



Article

Clinically Interpretable Machine Learning for Glycemic Risk Screening in Trauma Patients

Melike Tombaz ¹, Andreas K. Nüssler ¹, Engin Tercan ¹, Niklas R. Braun ², Andreas Fritsche ^{3,4,5}, Tina Histing ², Nico Pfeifer ⁶ and Sabrina Ehnert ^{1,*}

¹ Siegfried Weller Institute for Trauma Research, BG Clinic Tuebingen, University of Tuebingen, 72076 Tuebingen, Germany; melike.tombaz@gmail.com (M.T.); andreas.nuessler@gmail.com (A.K.N.); engintercan89@gmail.com (E.T.)

² Department for Traumatology and Reconstructive Surgery, BG Clinic Tuebingen, University of Tuebingen, 72076 Tuebingen, Germany; nbraun2@bgu-tuebingen.de (N.R.B.); thisting@bgu-tuebingen.de (T.H.)

³ Internal Medicine IV, Department of Diabetology, Endocrinology and Nephrology, University of Tuebingen, 72076 Tuebingen, Germany; andreas.fritsche@med.uni-tuebingen.de

⁴ German Center for Diabetes Research (DZD), 85764 Neuherberg, Germany

⁵ Institute for Diabetes Research and Metabolic Diseases, Helmholtz Center Munich, University of Tuebingen, 72076 Tuebingen, Germany

⁶ Institute for Bioinformatics and Medical Informatics, University of Tuebingen, 72076 Tuebingen, Germany; nico.pfeifer@uni-tuebingen.de

* Correspondence: sabrina.ehnert@med.uni-tuebingen.de; Tel.: +49-7071-606-1065

Abstract

Undiagnosed diabetes and prediabetes are common among trauma patients and increase perioperative complication rates, yet routine glycemic screening at hospital admission is rarely performed. We compared 10 supervised machine learning (ML) algorithms for early detection of impaired glucose metabolism in 1789 trauma surgery patients (1095 normal; 694 prediabetes/diabetes) using questionnaire data and preoperative laboratory values within a nested cross-validation framework with hyperparameter optimization. On the combined feature set, XGBoost achieved the highest area under the receiver operating characteristic curve (ROC-AUC; 0.815), followed by multilayer perceptron (0.814) and Gradient Boosting (0.813); adding blood parameters significantly improved performance for nine of ten algorithms (DeLong's test, $p < 0.05$). SHapley Additive exPlanations (SHAP) and XGBoost feature gains identified six overlapping predictors, yielding a simplified model with comparable mean discrimination. Inter-model differences were small, and a fully interpretable logistic regression model remained within one percentage point of XGBoost. The best-performing model demonstrated strong calibration and a higher net benefit than either universal testing or no testing across the evaluated threshold probabilities. In routine trauma practice, such a model could identify patients who would benefit from confirmatory glycated hemoglobin (HbA1c) testing, enabling earlier detection of dysglycemia, risk-stratified glucose management, and timely referral for lifestyle intervention in patients with prediabetes.



Academic Editors: Milan Toma and Constantinos Pattichis

Received: 14 July 2026

Revised: 17 August 2026

Accepted: 20 August 2026

Published: 22 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: supervised learning; diabetes mellitus; prediabetes; orthopedic trauma; XGBoost; SHAP

1. Introduction

Diabetes mellitus (DM) is a metabolic disorder defined by elevated blood sugar levels [1]. The Diabetes Atlas reports that 589 million individuals globally are affected by

diabetes, ranking it as the eighth leading cause of disease burden worldwide [2]. Type 2 diabetes accounts for more than 90% of all diabetes cases, and its precursor, prediabetes, is alarmingly prevalent, with 35–50% of the affected individuals progressing to type 2 diabetes within five years [3–5]. In Germany, 10% of the population had a confirmed diagnosis of diabetes, with total diabetes-related health expenditures reaching 40.4 billion dollars in 2024 [2].

The metabolic burden of diabetes is especially pronounced in trauma and orthopedic surgery [6]. In a tertiary trauma center, 10.3% and 12% of patients self-reported a diabetes diagnosis in 2016 and 2019, respectively, based on patient questionnaires without confirmatory testing. Pscherer et al. reported comparable rates using hospital documentation [6,7]. More recently, Maurer et al. observed that 23.4% of patients with septic surgical infections or chronic wounds in this setting had diabetes, highlighting the high prevalence of dysglycemia in high-risk trauma subgroups [6–8]. These patients experienced significantly higher complication rates and longer hospital stays compared to metabolically healthy individuals [7,9]. On average, prediabetic patients stayed 25% longer, while diabetic patients stayed nearly twice as long [7]. These patterns complicate treatment and place a substantial burden on healthcare resources.

Dysglycemia is particularly consequential in trauma and orthopedic surgery, where impaired glucose metabolism is associated with delayed fracture healing, postoperative infection, pseudoarthrosis, and prolonged hospitalization [10–12]. These complications occur more frequently not only in patients with diabetes but also in those with prediabetes, underscoring the value of identifying dysglycemia early [1,13]. Beyond the musculoskeletal system, untreated hyperglycemia contributes to renal, cardiovascular, and ophthalmic complications, reinforcing the need for screening methods suited for orthopedic trauma populations [2].

Despite the high prevalence and clinical significance, routine screening for diabetes and prediabetes is not standard practice in trauma clinics. The most widely used diagnostic criteria for diabetes require multiple tests, including the oral glucose tolerance test (OGTT) and glycated hemoglobin (HbA1c) test, which are rarely performed in routine clinical practice in trauma clinics [14]. As a result, many cases of prediabetes and diabetes remain undetected until complications develop. This gap underscores the need for practical, rapid screening tools to identify high-risk individuals at admission, prior to confirmatory laboratory testing. Early detection of impaired glucose metabolism (e.g., prediabetes and diabetes) may help delay the progression of cardiovascular, renal, neurological, and musculoskeletal complications [15].

Artificial intelligence (AI), particularly supervised machine learning (ML), offers a promising approach for clinical prediction. Supervised ML models are increasingly applied to identify patients at risk of adverse clinical outcomes and to support data-driven decision-making [16–20]. Various supervised learning models are specifically adapted for disease prediction, enabling timely identification of patients at elevated metabolic risk and facilitating early confirmatory testing and proactive perioperative management. Although ML methods are increasingly applied in general medical populations, few studies have assessed their effectiveness in trauma-specific populations, and even fewer have compared the performance of different ML algorithms in this context [21–24]. Existing machine learning applications relevant to this work fall into two largely separate streams. The first predicts diabetes or prediabetes, but almost exclusively in general or primary-care populations, where the objective is community screening or long-term risk stratification rather than perioperative decision-making. The second applies machine learning within trauma populations, but to predict trauma-related outcomes such as mortality, injury severity, or emergency triage, rather than metabolic risk [21–24]. Neither stream addresses glycemic

risk screening at trauma admission, which differs in case-mix, available admission data, and clinical objective: flagging existing but unrecognized impaired glucose metabolism before surgery.

To our knowledge, this is the first study to address that gap by systematically benchmarking 10 supervised ML algorithms for detecting impaired glucose metabolism in trauma surgery patients. Beyond discrimination, we evaluated model calibration and clinical utility and investigated whether a compact, clinically interpretable feature set could retain the predictive performance of the full feature set. This framework was designed to identify models that could support rapid prioritization of patients for confirmatory glycemic testing while relying primarily on information already collected during routine trauma admission.

2. Materials and Methods

2.1. Study Population and Data

This single-center observational study used prospectively collected data from trauma and orthopedic surgery patients recruited during three enrollment periods. Patients were enrolled at the preoperative anesthesia assessment unit of a tertiary trauma center, where patients scheduled for surgery undergo a standardized preoperative evaluation. Patients were recruited during three enrollment periods, each spanning six consecutive weeks (October–November 2021, October–November 2024, March–April 2025). Within each term, all eligible patients admitted for trauma or orthopedic surgery were enrolled consecutively, with identical eligibility criteria across periods and without further selection, which limits selection bias. This unit provides preoperative assessment for patients admitted for planned trauma and orthopedic surgery, including primary and revision surgery. Patients scheduled for surgery completed a questionnaire, and blood samples were collected for routine analyses before the surgery. To establish accurate ground-truth labels for the machine learning tasks, an HbA1c test was ordered for all enrolled patients strictly as part of this study. All patients provided written informed consent before participating in the study and completing the questionnaires, which were adapted from the Finnish Diabetes Risk Score (FINDRISK) and German Diabetes Risk Score (GDRS) [25,26]. Patients were excluded only if they declined participation, submitted incomplete questionnaires, or had insufficient information on sex or blood sugar levels (Figure 1).

Patients were classified as having diabetes if at least one of the following three criteria was met:

- (i) $\text{HbA1c} \geq 6.5\%$,
- (ii) random plasma glucose levels ≥ 200 mg/dL, or
- (iii) a previous diabetes diagnosis by a medical doctor.

Patients with an HbA1c level $\geq 5.7\%$ and $<6.5\%$ were diagnosed with prediabetes, as recommended by the American Diabetes Association (ADA) [14].

2.2. Data Preprocessing

The analyzed blood parameters were derived from the hospital's standard preoperative admission laboratory panel, encompassing hematological, biochemical, and metabolic markers. Categorical features were numerically encoded. Features missing in more than 25% of patients were excluded from the analysis. This threshold was selected empirically based on the observed distribution of missingness within the cohort and distinguished routinely available admission variables from nonroutine laboratory measurements with substantially higher missingness. The missingness threshold was applied during preprocessing, prior to model development, hyperparameter optimization, and performance evaluation. This approach was intended to retain clinically representative routine variables while limiting extensive imputation of sparsely measured parameters. Detailed missing-

ness percentages for all evaluated features are provided in Supplementary Table S1. For retained features, missing numerical values were imputed with the column median, while missing categorical values were imputed with the most frequent category (mode). The target variable, defined by patient blood glucose levels or previous diagnosis by a medical doctor, was encoded as a binary variable: 0 for normal blood glucose (Ctrl) and 1 for impaired glucose metabolism, including both prediabetes and diabetes (hereafter Pre-/DM). Detailed descriptions, measurement units, and data types of all features included in the analysis are provided in Supplementary Table S1.

