GALLSTONE PREDICTION

PRESENTED BY: GROUP 3

Project Team Overview

Team SignalX

Yuyang Chen
GitHub: yuyangchen11

Ruihe Zhang
GitHub: Mudkipython

Zhaihan Gong
GitHub: yuehuatong

Wendy Xu
GitHub: ZihanXu12

Serena Sun
GitHub: YS - 02

Name of the GitHub repository: Gallstone_Prediction

INDEX

01 Business Context

02 Hypothesis

03 Data

04 Modelling Approach

05 Model Evaluation

06 Results

07 Explainability of Results

08 Threats to Validity

09 Conclusion & Reccomendation

10 Lessons Learned & Next Step

01 Business Context

Background

- Gallstones affect 10–15% of the global population
- Delayed diagnosis lead to complications and higher healthcare costs.

Current Limitations

- Ultrasound is accurate but not scalable for population screening
- Limited imaging capacity leads to delayed evaluation

Key Opportunity

- Use routine clinical data (e.g., BMI, lab tests, demographics) to build a **low-cost, scalable** risk stratification tool before imaging

02 Hypothesis

Research Question

Can routine non-acute clinical indicators support early gallstone risk stratification?

Core Hypothesis

- Routine demographic and metabolic indicators contain measurable predictive signal for pre-imaging risk stratification.

Hypothesis Testing Strategy

- H0: Routine non-acute metabolic and demographic indicators have no predictive power beyond random chance.
- Evaluate multiple models under stratified cross-validation.
- Compare logistic regression against tree-based ensembles.
- **Primary metric: PR-AUC**

We use PR-AUC because it better reflects positive-class detection performance in clinical screening settings.

03 Data

Dataset Snapshot

- Publicly available clinical dataset obtained from Kaggle, originally collected at the Internal Medicine Outpatient Clinic of Ankara VM Medical Park Hospital.
- 319 patients
- 161 positive (49.5%)
- 38 structured clinical features
- June 2022 – June 2023
- Ethics approved (E2-23-4632)
- Balanced disease distribution
- Non-imaging features only
- Cross-sectional; clinically diagnosed gallstone status

Demographics & Comorbidities

- Age
- Gender
- Comorbidity
- Coronary Artery Disease (CAD)
- Hypothyroidism
- Hyperlipidemia
- Diabetes Mellitus (DM)

Anthropometrics

- Height
- Weight
- Body Mass Index (BMI)
- Obesity (%)

Bioimpedance – Body Composition

- Total Body Water (TBW)
- Extracellular Water (ECW)
- Intracellular Water (ICW)
- Extracellular Fluid / TBW (ECF/TBW)
- Total Body Fat Ratio (TBFR) (%)
- Total Fat Content (TFC)
- Lean Mass (LM) (%)
- Muscle Mass (MM)
- Bone Mass (BM)
- Body Protein (%)
- Visceral Fat Rating (VFR)
- Visceral Fat Area (VFA)
- Visceral Muscle Area (VMA)
- Hepatic Fat Accumulation (HFA)

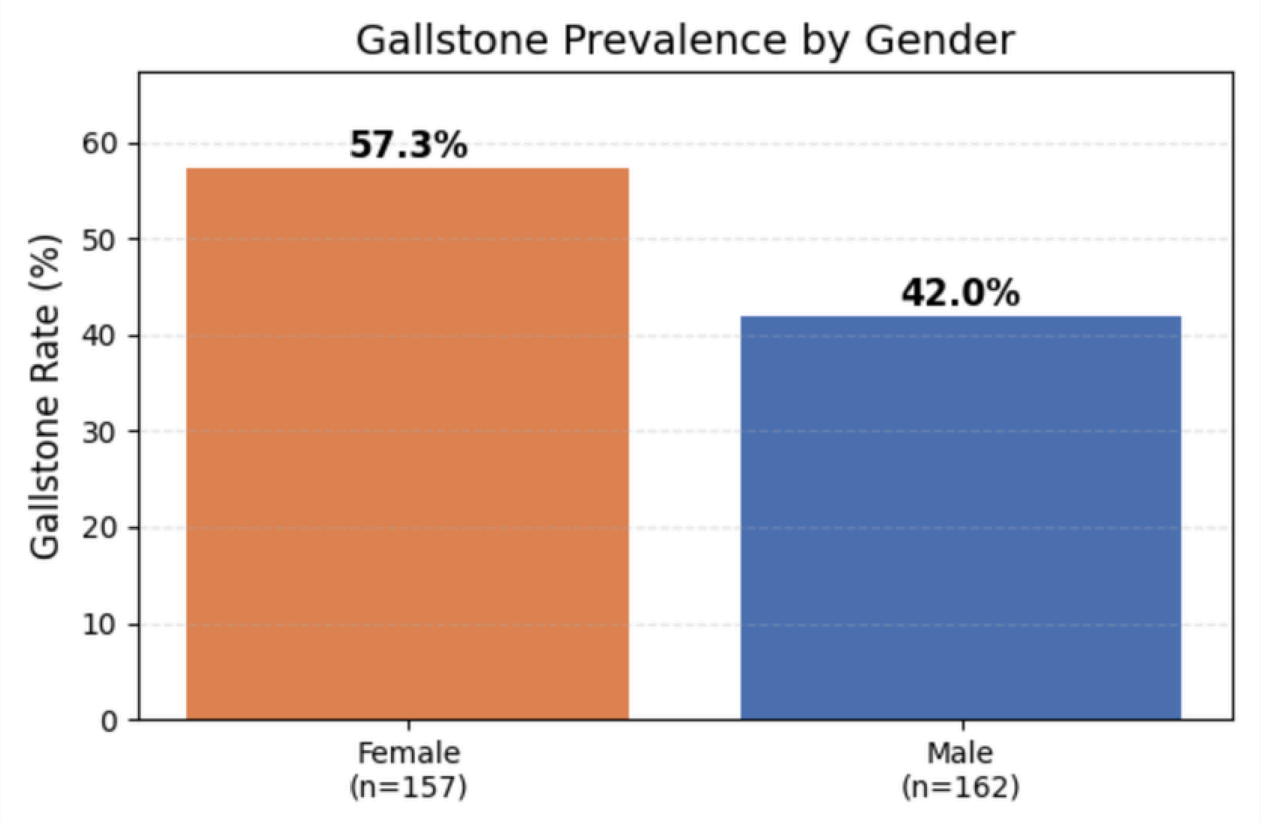
Laboratory Markers

- Glucose
- Total Cholesterol (TC)
- LDL
- HDL
- Triglyceride
- AST
- ALT
- ALP
- Creatinine
- Glomerular Filtration Rate (GFR)
- C-Reactive Protein (CRP)
- Hemoglobin (HGB)
- Vitamin D

