Shall You Get an Invasive Examination? An AI-Driven Risk Stratification Model for Individuals with Suboptimal Health Status

Online Appendix

1. Tables

Table A1. Feature Selection Results: Full Feature Set vs. Stepwise Logistic Regressio

Model	Accuracy	Precision	F1-score	Sensitivity	Specificity	AUC
Full features	0.708	0.320	0.322	0.325	0.812	0.63
Stepwise Logistic Regression	**0.632**	**0.297**	**0.38**	**0.528**	**0.660**	**0.64**

Table A2. All Features

Features	Standard Reference Ranges	
Age	52.2±11.2	
Sex	F/M	
Body Mass Index (BMI)	18.5-23.9	Kg/m2
Hypertension	Systolic ≥ 140 or Diastolic ≥ 90	mmHg
Alanine Transaminase (ALT)	0-36	U/L
Aspartate Aminotransferase (AST)	0-34	U/L
Total Bilirubin	0.1-1.2	mg/dL
Platelets Count	150-400	1000/uL
Calcium (Ca)	7.9-9.9	mg/dL
Triglyceride	<150	mg/dL
Total Cholesterol	<200	mg/dL
High-Density Lipoprotein (HDL)	M:>40,F:>50	mg/dL
Low-Density Lipoprotein (LDL)	<100	mg/dL
Diastolic Blood Pressure	< 80	mmHg
Systolic Blood Pressure	< 120	mmHg
Waist Circumference	M:<90 F:<80	cm
Blood Urea Nitrogen (BUN)	6-21	mg/dL
Creatinine	M:0.4-1.4,F:0.6-1.2	mg/dL
Sodium (Na)	135-145	mmol/L

Potassium (K)	3.5-5.1	mmol/L
White Blood Cells (WBC)	M:3.9-10.6,F:3.5-11	1000/uL
Hemoglobin (Hb)	M:13.5-17.5,F:12-16	g/dL
Mean Corpuscular Volume (MCV)	80-100	fL
Neutrophils	42-74	%
Lymphocytes	20-56	%
C-Reactive Protein (CRP)	<5	mg/L
Monocytes	0-12	%
Albumin	3.5-5.5	g/dL
Fasting Glucose (AC)	70-105	mg/dL
Postprandial Glucose (PC)	<200	mg/dL
Waist Circumference	M:≤90,F:≤80	CM
Uric Acid	<8	mg/dL
Fatty Liver Score	Mild/Moderate/Severe	
Dental Caries Count	<32	
Tooth Missing Count	<32	
Periodontitis	Absent/Present	
Gallstones	Absent/Present	
Dental Calculus	Absent/Present	
Gallbladder Polyps	Absent/Present	
Tea usage	Absent/Present	
Coffee usage	Absent/Present	
Areca usage	Absent/Present	
Alcohol usage	Absent/Present	
Systemic Immune-Inflammation Index (SII)	$\frac{\text{Platelets} \times \text{Neutrophil}}{\text{Lymphocyte}}$	
Systemic Inflammation Response Index (SIRI)	$\frac{\text{Monocyte} \times \text{Neutrophil}}{\text{Lymphocyte}}$	
International Prognosis Index (IPI)	$\frac{\text{CRP} \times \text{Neutrophil}}{\text{Lymphocyte} \times \text{Albumin}}$	
Neutrophil-to-Lymphocyte Ratio (NLR)	$\frac{\text{Neutrophil}}{\text{Lymphocyte}}$	
Lymphocyte-to-Monocyte Ratio (LMR)	$\frac{\text{Lymphocyte}}{\text{Monocyte}}$	
Atherogenic Index of Plasma (AIP)	$\text{Triglyceride} \times \text{HDL\_C}$	
Triglyceride-Glucose Index (TyG)	$\ln(\text{Triglyceride} \times \text{Fasting Glucose})$	

C-reactive protein-triglyceride-glucose index (CTI)	$0.412 \times \ln(\mathrm{CRP}) + \mathrm{TyG}$
Fibrosis-4 (FIB-4)	$\frac{\text{Age}\times\text{AST}}{\text{Platelets}\times\sqrt{\text{ALT}}}$
Adenomatous	Absent/Present

Table A3 Logistic Regression Regularization Parameter (α) Selection for the Proposed Method

