INTERNATIONAL JOURNAL OF
MEDICAL BIOCHEMISTRY

DOI: 10.14744/ijmb.2026.22599
Int J Med Biochem 2026;9(4):242–250

Research Article

Internal validation and explainability of machine learning models for abnormal liver biochemistry

Osman Demir[1], Muzaffer Katar[2], Omer Faruk Nasip[3]

[1]Department of Biostatistics, Tokat Gaziosmanpasa University Faculty of Medicine, Tokat, Türkiye
[2]Department of Biochemistry, Tokat Gaziosmanpasa University Faculty of Medicine, Tokat, Türkiye
[3]Department of Medical Education and Informatics, Tokat Gaziosmanpasa University Faculty of Medicine, Tokat, Türkiye

Abstract

Objectives: To develop, internally validate, and interpret machine learning models for classifying abnormal liver biochemistry using routine biochemical parameters.

Methods: This retrospective study included 335 individuals classified according to ICD-coded diagnoses retrieved from the hospital information system, comprising 175 individuals in the non-disorder group and 160 individuals in the abnormal liver biochemistry group. Aspartate aminotransferase, alanine aminotransferase, alkaline phosphatase, gamma-glutamyl transferase, total bilirubin, direct bilirubin, international normalized ratio, and albumin levels were used as predictors. Nine supervised learning models were evaluated using 10×5 repeated stratified cross-validation. The model with the numerically highest mean ROC-AUC was further examined using out-of-fold calibration analysis, exploratory decision curve analysis, and SHAP-based explanations.

Results: Aspartate aminotransferase showed the highest mean individual ROC-AUC (0.832), followed by alanine aminotransferase (0.796). XGBoost showed the numerically highest mean ROC-AUC (0.9640, 95% CI 0.9615–0.9664), with a mean accuracy of 0.9215 and F1-score of 0.9157. Mean sensitivity, specificity, positive predictive value, and negative predictive value were 0.9019, 0.9394, 0.9333, and 0.9161, respectively. Calibration analysis based on averaged repeated out-of-fold predictions yielded a Brier score of 0.056. SHAP analysis identified international normalized ratio, aspartate aminotransferase, and albumin as the most influential contributors to model predictions for the ICD-defined abnormal liver biochemistry outcome.

Conclusion: XGBoost showed high internally validated classification performance for the ICD-defined abnormal liver biochemistry outcome. These findings should be interpreted as internally validated classification performance rather than definitive clinical diagnostic accuracy. Although the outcome was based on ICD-coded diagnoses rather than biochemical threshold rules, external validation using independently adjudicated clinical outcomes remains necessary.

Keywords: Artificial intelligence, biomarkers, cross-validation studies, liver function tests, machine learning, ROC curve

How to cite this article: Demir O, Katar M, Nasip OF. Internal validation and explainability of machine learning models for abnormal liver biochemistry. Int J Med Biochem 2026;9(4):242–250.

The liver plays a central role in several essential physiological processes, including metabolism, detoxification, bile production, protein synthesis, and coagulation regulation [1, 2]. Abnormalities in liver-related biochemical parameters may indicate hepatocellular injury, cholestasis, impaired synthetic function, or mixed hepatic dysfunction [1]. As many liver disorders may remain asymptomatic or clinically nonspecific in their early stages, routine biochemical tests are widely used as first-line indicators for detecting liver-related abnormalities and guiding further clinical evaluations [1].

Routine biochemical biomarkers, such as aspartate aminotransferase, alanine aminotransferase, alkaline phosphatase, gamma-glutamyl transferase, bilirubin, international normalized ratio, and albumin, provide complementary information

Address for correspondence: Osman Demir, PhD. Department of Biostatistics, Tokat Gaziosmanpasa University Faculty of Medicine, Tokat, Türkiye
Phone: +90 506 381 72 74 **E-mail:** osmand.demir@gop.edu.tr **ORCID:** 0000-0002-1322-2716

Submitted: July 13, 2026 **Revised:** August 14, 2026 **Accepted:** August 18, 2026 **Available Online:** October 15, 2026

OPEN ACCESS This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

on different aspects of liver injury and function [1, 2]. Transaminases are commonly associated with hepatocellular injury, whereas alkaline phosphatase, gamma-glutamyl transferase, and bilirubin levels may reflect cholestatic or excretory abnormalities [1, 2]. Albumin and international normalized ratio are clinically relevant indicators of hepatic synthetic function and disease severity [2]. However, the interpretation of these biomarkers in combination may be challenging, as abnormal liver biochemistry may reflect complex and potentially nonlinear relationships among multiple biochemical parameters rather than an isolated abnormality in a single marker.

In recent years, machine learning methods have been increasingly applied to liver-related classification and computer-aided modeling tasks. Previous studies using clinical and biochemical datasets have compared algorithms such as logistic regression, k-nearest neighbors, decision tree, support vector machine, naïve Bayes, random forest, Extra Trees, and boosting-based models for classifying individuals with or without liver-related abnormalities [3, 4]. These studies commonly used biochemical variables such as bilirubin, alkaline phosphatase, aminotransferases, albumin, and albumin-to-globulin ratio, and reported model performance using accuracy, precision, recall, F1-score, specificity, and ROC-AUC [3, 4]. Systematic evidence also indicates that preprocessing, feature selection, outcome definition, and validation strategies are critical components of ML-based liver research [5].

Despite these advances, several methodological limitations persist. High apparent performance may occur when predictor variables overlap with the criteria used to define the outcome, creating a risk of incorporation bias or circular classification. Accuracy alone is insufficient for evaluating classification models, particularly when false-negative and false-positive classifications have different implications [6, 7]. A comprehensive evaluation should include discrimination, calibration, classification metrics, and internal validation [7–10]. Nevertheless, internal validation cannot establish transportability, and external validation is essential before broader clinical interpretation [8].

Another important issue is the interpretability of the results. Although complex machine learning algorithms may achieve high classification performance, their practical value is limited if the contribution of individual biomarkers to model predictions cannot be understood [11, 12]. Explainable artificial intelligence methods, particularly SHAP analysis, provide a framework for quantifying the relative contribution and direction of individual predictors [13].

Therefore, this study aimed to develop, internally validate, and interpret machine learning models for the classification of abnormal liver biochemistry using routine biochemical parameters. The study first evaluated the discriminative performance of individual biomarkers, then compared multiple supervised learning algorithms, and finally examined the calibration, exploratory decision-analytic net benefit, and SHAP-based explainability of the model with the numerically highest mean ROC-AUC. The analysis was intentionally positioned as an internally validated classification framework for an ICD-defined abnormal liver biochemistry outcome, rather than an independently validated clinical diagnostic tool. We hypothesized that multivariable ensemble-based models would show higher internal discrimination than individual biomarkers and conventional single-model classifiers.

Materials and Methods

Study design and dataset

This retrospective classification study analyzed an existing dataset of 335 individuals, comprising 175 individuals in the non-disorder group and 160 in the ICD-defined abnormal liver biochemistry group. The analysis was conducted using anonymized laboratory data of the patients. No independent clinical diagnosis, imaging confirmation, disease stage, or etiological classification was available in the dataset.