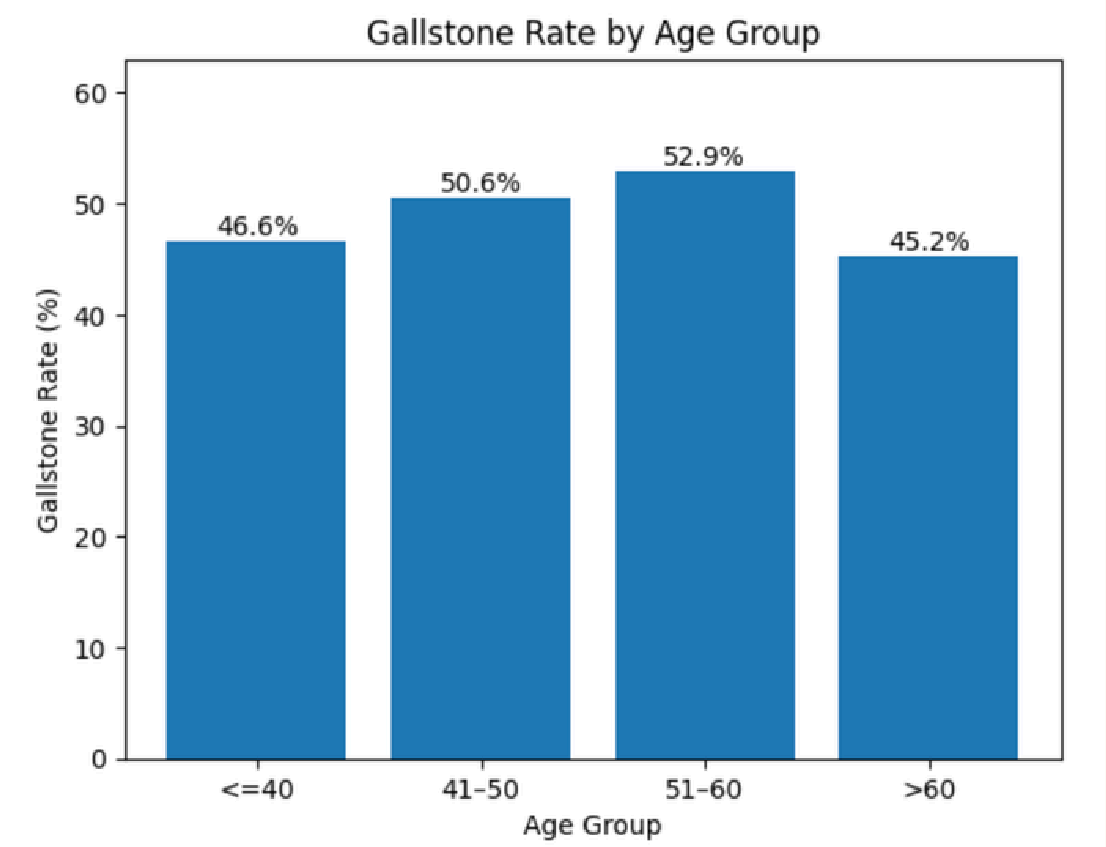
03

Data - Demographic Risk Patterns

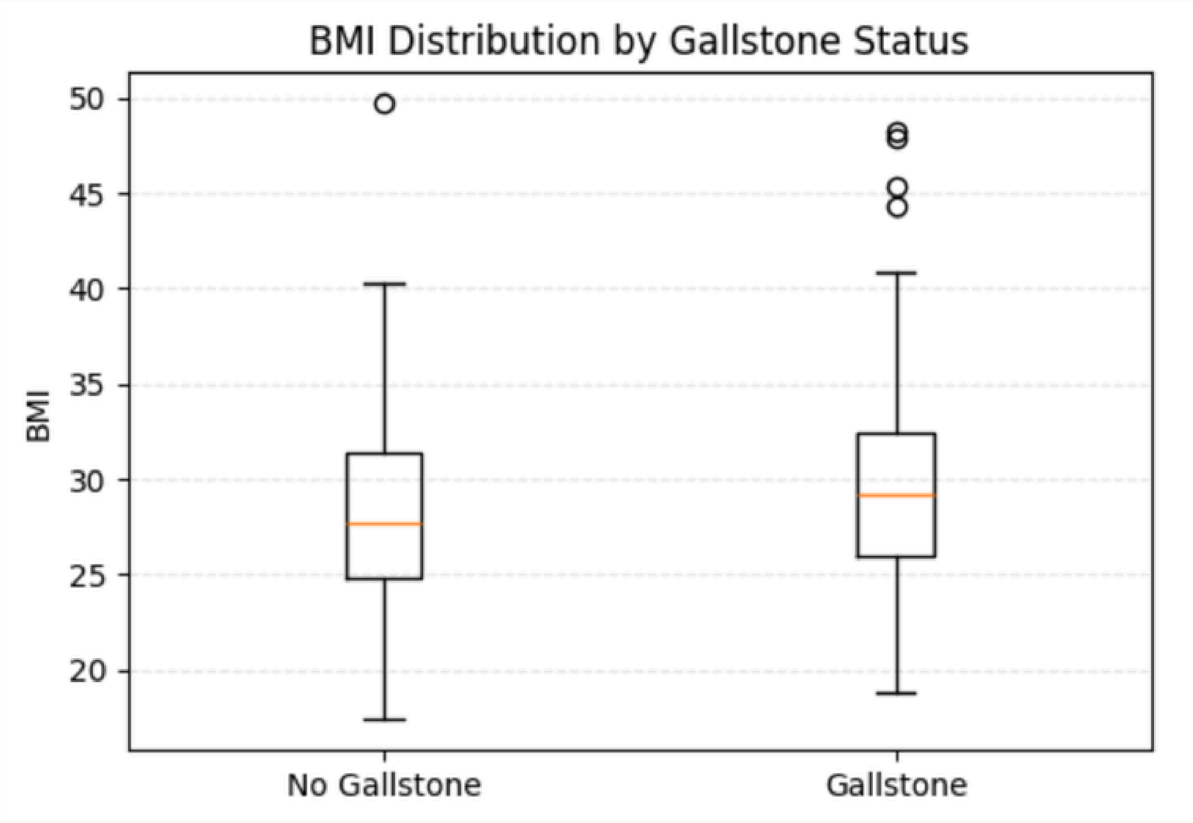
Gender vs Gallstone Rate



Age Group vs Gallstone Rate



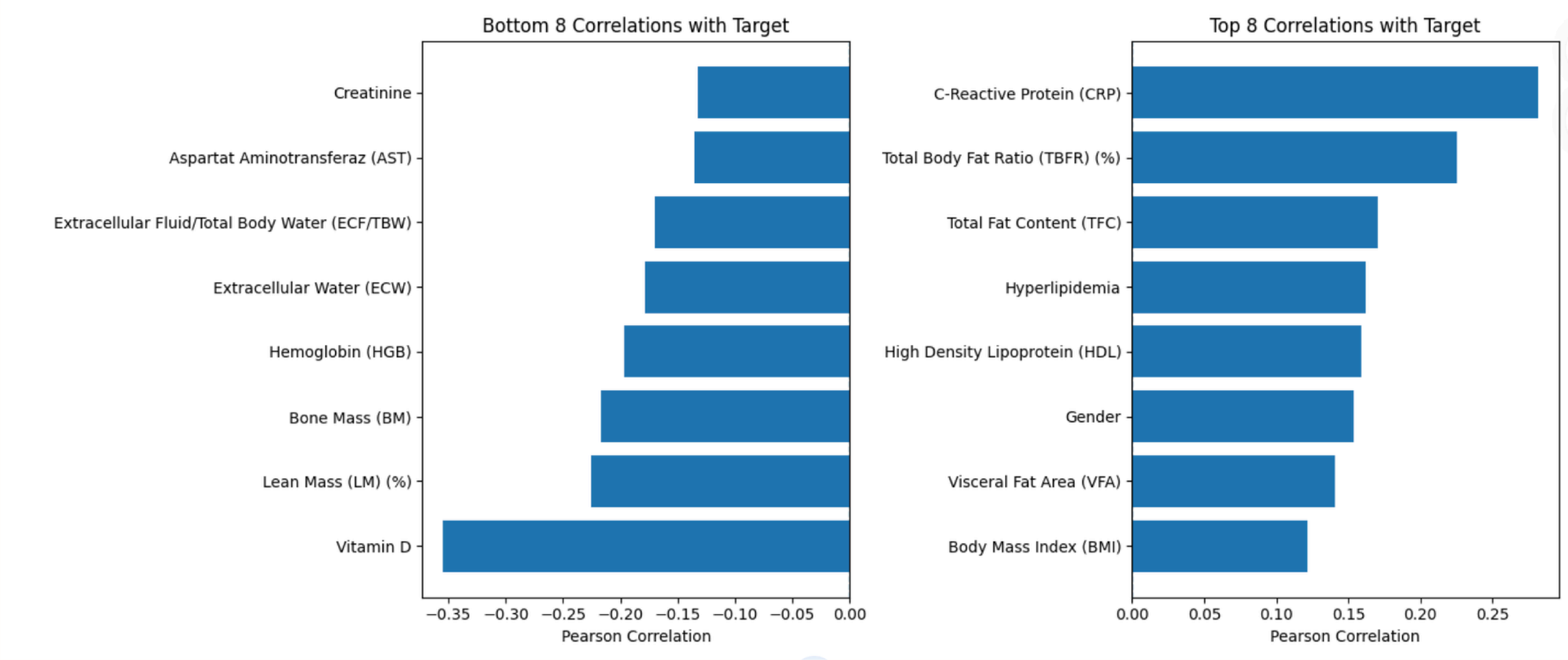
BMI Boxplot by Target



03

Data

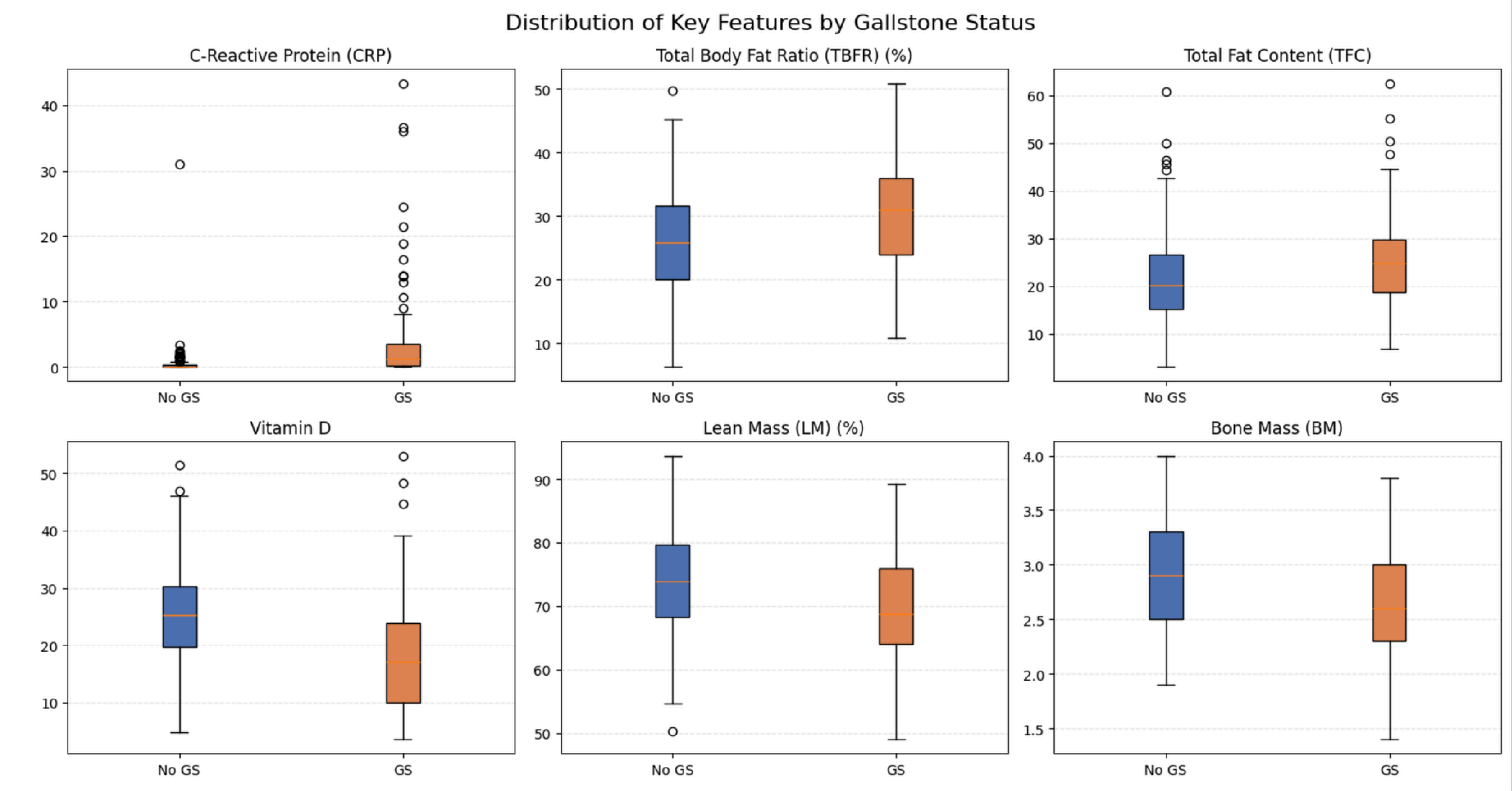
Top/Bottom Correlated Features



03

Data

Top 3 positive & top 3 negative correlated features



03

Data

Final Feature Selection Strategy

Leakage Control – Removed Features

To ensure early-stage risk prediction and avoid data leakage, we excluded acute diagnostic markers:

- C-Reactive Protein (CRP)
- AST
- ALT
- ALP
- AST/ALT ratio

These markers reflect acute inflammatory or hepatobiliary status and may capture post-diagnostic signals. Removing them ensures we are truly modeling early-stage risk rather than symptomatic status.