Model	Accuracy	Precision	Recall	F1-score	Specificity	AUC
α= 0.05	0.638 [0.631,0.645]	0.766 [0.748,0.783]	0.597 [0.572,0.622]	0.669 [0.658,0.681]	0.704 [0.666,0.741]	0.71
α= 0.10	0.639 [0.626,0.652]	0.763 [0.751,0.775]	0.599 [0.576,0.623]	0.671 [0.655,0.686]	0.702 [0.682,0.723]	0.72
α=0.2	**0.685 [0.673,0.696]**	**0.731 [0.716,0.745]**	**0.776 [0.751,0.800]**	**0.751 [0.7042,0.761]**	**0.539 [0.492,0.586]**	**0.72**

Table A4: Stepwise Logistic Regression Results ($\boldsymbol{\alpha = 0.2}$)

	Cofe.	std err	z	P > |z|	[0.025	0.975]
const	-7.2557	0.777	-9.342	0.000	-8.778	-5.733
Age	0.0777	0.004	19.276	0.000	0.070	0.086
Gender	0.6699	0.081	8.302	0.000	0.512	0.828
Triglyceride	0.2896	0.075	3.843	0.000	0.142	0.437
Postprandial Glucose (PC)	-0.083	0.001	-6.981	0.000	-0.011	-0.006
Fasting glucose	0.0105	0.002	4.420	0.000	0.006	0.015
Waist Circumference	0.0131	0.005	2.811	0.005	0.004	0.022
Potassium (K)	-0.3077	0.098	-3.135	0.002	-0.500	-0.115
White Blood Cells (WBC)	0.0568	0.023	2.424	0.015	0.011	0.103
Smoking	0.3486	0.108	3.228	0.001	0.137	0.560
Areca	-0.5290	0.182	-2.912	0.004	-0.885	-0.173
Hypertension	0.2356	0.115	2.040	0.041	0.009	0.462
Alanine Transaminase (ALT)	0.0034	0.002	1.876	0.061	-0.000	0.007

Gallbladder Polyps	0.1647	0.088	1.864	0.062	-0.008	0.338

2. Figures

Cohort Construction and Data Partitioning

Figure A1 illustrates the sequential inclusion and exclusion process applied to the initial annual health checkup cohort (N = 49,522). Patients were excluded stepwise if they (1) did not meet fatty liver diagnostic criteria, (2) did not satisfy MASLD definition, (3) had incomplete clinical, biochemical, or lifestyle variables required for model construction, or (4) had inadequate bowel preparation that could compromise colonoscopy-based adenoma labeling.

Importantly, patients with inadequate bowel preparation during colonoscopy were excluded to ensure reliable outcome labeling. Suboptimal bowel preparation may impair mucosal visualization and increase the risk of missed adenomas, introducing potential outcome misclassification. Exclusion of patients with inadequate bowel preparation was performed to reduce outcome misclassification. Poor bowel preparation is known to increase the risk of missed adenomas, thereby introducing label noise into negative classifications.

Absolute sample counts remaining after each exclusion step are reported to ensure transparency. The final analytic cohort comprised 3,971 patients (Adenoma: 2,443 [61.5%]; Non-adenoma: 1,528 [38.5%]). The dataset was subsequently evaluated using stratified 10-fold cross-validation, preserving class proportions across folds. The final analytic cohort (N = 3,971) was partitioned into ten approximately equal folds (≈397 samples per fold), maintaining the overall adenoma prevalence of 61.5% within each fold. In each iteration: Eight folds were used for training (≈3,176 samples), One fold was used for validation (≈397 samples), One fold was used for testing (≈397 samples).

