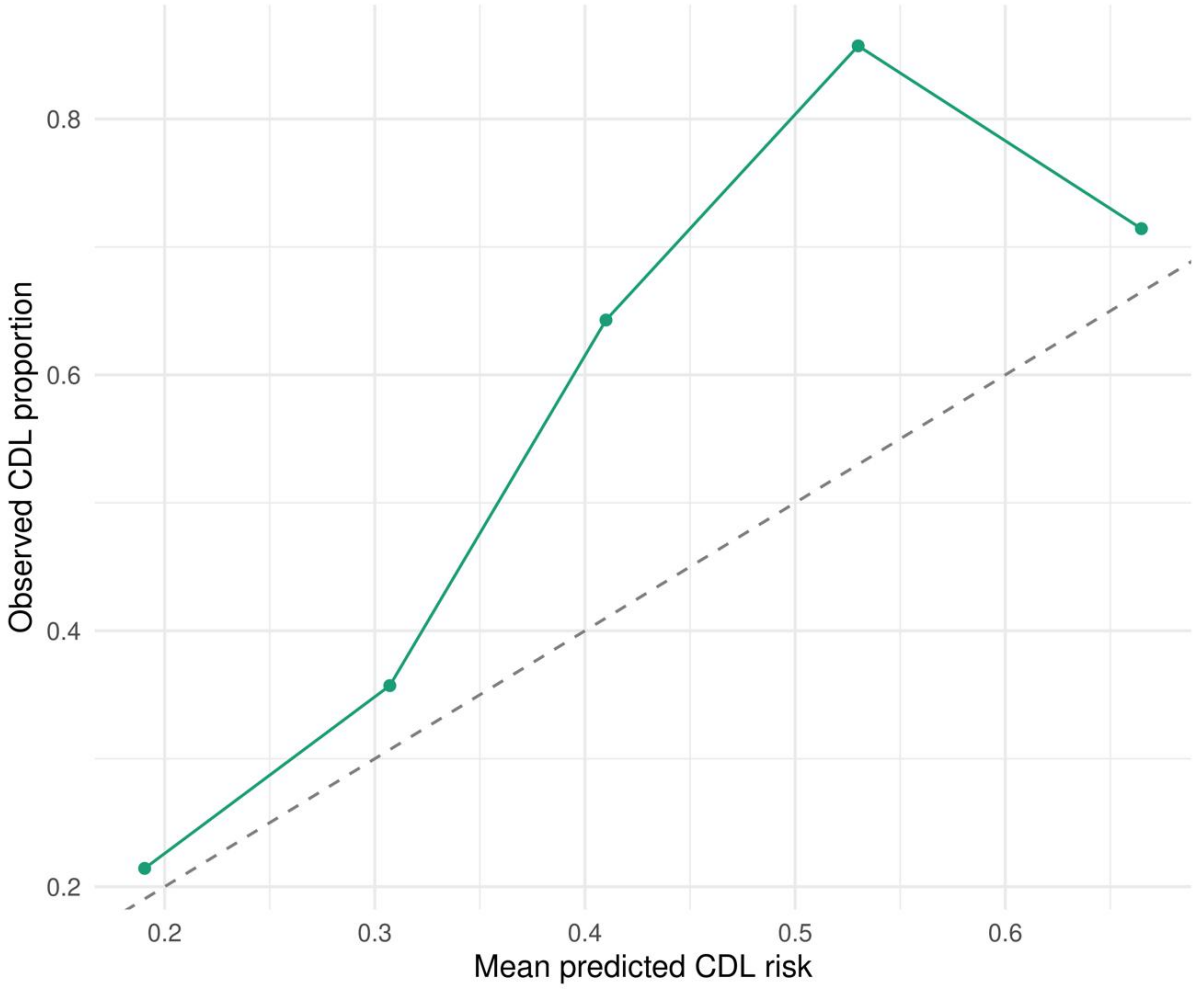