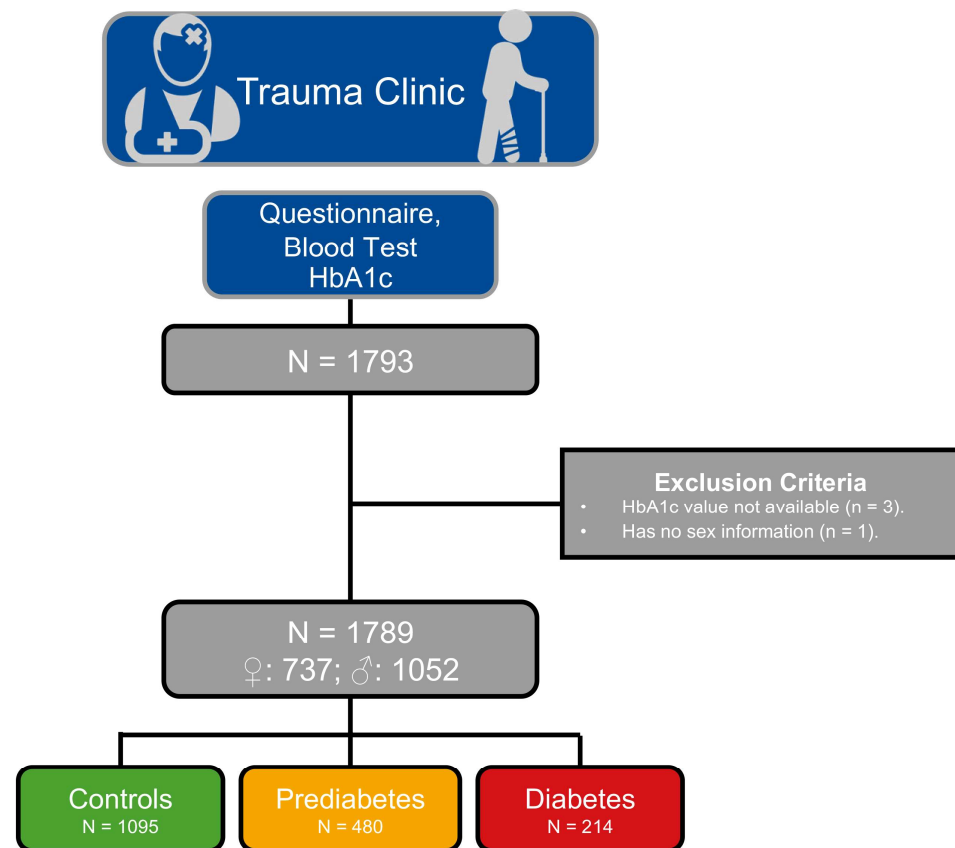


Figure 1. Flow diagram of patient selection and stratification.

2.3. Multicollinearity Assessment

To mitigate multicollinearity, we employed an iterative backward elimination procedure based on Variance Inflation Factors (VIF). In each step, VIF scores were recalculated for all features, and the single feature with the highest VIF exceeding a threshold of 10 was removed. This cycle continued until all remaining features satisfied the VIF criteria. Features that were retained after multicollinearity filtering, along with their final variance inflation factors, are listed in Supplementary Table S2.

2.4. Algorithm Selection and Performance Evaluation

A nested cross-validation (Nested CV) framework was implemented to provide robust performance estimates and reduce the risk of overfitting. This framework separates model selection and hyperparameter optimization from performance evaluation on the held-out outer folds. The outer loop employed five-fold stratified cross-validation to evaluate generalizability, while the inner loop, also five-fold, was used for hyperparameter optimization via Optuna. Numerical feature standardization (using *StandardScaler* to achieve a mean of 0 and a standard deviation of 1) was performed within the cross-validation folds. During the inner loop, the scaler was fit exclusively on the inner training partition and applied to

the inner validation partition. Similarly, in the outer loop, a new scaler was fit exclusively on the outer training partition and applied to the held-out outer test partition.

The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) was the primary metric for model evaluation. Additional performance metrics included accuracy, precision, recall, F1-score, specificity, and Matthews correlation coefficient (MCC). Precision, recall, and F1-score refer to the positive Pre-/DM class. Metrics were calculated independently in each outer test fold and are reported as mean $\pm$ standard deviation (SD) across five folds. Binary class predictions were generated using each estimator's default decision rule. For classifiers providing predicted probabilities, a probability threshold of 0.5 was used to assign the positive Pre-/DM class.

Ten supervised machine learning algorithms were evaluated: Support Vector Machine (SVM) [27], K-Nearest Neighbors (KNN) [28], Multilayer Perceptron (MLP) [29,30], Decision Tree [31], Random Forest [32], ExtraTrees [33], AdaBoost [34], Gradient Boosting [35], Logistic Regression [36], and Extreme Gradient Boosting (XGBoost) [37]. The algorithm with the highest mean ROC-AUC across the outer folds was selected for final model development and interpretability analysis. The complete machine learning pipeline is illustrated in Figure 2.

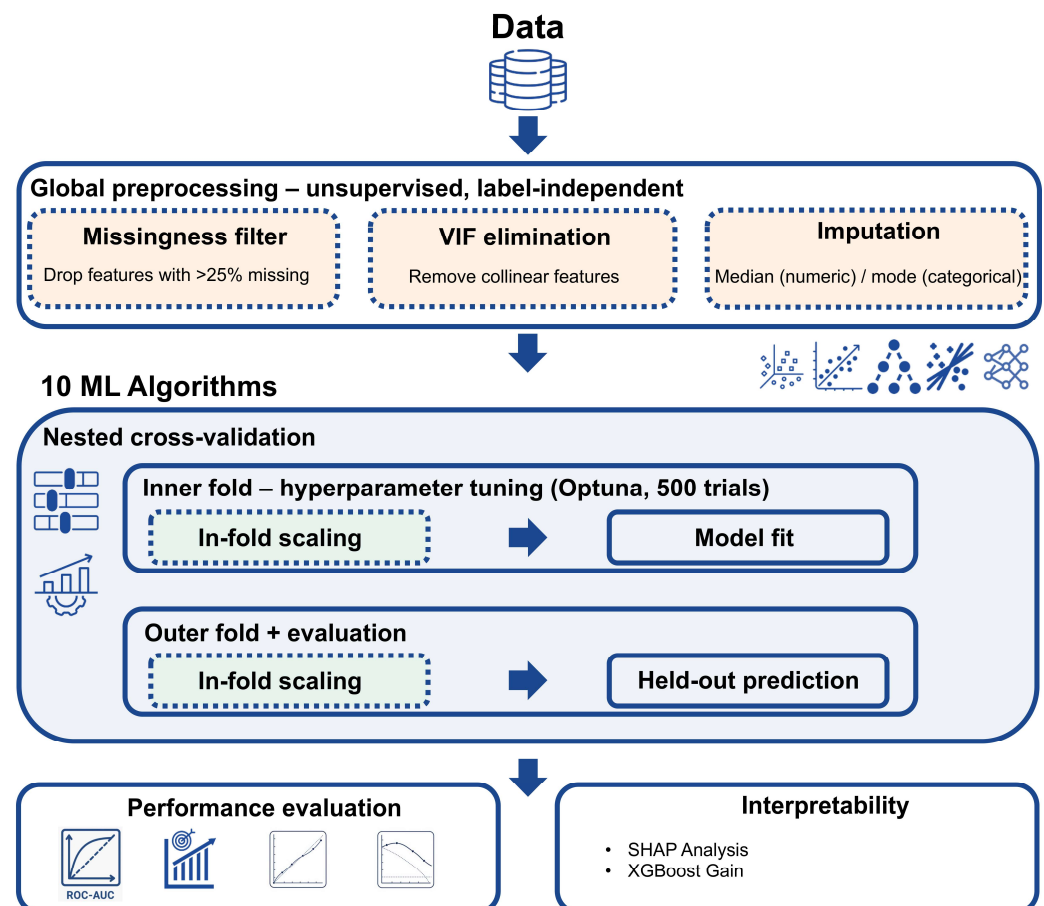


Figure 2. Machine learning pipeline. Data underwent preprocessing (missingness filtering and missing value imputation) followed by variance inflation factor (VIF) analysis to remove multicollinear features. Ten supervised algorithms were trained using nested five-fold cross-validation with Optuna hyperparameter tuning. Numerical standardization was fitted within the corresponding training folds and subsequently applied to held-out validation or test data. Model performance was evaluated primarily using Receiver Operating Characteristic–Area Under the Curve (ROC-AUC), with additional classification metrics. The highest-performing model was subsequently subjected to global interpretability analysis and consensus feature selection.

All algorithms were implemented using their standard reference implementations in scikit-learn (v1.4.2), except XGBoost, which used the xgboost package (v2.1.3). Hyperparameter optimization was performed with Optuna (v3.6.1). Within each outer-training fold, 500 Optuna trials were performed, with each trial evaluated using stratified five-fold inner cross-validation. To ensure determinism, a single random seed (42) was applied across Python, NumPy (v2.4.4), and all estimators, and numerical libraries (OpenMP, MKL, OpenBLAS, BLIS, NumExpr) were constrained to a single thread. The complete hyperparameter search spaces are reported in Supplementary Table S3.