Cross-Validated Performance (PR-AUC)

Method	# Features	PR-AUC
Mutual Information	20	0.829
RFECV	33	0.842
LASSO (Selected)	**28**	**0.843**

Final Decision

- LASSO achieved highest PR-AUC
- Provided automatic sparsity
- Reduced dimensionality (33 → 28)
- Maintained interpretability

04 Model

Data → Baseline Models → Cross-Validation → Feature Refinement → Final Model Selection

Modeling Strategy

- Supervised learning
- Binary classification (Gallstone: Yes/No)
- Stratified K-fold cross-validation
- Out-of-fold prediction for unbiased evaluation

Model Evaluation

- 319 samples, 5-fold stratified cross-validation
- 49.5% positive
- Hyperparameter tuning via cross-validation

Benchmark Models

- Logistic Regression
- Random Forest
- Histogram-based Gradient Boosting
- Light Gradient Boosting Machine
- XGBoost

Evaluation Metrics

- Primary: PR-AUC
    - precision–recall trade-off
- ROC-AUC
- Recall
- Calibration: Brier

05 Model Validation

Model Benchmark Results

Model Comparison

Model	ROC-AUC	PR-AUC	Recall	Brier
Logistic Regression	**0.837**	**0.843**	0.829	**0.163**
LightGBM	0.789	0.788	0.766	0.209
XGBoost	0.793	0.783	0.778	0.192
Random Forest	0.777	0.752	0.835	0.195
HistGB	0.789	0.789	**0.937**	0.199

Model Selection Strategy

Final Selection - Logistic Regression

Performance

- Best ROC-AUC (0.837) and PR-AUC (0.843)
- Lowest Brier Score → well-calibrated probabilities