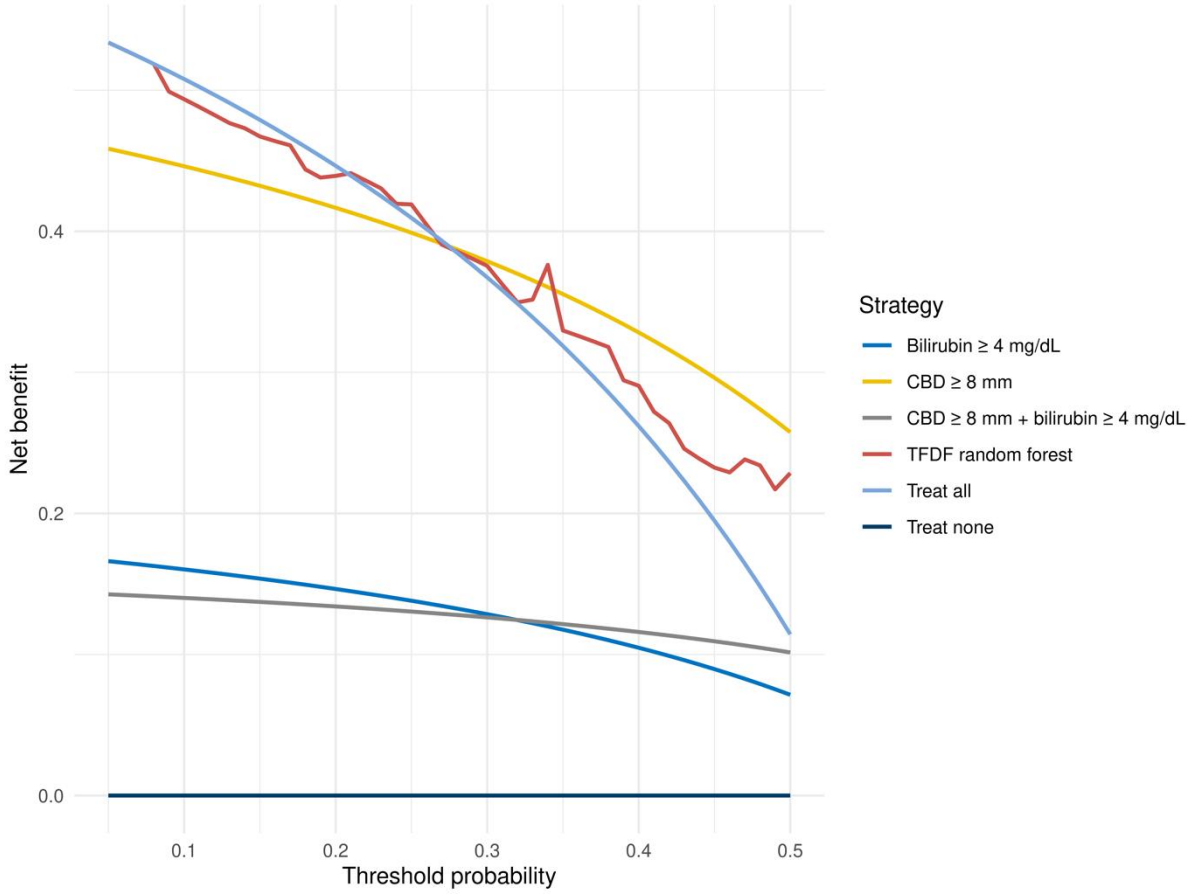
Supplemental material