2.5. Global Optimization and Final Model Configuration

After the algorithm with the highest mean ROC-AUC was identified through nested cross-validation, hyperparameters were re-optimized on the complete VIF-reduced dataset using a stratified five-fold cross-validation procedure within the Optuna framework (500 trials) [38]. The resulting model was fit to the full VIF-reduced dataset and used for SHapley Additive exPlanations (SHAP) and Gain-based feature importance analyses. This full-dataset reference model was used only for interpretability and feature selection and was not used to derive the primary nested cross-validation performance estimates reported in Section 2.4. The globally optimized hyperparameters of the final model are reported in Supplementary Table S4.

2.6. Benchmarking Against Established Diabetes Risk Scores

To provide a clinically established questionnaire-based benchmark, FINDRISK and GDRS were reconstructed from the corresponding variables available in the study questionnaire. Their discriminative performance for distinguishing Pre-/DM from normal glycemic status was assessed using receiver operating characteristic analysis. ROC-AUC values and 95% confidence intervals were calculated, with confidence intervals estimated using 2000 bootstrap resamples with replacement. Bootstrap samples containing only one outcome class were excluded. Because the study questionnaire was adapted from these instruments rather than administered as the original stand-alone questionnaires, these analyses were considered reconstructed benchmark comparisons. The reconstructed score ROC curves are presented in Supplementary Figure S1.

2.7. Diagnosis-Naïve Sensitivity Analysis

To evaluate model performance specifically for opportunistic detection of previously unrecognized dysglycemia, a sensitivity analysis was performed in a diagnosis-naïve subset. Patients with a previously documented diagnosis of diabetes were excluded. In addition, the self-reported questionnaire variable “Previous high blood sugar levels” was removed from the predictor set to prevent reliance on prior glycemic information. The nested cross-validation pipeline was then repeated in the resulting subset using the same model-development and evaluation framework as in the primary analysis. ROC curves for the diagnosis-naïve sensitivity analysis are shown in Supplementary Figure S2.

2.8. Model Calibration and Clinical Utility Analysis

Calibration and clinical utility were evaluated using out-of-fold predictions from the outer-fold models of the nested cross-validation procedure. For each outer fold, the *StandardScaler* was fitted exclusively to the outer-training partition and applied to the corresponding held-out test partition before prediction. Calibration was summarized by binning held-out predicted probabilities into deciles within each fold, averaging predicted probabilities and observed event frequencies across folds, and displaying variability as ± 1 SD relative to the line of perfect calibration.

To assess clinical utility, we conducted Decision Curve Analysis (DCA) on pooled out-of-fold predictions, calculating net benefit across threshold probabilities from 0.01 to 0.99. In this screening context, a positive model result was interpreted as prompting confirmatory HbA1c testing rather than directly initiating treatment. The model was compared with strategies of testing all patients and testing no patients.

2.9. Learning Curve Analysis

Learning curves were generated for all ten algorithms to assess convergence behavior and potential overfitting. For each outer fold, the hyperparameter configuration selected during the corresponding inner cross-validation optimization was retained and refitted on eight progressively increasing subsets of the outer-training data, ranging from 10% to 100%. Numerical standardization was fitted separately on each training subset and subsequently applied to the unchanged held-out outer-test fold. ROC-AUC was calculated for both the training subset and held-out test fold, and mean values with variability across the five outer folds were plotted as a function of training-set size. Hyperparameter optimization was not repeated at each training-set size. Complete learning curves for all ten algorithms are provided in Supplementary Figures S3 and S4 for the questionnaire-only and combined feature sets, respectively.

2.10. Interpretability and Consensus Feature Selection

Feature importance was analyzed using two complementary metrics to identify the clinical factors influencing model predictions. Global feature contributions were quantified by aggregating SHAP (SHapley Additive exPlanations, v0.46.0) values across the dataset, and a beeswarm plot was generated to visualize the magnitude and direction of each variable's effect on predicted risk [39]. XGBoost gain-based feature importance was also calculated. Gain represents the average improvement in the model's optimization objective attributable to splits involving a given feature. Features were ranked separately according to mean absolute SHAP value and XGBoost gain. The ten highest-ranked features from each method were identified, and the intersection of these two top-ten lists was retained as the consensus feature set. This procedure resulted in six overlapping features. Average ranks were used only to order the retained features and did not determine whether a feature was included. To assess whether model discrimination could be retained using this compact feature set, a reduced model was trained using this subset and evaluated against the full-feature model with identical five-fold stratified cross-validation splits. Within each outer-training fold, hyperparameters were optimized independently for the full and reduced models using the inner cross-validation procedure. Both optimized models were subsequently evaluated on the same held-out outer-test partition. Because the consensus features were identified using the complete dataset before this comparison, the reduced-model analysis was considered a *post hoc* internal assessment rather than an independent validation.

2.11. Prototype Clinical Screening Interface

To demonstrate practical application, a lightweight prototype interface was developed using *Streamlit* (v1.57.0). This tool uses the reduced-feature-set model, allowing users to enter clinical features and obtain an estimated probability of impaired glucose metabolism for an individual patient. The prototype is a technical feasibility demonstration only: it outputs a continuous predicted probability and does not embed a fixed decision threshold or issue an automated testing recommendation.

2.12. Statistical Analysis

Continuous quantitative features were reported as medians and interquartile ranges (IQRs), while categorical features were summarized as counts and percentages. Group comparisons were performed using the Wilcoxon rank-sum test or Kruskal–Wallis test for continuous variables, depending on data distribution. Categorical variables were compared using the Chi-square test; however, Fisher’s exact test was employed when the expected frequency in any cell of the contingency table was less than 5.

To assess whether the addition of routine laboratory parameters improved discrimination, the ROC-AUC of each algorithm trained on questionnaire data alone was compared with that of the same algorithm trained on combined questionnaire and laboratory data using DeLong’s test for two correlated ROC curves. For each algorithm, out-of-fold predictions were pooled across the five outer folds, generated using identical stratified cross-validation splits so that predictions were aligned to the same patients, and the resulting paired ROC curves were compared. Because out-of-fold predictions were pooled across the five outer folds prior to testing, each condition’s pooled ROC curve reflects five distinct fold-specific models rather than a single fixed classifier evaluated once; this pooling approach is standard practice for cross-validated model comparison but represents a simplification relative to DeLong’s original single-classifier framework. Statistical significance was defined as $p < 0.05$. Analyses were performed with Python (3.12, Python Software Foundation, Wilmington, DE, USA).

3. Results

3.1. Baseline Patient Characteristics

A total of 1789 trauma patients were included in the study. Of these, 1095 (61.2%) exhibited normal blood glucose levels, while 694 (38.8%) were classified as either having prediabetes or diabetes (Pre-/DM). Patients in the Pre-/DM group were significantly older (63.0 vs. 49.0 years), had a higher body mass index (BMI) (27.78 kg/m² vs. 25.98 kg/m²), and a greater waist circumference (101.0 cm vs. 93.0 cm) compared to patients with normal glucose levels (Table 1).

Table 1. Baseline characteristics of patients.

Characteristic	Normal Blood Glucose (<i>n</i> = 1095)	Pre-/DM (<i>n</i> = 694)	<i>p</i> -Value
Demographics			
Age (years)	49.0 [31.0, 62.0]	63.0 [55.0, 73.4]	<0.001
Sex			0.001
Male	678 (61.9%)	374 (53.9%)	
Female	417 (38.1%)	320 (46.1%)	
Anthropometrics			
Height (m)	1.75 [1.67, 1.80]	1.71 [1.65, 1.79]	<0.001
Weight (kg)	80.0 [68.0, 91.0]	82.0 [70.0, 95.75]	<0.001
BMI (kg/m ²)	25.98 [23.03, 29.32]	27.78 [24.39, 31.86]	<0.001
Waist Circumference (cm)	93.0 [84.0, 103.0]	101.0 [92.0, 112.0]	<0.001
Laboratory Values			
MCHC	33.90 [33.2, 34.5]	33.30 [32.8, 34.0]	<0.001
Hematocrit	42.60 [39.80, 45.20]	42.0 [39.0, 45.0]	0.009
Leukocytes	6.62 [5.55, 8.00]	7.20 [5.90, 8.60]	<0.001
Prothrombin Time (QUICK)	100.0 [93.0, 109.0]	105.0 [96.25, 112.0]	<0.001
APTT	24.0 [23.0, 25.0]	23.0 [22.0, 25.0]	<0.001
Potassium	4.10 [3.80, 4.30]	4.20 [3.90, 4.50]	<0.001
Sodium	141.00 [139.00, 142.00]	140.00 [139.00, 142.00]	<0.001

Table 1. Cont.

Characteristic	Normal Blood Glucose (n = 1095)	Pre-/DM (n = 694)	p-Value
Medical History			
Antihypertensive medication			<0.001
Yes	275 (25.1%)	367 (52.9%)	
No	819 (74.9%)	327 (47.1%)	
Previous high blood sugar levels			<0.001
Yes	43 (3.9%)	220 (31.7%)	
No	1051 (96.1%)	474 (68.3%)	
Biological mother has diabetes			<0.001
Yes	102 (9.3%)	137 (19.7%)	
No	993 (90.7%)	557 (80.3%)	
Biological father has diabetes			0.025
Yes	120 (11.0%)	101 (14.6%)	
No	974 (89.0%)	593 (85.4%)	
Family History of Diabetes			<0.001
First Degree Relative	281 (25.7%)	257 (37.0%)	
Second Degree Relative	100 (9.1%)	42 (6.1%)	
No history	714 (65.2%)	395 (56.9%)	
Lifestyle Factors			
At least 30 min of daily exercise			0.056
Yes	1031 (94.2%)	638 (91.9%)	
No	63 (5.8%)	56 (8.1%)	
Smoking Status			<0.001
Non-smoker	607 (55.4%)	324 (46.7%)	
Ex-smoker	248 (22.6%)	225 (32.4%)	
Smoker	240 (21.9%)	145 (20.9%)	
Department			
Arthroscopy	128 (11.7%)	65 (9.4%)	<0.001
Endoprosthetics	121 (11.1%)	155 (22.4%)	
Hand and Plastic Surgery	233 (21.3%)	152 (21.9%)	
Oral and Maxillofacial Surgery	94 (8.6%)	37 (5.3%)	
Rehabilitation and Treatment	1 (0.1%)	2 (0.3%)	
Coordination			
Septic Surgery	53 (4.8%)	56 (8.1%)	
Spine and Spinal Cord Surgery	21 (1.9%)	8 (1.2%)	
Trauma Surgery	443 (40.5%)	218 (31.5%)	

Continuous data are presented as median [interquartile range]; categorical variables as n (%). Percentages were calculated using the number of patients with available data for each characteristic. Therefore, denominators may differ from the total group sizes. APTT: activated partial thromboplastin time; MCHC: mean corpuscular hemoglobin concentration.