The predictor variables were aspartate aminotransferase (AST; EC 2.6.1.1), alanine aminotransferase (ALT; EC 2.6.1.2), alkaline phosphatase (ALP; EC 3.1.3.1), gamma-glutamyl transferase (GGT; EC 2.3.2.2), total bilirubin, direct bilirubin, international normalized ratio (INR), and albumin levels. Enzyme activities were expressed as U/L, bilirubin concentrations as mg/dL, albumin as g/dL, and INR as a dimensionless ratio.

Group assignment was based on ICD-coded diagnoses retrieved from the hospital information system. Individuals were classified into the abnormal liver biochemistry group or the non-disorder group according to the diagnostic coding recorded in the hospital database. The outcome label was not generated by applying threshold-based rules to AST, ALT, ALP, GGT, total bilirubin, direct bilirubin, INR, or albumin values. Therefore, the binary outcome was not a deterministic function of the biochemical predictors evaluated in the present study.

Ethical considerations

The study was approved by the Tokat Gaziosmanpaşa University Non-Interventional Clinical Research Ethics Committee. (Date: 20.02.2026, Decision no: 26-MOBAEK-063). The study was conducted in accordance with the principles of the Declaration of Helsinki. The requirement for informed consent was waived because of the retrospective nature of the study. No other interventions were performed.

Definition of the abnormal liver biochemistry outcome

The binary outcome was based on ICD-coded diagnoses retrieved from the hospital information system and indicated membership in either the abnormal liver biochemistry group or the non-disorder group. Group assignment was therefore based on diagnostic coding rather than on threshold-based rules derived from the biochemical predictor variables. No deterministic combination of AST, ALT, ALP, GGT, total bilirubin, direct bilirubin, INR, or albumin was used to define the outcome. However, because laboratory findings may have contributed to the original clinical diagnostic process underlying ICD coding, residual incorporation bias cannot be completely excluded.

Data preprocessing

Before statistical analysis and model development, the dataset was examined for quality and consistency. Missing values, duplicate observations, outliers, variable formats, and class distributions were also assessed.

Missing values were identified only for INR (93/335, 27.8%), whereas the remaining biochemical predictors had complete data. Missing INR values were handled within the cross-validation procedure using median imputation. For each fold, the median was estimated exclusively from the training data and subsequently applied to the corresponding validation data to prevent information leakage. Continuous biochemical parameters, including AST, ALT, ALP, GGT, total bilirubin, direct bilirubin, INR, and albumin, were retained as numerical predictor variables. The outcome variable was encoded as a binary target variable.

The outcome classes were relatively balanced, with 175 and 160 individuals in the non-disorder and abnormal liver biochemistry groups, respectively. Therefore, no oversampling or synthetic data generation methods, such as SMOTE, were applied. To reduce the influence of scale differences among biochemical parameters, continuous predictors were standardized when required, particularly for scale-sensitive algorithms such as Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Multilayer Perceptron. Scaling parameters were estimated exclusively from the training data within each cross-validation fold and subsequently applied to the corresponding validation data to prevent information leakage.

Descriptive and comparative statistical analysis

Descriptive statistics were calculated for all biochemical parameters according to the study group. The distributional characteristics of continuous variables were assessed before group comparisons. Because the biochemical variables showed non-normal distributions, they were summarized as median and interquartile range (IQR). The non-disorder and abnormal liver biochemistry groups were compared in terms of AST, ALT, ALP, GGT, total bilirubin, direct bilirubin, INR, and albumin using the Mann–Whitney U test. A two-sided p-value of <0.05 was considered statistically significant.

Discriminative performance of individual biomarkers

The individual discriminative performance of each biochemical biomarker was evaluated using receiver operating characteristic (ROC) curve analysis. ROC analyses were performed separately for AST, ALT, ALP, GGT, total bilirubin, direct bilirubin, INR, and albumin to determine their ability to distinguish between the ICD-defined abnormal liver biochemistry and non-disorder groups.

Biomarker-specific performance was evaluated using the same 10×5 repeated stratified cross-validation framework. Within each fold, the optimal cutoff was determined from the training data using the Youden index (sensitivity + specificity – 1) and subsequently applied to the corresponding validation data. ROC-AUC, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and the Youden index were calculated for the validation data and summarized across the repeated cross-validation procedure. For each biomarker, ROC-AUC was reported as the mean with a 95% confidence interval calculated from the 10 repetition-level mean ROC-AUC estimates using the t-distribution. For the individual INR analysis, only observations with available INR measurements were included, and no imputation was performed.

Conventional statistical modeling

Multivariable logistic regression analysis was performed using a conventional statistical reference model. The ICD-defined abnormal liver biochemistry outcome was used as the dependent variable. Because of substantial missingness in INR, the primary multivariable model included AST, ALT, ALP, GGT, total bilirubin, direct bilirubin, and albumin as independent variables. AST, ALT, ALP, GGT, total bilirubin, and direct bilirubin were transformed using log(1+x) because of skewed distributions and subsequently standardized, whereas albumin was standardized without transformation. Odds ratios (ORs) were therefore reported per one-standard-deviation increase, together with 95% confidence intervals and p-values. Multicollinearity was assessed using variance inflation factors (VIFs). Sensitivity analyses were performed by excluding either total or direct bilirubin and by fitting a complete-case model including INR.

Model discrimination was assessed using ROC-AUC, overall probabilistic prediction error using the Brier score, and goodness of fit using the Hosmer–Lemeshow test. Nagelkerke R^2 was also calculated as a measure of model explanatory performance.

Machine learning model development and internal validation

Several supervised machine learning algorithms were developed to classify individuals according to the ICD-defined abnormal liver biochemistry outcome. The evaluated models included logistic regression, K-nearest neighbors, support vector machines, decision trees, random forests, gradient boosting, XGBoost, Naïve Bayes, and multilayer perceptron. No data-driven hyperparameter optimization was performed. Models were evaluated using fixed parameter settings within a uniform cross-validation framework to facilitate consistent model comparison and to limit optimization-related overfitting in the relatively small dataset. Detailed model-specific parameter settings are provided in Appendix 1.

Model performance was internally validated using 10×5 repeated stratified cross-validation (10 repetitions of five-fold cross-validation). Stratification was applied to preserve the distribution of the outcome classes in each fold. Performance metrics were averaged across the repeated cross-validation procedure. ROC-AUC was reported as the mean with standard deviation (SD) and 95% confidence interval (CI). For ROC-AUC, 95% confidence intervals were calculated from the 10 repe-

tition-level mean estimates using the t-distribution. Pairwise comparisons among the three models with the highest mean ROC-AUC values were performed using paired t-tests on the 10 matched repetition-level mean ROC-AUC estimates obtained from the repeated cross-validation procedure. Differences in mean ROC-AUC were reported with 95% confidence intervals and two-sided p-values.

Performance metrics

The primary performance metric was the ROC-AUC, which was used to assess the discriminatory ability of each model.

The secondary performance metrics included accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1-score. The abnormal liver biochemistry group was defined as the positive class for the calculation of these classification metrics. Confusion matrix-derived indices were calculated to provide an interpretable assessment of true-positive, true-negative, false-positive, and false-negative classifications.