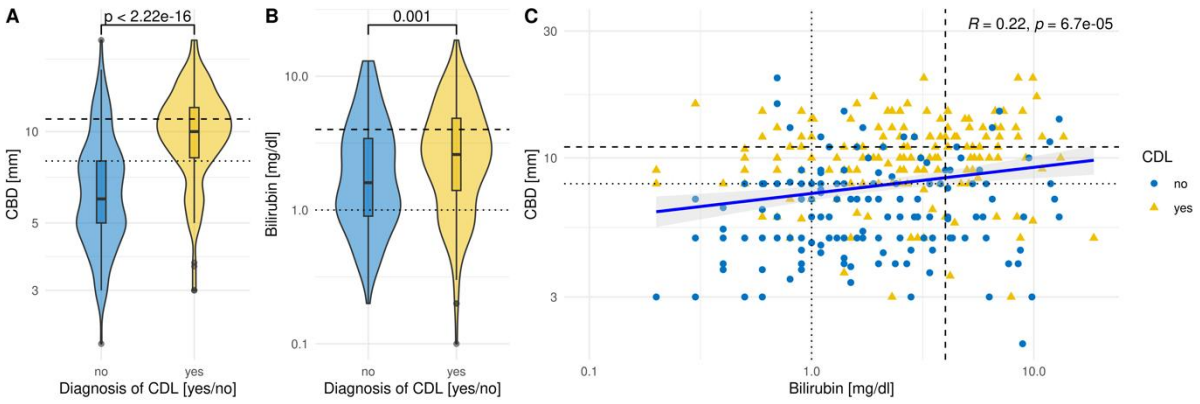
Figures



Supplementary Figure 1: Calibration plot (reliability plot) of the TFDF RF model. Each point represents one quintile of predicted CDL probability. The dashed diagonal line indicates perfect calibration. Points above the diagonal indicate underprediction; points below indicate overprediction. Point size is proportional to the number of patients in each bin. The Brier score, calibration intercept, and calibration slope are reported in the primary model performance table above.