Several clinical and laboratory characteristics also differed between groups. Previous high blood sugar levels were reported by 31.7% of patients in the Pre-/DM group compared with 3.9% of those with normal glucose metabolism. Antihypertensive medication use and a family history of diabetes were also more frequent in the Pre-/DM group. Differences were additionally observed in several routinely measured laboratory parameters, including mean corpuscular hemoglobin concentration (MCHC), prothrombin time (QUICK value), activated partial thromboplastin time (APTT), leukocyte count, potassium, and sodium (Table 1).

3.2. Model Evaluation of Questionnaire-Only and Combined Data

After pre-processing, ten supervised machine learning models were trained using either questionnaire data alone or in combination with preoperative blood values. Features

retained after multicollinearity filtering, with their final variance inflation factors, are listed in Supplementary Table S2, and the hyperparameter search spaces explored for each algorithm are given in Supplementary Table S3. Logistic Regression achieved the highest ROC-AUC (0.797 ± 0.024) among questionnaire-only models (Figure 3a), whereas MLP demonstrated the highest F1-score (61.3 ± 4.0). Performance across questionnaire-only models was relatively similar, with mean ROC-AUC ranging from 0.764 to 0.797. A detailed summary of the performance metrics is provided in Table 2.

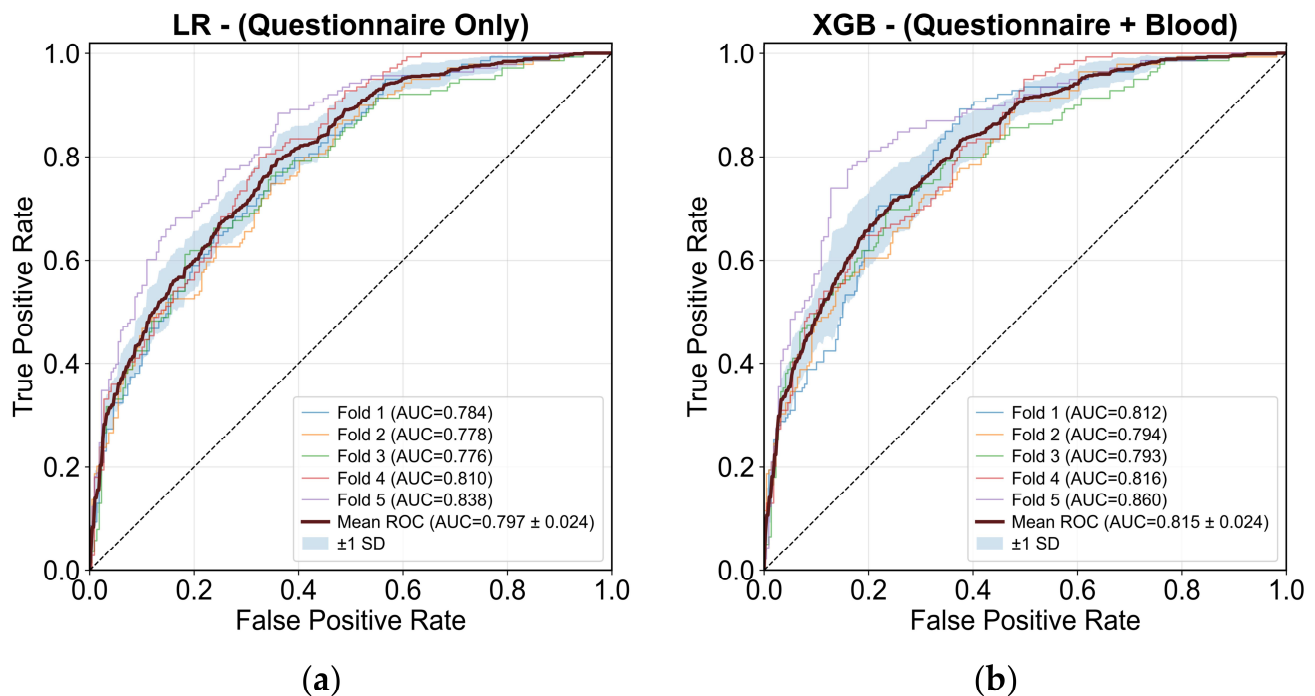


Figure 3. Receiver operating characteristic (ROC) curves across outer folds for two representative models. (a) Logistic Regression trained on questionnaire-only features. (b) Extreme Gradient Boosting (XGBoost) was trained on the combined questionnaire and blood features. Individual fold ROC curves are shown in different colors, with the mean ROC curve in bold, the shaded area indicating ± 1 standard deviation (SD), and the mean area under the curve (AUC) reported with SD.

Table 2. Performance metrics for all algorithms with questionnaire-only data.

Model	AUC-ROC	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Specificity (%)	MCC
LR	0.797 $\pm$ 0.024	73.2 $\pm$ 2.4	69.6 $\pm$ 3.2	54.8 $\pm$ 4.9	61.2 $\pm$ 4.0	84.8 $\pm$ 1.8	0.419 $\pm$ 0.053
RF	0.796 $\pm$ 0.026	72.4 $\pm$ 1.6	68.7 $\pm$ 2.5	53.2 $\pm$ 3.0	59.9 $\pm$ 2.7	84.7 $\pm$ 1.4	0.402 $\pm$ 0.036
MLP	0.796 $\pm$ 0.024	72.6 $\pm$ 2.1	67.8 $\pm$ 3.0	56.2 $\pm$ 6.1	61.3 $\pm$ 4.0	83.0 $\pm$ 2.8	0.410 $\pm$ 0.047
GB	0.795 $\pm$ 0.025	73.2 $\pm$ 2.2	70.3 $\pm$ 3.6	53.5 $\pm$ 3.4	60.7 $\pm$ 3.3	85.7 $\pm$ 2.1	0.418 $\pm$ 0.048
XGB	0.793 $\pm$ 0.031	71.0 $\pm$ 1.6	66.1 $\pm$ 3.3	52.5 $\pm$ 2.0	58.4 $\pm$ 1.9	82.8 $\pm$ 2.4	0.373 $\pm$ 0.034
ET	0.792 $\pm$ 0.023	71.8 $\pm$ 1.9	69.9 $\pm$ 2.5	47.7 $\pm$ 5.4	56.6 $\pm$ 4.4	87.0 $\pm$ 1.5	0.384 $\pm$ 0.045
AB	0.790 $\pm$ 0.026	71.9 $\pm$ 2.1	69.2 $\pm$ 1.9	49.4 $\pm$ 6.2	57.5 $\pm$ 4.7	86.1 $\pm$ 1.0	0.387 $\pm$ 0.049
SVM	0.783 $\pm$ 0.027	71.2 $\pm$ 1.8	82.0 $\pm$ 2.6	33.0 $\pm$ 5.0	46.9 $\pm$ 5.2	95.4 $\pm$ 0.8	0.381 $\pm$ 0.043
KNN	0.768 $\pm$ 0.027	69.8 $\pm$ 2.4	65.9 $\pm$ 4.7	46.1 $\pm$ 2.9	54.3 $\pm$ 3.5	84.8 $\pm$ 2.4	0.339 $\pm$ 0.055
DT	0.764 $\pm$ 0.036	72.0 $\pm$ 2.4	73.1 $\pm$ 5.4	44.5 $\pm$ 5.0	55.1 $\pm$ 4.5	89.4 $\pm$ 3.2	0.390 $\pm$ 0.057

For each algorithm, performance metrics are given as the mean $\pm$ SD of the 5 outer folds. Bold values indicate the highest performance for each metric across the compared models. Precision, recall, and F1-score refer to the positive Pre-/DM class. AB: AdaBoost, DT: Decision Tree, ET: Extra-Trees, GB: Gradient Boosting, KNN: k-Nearest Neighbors, LR: Logistic Regression, MLP: Multilayer Perceptron, RF: Random Forest, SVM: Support Vector Machine, XGB: eXtreme Gradient Boosting, ROC-AUC: Receiver Operating Characteristic–Area Under the Curve, MCC: Matthews Correlation Coefficient.

When questionnaire variables were combined with routine laboratory parameters, XGBoost achieved the numerically highest mean ROC-AUC (0.815 ± 0.024) (Figure 3b), followed by MLP (0.814 ± 0.026) and GB (0.813 ± 0.026). The highest mean F1-score was observed for MLP (66.1 ± 3.8), whereas KNN showed the highest specificity (90.3 ± 1.7). A detailed summary of the performance metrics is provided in Table 3. Per-fold confusion matrices for all algorithms using the full-feature configuration are provided in Supplementary Figure S5.

Table 3. Performance metrics of the questionnaire + blood values.