Robustness

- Stable after removing confounders
- Less overfitting than tree-based models

Interpretability

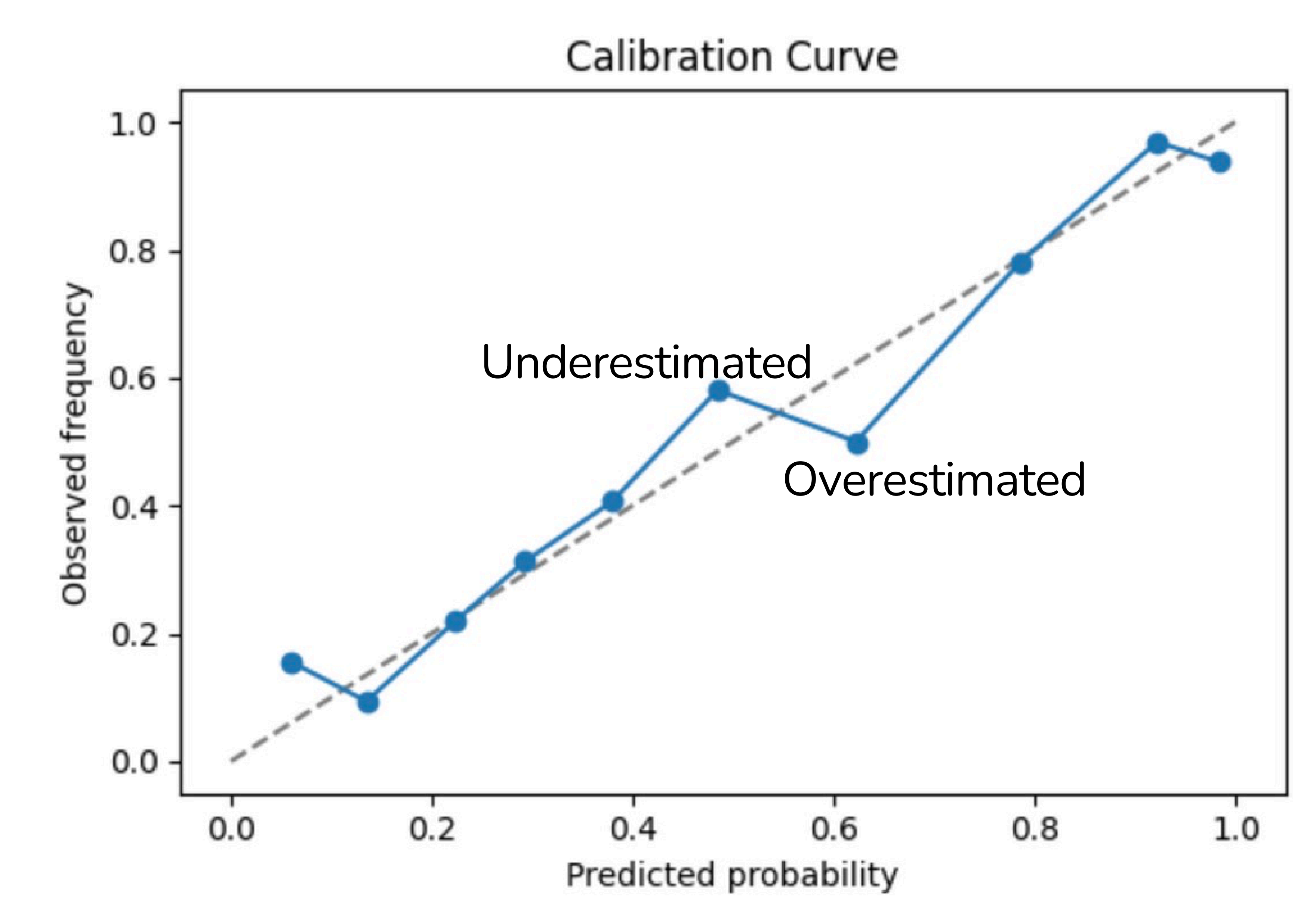
- Clinicians can understand and trust the prediction
- Enables a simple risk calculator for routine use

Deployment

- No complex infrastructure needed
- Easy integration into existing LIS/EMR systems