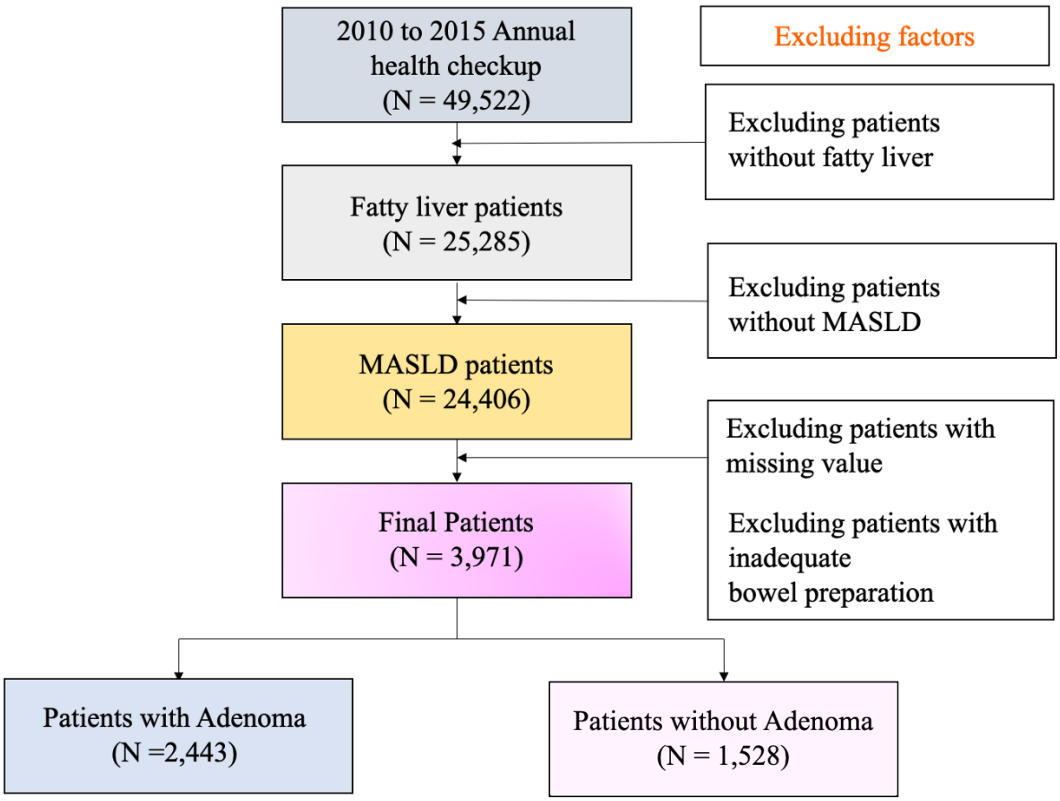


Figure A1 Overview of the Data Collection Process

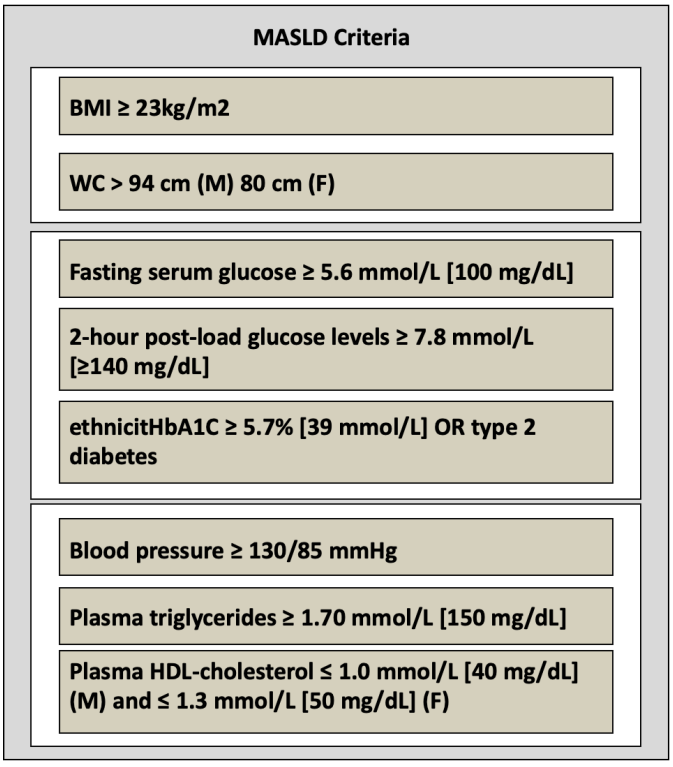