Accuracy was defined as the proportion of correctly classified individuals among all individuals. Sensitivity represented the proportion of individuals in the abnormal liver biochemistry group who were correctly identified by the model, whereas specificity represented the proportion of individuals in the non-disorder group who were correctly classified. Positive predictive value (PPV) represented the proportion of predicted positive cases that were truly positive, whereas negative predictive value (NPV) represented the proportion of predicted negative cases that were truly negative. The F1-score was calculated as the harmonic mean of PPV and sensitivity.

Calibration, exploratory decision analysis, and explainability

For the model with the numerically highest mean ROC-AUC, additional analyses were performed to evaluate calibration, exploratory decision-analytic performance, and interpretability. Calibration was assessed using out-of-fold predicted probabilities generated during 10×5 repeated stratified cross-validation. For each individual, the out-of-fold predicted probabilities obtained across the 10 repetitions were averaged before calculation of the calibration curve and Brier score. The calibration curve was used for visual assessment, and the Brier score was calculated as a quantitative measure of overall probabilistic prediction error. Lower Brier scores indicate better overall probabilistic prediction accuracy.

Decision curve analysis was performed using the same averaged repeated out-of-fold predicted probabilities as an exploratory assessment of net benefit across a range of threshold probabilities. The model was compared with the default classify-all and classify-none strategies. Because the outcome was ICD-defined and no independent clinical endpoint was available, the decision curve analysis was considered exploratory and was not interpreted as evidence of established clinical utility.

To improve interpretability, SHAP analysis was performed for the model with the numerically highest mean ROC-AUC. SHAP values were calculated for validation observations within each cross-validation fold. For each individual, SHAP values obtained across the 10 repetitions were averaged to derive repeated out-of-fold feature contributions. A SHAP summary plot was generated to identify the most influential predictors and illustrate the direction and magnitude of individual biomarker contributions to the classification of the ICD-defined abnormal liver biochemistry outcome.

Statistical analysis

All analyses were performed using Python version 3.11. Data management was conducted using Pandas and NumPy, statistical analyses using SciPy and Statsmodels, machine learning using Scikit-learn and XGBoost, and explainability analysis using SHAP. Graphical outputs were generated using Matplotlib. Two-sided p-values <0.05 were considered statistically significant. Classification performance metrics were summarized as mean values across the repeated cross-validation procedure, while ROC-AUC was additionally reported with the standard deviation and 95% confidence interval.

The study was reported in accordance with the TRIPOD+AI reporting guidance [14], and the completed TRIPOD+AI checklist is provided as Appendix 2.

Results

Study population and dataset characteristics

The final dataset comprised 335 individuals, including 175 in the non-disorder group and 160 in the abnormal liver biochemistry group. These groups accounted for 52.2% and 47.8% of the dataset, respectively. The predictor variables included AST, ALT, ALP, GGT, total bilirubin, direct bilirubin, INR, and albumin.

The biochemical characteristics of the study groups are shown in Table 1.

All evaluated biochemical parameters differed significantly between the groups. The abnormal liver biochemistry group had higher median AST, ALT, ALP, GGT, total bilirubin, direct bilirubin, INR, and albumin values than the non-disorder group. INR comparisons were based on available cases without imputation.

Discriminative performance of individual biochemical biomarkers

The discriminative performance of individual biochemical biomarkers in distinguishing the abnormal liver biochemistry and non-disorder groups is presented in Table 2. Among the evaluated biomarkers, AST showed the highest mean ROC-AUC (0.832), followed by ALT (0.796). Across repeated cross-validation, the mean training-derived cutoff for AST was ≥23.32, with a mean sensitivity of 77.4% and specificity of 78.4%. For ALT, the corresponding mean cutoff was ≥27.16, with a mean sensitivity of 73.4% and specificity of 72.9%.

Table 1. Biochemical characteristics of the study groups

Variable	Non-disorder group (n=175)	Abnormal liver biochemistry (n=160)	p
AST	18.60 (15.00–22.45)	41.00 (24.00–66.78)	<0.001
ALT	16.50 (10.50–27.05)	51.80 (25.95–92.10)	<0.001
ALP	79.00 (64.00–106.00)	108.00 (81.00–176.00)	<0.001
GGT	21.00 (13.00–41.00)	42.00 (21.75–128.50)	<0.001
Total bilirubin	0.46 (0.30–0.69)	0.54 (0.36–0.81)	0.006
Direct bilirubin	0.16 (0.11–0.27)	0.21 (0.15–0.39)	<0.001
INR	1.04 (0.97–1.15)	1.08 (1.02–1.22)	0.009
Albumin	4.14 (3.24–4.58)	4.50 (4.32–4.70)	<0.001

Values are presented as median (interquartile range). Between-group comparisons were performed using the Mann-Whitney U test. Bilirubin conversion to SI units: mg/dL ×17.1=μmol/L; albumin conversion: g/dL ×10=g/L. INR descriptive statistics and between-group comparison were based on available cases without imputation. ALP: Alkaline phosphatase; ALT: Alanine aminotransferase; AST: Aspartate aminotransferase; GGT: Gamma-glutamyl transferase; INR: International normalized ratio.

Table 2. Discriminative performance of individual biochemical biomarkers

Biomarker	ROC-AUC	95% CI	Mean cut-off	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Youden index
AST	0.832	0.830–0.835	≥23.32	77.4	78.4	77.0	79.4	0.558
ALT	0.796	0.794–0.798	≥27.16	73.4	72.9	71.6	75.4	0.464
ALP	0.685	0.683–0.687	≥85.76	70.4	55.8	59.5	68.7	0.262
GGT	0.677	0.675–0.680	≥34.16	63.1	64.0	61.9	66.0	0.271
Total bilirubin	0.587	0.584–0.590	≥0.55	53.4	62.8	56.5	59.8	0.162
Direct bilirubin	0.638	0.635–0.641	≥0.19	65.4	58.8	59.1	65.5	0.242
INR	0.610	0.605–0.615	≥1.02	73.1	43.1	32.9	81.6	0.162
Albumin	0.658	0.656–0.660	≥4.45	69.4	64.9	64.6	70.1	0.343

Values represent mean performance across 10×5 repeated stratified cross-validation. Biomarker-specific cutoff values were determined in each training fold using the Youden index and evaluated in the corresponding validation fold. The positive outcome was the abnormal liver biochemistry group. INR analysis was based on available cases without imputation. ALP: Alkaline phosphatase; ALT: Alanine aminotransferase; AST: Aspartate aminotransferase; CI: Confidence interval; GGT: Gamma-glutamyl transferase; INR: International normalized ratio; NPV: Negative predictive value; PPV: Positive predictive value; ROC-AUC: Receiver operating characteristic area under the curve.

Compared with AST and ALT, the remaining biomarkers showed lower individual discriminative performance, with mean ROC-AUC values ranging from 0.587 to 0.685. Overall, individual biomarkers showed lower discrimination than the best-performing multivariable machine learning models.

Multivariable logistic regression analysis

Multivariable logistic regression analysis was performed to evaluate the associations between routine biochemical biomarkers and the ICD-defined abnormal liver biochemistry outcome. The results are presented in Table 3. AST, ALP, and albumin levels were significantly associated with group membership. In contrast, ALT, GGT, total bilirubin, and direct bilirubin were not statistically significant in the multivariable model.