06

Model Performance Results

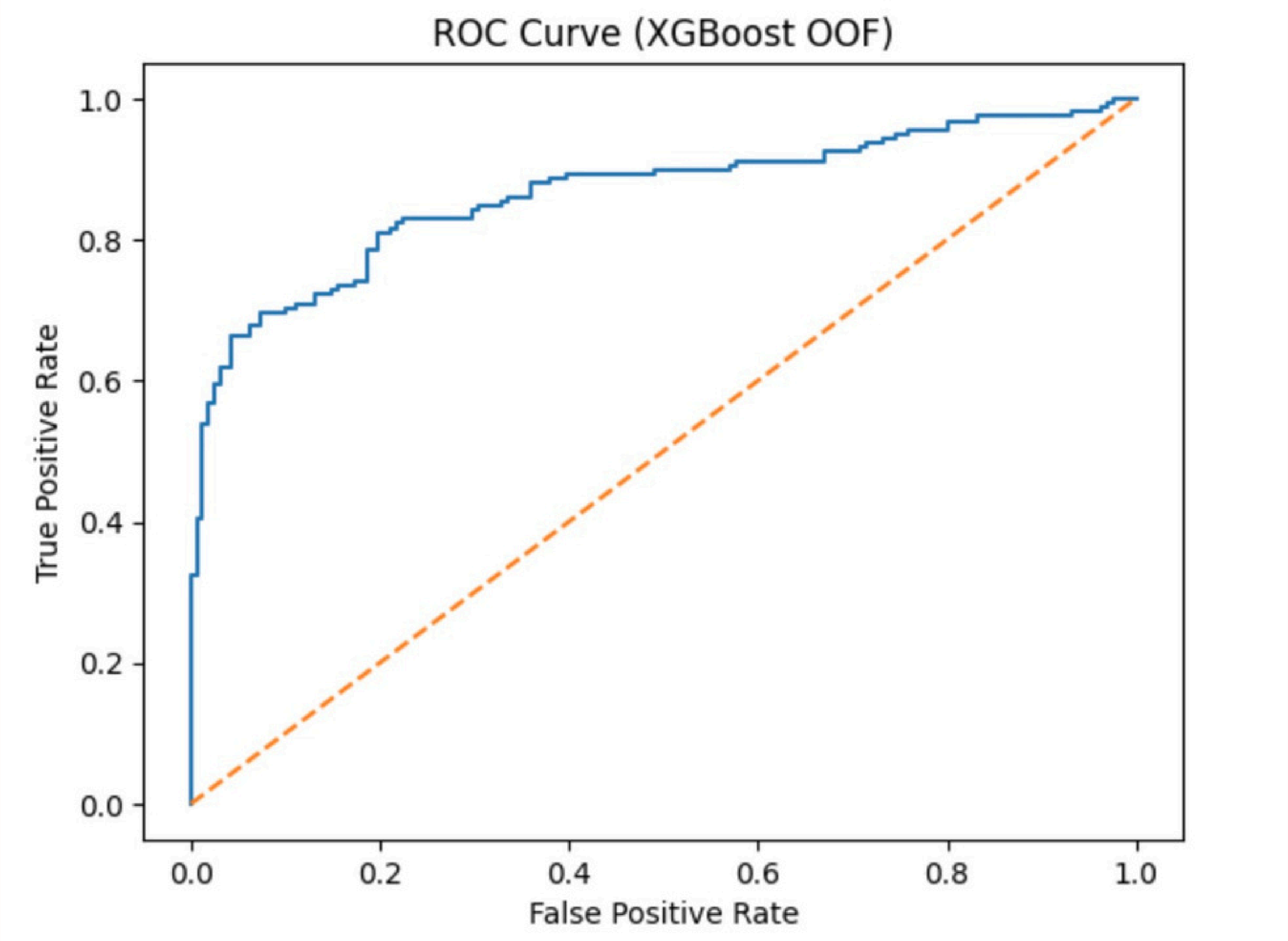


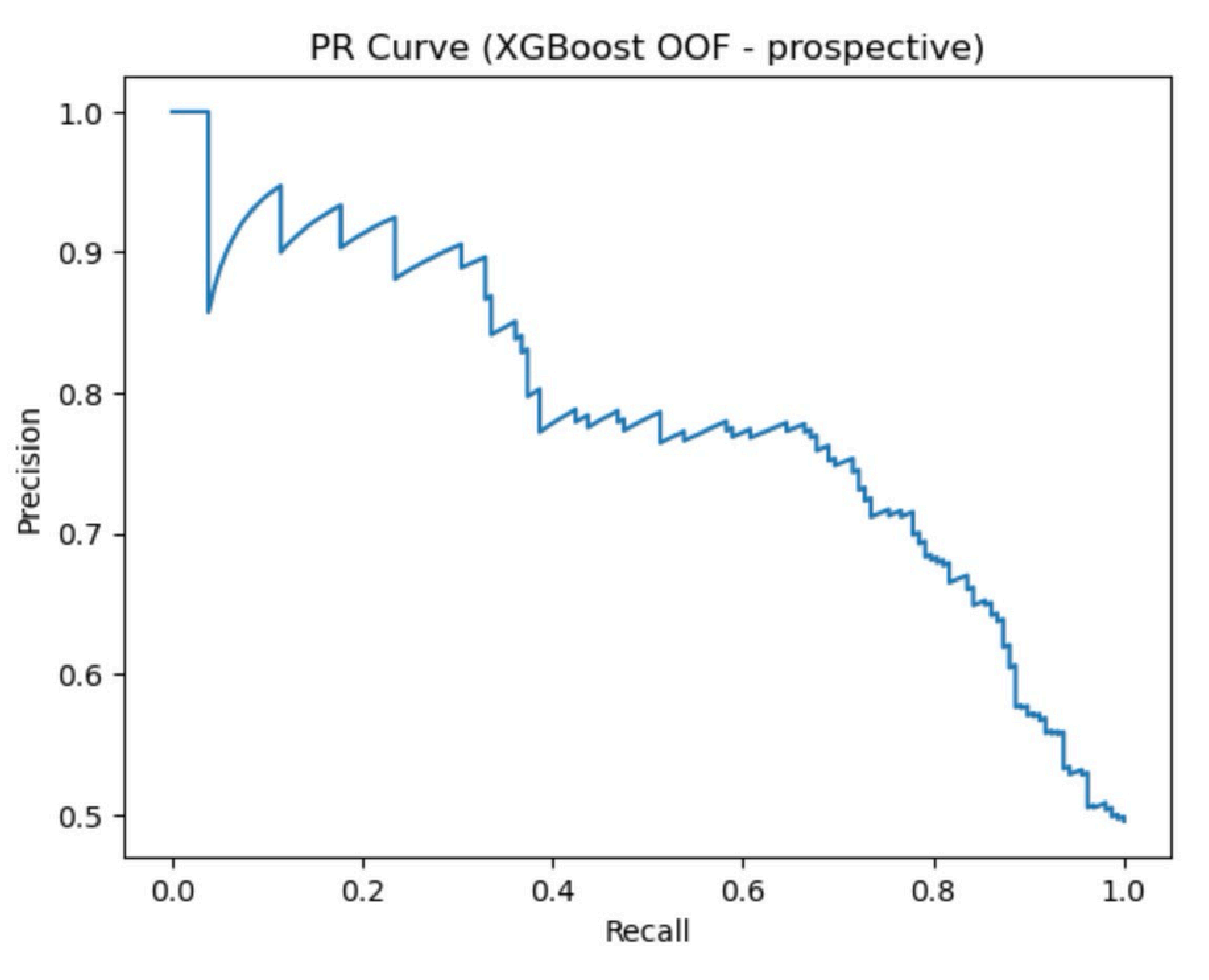
Compare with SOTA / Previous Work

- Traditional clinical risk assessment relies on manual evaluation of metabolic indicators
- Modern SOTA methods for structured medical data use tree-based ensemble models
- Consistent with ensemble learning theory, boosting models in our experiments achieved strong positive-class discrimination
- Our final model (Logistic Regression) achieves comparable performance (PR-AUC = 0.843) while maximizing clinical interpretability for screening deployment

OUR MODELLING STRATEGY BALANCES SOTA PERFORMANCE WITH CLINICAL ACTIONABILITY

06 Results





ROC / PR evaluate performance across all thresholds, and confirm strong discrimination
Clinical deployment requires choosing a specific decision threshold

06 Results - Threshold Selection

Why Threshold Matters

- Model outputs probabilities (0–1)
- Clinical decision requires binary classification (High / Low risk)