Model	AUC-ROC	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Specificity (%)	MCC
XGB	0.815 ± 0.024	74.6 ± 2.5	70.8 ± 3.7	58.9 ± 4.4	64.3 ± 3.8	84.6 ± 2.2	0.453 ± 0.055
MLP	0.814 ± 0.026	74.9 ± 2.4	69.3 ± 2.3	63.3 ± 5.1	66.1 ± 3.8	82.3 ± 0.7	0.464 ± 0.053
GB	0.813 ± 0.026	74.1 ± 2.4	70.7 ± 3.4	56.8 ± 4.3	62.9 ± 3.9	85.1 ± 1.6	0.441 ± 0.054
AB	0.812 ± 0.029	74.4 ± 2.8	71.8 ± 3.8	55.8 ± 4.9	62.8 ± 4.5	86.2 ± 1.5	0.446 ± 0.062
RF	0.811 ± 0.019	74.3 ± 3.1	72.5 ± 5.4	54.3 ± 4.4	62.1 ± 4.6	86.9 ± 2.7	0.443 ± 0.069
LR	0.810 ± 0.026	73.3 ± 2.5	68.8 ± 3.2	57.2 ± 5.0	62.4 ± 4.1	83.6 ± 1.8	0.425 ± 0.056
ET	0.809 ± 0.021	74.3 ± 1.9	72.6 ± 2.2	54.0 ± 5.1	61.9 ± 3.9	87.1 ± 1.4	0.443 ± 0.044
SVM	0.808 ± 0.030	72.8 ± 3.1	74.7 ± 9.5	49.6 ± 15.9	56.7 ± 12.8	87.5 ± 6.0	0.416 ± 0.063
KNN	0.781 ± 0.018	69.1 ± 1.2	70.1 ± 3.5	35.6 ± 2.4	47.1 ± 2.4	90.3 ± 1.7	0.318 ± 0.030
DT	0.775 ± 0.037	70.6 ± 3.2	66.0 ± 10.1	56.5 ± 9.4	59.6 ± 3.2	79.5 ± 9.7	0.381 ± 0.067

For each algorithm, performance metrics are given as the mean $\pm$ SD of the 5 outer folds. Bold values indicate the highest performance for each metric across the compared models. AB: AdaBoost, DT: Decision Tree, ET: Extra-Trees, GB: Gradient Boosting, KNN: k-Nearest Neighbors, LR: Logistic Regression, MLP: Multilayer Perceptron, RF: Random Forest, SVM: Support Vector Machine, XGB: eXtreme Gradient Boosting, ROC-AUC: Receiver Operating Characteristic–Area Under the Curve, MCC: Matthews Correlation Coefficient.

The incremental effect of adding routine laboratory parameters was assessed using paired DeLong comparisons of the pooled out-of-fold predictions. ROC-AUC increased for all 10 algorithms, and the improvements reached statistical significance for 9 of the 10 models. The comparison for k-Nearest Neighbors did not reach statistical significance (Table 4).

Table 4. DeLong test results comparing the ROC-AUCs between questionnaire-only and questionnaire + blood values.

Model	ROC-AUC Questionnaire + Blood	ROC-AUC Questionnaire Only	Δ -AUC	p -Value
AB	0.8079	0.7875	0.0203	0.0005
DT	0.7770	0.7584	0.0186	0.0251
ET	0.8097	0.7910	0.0187	0.0000
GB	0.8133	0.7958	0.0176	0.0017
KNN	0.7809	0.7691	0.0119	0.0637
LR	0.8109	0.7976	0.0133	0.0045
MLP	0.8139	0.7964	0.0175	0.0023
RF	0.8118	0.7970	0.0147	0.0077
SVM	0.7888	0.7258	0.0630	0.0000
XGB	0.8159	0.7927	0.0232	0.0001

3.3. Comparison with Established Diabetes Risk Scores

The reconstructed established diabetes risk scores showed moderate discrimination for Pre-/DM in the study cohort (Supplementary Figure S1). FINDRISK achieved a ROC-AUC of 0.749 (95% CI, 0.725–0.772), while GDRS achieved a ROC-AUC of 0.756 (95% CI, 0.733–0.778). Both values were numerically lower than the best-performing questionnaire-only machine-learning model, Logistic Regression, which achieved a mean outer-fold ROC-AUC of 0.797 ± 0.024 .

3.4. Performance in Diagnosis-Naïve Cohort

After excluding patients with previously diagnosed diabetes, the diagnosis-naïve sensitivity analysis included 1610 patients, comprising 1095 with normal glucose metabolism and 515 with Pre-/DM. Within the diagnosis-naïve cohort, discrimination was lower than in the primary analysis but remained informative (Supplementary Figure S2). XGBoost trained using the combined questionnaire and laboratory feature set achieved a mean ROC-AUC of 0.764 ± 0.012 , while Logistic Regression achieved 0.734 ± 0.023 for questionnaire variables alone. These results indicate that predictive information remained after exclusion of known diabetes and historical self-reported hyperglycemia, although performance was attenuated relative to the complete cohort.

3.5. Model Calibration and Clinical Utility

Calibration analysis of the model with the highest ROC-AUC (XGBoost) demonstrated close agreement between predicted probabilities and observed event frequencies across folds (Figure 4a). The mean calibration curve closely followed the diagonal reference line, indicating well-calibrated risk estimates. Decision curve analysis showed a higher net benefit for model-guided confirmatory testing than for testing all patients or testing no patients across a broad range of threshold probabilities (Figure 4b). In this analysis, a positive result represented prioritization for confirmatory HbA1c testing rather than initiation of treatment. Calibration curves for all ten algorithms are provided in Supplementary Figure S6, while corresponding decision curves are shown in Supplementary Figure S7. Numerical expected calibration error (ECE), Brier score, and net-benefit summaries are provided in Supplementary Table S5.

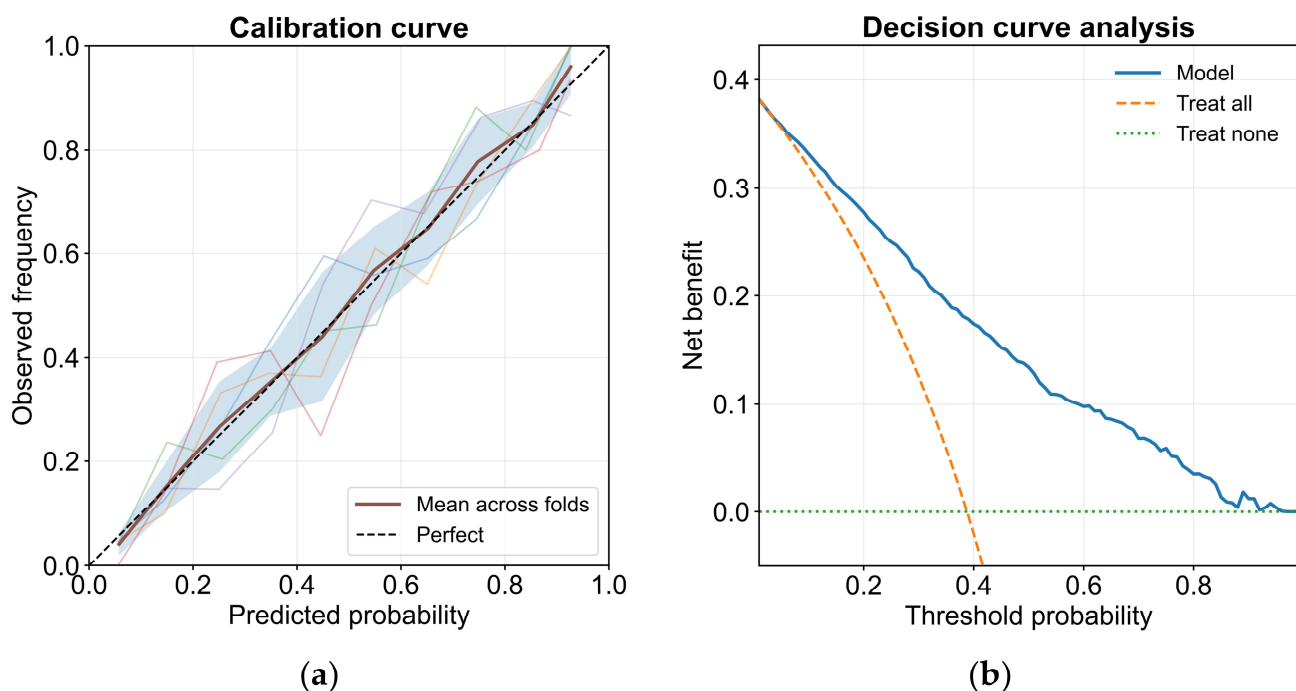


Figure 4. Calibration and decision curve analysis (DCA) for the best-performing XGB model trained on the combined questionnaire and blood features. (a) Calibration plot showing observed versus predicted probabilities across outer folds. Individual colored lines represent the outer cross-validation folds. The solid brown line represents the mean calibration curve, the shaded area indicates ± 1 standard deviation (SD), and the dashed black line represents perfect calibration. (b) Decision curve showing the net benefit of model-guided confirmatory testing compared with testing all patients and testing no patients across a range of threshold probabilities.

3.6. Learning Curve Results

Learning-curve analysis showed heterogeneous generalization behavior across algorithms. For XGBoost and MLP using the combined questionnaire and laboratory feature set, held-out ROC-AUC generally increased with increasing training-set size, while training performance remained higher than held-out performance across the evaluated range (Figure 5). A persistent training-validation gap therefore remained visible for both models at the largest training-set size. The complete learning curves showed algorithm-specific differences in generalization behavior, with Logistic Regression demonstrating comparatively close alignment between training and held-out performance and larger persistent gaps observed for several more flexible models (Supplementary Figures S3 and S4).