Although VIF analysis indicated notable multicollinearity between total and direct bilirubin, sensitivity analyses excluding either bilirubin variable did not materially alter the associations of AST, ALP, and albumin with the outcome.

In a complete-case sensitivity analysis including INR (n=242), INR was not independently associated with the outcome (OR per 1-SD increase=1.046, 95% CI 0.699–1.565; p=0.828), while the principal associations for AST, ALP, and albumin were retained.

Table 3. Multivariable logistic regression analysis for the abnormal liver biochemistry outcome

Variable	OR per 1-SD increase	95% CI	p	VIF
AST	6.913	3.690–12.949	<0.001	2.42
ALT	1.519	0.919–2.513	0.103	2.94
ALP	1.796	1.166–2.767	0.008	1.67
GGT	0.999	0.630–1.585	0.998	2.03
Total bilirubin	1.027	0.492–2.143	0.944	7.63
Direct bilirubin	1.094	0.465–2.573	0.837	8.73
Albumin	2.600	1.765–3.829	<0.001	1.32

AST, ALT, ALP, GGT, total bilirubin, and direct bilirubin were transformed using log(1+x) because of skewed distributions and subsequently standardized. Albumin was standardized without transformation. Odds ratios therefore represent the change in odds associated with a one-standard-deviation increase in each predictor. INR was not included in the primary multivariable model because of substantial missingness. ALP: Alkaline phosphatase; ALT: Alanine aminotransferase; AST: Aspartate aminotransferase; CI: Confidence interval; GGT: Gamma-glutamyl transferase; OR: Odds ratio; VIF: Variance inflation factor.

The conventional logistic regression model showed an apparent ROC-AUC of 0.8854 and a Brier score of 0.1321 when fitted to the full analytical dataset. The Hosmer–Lemeshow test did not indicate significant lack of fit (χ^2=11.509, df=8; p=0.175), and the Nagelkerke R^2 was 0.550.

Table 4. Cross-validated performance of machine learning models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1-score	ROC-AUC, mean±SD (95% CI)
Logistic regression	0.7684	0.6937	0.8366	0.8000	0.7514	0.7401	0.8445±0.0066 (0.8397–0.8492)
KNN	0.7919	0.7669	0.8149	0.7934	0.7953	0.7778	0.8573±0.0118 (0.8488–0.8657)
SVM	0.7719	0.6812	0.8549	0.8128	0.7490	0.7383	0.8602±0.0057 (0.8561–0.8643)
Decision tree	0.8382	0.8400	0.8366	0.8297	0.8546	0.8322	0.8383±0.0149 (0.8276–0.8489)
Random forest	0.8979	0.8456	0.9457	0.9354	0.8727	0.8868	0.9539±0.0061 (0.9495–0.9583)
Gradient boosting	0.9185	0.8956	0.9394	0.9333	0.9102	0.9126	0.9615±0.0032 (0.9592–0.9639)
XGBoost	0.9215	0.9019	0.9394	0.9333	0.9161	0.9157	0.9640±0.0034 (0.9615–0.9664)
Naive bayes	0.6743	0.3956	0.9291	0.8407	0.6281	0.5342	0.8303±0.0040 (0.8274–0.8331)
Multilayer perceptron	0.8170	0.8006	0.8320	0.8176	0.8243	0.8061	0.8928±0.0076 (0.8873–0.8982)

KNN: K-nearest neighbors; MLP: Multilayer perceptron; NPV: Negative predictive value; PPV: Positive predictive value; ROC-AUC: Receiver operating characteristic area under the curve; SD: Standard deviation; SVM: Support vector machine; XGBoost: Extreme gradient boosting. Performance estimates were obtained using 10×5 repeated stratified cross-validation. Accuracy, sensitivity, specificity, PPV, NPV, and F1-score are presented as mean values, whereas ROC-AUC is presented as mean±SD (95% CI). Sensitivity, specificity, PPV, NPV, and F1-score were calculated with the abnormal liver biochemistry group as the positive class. Missing INR values were imputed within each training fold using the median estimated from that training fold and subsequently applied to the corresponding validation fold.

Cross-validated machine learning model performance

The cross-validated performance of the evaluated machine learning models is presented in Table 4. Among the evaluated models, XGBoost showed the numerically highest mean ROC-AUC (0.9640, 95% CI 0.9615–0.9664), with a mean accuracy of 0.9215 and F1-score of 0.9157. Gradient Boosting and Random Forest also showed high mean ROC-AUC values of 0.9615 and 0.9539, respectively. These estimates reflect internal classification performance for the abnormal liver biochemistry outcome.

Pairwise comparisons showed no statistically significant difference in mean ROC-AUC between XGBoost and Gradient Boosting (ΔROC-AUC=0.0024, 95% CI −0.0002 to 0.0051; p=0.069). XGBoost showed a significantly higher mean ROC-AUC than Random Forest (ΔROC-AUC=0.0101, 95% CI 0.0066–0.0135; p<0.001), while Gradient Boosting also showed a significantly higher mean ROC-AUC than Random Forest (ΔROC-AUC=0.0077, 95% CI 0.0036–0.0117; p=0.002).

Tree-based ensemble and boosting algorithms generally showed higher observed classification performance than conventional and single-model classifiers. However, these findings represent internal discrimination of the ICD-defined abnormal liver biochemistry outcome and do not establish independent clinical diagnostic validity.

ROC-AUC performance of the machine learning models

Figure 1 shows the mean receiver operating characteristic (ROC) curves of the evaluated machine learning models across 10×5 repeated stratified cross-validation. Consistent with the repeated cross-validation results presented in Table 4, XGBoost showed the numerically highest mean ROC-AUC (0.9640), followed by Gradient Boosting (0.9615) and Random Forest (0.9539). Among the remaining models, Multilayer Perceptron showed the highest mean ROC-AUC (0.8928), followed by SVM (0.8602) and KNN (0.8573).

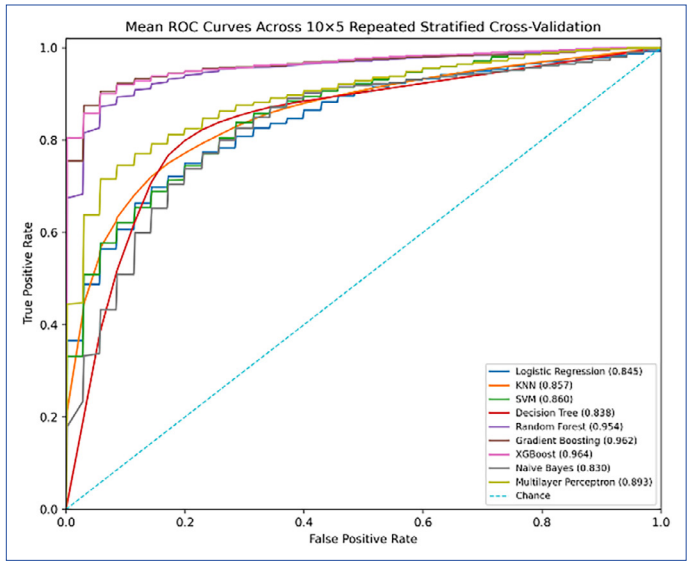


Figure 1. Mean receiver operating characteristic curves of the evaluated machine learning models across 10×5 repeated stratified cross-validation. XGBoost showed the numerically highest mean ROC-AUC, followed by Gradient Boosting and Random Forest.