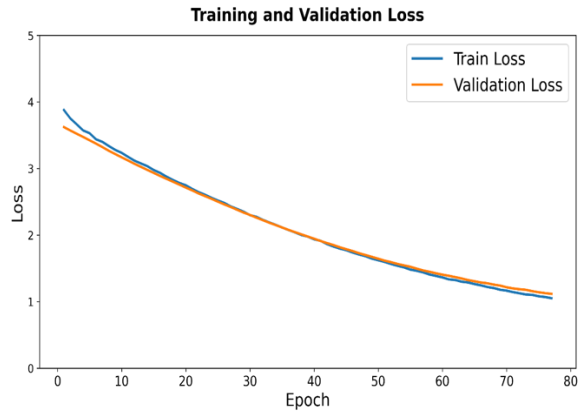
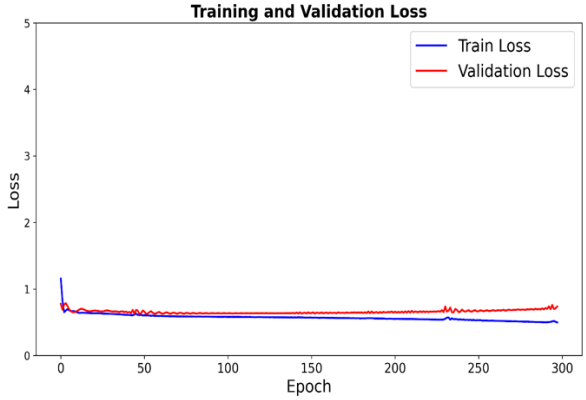
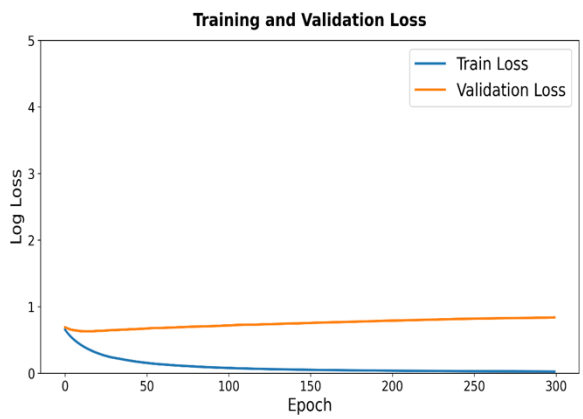
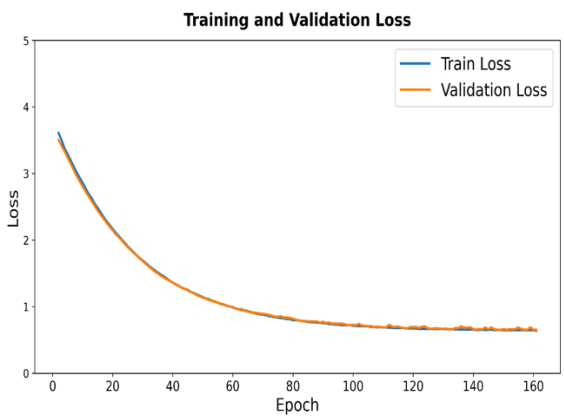