Threshold Selection Strategy

- Selected using OOF predictions
- Optimized by maximizing F1 score
- Balances precision and recall

Final Operating Threshold

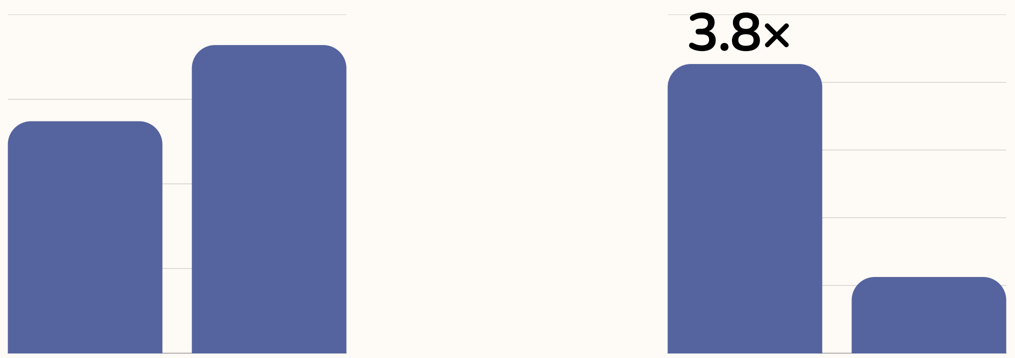
LR optimal threshold (OOF): 0.372

Decision rule:

- If predicted probability ≥ 0.372 → High Risk
- Otherwise → Low Risk

Risk Seperation Result Under Threshold of 0.597

Risk Tire	Population	Observed Gallstone Rate
High Risk	175	74.9%
Low RIsk	144	18.8%



06

Results - Error Analysis

Confusion Matrix (at threshold of 0.372)

	Predicted Low	Predicted High
Actual Low	136 (TN)	25 (FP)
Actual High	38 (FN)	120 (TP)

Error Structure

- False Negatives: 38 patients (24.1%)
- False Positives: 25 patients (15.5%)
- Model favors precision over recall

Key Takeaways

- High recall (76%) supports use as a screening tool
- Moderate false positives acceptable for early detection
- Remaining 24.1% FN highlights need for physician oversight

07

Application

How should strategy adapt to hospital capacity?

Total: 319

Policy	Patient Flagged	Missed Cases	Extra Follow-ups (FP)	F1
Balanced Threshold 0.441	175	27	44	0.787
Recall Focused (reduce missed diagnoses) Threshold 0.54	254	8	104	0.728
Top 50 (resource-limited)	50	110	2	0.462

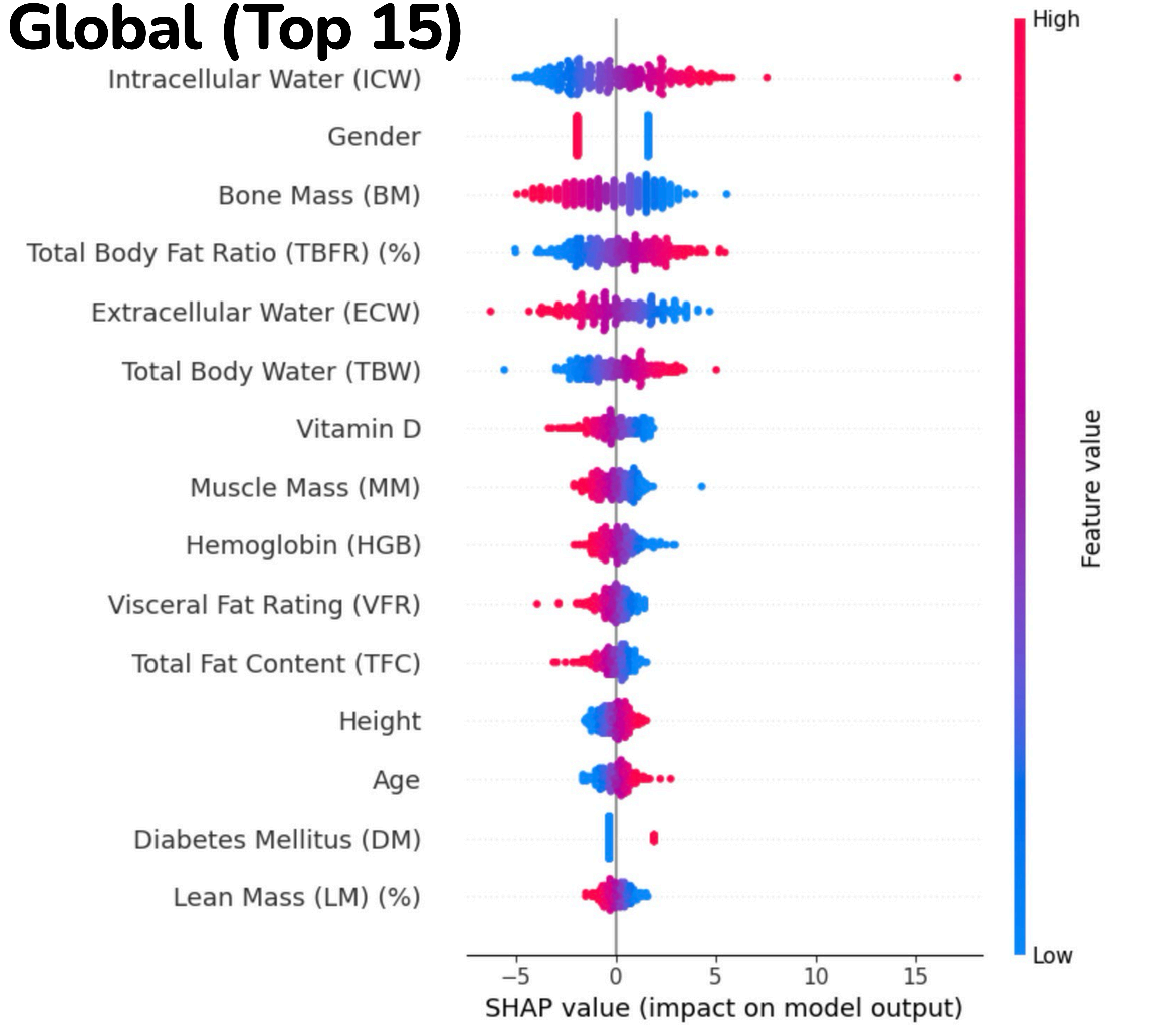
- F1-opt balances precision and recall
- Cost-sensitive maximizes recall, minimizing missed cases
- Top-k approach prioritizes highest-risk individuals

Different operating thresholds enable flexible clinical deployment depending on resource availability.