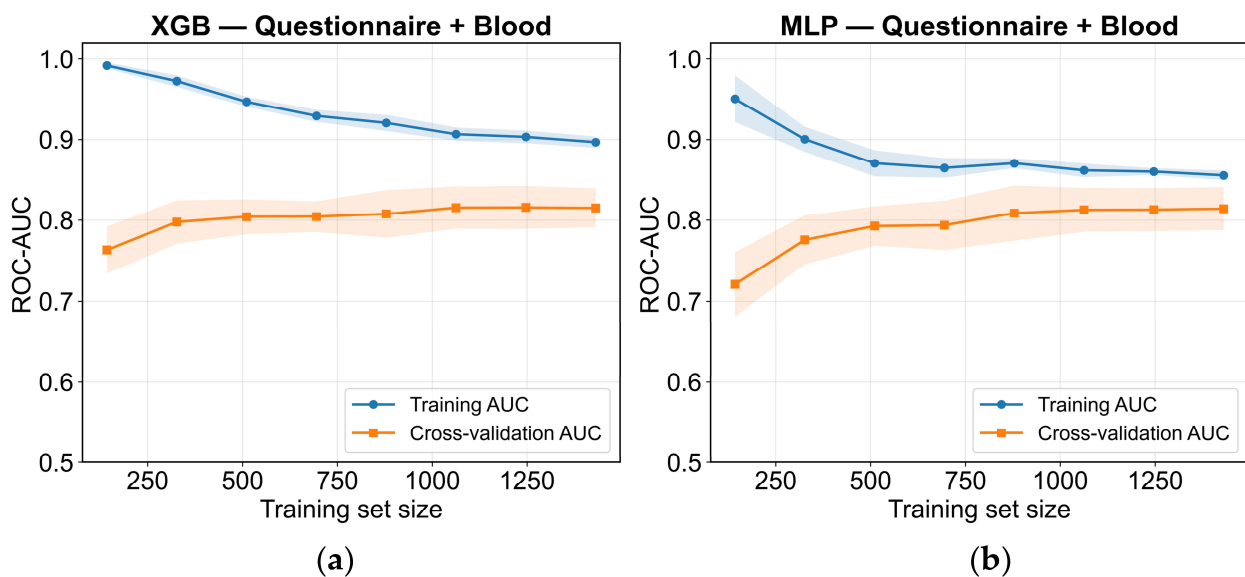


Figure 5. Learning curves for (a) XGBoost and (b) Multilayer Perceptron (MLP) using questionnaire + blood variables. Training ROC-AUC and held-out cross-validation ROC-AUC are shown across progressively increasing training-set sizes. Shaded areas indicate ± 1 standard deviation across outer folds.

3.7. Feature Importance and Reduced Model

The top ten predictors identified by SHAP (Figure 6a) and the top ten predictors identified by XGBoost feature gains (Figure 6b) were compared, resulting in six overlapping features retained for further evaluation: previous high blood sugar levels, age, waist circumference, MCHC, APTT, and QUICK value. The six-feature XGBoost model was subsequently compared with the full-feature XGBoost model using identical outer folds and independent inner-loop hyperparameter optimization. Across the five outer folds, the full-feature XGBoost model achieved a mean ROC-AUC of 0.815 ± 0.024 , while the reduced six-feature model achieved 0.819 ± 0.025 . The reduced model achieved a numerically higher ROC-AUC in four of the five outer folds, whereas the full model performed slightly better in one fold. These findings indicate comparable mean discrimination across the matched outer folds in this post hoc internal analysis. Corresponding per-fold confusion matrices for the reduced six-feature configuration are provided in Supplementary Figure S8.

3.8. Prototype Screening Interface

To illustrate the technical feasibility of implementation, a simplified screening prototype was developed in *Streamlit* using the six selected features (Figure 7). The interface enables rapid manual entry of routinely available admission variables and provides an immediate estimate of the probability of impaired glucose metabolism. This design reflects

a workflow-compatible triage approach that is rapid, low burden, and relies solely on information already collected during standard preoperative trauma assessment. Although not publicly deployed, this prototype illustrates a potential route for future workflow integration and requires prospective usability and clinical-impact evaluation.

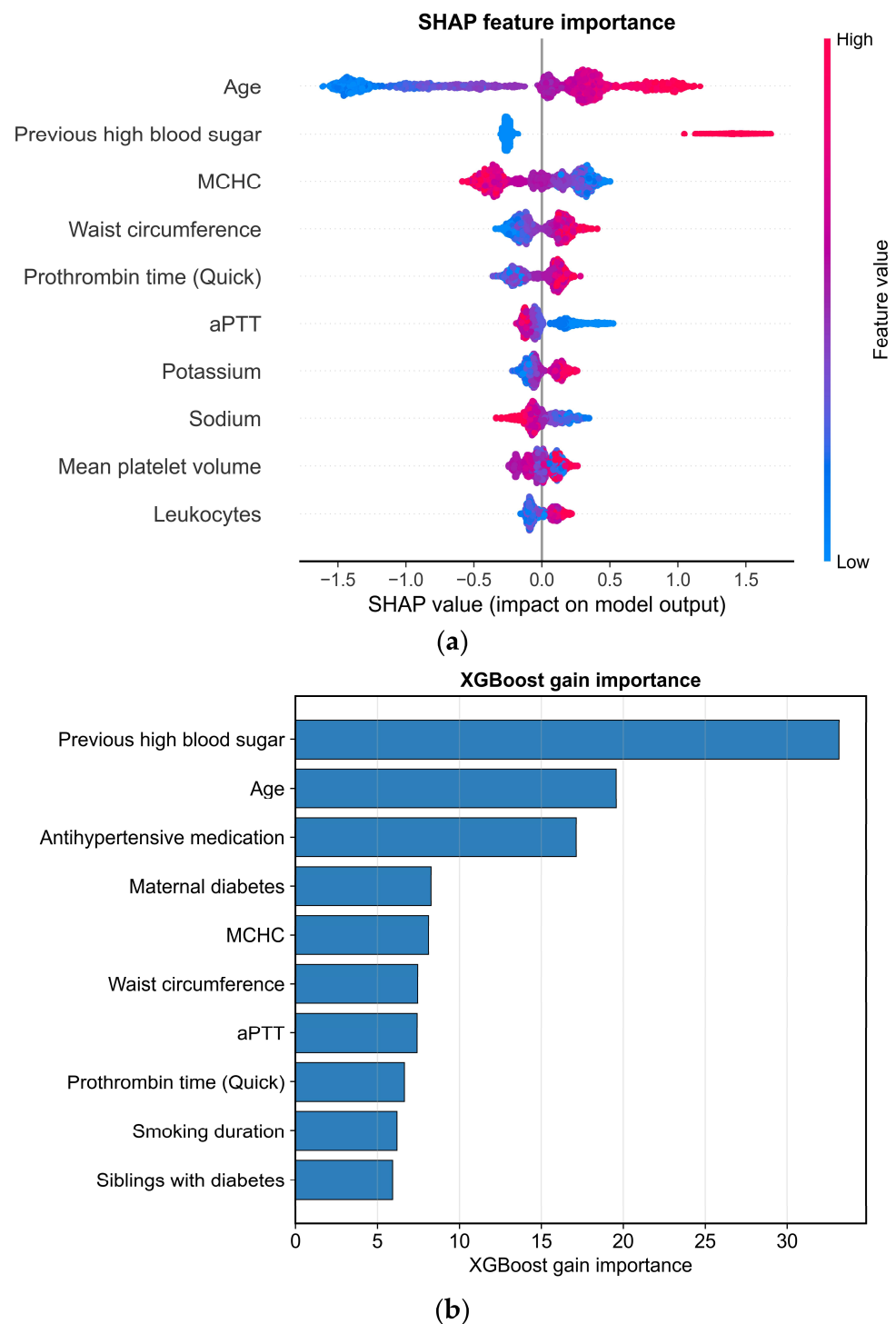


Figure 6. Model interpretability using Shapley Additive exPlanations (SHAP) and XGBoost feature importance. SHAP analysis of the best-performing XGBoost model trained on the combined questionnaire and blood features. **(a)** SHAP beeswarm plot showing the direction and magnitude of each feature's effect on the model's predictions; color represents the original feature value (blue = low, red = high). **(b)** XGBoost gain-based feature importance ranking highlighting features that most improved model performance during training.

AIDMT AI-based Diabetes Mellitus Prediction Tool for Trauma Clinics

About AIDMT

AIDMT is intended for use in **trauma clinics** as a decision-support tool to estimate the risk of **prediabetes and diabetes** based on routinely collected data. It does not provide a definitive diagnosis, and all results should be interpreted by qualified clinicians.

[Home](#) [Info](#)

Patient values

Age (years): 45

QUICK (%): 0.00

Previous high blood sugar levels? ☒ No ☐ Yes

APTT (s): 0.00

Waist circumference (cm): 0.00

MCHC (g/dL): 0.00

Calculate Risk

Figure 7. Prototype screening tool. *Streamlit*-based demonstration tool using the six selected predictors to provide a rapid estimated risk of impaired glucose metabolism at trauma admission. The interface allows manual entry of routine variables and displays a single predicted probability, illustrating the technical feasibility and potential workflow integration of a compact machine learning (ML) model.

4. Discussion

In this study, we systematically compared ten supervised machine learning algorithms for the early detection of prediabetes and diabetes in trauma surgery patients using data collected at hospital admission. Using a nested cross-validation framework with hyperparameter optimization, XGBoost achieved the numerically highest mean ROC-AUC (0.815 ± 0.024) on the combined questionnaire and laboratory feature set, followed closely by MLP (0.814 ± 0.026) and Gradient Boosting (0.813 ± 0.026). However, the fully interpretable Logistic Regression model remained within one percentage point (0.810 ± 0.026), and no pairwise statistical testing of between-model differences was performed. The addition of routine blood parameters significantly improved discrimination for nine of ten algorithms (DeLong's test, $p < 0.05$), yet the tight clustering of performance across all models indicates that the predictive ceiling in this setting is driven primarily by feature informativeness rather than model complexity. This finding carries direct practical implications: for deployment contexts where interpretability, computational simplicity, or regulatory transparency are prioritized, Logistic Regression represents a well-supported default choice rather than a compromise. Evaluation using additional metrics revealed complementary strengths across models: MLP achieved the highest F1-score, while KNN showed the highest specificity, confirming that no single algorithm excels uniformly across all evaluation criteria and reinforcing the importance of assessing multiple dimensions of predictive performance when selecting models for clinical deployment.