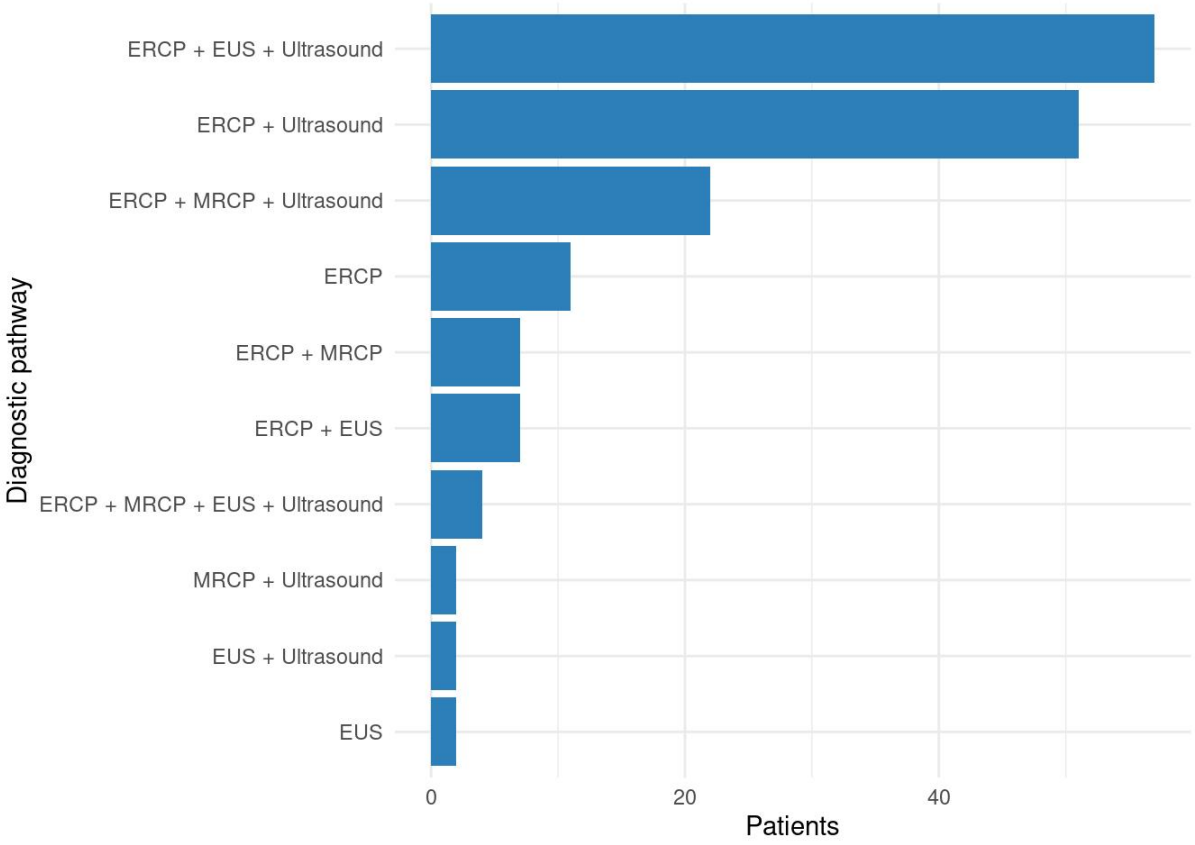


Supplementary Figure 2: Decision curve analysis. Net benefit of the TFDF random-forest model across clinically plausible threshold probabilities, shown against simple diameter- and bilirubin-based decision rules as well as treat-all and treat-none strategies. A positive separation above the treat-none line indicates clinical utility within that threshold range.

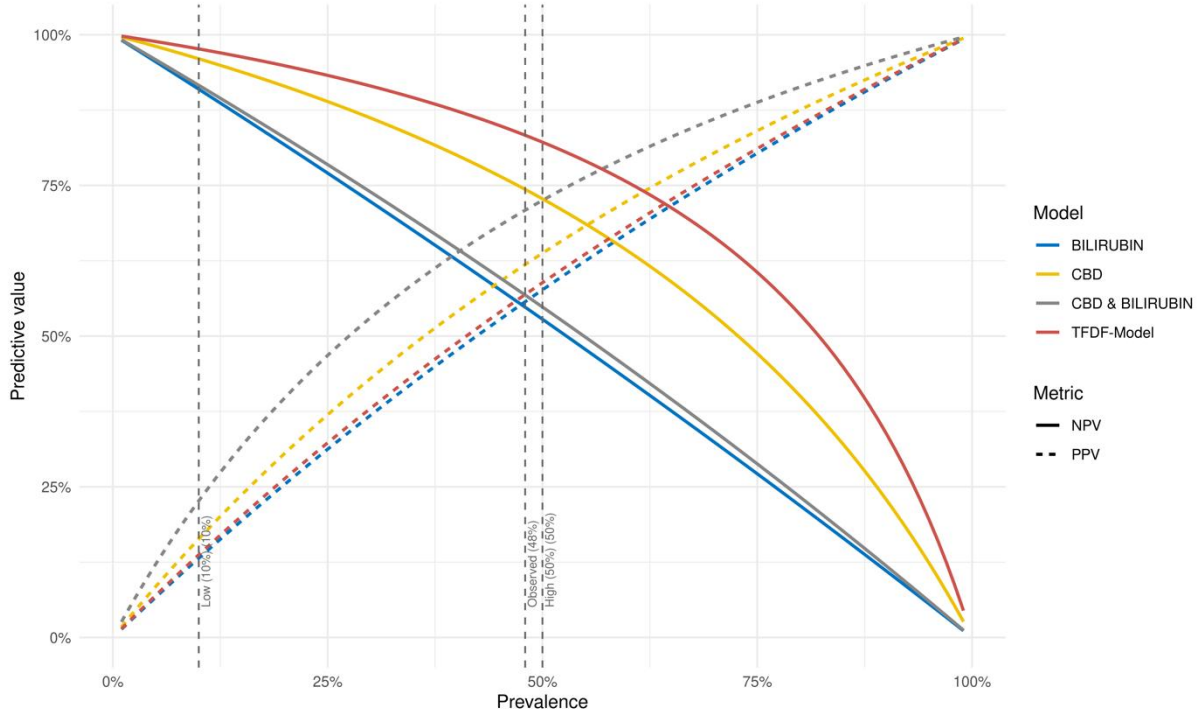


Supplementary Figure 3: Association of CBD width and bilirubin with CDL. A. Violin plot of CBD diameter in patients with (blue) and without (yellow) CDL.