GBM: Gradient boosting machine; KNN: K-nearest neighbors; LR: Logistic regression; MLP: Multilayer perceptron; NB: Naïve Bayes; RF, random forest; ROC-AUC: Receiver operating characteristic area under the curve; SVM: Support vector machine; XGBoost: Extreme gradient boosting.

Overall, the ensemble-based tree models showed higher observed internal discrimination than the conventional and single-classifier models. Therefore, XGBoost was selected for additional calibration, exploratory decision curve, and explainability analyses.

Cross-validated classification performance of the selected XGBoost model

Based on its numerically highest mean ROC-AUC among the evaluated models, XGBoost was selected for additional calibration, exploratory decision curve, and explainability analyses. Across 10×5 repeated stratified cross-validation, XGBoost

achieved a mean accuracy of 0.9215, sensitivity of 0.9019, specificity of 0.9394, positive predictive value of 0.9333, negative predictive value of 0.9161, F1-score of 0.9157, and ROC-AUC of 0.9640 (95% CI 0.9615–0.9664).

For descriptive visualization of classification errors, an aggregate out-of-fold confusion matrix was constructed by averaging each individual's predicted probabilities across the 10 cross-validation repetitions and applying a classification threshold of 0.50. The resulting matrix included 166 true-negative, 9 false-positive, 13 false-negative, and 147 true-positive classifications (Appendix 3). Because this aggregate matrix was based on patient-level averaged out-of-fold predictions, its derived classification metrics are not directly equivalent to the mean fold-level performance estimates reported in Table 4.

These estimates represent internally validated classification performance for the ICD-defined abnormal liver biochemistry outcome and should not be interpreted as independent clinical diagnostic accuracy.

Calibration, exploratory decision analysis, and explainability

Calibration analysis based on averaged repeated out-of-fold predictions yielded a Brier score of 0.056, indicating relatively low overall probabilistic prediction error. Exploratory decision curve analysis based on the same averaged repeated out-of-fold predictions showed that the model provided a higher net benefit than the classify-all and classify-none strategies across a broad range of threshold probabilities. Because the outcome was ICD-defined and no independent clinical endpoint was available, these findings should be interpreted as exploratory decision-analytic results rather than evidence of established clinical utility (Appendix 4, 5).

SHAP analysis was used to evaluate the contribution and direction of biochemical variables in the XGBoost model. The SHAP summary plot is shown in Figure 2. INR had the highest mean absolute SHAP value, followed closely by AST and then albumin, indicating that these variables contributed most strongly to model predictions. These findings reflect the relative contribution of the variables to the classification of the ICD-defined abnormal liver biochemistry outcome but do not establish causal or pathophysiological importance.

Discussion

This study developed and internally validated machine learning models for classifying abnormal liver biochemistry using routine biochemical parameters. Internal validation was performed using 10×5 repeated stratified cross-validation. Individual biomarkers showed modest discriminatory performance, whereas multivariable machine learning models showed higher internal classification performance. Among the evaluated models, XGBoost showed the numerically highest mean ROC-AUC and was selected for calibration, exploratory decision analysis, and SHAP-based interpretation. These findings reflect the model's ability to classify an ICD-defined

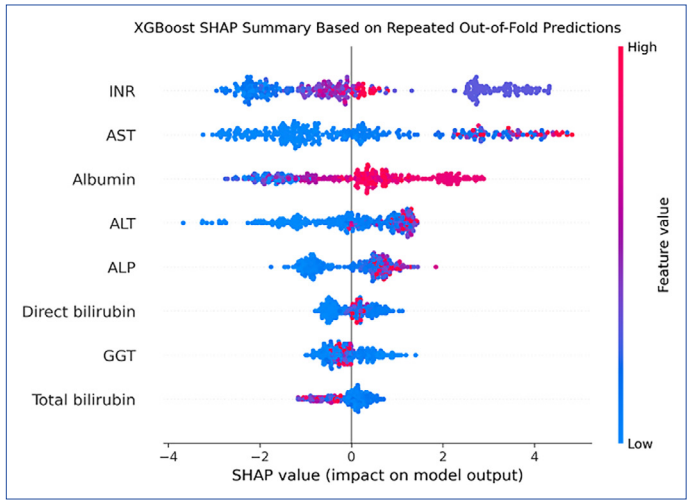


Figure 2. Repeated out-of-fold SHAP summary plot of the XGBoost model. SHAP values were calculated for validation observations within each cross-validation fold and averaged across the 10 repetitions for each individual. Each point represents an individual observation. Features are ranked according to their mean absolute SHAP values. Higher feature values are shown in red and lower values in blue. Positive SHAP values increase, whereas negative SHAP values decrease, the predicted probability of the ICD-defined abnormal liver biochemistry outcome.

ALP: Alkaline phosphatase; ALT: Alanine aminotransferase; AST: Aspartate aminotransferase; GGT: Gamma-glutamyl transferase; INR: International normalized ratio; SHAP: SHapley Additive exPlanations.

outcome using routine biochemical predictors and should not be interpreted as evidence of independently validated clinical diagnostic performance [14].

The between-group differences in AST, ALT, ALP, GGT, bilirubin, INR, and albumin levels are consistent with the established clinical roles of these parameters in characterizing liver-related laboratory abnormalities [1, 2]. AST and ALT showed the highest individual ROC-AUC values, indicating greater individual discrimination between the ICD-defined abnormal liver biochemistry and non-disorder groups than the other evaluated biomarkers. However, although the group label was derived from ICD-coded diagnoses rather than directly from biochemical thresholds, laboratory abnormalities may have contributed to the original diagnostic coding process; therefore, some degree of incorporation bias cannot be entirely excluded.

Although AST and ALT showed the strongest individual discrimination, their ROC-AUC values were lower than those of the multivariable machine learning models. This finding suggests that combining information across multiple biochemical parameters may improve discrimination compared with individual biomarkers alone. Similar ensemble-based approaches have demonstrated strong performance in liver-related datasets [15, 16]. Nevertheless, the apparent advantage of multivariable modeling in the present study may partly reflect associations between the biochemical predictors and the ICD-defined outcome, particularly because laboratory findings may have contributed to the original diagnostic coding process.

Among the evaluated algorithms, XGBoost showed the numerically highest mean ROC-AUC and balanced classification metrics. However, the difference in mean ROC-AUC between XGBoost and Gradient Boosting was not statistically significant, indicating that the observed numerical advantage of XGBoost over Gradient Boosting should not be interpreted as evidence of clear statistical superiority. Its performance may be related to its ability to model nonlinear relationships and interactions through regularized gradient boosting [17]. Nevertheless, increasing algorithmic complexity cannot overcome bias introduced by the definition of the outcome variable. Accordingly, the high performance observed in this study should be interpreted as internally validated classification performance for an ICD-defined outcome rather than as evidence that XGBoost can independently diagnose liver disease in other clinical populations.