07

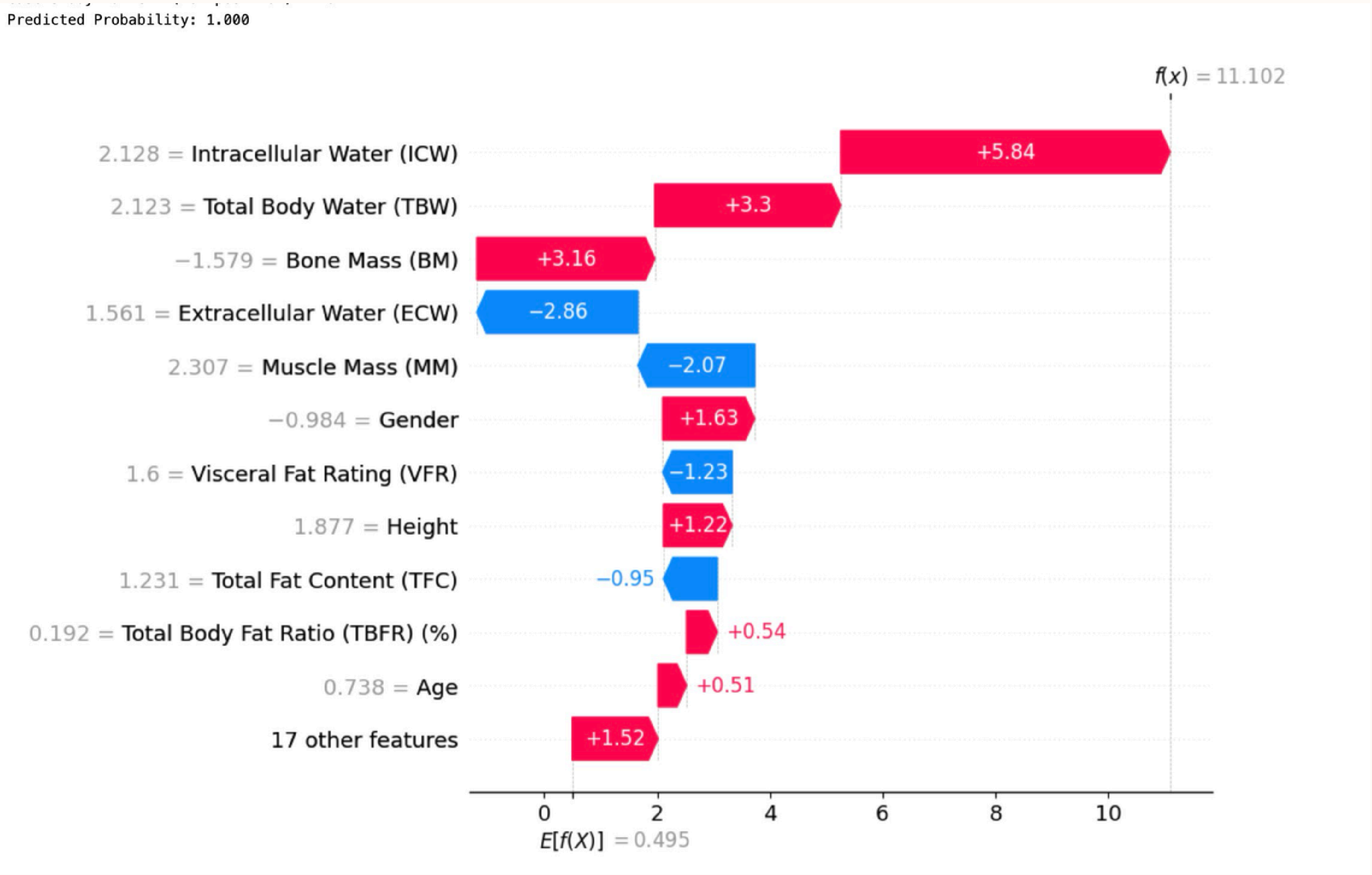
Feature Importance & Case Study

Global (Top 15)



High-Risk Patient Example

Patient’s predicted risk = 1.00 (> 0.372 threshold) → flagged as high-risk



support hypothesis: routine demographic and metabolic indicators predict gallstone risk

08 Threats to Validity

01 Sample Size & Generalizability

Model was trained on a relatively small sample (N=319)

02 Lack of External Validation

Not yet validated on an independent external cohort

03 Implementation

Real-world implementation needs system integration, governance approval, and monitoring

09

Conclusions & Recommendations

We recommend pilot deployment under controlled settings

SCREENING CENTERS

- Risk-based screening allocation

CLINICIANS

- SHAP explanations enhance interpretability
- Designed as decision-support, not replacement

POPULATION HEALTH

- Early risk identification using routine data

The proposed model demonstrates strong discriminative ability (ROC-AUC ≈ 0.85) and clinically meaningful risk stratification

10 Lessons Learned & Next Steps

Lessons Learned

- Model performance alone is insufficient
- deployment strategy and threshold design determine impact

Next Steps

01 Validate on an independent external cohort

02 Collect longitudinal data for stronger evidence

03 Conduct a prospective pilot study in screening settings

04 Gather clinician feedback for iterative refinement

05 Establish monitoring framework to track model performance and data drift after deployment.

THANK YOU!