The clinical rationale for this work rests on a gap between disease prevalence and screening practice in trauma settings. HbA1c testing is individually inexpensive, yet it is

not part of the standard preoperative workup in trauma clinics, where clinical attention is directed toward surgical planning and acute injury management rather than metabolic screening. No established protocol mandates universal glycemic testing at trauma admission, and adding it to an already time-pressured workflow introduces logistical burden without existing evidence of cost-effectiveness in this population. Our study design required HbA1c for all participants to establish a reliable ground truth, but the proposed model is not intended to replicate universal testing. Instead, it uses information already collected during routine preoperative assessment, the anesthesia questionnaire, and the standard admission laboratory panel to identify patients who would benefit most from targeted confirmatory HbA1c testing. This positions the model as a pre-screening triage layer consistent with American Diabetes Association recommendations [14], rather than an alternative to biochemical diagnosis. To benchmark this approach against established screening instruments, we reconstructed FINDRISK and GDRS scores from the corresponding variables available in the study questionnaire. FINDRISK achieved a ROC-AUC of 0.749 (95% CI, 0.725–0.772) and GDRS achieved 0.756 (95% CI, 0.733–0.778), both numerically lower than the best questionnaire-only machine-learning model (Logistic Regression, 0.797 ± 0.024). This numerical difference suggests that multivariable modeling of the available questionnaire information may capture additional discriminatory information beyond conventional point-based scoring in this trauma population, although this comparison should be regarded as a clinical benchmark rather than a formal external validation, since the questionnaire was adapted from these instruments rather than administered as the original stand-alone tools. In the proposed workflow, risk estimation could be performed automatically from preoperative assessment data, with higher-risk patients prioritized for HbA1c testing and early consultation with internal medicine [40]. Patients with confirmed diabetes would benefit from intensified perioperative glucose monitoring and co-management by endocrinology, in alignment with current perioperative guidance recommending blood glucose maintenance within 80–180 mg/dL [41,42]. Patients identified as prediabetic, who might otherwise go undetected during admission, present a particularly valuable detection opportunity, as structured lifestyle interventions can reduce progression to type 2 diabetes by up to 58% [43,44]. The prototype screening interface developed in *Streamlit* is a technical feasibility demonstration only: it outputs a continuous predicted probability and does not embed a fixed decision threshold or issue an automated testing recommendation. Any future clinical deployment would require prospective threshold calibration within the implementing institution's workflow.

A critical methodological concern is the potential for algorithmic reliance on established clinical histories to inflate performance metrics. "Previous high blood sugar levels" ranked among the strongest predictors, and its relationship to the outcome definition warranted careful scrutiny. To address this, we performed a diagnosis-naïve sensitivity analysis excluding patients with previously documented diabetes and removing the self-reported glycemic history variable from the predictor set. Within this stricter screening scenario, XGBoost trained on the combined feature set achieved a mean ROC-AUC of 0.764 ± 0.012 , while Logistic Regression trained on the questionnaire only achieved 0.734 ± 0.023 . Although discrimination was attenuated relative to the complete cohort, these results indicate that the model's predictive capacity was not solely attributable to prior diagnostic information and instead retained information from biochemical and anthropometric patterns, supporting its potential utility for detecting previously unrecognized dysglycemia at admission.

The most influential predictors identified in our study are well supported by biological and epidemiological evidence. Age and waist circumference are established risk factors for impaired glucose metabolism [45–49]. MCHC, a marker not typically emphasized in dia-

betes screening, has been independently associated with diabetes mellitus in hematological profiling studies [50,51]. Notably, APTT and prothrombin time (QUICK), which are routine coagulation parameters collected as part of the standard preoperative panel, emerged among the top-ranked features. Diabetes induces quantitative increases in intrinsic and extrinsic coagulation factors, shortening APTT and enhancing tissue-factor-driven activity, resulting in a prothrombotic shift that is particularly evident in surgical and perioperative populations [52,53]. These coagulation markers represent a trauma-specific advantage: they are already measured for surgical planning purposes and provide metabolic screening information at no additional cost or effort. Antihypertensive medication use ranked among the strongest predictors by gain-based importance but was not retained in the consensus feature set, as it did not appear among the top ten by SHAP. Its exclusion reflects the consensus selection strategy, which prioritized features ranked highly by both metrics, even though the higher prevalence of antihypertensive treatment in the Pre-/DM group (52.9% vs. 25.1%; $p < 0.001$) is consistent with the well-established clustering of dysglycemia and hypertension within the metabolic syndrome.

The six-feature consensus model, retaining previous high blood sugar levels, age, waist circumference, MCHC, APTT, and QUICK, achieved a mean ROC-AUC of 0.819 ± 0.025 compared to 0.815 ± 0.024 for the full 47-feature model, with the reduced model achieving numerically higher discrimination in four of five outer folds. However, because the consensus features were identified from SHAP and gain rankings of a model fitted to the complete dataset before this comparison, this analysis is *post hoc* in nature, and its performance estimate should be regarded as optimistic pending external validation. Beyond discrimination, model calibration and clinical utility were evaluated to determine whether the predicted probabilities are reliable enough to guide triage decisions. Calibration analysis of XGBoost yielded an expected calibration error (ECE) of 0.011 and a Brier score of 0.168, both of which were the lowest among the evaluated algorithms (Supplementary Table S5). An ECE of 0.011 is well below the 0.05 threshold generally considered adequate for clinical decision support, indicating that predicted probabilities closely match observed event rates: in practical terms, a patient receiving a predicted risk of 0.6 can be expected to have approximately a 60% probability of impaired glucose metabolism [54]. Notably, calibration differentiated the models more meaningfully than discrimination; while ROC-AUC ranged narrowly from 0.775 to 0.815, ECE ranged from 0.011 (XGBoost) to 0.158 (AdaBoost), indicating that comparable discrimination does not guarantee comparable probability reliability. Logistic Regression also demonstrated strong calibration (ECE 0.026), further supporting its viability as a deployment candidate. Decision curve analysis demonstrated that model-guided selective testing provided higher net benefit than both testing all patients and testing no patients across a broad range of threshold probabilities. In the screening context of this study, where a positive model result prompts a low-cost, low-harm confirmatory HbA1c test, the clinically relevant threshold range is approximately 0.10 to 0.30. Across this range, XGBoost maintained consistent net benefit (0.332 at threshold 0.10; 0.278 at 0.20; 0.222 at 0.30), as did Logistic Regression (0.329; 0.273; 0.228). The finding that model-guided screening outperformed the ‘test all’ strategy is particularly relevant, as it demonstrates from a decision-theoretic standpoint that targeted screening provides greater clinical value than universal HbA1c testing, directly addressing the question of whether routine testing alone would be a simpler alternative. These findings, comparable discrimination with a compact feature set, well-calibrated probability outputs, and demonstrable net clinical benefit over default strategies, collectively support the practical feasibility of integrating such a model into time-sensitive trauma admission workflows.

To our knowledge, this is the first study to benchmark ten supervised machine learning algorithms for glycemic risk prediction in a trauma surgery population, combining nested

cross-validation, consensus feature selection, and decision-curve-based clinical utility assessment within a single, consistently applied framework. A recent comprehensive review of over 2300 studies on ML-based diabetes prediction confirmed that the field remains dominated by general-population screening applications, with trauma and surgical populations largely underrepresented [55]. Previous diabetes prediction models have primarily focused on general population screening or chronic disease monitoring [56–64], with reported ROC-AUC values ranging from 0.67 to 0.87 depending on dataset composition [57,59,62,65]. This performance range is also consistent with broader ML-benchmarking evidence outside the trauma setting, where more complex algorithms have often failed to outperform logistic regression once an adequately informative feature set is available [62]. Consistent with these findings, recent work by Idris et al. demonstrated that an interpretable XGBoost model trained on routine clinical variables can achieve effective risk stratification in primary screening workflows, underscoring the clinical feasibility of lightweight feature sets [66]. Similarly, Yoo et al. developed and externally validated an ML-based self-assessment system for prediabetes and diabetes using seven non-invasive parameters, demonstrating that compact, interpretable models can outperform conventional screening guidelines across diverse populations [67]. Although comparable ML approaches have recently been applied to predict postoperative hyperglycemia in non-diabetic cardiac surgery patients [68], glycemic risk screening at trauma admission remains largely unexplored. This represents a clinically relevant gap, given the established adverse effects of diabetes on bone and wound healing. In the musculoskeletal domain more broadly, interpretable machine learning approaches combining compact feature sets with model-agnostic explanations have been applied for automated diagnostic tasks, illustrating the broader utility of transparent ML pipelines in orthopedic and trauma-adjacent applications [68].

Several methodological features of the present study strengthen the reliability of the reported findings: prospective data collection with confirmatory HbA1c for all participants provided a robust ground truth independent of clinical suspicion; the nested cross-validation framework separated hyperparameter tuning from performance evaluation; learning curves for all ten algorithms demonstrated convergence behavior and characterized the training-validation gap for each model class; per-fold confusion matrices increased transparency regarding classification performance, consensus feature selection using two independent importance metrics reduced the risk of method-specific ranking artifacts; model evaluation encompassed discrimination, calibration, and decision-curve analysis rather than relying on AUC alone; the diagnosis-naïve sensitivity analysis supported the model's capacity for opportunistic screening beyond recapitulation of known diagnoses; and the analysis code has been made publicly available to support reproducibility.

