IV) PDAC from routine variables.

1. Establish the key predictors of stage
2. Develop and compare three tuned models
3. Evaluate on a held-out test set

Related Work

Study	Data used	Finding
Schneider et al. [5]	Clinical + lab	AUROC 0.85
Kaissis et al. [7]	Radiomic + molecular	Chemo response
Placido et al. [9]	Disease-code records	AUROC 0.88
Večurkovská et al. [8]	Review, 13 studies	AUROC 0.84–0.97

Data and Study Design

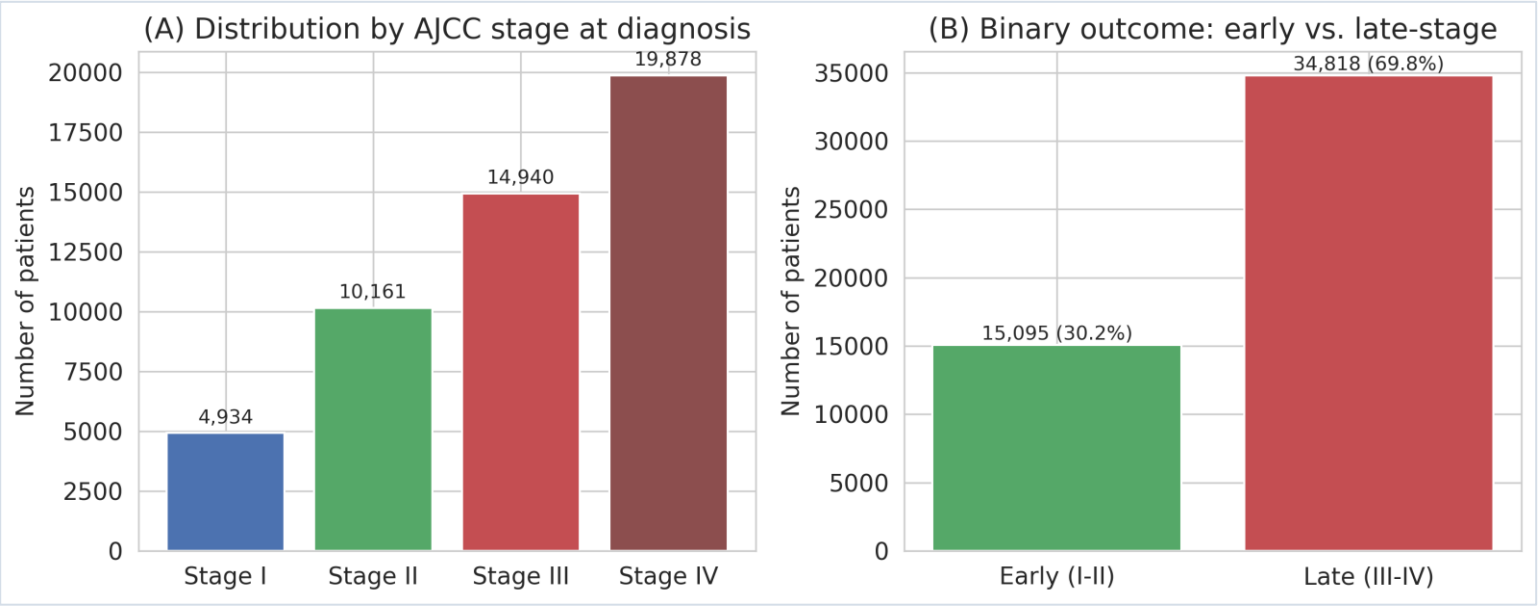


Data	N = 49,913
Inputs	38 predictors
Outcome	Late (III/IV): 69.8%
Source	Kaggle, likely simulated

Methods

1	**Selection**	Correlation, LASSO, RFE
2	**Models**	LR, Random Forest, XGBoost
3	**Validation**	Nested CV; 25% test set
4	**Explain**	SHAP, permutation importance

Results: Cohort (Figure 2)



69.8% late-stage

64.5 y mean age

Figure 2

N = 49,913 patients; 51.9% male

Results: Feature Screening

≤ 0.010	max |r| with stage
2 of 38	$p < 0.05$ (≈ 1.9 by chance)
13 of 38	kept by LASSO, all small