The dotted and dashed horizontal lines indicate CBD diameters of 8 mm and 11 mm, respectively. **B.** Violin plot of bilirubin levels in patients with and without CDL. The dotted and dashed horizontal lines indicate bilirubin levels of 1.0 mg/dL and 4.0 mg/dL, respectively. For panels **A** and **B**, group comparisons were performed using a two-sided Mann–Whitney U test. **C.** Scatter plot showing the relationship between bilirubin and CBD diameter, stratified by CDL status, with Spearman correlation analysis. The dotted and dashed vertical lines indicate bilirubin levels of 1.0 mg/dL and 4.0 mg/dL, respectively, and the dotted and dashed horizontal lines indicate CBD diameters of 8 mm and 11 mm, respectively. Bilirubin is displayed on a logarithmic scale in panels **B** and **C**.



Supplementary Figure 4: Overview of the main diagnostic routes of CDL verification within the included CDL-positive cases.



Supplementary Figure 5: Prevalence vs PPV and NPV. PPV (solid) and NPV (dashed) of the four main decision rules across a range of CDL prevalences (1–99%). Sensitivity and specificity are held fixed at the values estimated from the held-out test set. Vertical dashed lines mark the three illustrative scenarios: low (10%), observed sample (48%), and high (50%) prevalence.

Tables

Supplementary Table 1. Model development and implementation details

Item	Specification
Model family	TensorFlow Decision Forests random forest
Implemented estimator	tfdf.keras.RandomForestModel()
Historical notebook alternative	A GradientBoostedTreesModel() call is present only as a commented-out alternative
Hyperparameters explicitly set	compute_oob_variable_importances = TRUE; otherwise TFDF defaults
Feature selection strategy	Pre-specified admission demographics/laboratories; variables with substantial missingness were excluded before fitting; no univariate p-value filtering
Missing-data handling	IterativeImputer(random_state = 41) fitted on the training data
Train/test split	Single 80/20 random split with random_state = 42

Oversampling	SMOTE (random_state = 42) applied in the training pipeline to the sludge/microlithiasis subgroup labels rather than to the CDL outcome

Supplementary Table 2: Sample size and training set characteristics for model development

Metric	Value
Total cohort size	348.00
Hold-out test set size	70.00
Training set size	278.00
CDL events in training set	128.00
CDL non-events in training set	150.00
Predictors	14.00
Events per variable	9.14
Training-set minority subgroup before SMOTE	126.00
Training-set majority subgroup before SMOTE	152.00
Synthetic cases added by SMOTE	26.00
Training-set size after SMOTE	304.00

Supplementary Table 3: Internal validation

Validation strategy	AUC	Sensitivity	Specificity	NPV
Single 80/20 split (as published)	0.739	92.3%	35.5%	78.6%
10 × 10-fold cross-validation	0.655	92.5%	19.4%	73.8%
Nested cross-validation (10 outer / 3 inner)	0.648	92.8%	22.7%	77.4%
Out-of-bootstrap (100 resamples)	0.632	89.5%	20.7%	69.2%
Temporal (dev 2005–2016 → val 2017–2021)	0.644	75.3%	31.5%	51.1%