Calibration and decision-analytic assessments provided complementary information beyond discrimination. Calibration based on averaged repeated out-of-fold predictions yielded a Brier score of 0.056, although the calibration curve deviated from the ideal line in the intermediate probability ranges. Exploratory decision curve analysis, based on the same out-of-fold predictions, showed a higher net benefit than the classify-all and classify-none strategies across a broad range of threshold probabilities. Because the endpoint was based on ICD-coded diagnoses and external validation was not performed, these results should not be interpreted as proof of clinical usefulness; rather, they describe potential decision-analytic behavior within the current dataset [7, 9, 10].

SHAP analysis improved the model transparency by showing how each biochemical parameter contributed to individual predictions. INR had the highest mean absolute SHAP value, followed closely by AST and then albumin, indicating that these variables had the largest overall contributions to the XGBoost predictions. These results help explain the internal logic of the model; however, SHAP importance should not be interpreted as causality or independent clinical importance, particularly because laboratory findings may have contributed to the clinical diagnostic process underlying the ICD-coded outcome [11, 13].

This study had several strengths. We compared individual biomarker discrimination with multivariable machine learning classification, used 10×5 repeated stratified cross-validation for internal validation, and evaluated discrimination, calibration, exploratory decision-analytic net benefit, and explainability. The study also provides a transparent account of the limitations imposed by an ICD-defined abnormal liver biochemistry outcome and positions the analysis as an exploratory, internally validated classification framework.

This study has some important limitations. An important limitation is the potential for residual incorporation bias. Although the outcome label was based on ICD-coded diagnoses retrieved from the hospital information system and was not generated directly from the biochemical predictors, laboratory findings may have contributed to the original clinical diagnostic process and subsequent ICD coding. Moreover, the analytical dataset did not contain independently adjudicated clinical outcomes or detailed information on the diagnostic process underlying each ICD-coded classification. Consequently, the observed model performance should be interpreted as internally validated classification performance for an ICD-defined outcome and not as validation against an independently adjudicated clinical endpoint.

The dataset was relatively small, retrospective, and derived from a single center, and external validation was not performed. The relatively small sample size in relation to the number and complexity of the evaluated algorithms may also increase the risk of model-selection optimism, despite the use of repeated cross-validation. Although 10×5 repeated stratified cross-validation was used to improve the robustness of internal performance estimates, this procedure does not substitute for validation in an independent external population. No data-driven hyperparameter optimization was performed; therefore, the reported results reflect the performance of the prespecified model configurations rather than optimized versions of the evaluated algorithms. Detailed clinical information on etiology, disease stage, imaging findings, comorbidities, medication use, and longitudinal outcomes was unavailable. Albumin-related findings should be interpreted cautiously because albumin levels were unexpectedly higher in the abnormal liver biochemistry group, which may reflect characteristics of the study population or outcome definition rather than a typical disease-related pattern. The non-disorder group should therefore not be interpreted as a prospectively recruited healthy reference population, and unmeasured differences in clinical setting or patient selection may have contributed to the observed albumin distribution. Finally, the calibration curve showed some deviation from ideal calibration, particularly in the intermediate probability range. Therefore, external validation using independently and clinically adjudicated outcomes is essential before considering clinical application.

Conclusion

In conclusion, routine biochemical parameters supported the internal classification of the ICD-defined abnormal liver biochemistry outcome. Among the evaluated models, XGBoost showed the numerically highest mean ROC-AUC and was selected for additional calibration, exploratory decision analysis, and explainability analyses. Repeated out-of-fold SHAP analysis identified INR, followed closely by AST and albumin, as the most influential contributors to XGBoost predictions.

This study should be interpreted as an exploratory, internally validated classification analysis of an ICD-defined outcome rather than as a clinical diagnostic model. The possibility of incorporation bias, absence of independent outcome adjudication, limited sample size, and lack of external validation restrict clinical interpretation. Future studies should evaluate the model in larger multicenter cohorts using independently defined clinical outcomes before considering any diagnostic or decision-support applications.

Disclosures

Ethics Committee Approval: The study was approved by the Tokat Gaziosmanpaşa University Non-Interventional Clinical Research Ethics Committee (no: 26-MOBAEK-063, date: 20/02/2026).

Informed Consent: The requirement for informed consent was waived by the ethics committee because of the retrospective study design and the use of existing anonymized data.

Conflict of Interest Statement: None declared.

Funding: The author declares that this study has received no financial support.

Use of AI for Writing Assistance: OpenAI ChatGPT was used for language editing, wording refinement, and formatting assistance. The authors critically reviewed and revised all AI-assisted content and take full responsibility for the accuracy, integrity, and final content of the manuscript.

Authorship Contributions: Concept – O.D., M.K.; Design – O.D., M.K., O.F.N.; Supervision – O.D.; Funding – M.K.; Materials – M.K.; Data collection and/or processing – M.K.; Data analysis and/or interpretation – O.D., M.K., O.F.N.; Literature search – O.D., O.F.N.; Writing – O.D.; Critical review – M.K., O.F.N.

Peer-review: Externally peer-reviewed.

References

1. Kwo PY, Cohen SM, Lim JK. ACG clinical guideline: evaluation of abnormal liver chemistries. Am J Gastroenterol 2017;112:18–35. [CrossRef]
2. Newsome PN, Cramb R, Davison SM, Dillon JF, Foulerton M, Godfrey EM, et al. Guidelines on the management of abnormal liver blood tests. Gut 2018;67:6–19. [CrossRef]
3. Mostafa F, Hasan E, Williamson M, Khan H. Statistical machine learning approaches to liver disease prediction. Livers 2021;1(4):294–312. [CrossRef]
4. Rabbi MF, Mahedy Hasan SM, Champa AI, Asifzaman M, Hasan MK. Prediction of liver disorders using machine learning algorithms: a comparative study. Proc Int Conf Artificial Intelligence and Computational Technologies 2020:111–6. [CrossRef]
5. Khan RA, Luo Y, Wu FX. Machine learning based liver disease diagnosis: a systematic review. Neurocomputing 2022;468:492–509. [CrossRef]
6. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27:861–74. [CrossRef]
7. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21(1):128–38. [CrossRef]
8. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594. [CrossRef]
9. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17(1):230. [CrossRef]
10. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565–74. [CrossRef]
11. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence: concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82–115. [CrossRef]
12. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. Proc ACM SIGKDD Int Conf Knowl Discov Data Min 2016:1135–44. [CrossRef]
13. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30:4765–74.
14. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378. [CrossRef]
15. Dalal S, Onyema EM, Malik A. Hybrid XGBoost model with hyperparameter tuning for prediction of liver disease with better accuracy. World J Gastroenterol 2022;28(46):6551–63. [CrossRef]
16. Lu CH, Wang W, Li YJ, Chang IW, Chen CL, Su CW, et al. Machine learning models for predicting significant liver fibrosis in patients with severe obesity and nonalcoholic fatty liver disease. Obes Surg 2024;34(12):4393–404. [CrossRef]
17. Chen T, Guestrin X. XGBoost: a scalable tree boosting system. Proc ACM SIGKDD Int Conf Knowl Discov Data Min 2016:785–94. [CrossRef]