Figure A2 Diagnostic Criteria for MASLD

Model Convergence and Early Stopping Strategy

Figure A3 illustrates the training and validation loss curves for CNN, Transformer, XGBoost, LSTM, LSTM with Attention, and our proposed guided attention model. The loss curves provide a visual assessment of model convergence and generalization. In general, stable learning is indicated when both training and validation losses decrease smoothly and remain close to each other. In contrast, overfitting is suggested when the training loss continues to decline but the validation loss plateaus or increases. As shown in Figure A3, XGBoost exhibits a clear overfitting tendency. Its training loss decreases rapidly and approaches a very low level, while the validation loss begins to increase after the early iterations. This widening gap between training and validation loss suggests that the model increasingly fits the training data but does not generalize equally well to unseen validation data. The Transformer model also shows unstable validation loss fluctuations, indicating less stable convergence. In contrast, the LSTM-based models show smoother convergence patterns. The proposed guided attention model demonstrates a relatively small gap between training and validation loss, indicating stable learning and improved generalization. Early stopping was applied based on validation loss to reduce overfitting and prevent unnecessary training beyond the point of best validation performance.

CNN	Transformer
	
XGBoost	LSTM
	

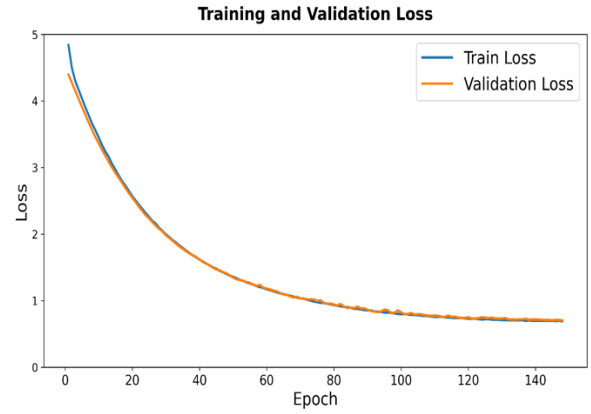
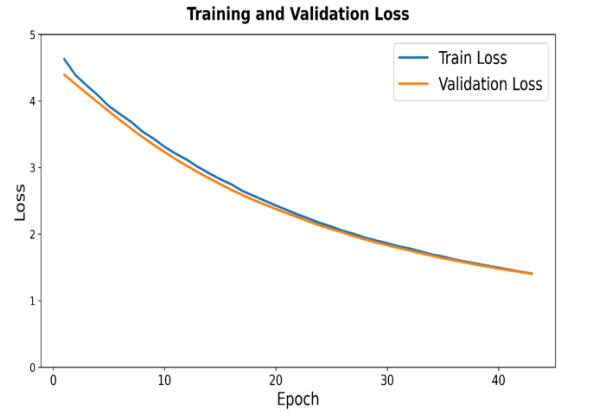
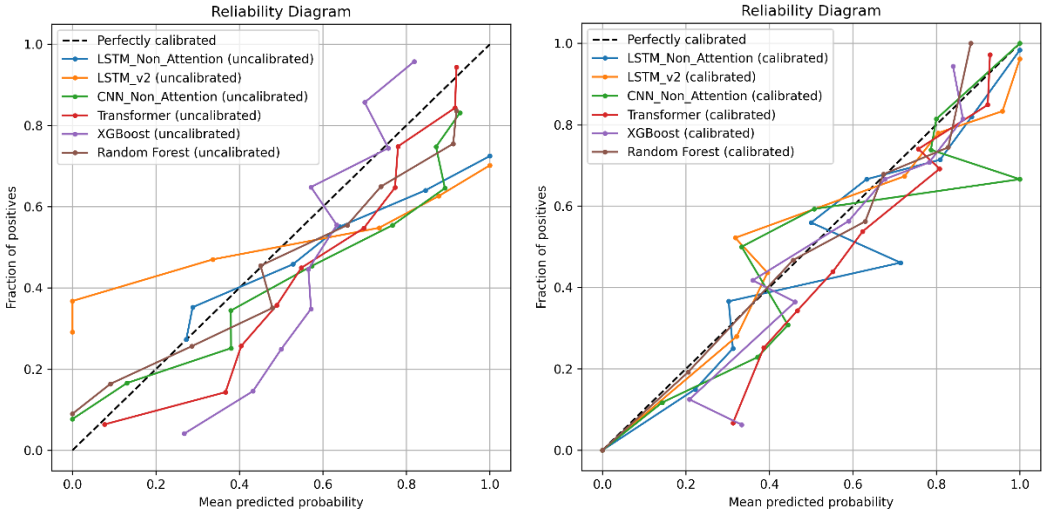
LSTM with Attention	Our proposed method
	

Figure A3. Training and Validation Loss Curves With Early Stopping Across Models.

Calibration Analysis

Calibration curves before and after probability scaling are shown in Figures A3a and A3b. The dashed diagonal represents perfect calibration, where predicted probabilities equal observed event frequencies. Prior to calibration, several models exhibited systematic deviations from the ideal line, particularly in intermediate and higher probability ranges, indicating over- or underestimation of risk despite comparable discrimination performance. This highlights that similar AUC values do not guarantee reliable probability estimates.

After probability scaling, the calibration curves demonstrate improved alignment with the diagonal across most probability bins, with reduced overconfidence and greater stability in mid-risk regions. Discrimination performance remained largely unchanged, indicating that probability scaling improved risk reliability rather than ranking behavior. Because screening decisions are threshold-based, improved calibration enhances the validity of referral policies and reduces the risk of inappropriate over- or under-referral in clinical deployment.



A4a **A4b**

Figure A4 Calibration Analysis

SHAP attention heatmap

To provide concrete evidence of interpretive behavior beyond global SHAP summaries, we visualize feature-level attribution patterns using an attention heatmap (Figure A4). The upper panel shows the aggregated risk trajectory, while the lower panel illustrates feature-wise contribution intensities across samples.

The guided attention mechanism concentrates weight on metabolically relevant variables in high-risk individuals, particularly within lipid and glycemic domains. Compared with baseline attention-free models, the attribution distribution is less diffuse and exhibits greater stability across adjacent samples. Notably, subgroup inspection of individuals with confirmed advanced pathology reveals consistent amplification of clinically coherent features rather than isolated spurious signals.