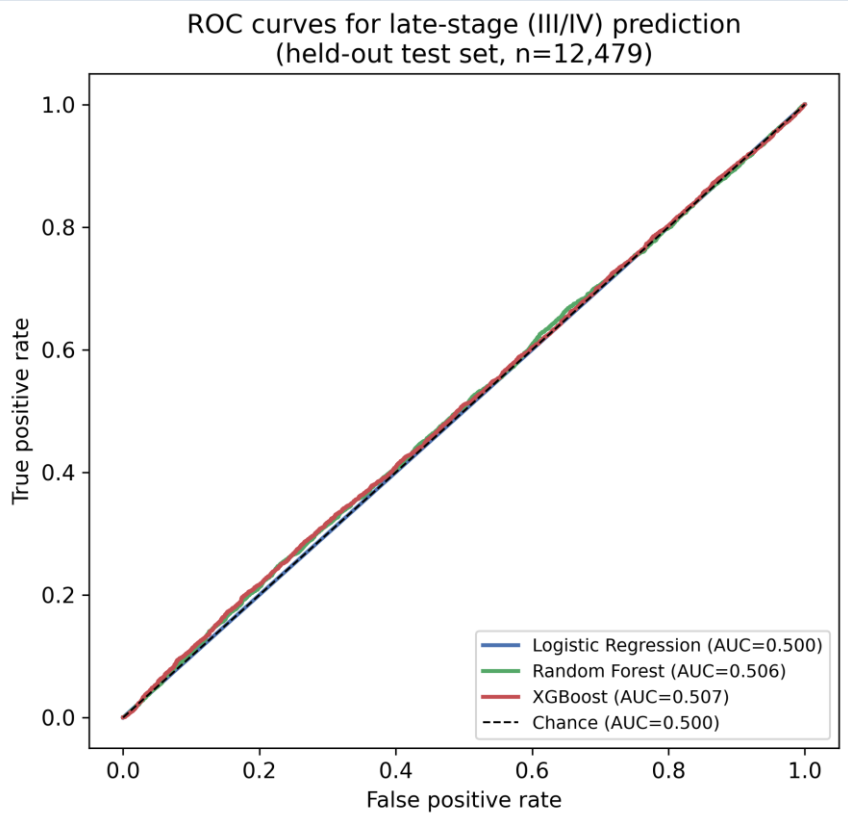
No feature is a meaningful predictor

Results: Model Performance

Model	Test AUC	Nested CV	Accuracy
Logistic Regression	0.500	0.498	0.698
Random Forest	0.506	0.498	0.504
XGBoost	0.507	0.497	0.506

LR labelled everyone "late-stage": its 69.8% accuracy only mirrors prevalence

Results: ROC Curves (Figure 3)



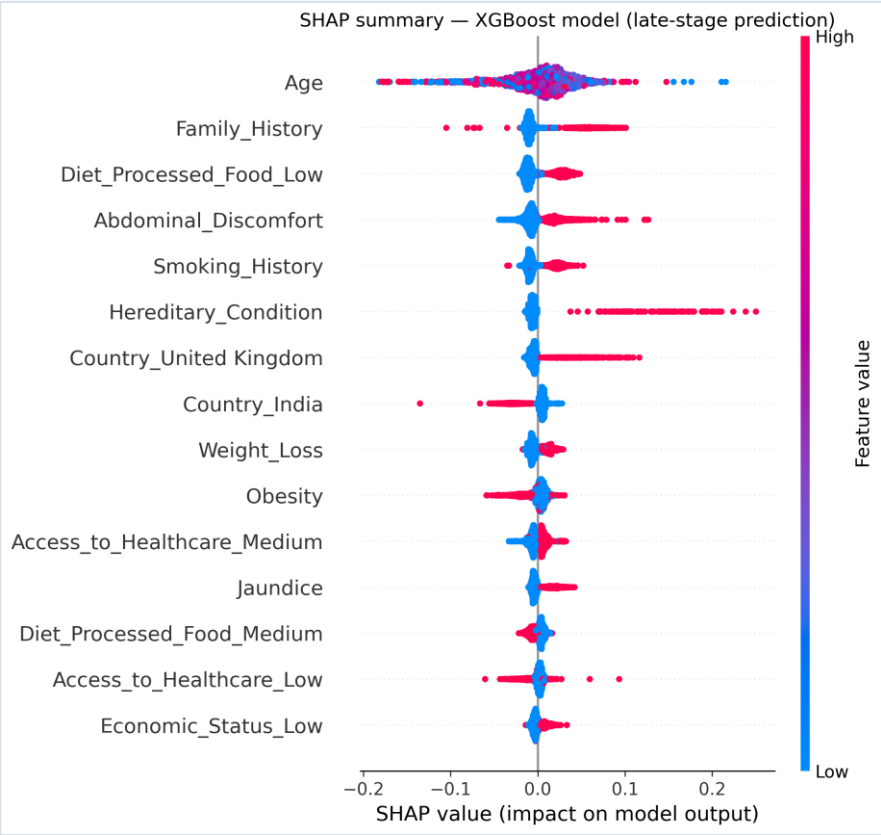
0.500 test AUC: LR

0.506 test AUC: RF

0.507 test AUC: XGBoost

Curves lie on the chance diagonal

Results: Interpretation (Figure 4)



0.0288 largest SHAP value (Age)

0.0033 largest permutation drop in AUC

No feature moves risk far from base rate

Discussion: Why Chance-Level?

1	**No biomarkers**	CA19-9, ECOG, blood count absent
2	**Synthetic data**	Stage may not depend on inputs
3	**Tiny true effect**	AUC at most 0.49–0.52

The limit is the data, not the algorithm

Conclusion

Demographic and lifestyle data alone appear insufficient to predict late-stage PDAC.

- Three models, two validations agree
- H_0 could not be rejected
- Caveat: dataset likely synthetic

Implications and Next Steps

Practice

- Capture blood count, CA19-9, ECOG
- Decision support only

Research

- Repeat on real local data
- Try federated learning
- Report null results

Selected References

[1] F. Schneider, H. Chen, U. Pelzer, R. G. Compton, et al., “The basis for future personalized therapy approaches – Machine learning-generated 1-year survival rate, metastatic status and therapy-dependent survival in pancreatic cancer patients,” Eur. J. Cancer, vol. 234, Art. no. 116189, 2026.

[2] D. Placido, B. Yuan, J. X. Hjaltelin, C. Zheng, et al., “A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories,” Nat. Med., vol. 29, pp. 1113–1122, 2023.

Selected References (2/2)

[3] G. Kaissis, S. Ziegelmayer, F. Lohöfer, K. Steiger, et al., “A machine learning algorithm predicts molecular subtypes in pancreatic ductal adenocarcinoma with differential response to gemcitabine-based versus FOLFIRINOX chemotherapy,” PLoS ONE, vol. 14, no. 10, Art. no. e0218642, 2019.

[4] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 4765–4774.

Thank You

Questions and Discussion

Sunny O. Agbai | agbaisunny@outlook.com

Co-author: Solomon C. Nwaneri | snwaneri@unilag.edu.ng