Appendix 1. Model-specific parameter settings

Model	Preprocessing within each cross-validation fold	Model-specific settings
Logistic regression	Median imputation; StandardScaler	C=1.0; solver=lbfgs; max_iter=1000; class_weight=None; random_state=42; penalty was not explicitly set in the estimator call (scikit-learn default behavior)
K-nearest neighbors (KNN)	Median imputation; StandardScaler	n_neighbors=5; weights=uniform; algorithm=auto; metric=Minkowski; p=2
Support vector machine (SVM)	Median imputation; StandardScaler	C=1.0; kernel=RBF; gamma=scale; probability=True; class_weight=None; random_state=42
Decision tree	Median imputation; no scaling	criterion=Gini; splitter=best; max_depth=None; min_samples_split=2; min_samples_leaf=1; class_weight=None; random_state=42
Random forest	Median imputation; no scaling	n_estimators=100; criterion=Gini; max_depth=None; min_samples_split=2; min_samples_leaf=1; max_features=sqrt; bootstrap=True; class_weight=None; random_state=42
Gradient boosting	Median imputation; no scaling	n_estimators=100; learning_rate=0.1; max_depth=3; subsample=1.0; max_features=None; random_state=42
XGBoost	Median imputation; no scaling	booster=gbtree; n_estimators=100; learning_rate=0.3; max_depth=6; min_child_weight=1; subsample=1.0; colsample_bytree=1.0; gamma=0; reg_alpha=0; reg_lambda=1; objective=binary:logistic; eval_metric=logloss; random_state=42
Naïve bayes	Median imputation; no scaling	GaussianNB; var_smoothing=1×10^{-9}
Multilayer perceptron (MLP)	Median imputation; StandardScaler	hidden_layer_sizes=(100,); activation=ReLU; solver=Adam; alpha=0.0001; learning_rate=constant; learning_rate_init=0.001; max_iter=5000; early_stopping=-False; random_state=42

No data-driven hyperparameter tuning was performed. Missing INR values were imputed using the median estimated exclusively from each training fold and applied to the corresponding validation fold. Standardization, where indicated, was likewise fitted only on the training fold. Internal validation used RepeatedStratifiedKFold with 5 folds, 10 repeats, and random_state=42. The settings shown are the exact estimator arguments used in the verified rerun that reproduced the reported Table 4 mean performance values. Software versions in the verified rerun were scikit-learn 1.8.0 and XGBoost 3.1.3.

Appendix 2. TRIPOD+AI reporting checklist [14]

Item	Topic	How addressed in this study	Reported on page/section
1	Title	Development and internal validation, target context, and abnormal liver biochemistry outcome are identified.	Title
2	Abstract	Objectives, retrospective data source, predictors, internal validation, performance, calibration, and interpretation are summarized.	Abstract
3a	Background	Clinical context, rationale, and previous machine-learning studies in liver-related classification are described.	Introduction
3b	Target population/ intended purpose	The study is positioned as an internally validated laboratory classification framework rather than an independently validated clinical diagnostic tool.	Introduction
3c	Health inequalities	Sociodemographic variables required for assessment of health inequalities were not available in the dataset.	Not applicable
4	Objectives	Development, internal validation, model comparison, calibration, exploratory decision analysis, and explainability objectives are stated.	Introduction
5a	Data source	An existing retrospective anonymized laboratory dataset and ICD-coded diagnoses from the hospital information system were used.	Methods: Study design and dataset
5b	Data dates	The retrospective dataset was obtained from the institutional hospital information system; specific accrual dates were not available in the analytical dataset.	Methods: Study design and dataset
6a	Setting / centres	The study used retrospective data derived from a single institutional setting.	Methods; Discussion: Limitations
6b	Eligibility criteria	The final analytical dataset comprised 335 individuals classified into non-disorder and abnormal liver biochemistry groups according to recorded ICD-coded diagnoses.	Methods: Study design and dataset
6c	Treatments	Treatment and medication information was unavailable and was not used for model development.	Discussion: Limitations

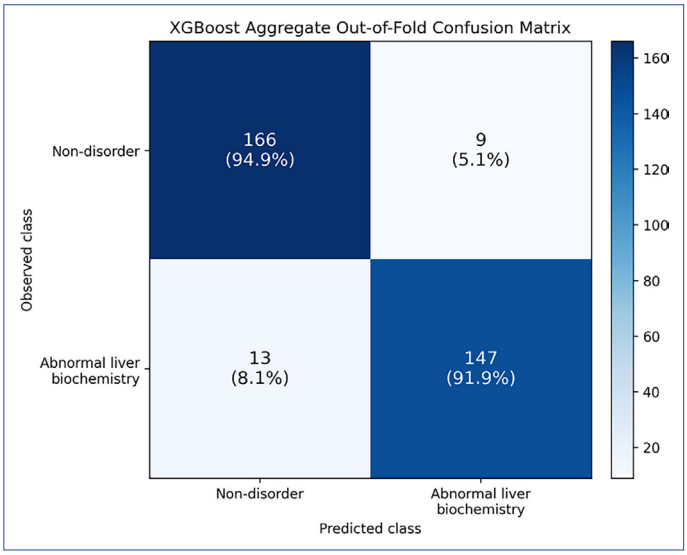
Appendix 2. Cont.

Item	Topic	How addressed in this study	Reported on page/section
7	Data preparation	Data-quality checks, missing-data handling, fold-wise imputation, scaling, class distribution, and procedures to prevent information leakage are described.	Methods: Data preprocessing
8a	Outcome	The binary outcome was based on ICD-coded diagnoses and was not generated using deterministic biochemical threshold rules.	Methods: Definition of the abnormal liver biochemistry outcome
8b	Outcome assessor subjectivity	Outcome labels were based on pre-existing ICD coding; independent clinical adjudication was unavailable.	Methods; Discussion
8c	Outcome blinding	No study-level outcome adjudication was performed because outcome labels pre-existed in the hospital information system.	Methods
9a	Predictor selection	Eight routinely available biochemical parameters were evaluated as predictors.	Introduction; Methods
9b	Predictor definition / measurement	Predictors, measurement units, and handling procedures are described.	Methods: Study design and dataset
9c	Subjective predictor assessment	Predictors consisted of objective laboratory measurements and did not require subjective study-level assessment.	Not applicable
10	Sample size	The analysis used the complete available analytical dataset of 335 individuals; no additional sampling was performed.	Methods; Results
11	Missing data	Missingness occurred only for INR; fold-wise median imputation was used for machine-learning analyses, whereas individual INR analyses used available cases without imputation.	Methods: Data preprocessing; individual biomarker analysis
12a	Use/partition of data	Internal validation used 10×5 repeated stratified cross-validation, with preprocessing performed within training folds.	Methods: Machine learning model development and internal validation
12b	Predictor handling	Fold-wise standardization was used for scale-sensitive algorithms; transformations and standardization for conventional regression are specified.	Methods; Table 3
12c	Model building/ hyperparameters/ internal validation	Nine supervised algorithms with fixed parameter settings were evaluated without data-driven hyperparameter optimization. Full settings are provided in Supplementary Table S1.	Methods; Supplementary Table S1
12d	Cluster heterogeneity	The dataset was derived from a single centre; cluster-level heterogeneity analysis was therefore not applicable.	Not applicable
12e	Performance/ model comparison	ROC-AUC, classification metrics, calibration, Brier score, exploratory decision curve analysis, SHAP, and paired ROC-AUC comparisons were evaluated.	Methods
12f	Model updating	No external model updating or recalibration was performed.	Not applicable
12g	Predictions in external evaluation	No independent external evaluation dataset was available.	Not applicable
13	Class imbalance	Outcome classes were relatively balanced (175 vs 160); oversampling and synthetic data generation were not used.	Methods: Data preprocessing
14	Fairness	Fairness analyses could not be performed because relevant sociodemographic variables were unavailable.	Not applicable
15	Model output/ thresholds	The positive class and classification metrics are defined; a 0.50 threshold was used for the aggregate OOF confusion matrix, and biomarker cutoffs were derived from training folds using the Youden index.	Methods; Results
16	Development vs evaluation data	Model development and internal validation were performed within the same dataset using repeated cross-validation; no independent external dataset was used.	Methods; Discussion
17	Ethics	Ethics approval, approval number/date, Declaration of Helsinki compliance, and waiver of informed consent are reported.	Methods; Ethical considerations
18a	Funding	No external funding was received.	Funding
18b	Conflicts of interest	No financial or non-financial conflicts of interest were declared.	Conflict of interest
18c	Protocol	No separate prospective study protocol was registered for this retrospective analysis.	Not applicable
18d	Registration	The retrospective study was not prospectively registered.	Not applicable
18e	Data sharing	Data are available from the corresponding author upon reasonable request.	Data availability
18f	Code sharing	Analytical code was not deposited in a public repository.	Not publicly available