These findings suggest that guided attention refines representational focus toward clinically meaningful feature clusters, supporting interpretive stability rather than merely replicating standard SHAP importance rankings.

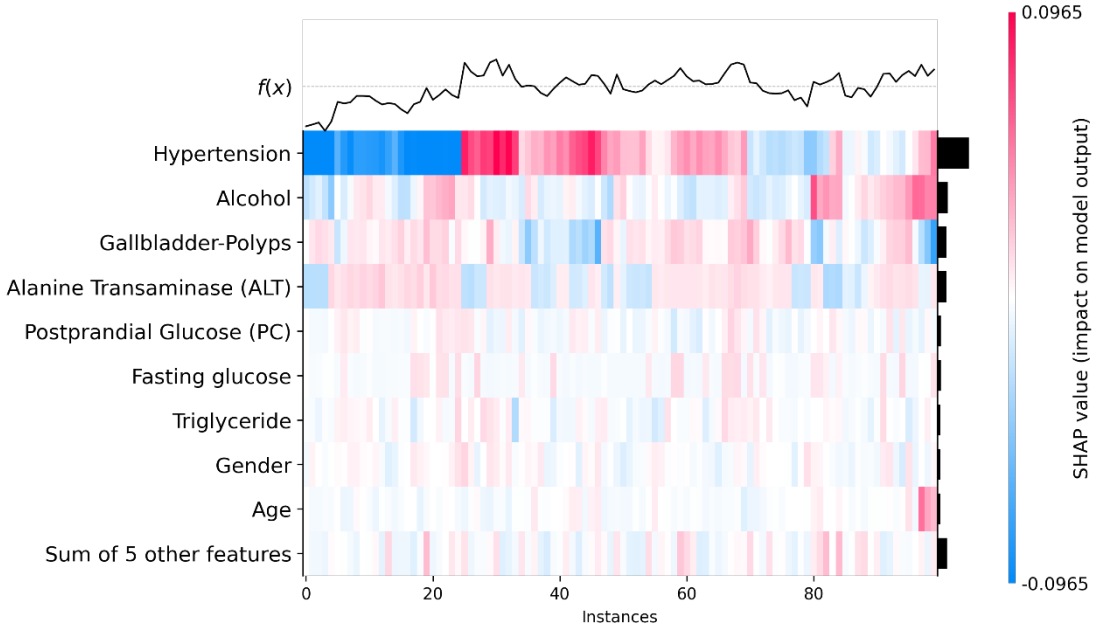


Figure A5 Visualization of guided attention attribution patterns.

Patient-level Examples of AI-driven Risk Stratification

To further substantiate the interpretability of the proposed framework, we present two patient-level case studies demonstrating the alignment between guided attention patterns, SHAP-based local explanations, and independent clinician evaluation.

- **Case 1: High-Risk Patient Prediction (Figure A6)**

The first case involves a 59-year-old male patient classified as high risk, with a predicted probability of 0.55, exceeding the operational threshold (0.45). The guided attention heatmap shows concentrated weights on metabolically relevant features, particularly ALT, postprandial glucose, and age. The corresponding SHAP local explanation indicates that these variables contribute positively to the predicted risk estimate. Independent clinician review confirmed that the highlighted features are clinically coherent with established risk factors for adenomatous polyps. Based on expert feedback, the guide vector was refined (from v1.0 to v2.0), demonstrating the feasibility of iterative expert-in-the-loop calibration.