Supplementary Table 4: Performance and calibration metrics of the final RF model

Metric	Value
AUC	0.739
AUC 95% CI	0.624 to 0.854

Sensitivity	0.923
Sensitivity 95% CI	0.797 to 0.973
Specificity	0.355
Specificity 95% CI	0.211 to 0.531
PPV	0.643
PPV 95% CI	0.512 to 0.755
NPV	0.786
NPV 95% CI	0.524 to 0.924
Accuracy	0.671
Accuracy 95% CI	0.555 to 0.77
Brier score	0.229
Calibration intercept	0.628
Calibration slope	1.083

Supplementary Table 5: Threshold table

Threshold	Sensitivity	Specificity	PPV	NPV	Accuracy
0.10	0.974	0.000	0.551	0.000	0.543
0.15	0.974	0.032	0.559	0.500	0.557
0.20	0.949	0.194	0.597	0.750	0.614
0.25	0.923	0.355	0.643	0.786	0.671
0.30	0.872	0.419	0.654	0.722	0.671
0.35	0.744	0.645	0.725	0.667	0.700
0.40	0.692	0.677	0.730	0.636	0.686
0.50	0.513	0.871	0.833	0.587	0.671

Supplementary Table 6: Highest pairwise correlations among candidate predictor variables

Variable 1	Variable 2	Spearman rho	Absolute rho
lab_values_admission_gpt_alt_in_u_l	lab_values_admission_got_ast_in_u_l	0.811	0.811
lab_values_admission_alkaline_phosphatase_in_u_l	lab_values_admission_gamma.gt_in_u_l	0.645	0.645
lab_values_admission_ga	lab_values_admission_gp	0.613	0.613

mma.gt_in_u_l	t_alt_in_u_l		
lab_values_admission_alkaline_phosphatase_in_u_l	lab_values_admission_bilirubin_in_mg_dl	0.550	0.550
lab_values_admission_alkaline_phosphatase_in_u_l	lab_values_admission_gpt_alt_in_u_l	0.537	0.537
lab_values_admission_gamma.gt_in_u_l	lab_values_admission_bilirubin_in_mg_dl	0.537	0.537
lab_values_admission_gamma.gt_in_u_l	lab_values_admission_got_ast_in_u_l	0.535	0.535
lab_values_admission_gpt_alt_in_u_l	lab_values_admission_bilirubin_in_mg_dl	0.497	0.497
baseline_characteristics_weight_in_kg	baseline_characteristics_height_in_cm	0.432	0.432
lab_values_admission_creatinine_in_mg_dl	baseline_characteristics_age	0.421	0.421

Supplementary Table 7: Prevalence-dependent predictive values. PPV and NPV are computed under three hypothetical prevalence scenarios using the Bayes-theorem formulas: PPV(p) = Sens × p / (Sens × p + (1 − Spec) × (1 − p)); NPV(p) = Spec × (1 − p) / ((1 − Sens) × p + Spec × (1 − p)). Sensitivity and Specificity are fixed at the values estimated from the test set.

Model	Scenario	Prevalence	Sensitivity	Specificity	PPV	NPV
RF	Low (10%)	0.10	0.923	0.355	0.137	0.976
RF	Observed	0.48	0.923	0.355	0.569	0.833
RF	High (50%)	0.50	0.923	0.355	0.589	0.822
CBD & Bilirubin	Low (10%)	0.10	0.256	0.903	0.227	0.916
CBD & Bilirubin	Observed	0.48	0.256	0.903	0.709	0.568
CBD & Bilirubin	High (50%)	0.50	0.256	0.903	0.725	0.548
CBD	Low	0.10	0.795	0.548	0.163	0.960

	(10%)					
CBD	Observed	0.48	0.795	0.548	0.619	0.743
CBD	High	0.50	0.795	0.548	0.638	0.728
	(50%)					
Bilirubin	Low	0.10	0.308	0.774	0.132	0.910
	(10%)					
Bilirubin	Observed	0.48	0.308	0.774	0.557	0.548
Bilirubin	High	0.50	0.308	0.774	0.577	0.528
	(50%)					