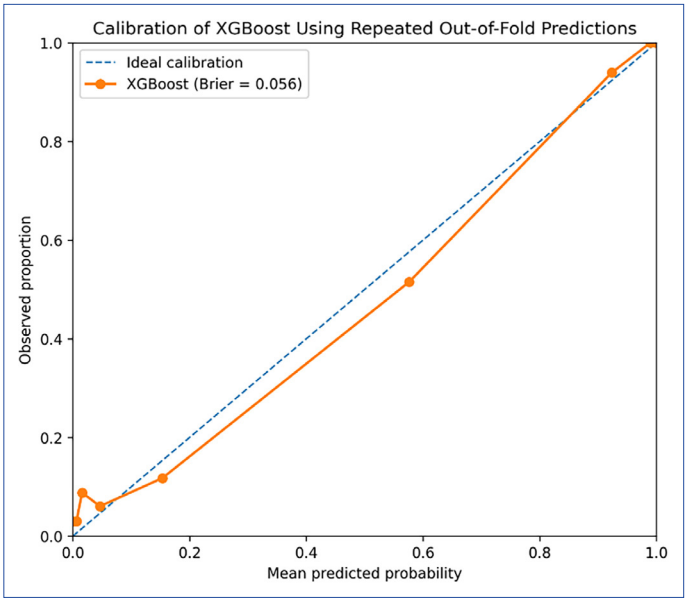
Appendix 2. Cont.

Item	Topic	How addressed in this study	Reported on page/section
19	Patient/public involvement	Patients or members of the public were not involved in the design, conduct, reporting, or dissemination planning of this retrospective analysis.	Not applicable
20a	Participant flow	The final analytical sample and outcome-group counts are reported. No longitudinal follow-up was involved.	Results: Study population and dataset characteristics
20b	Participant characteristics	Available biochemical characteristics are presented by outcome group; additional demographic and clinical characteristics were unavailable.	Table 1; Discussion: Limitations
20c	Comparison with development data	No separate external evaluation dataset was used.	Not applicable
21	Numbers in analyses	Overall sample size and group counts are reported; the INR complete-case sensitivity analysis reports the corresponding analytical sample.	Results
22	Full model specification	Model-specific estimator settings and preprocessing procedures are provided in Supplementary Table S1.	Supplementary Table S1
23a	Model performance	Repeated-CV discrimination and classification metrics, ROC-AUC uncertainty, calibration, exploratory decision analysis, and explainability results are reported.	Results; Table 4; Figures; Supplementary Figures
23b	Cluster heterogeneity	No multicentre or clustered evaluation was performed.	Not applicable
24	Model updating	No model updating following external evaluation was performed.	Not applicable
25	Interpretation	Findings are interpreted as internal classification performance rather than independent clinical diagnostic validity; incorporation bias and generalizability are discussed.	Discussion
26	Limitations	Outcome definition, residual incorporation bias, sample size, single-centre design, missing clinical information, lack of external validation, and calibration limitations are discussed.	Discussion: Limitations
27a	Unavailable/ poor-quality inputs	INR missingness and its handling are described; the model is not proposed for clinical deployment.	Methods; Discussion
27b	User interaction / expertise	Clinical implementation and user requirements were not defined because the model is exploratory and not proposed for deployment.	Not applicable
27c	Future research / generalizability	External validation in larger multicentre cohorts with independently adjudicated outcomes is recommended before clinical application.	Discussion; Conclusions

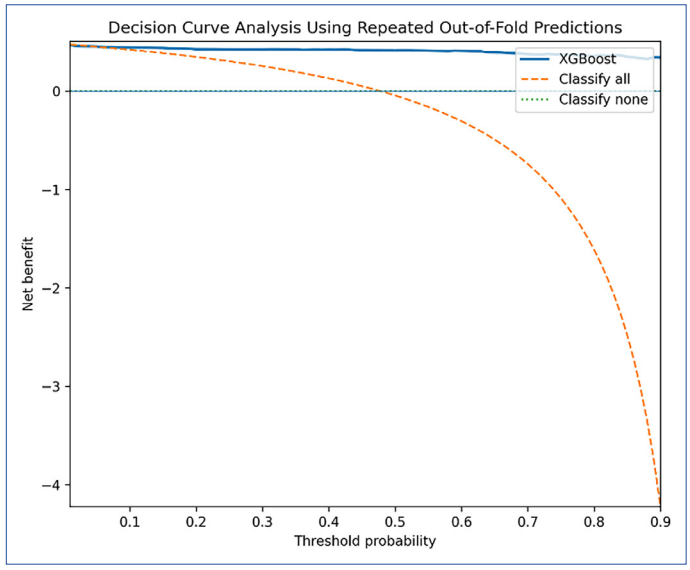
Checklist prepared against the 27-item TRIPOD+AI framework. Item descriptions are abbreviated. Page/section locations refer to the revised manuscript structure; final page numbers should be updated after the tracked-changes version is paginated.



Appendix 3. Aggregate out-of-fold confusion matrix of the XGBoost model based on 10×5 repeated stratified cross-validation. For each individual, predicted probabilities obtained across the 10 repetitions were averaged and converted to class predictions using a threshold of 0.50. Cell values show counts and row percentages.



Appendix 4. Calibration curve of the XGBoost model based on averaged repeated out-of-fold predictions from 10×5 repeated stratified cross-validation. Predicted probabilities obtained for each individual across the 10 repetitions were averaged before calibration assessment. The diagonal line represents perfect calibration. The aggregate out-of-fold Brier score was 0.056.



Appendix 5. Decision curve analysis of the XGBoost model based on averaged repeated out-of-fold predictions from 10×5 repeated stratified cross-validation. Net benefit is shown across threshold probabilities for the XGBoost model and the classify-all and classify-none strategies. The analysis was considered exploratory because the outcome was ICD-defined and no independent clinical endpoint was available.