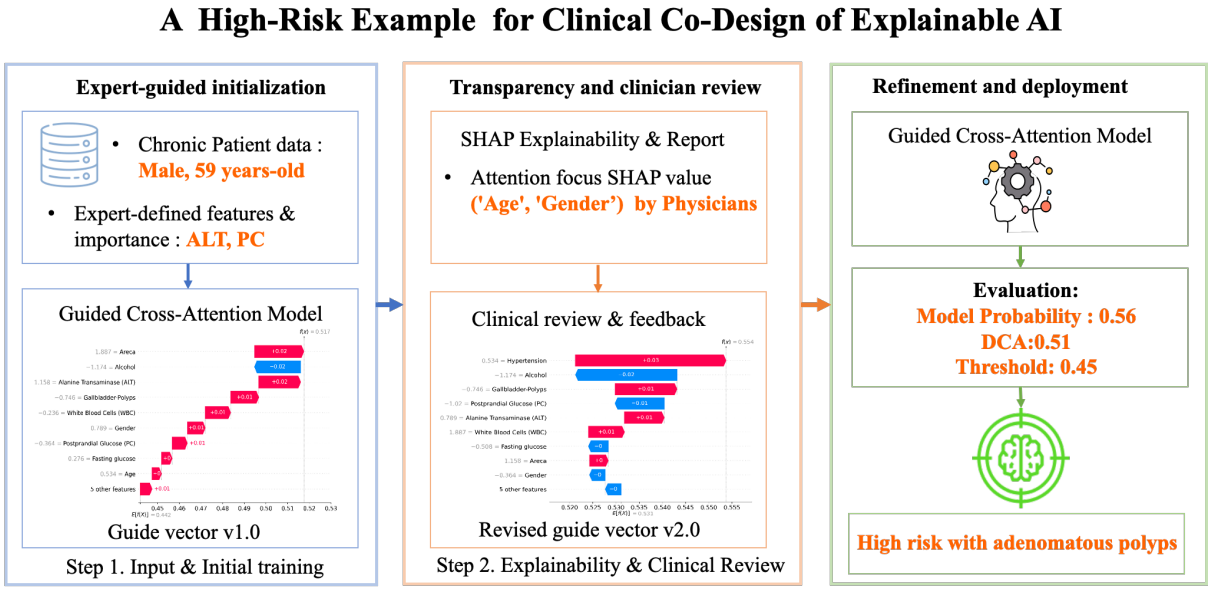


Figure A6 High-risk Example of an AI-driven risk stratification

- **Case 2: Low-Risk Patient Prediction (Figure A7)**

The second case involves a 35-year-old male patient classified as low risk, with a predicted probability below the operational threshold (0.45). The guided attention heatmap allocates greater weight to relatively normal metabolic indicators and younger age, reflecting protective patterns in the representation. The corresponding SHAP local explanation indicates that these factors contribute negatively to the risk estimate. Independent clinician review confirmed that the attribution pattern is consistent with standard clinical risk assessment principles.

A Low-Risk Example for Clinical Co-Design of Explainable AI

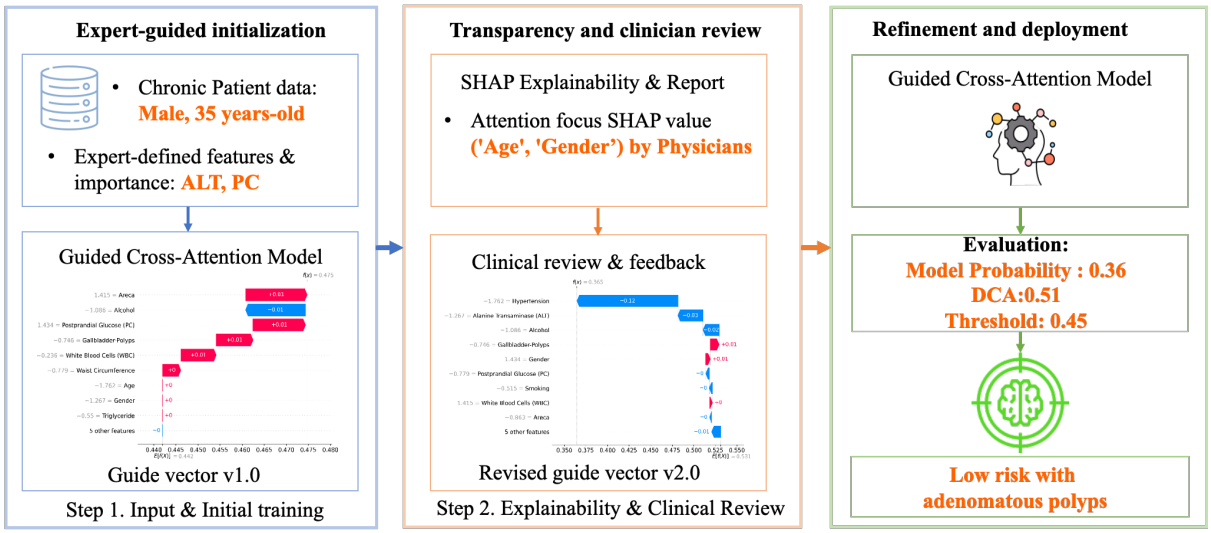


Figure A7 Low-risk Example of an AI-driven risk stratification