Despite these strengths, several limitations warrant consideration. The study was conducted at a single tertiary trauma center in Germany, and the machine learning analyses were performed on the completed dataset with internal validation only. Differences in patient demographics, referral patterns, laboratory workflows, questionnaire completion, and prevalence of dysglycemia may affect model performance at other institutions. The study population was recruited through a preoperative anesthesia assessment pathway during three discrete six-week enrollment periods and did not include acute polytrauma patients managed through the emergency pathway; although consecutive enrollment within each period limits selection bias, the temporal sampling may not capture seasonal or secular trends. The available dataset lacked structured measures of injury type and trauma severity. The available sample size may have limited the generalizability of the models, as indicated by the learning-curve analysis. Logistic Regression demonstrated the closest alignment between training and validation performance, whereas more flexible algorithms, particularly tree-based ensemble methods and KNN, exhibited larger generalization gaps.

Importantly, validation performance remained stable or continued to improve at the largest training-set sizes rather than declining, suggesting that the observed predictive performance reflected genuine signal rather than progressive overfitting. However, the absence of a clear performance plateau for several models indicates that their generalization capacity may not have been fully realized within the present cohort. These findings further underscore the importance of validation in larger, independent, multicenter populations. Within this context, Logistic Regression's combination of the tightest convergence behavior, near-equivalent discrimination (ROC-AUC 0.810), and strong calibration (ECE 0.026) provides the strongest overall profile for further prospective clinical evaluation among the evaluated algorithms. The present work reports global SHAP-based feature attributions; per-patient local explanations, which could enhance clinician trust in individual predictions, represent a natural extension for a deployed tool. Prospective external validation across multiple trauma centers and healthcare settings is required before the model, or the reduced feature configuration, can be considered for routine implementation. Future research should prioritize multicenter validation, integration into real-time hospital information systems, and evaluation of whether model-guided screening changes clinician testing behavior and patient outcomes.

5. Conclusions

In summary, machine learning models trained on routinely collected admission data can detect glycemic risk in trauma surgery patients with moderate-to-good discrimination, strong calibration, and consistent net benefit over default screening strategies. XGBoost achieved the numerically highest mean ROC-AUC, but differences among the leading algorithms were small relative to cross-validation variability, and the fully interpretable Logistic Regression model performed within one percentage point while demonstrating the most stable convergence behavior, supporting its prioritization for prospective clinical evaluation in settings where transparency is particularly important. A compact six-feature model retained comparable discrimination, suggesting that effective screening does not require extensive variable collection beyond information already available at preoperative assessment. However, as this study was conducted at a single tertiary trauma center with internal validation only, these findings should be regarded as a hypothesis-generating proof of concept. Prospective multicenter external validation across diverse trauma populations and subsequent evaluation of the clinical impact on testing behavior and patient outcomes are required before ML-based glycemic screening can be recommended for routine trauma care.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/make8090254/s1>. Table S1: Details of features; Table S2: Variance inflation factor (VIF) values for questionnaire + blood features retained for model training; Table S3: Hyperparameter search spaces; Table S4: Globally optimized hyperparameters of the final XGBoost model; Table S5: Calibration and Decision Curve Analysis metrics across all evaluated models; Figure S1: Discriminative performance of FINDRISK and GDRS; Figure S2: Diagnosis-naïve sensitivity analysis; Figure S3: Learning curves using questionnaire-only variables; Figure S4: Learning curves using combined features; Figure S5: Per-fold confusion matrices for full-feature models; Figure S6: Calibration curves for all ten algorithms; Figure S7: Decision curve analysis for all ten algorithms; Figure S8: Per-fold confusion matrices for reduced six-feature models.

Author Contributions: Conceptualization, M.T., A.K.N., N.P. and S.E.; methodology, M.T., N.P. and S.E.; software, M.T.; validation, M.T. and N.P.; formal analysis, M.T.; investigation, M.T. and S.E.; resources, A.K.N., T.H. and S.E.; data curation, M.T., N.R.B. and S.E.; writing—original draft preparation, M.T.; writing—review and editing, M.T., A.K.N., E.T., N.R.B., A.F., T.H., N.P. and

S.E.; visualization, M.T.; supervision, A.K.N., N.P. and S.E.; project administration, S.E.; funding acquisition, A.F. and S.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the German Research Council, grant number EH471/5-1, and by the Bundesministerium für Bildung und Forschung, grant number 13GW0661D. We acknowledge support from the Open Access Publication Fund of the University of Tübingen.

Informed Consent Statement: The study was approved by the ethics committee of the University of Tübingen on 20 September 2021, with project number 454/2021BO2. Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets generated in the current study are not publicly available due to participant privacy and ethical concerns. The analysis code required to reproduce the machine learning workflow is publicly available at <https://github.com/meltombaz/trauma-glycemia-ml> (accessed on 19 August 2026).

Acknowledgments: The authors acknowledge Philomena Huber, Emely Dobesch, Sara T. Feistle, Viktoria Werner, and Adriana Converso for their assistance in collecting blood samples and administering patient questionnaires. Their contributions to data acquisition were instrumental in completing this study. During the preparation of this manuscript, the authors used Grammarly (version 1.172.1.0, Grammarly Inc., New York, NY, USA) for grammar and language checking. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

AB	AdaBoost
ADA	American Diabetes Association
AI	Artificial Intelligence
APTT	Activated Partial Thromboplastin Time
BMI	Body Mass Index
DCA	Decision Curve Analysis
DM	Diabetes Mellitus
DT	Decision Tree
ECE	Expected Calibration Error
ET	Extra-Trees
GB	Gradient Boosting
HbA1c	Glycated Hemoglobin
IQR	Interquartile Range
KNN	K-Nearest Neighbors
LR	Logistic Regression
MCHC	Mean Corpuscular Hemoglobin Concentration
ML	Machine Learning
MLP	Multilayer Perceptron
Nested CV	Nested Cross-Validation
RF	Random Forest
ROC-AUC	Receiver Operating Characteristic–Area Under the Curve
SD	Standard Deviation
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
VIF	Variance Inflation Factor
XGB	XGBoost, eXtreme Gradient Boosting

References

1. Buysschaert, M.; Medina, J.L.; Buysschaert, B.; Bergman, M. Definitions (and Current Controversies) of Diabetes and Prediabetes. *Curr. Diabetes Rev.* **2016**, *12*, 8–13. [\[CrossRef\]](#) [\[PubMed\]](#)
2. International Diabetes Federation. *IDF Diabetes Atlas 11th Edition*; International Diabetes Federation: Brussels, Belgium, 2025.
3. Tabák, A.G.; Herder, C.; Rathmann, W.; Brunner, E.J.; Kivimäki, M. Prediabetes: A high-risk state for diabetes development. *Lancet* **2012**, *379*, 2279–2290. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Magliano, D.J.; Boyko, E.J.; IDF Diabetes Atlas 10th Edition Scientific Committee. *IDF Diabetes Atlas*; International Diabetes Federation: Brussels, Belgium, 2021.
5. Hauguel-Moreau, M.; Hergault, H.; Cazabat, L.; Pépin, M.; Beauchet, A.; Aïdan, V.; Ouadahi, M.; Josseran, L.; Hage, M.; Rodon, C.; et al. Prevalence of prediabetes and undiagnosed diabetes in a large urban middle-aged population: The CARVAR 92 cohort. *Cardiovasc. Diabetol.* **2023**, *22*, 31. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Wintermeyer, E.; Ihle, C.; Ehnert, S.; Schreiner, A.J.; Stollhof, L.; Stöckle, U.; Nussler, A.; Fritsche, A.; Pscherer, S. Assessment of the Influence of Diabetes mellitus and Malnutrition on the Postoperative Complication Rate and Quality of Life of Patients in a Clinic Focused on Trauma Surgery. *Z. Orthop. Unfall.* **2019**, *157*, 173–182. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Pscherer, S.; Nüssler, A.; Bahrs, C.; Reumann, M.; Ihle, C.; Stöckle, U.; Ehnert, S.; Freude, T.; Ochs, B.G.; Flesch, I.; et al. Retrospective Analysis of Diabetics with Regard to Treatment Duration and Costs. *Z. Orthop. Unfall.* **2017**, *155*, 72–76. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Maurer, E.; Wallmeier, V.; Reumann, M.; Ehnert, S.; Ihle, C.; Schreiner, A.J.; Flesch, I.; Stollhof, L.E.; Histing, T.; Nüssler, A.K. Increased Age, Cardiovascular Comorbidities, COPD, and Diabetes mellitus Determine Excess Mortality in Trauma Patients with SSI or Chronic Wounds. *Z. Orthop. Unfall.* **2023**, *161*, 412–421. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Tercan, E.; Tombaz, M.; Huber, P.; Rollmann, M.F.; Nussler, A.K.; Ehnert, S. The Financial Burden of Prediabetes and Diabetes in Trauma Care Under the G-DRG System. *Z. Orthop. Unfall.* **2026**, *Online ahead of print*. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Stryker, L.S.; Abdel, M.P.; Morrey, M.E.; Morrow, M.M.; Kor, D.J.; Morrey, B.F. Elevated postoperative blood glucose and preoperative hemoglobin A1C are associated with increased wound complications following total joint arthroplasty. *J. Bone Jt. Surg. Am.* **2013**, *95*, 808–814. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Chrastil, J.; Anderson, M.B.; Stevens, V.; Anand, R.; Peters, C.L.; Pelt, C.E. Is Hemoglobin A1c or Perioperative Hyperglycemia Predictive of Periprosthetic Joint Infection or Death Following Primary Total Joint Arthroplasty? *J. Arthroplast.* **2015**, *30*, 1197–1202. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Cancienne, J.M.; Cooper, M.T.; Laroche, K.A.; Verheul, D.W.; Werner, B.C. Hemoglobin A(1c) as a Predictor of Postoperative Infection Following Elective Forefoot Surgery. *Foot Ankle Int.* **2017**, *38*, 832–837. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Schlesinger, S.; Neuenschwander, M.; Barbaresko, J.; Lang, A.; Maalmi, H.; Rathmann, W.; Roden, M.; Herder, C. Prediabetes and risk of mortality, diabetes-related complications and comorbidities: Umbrella review of meta-analyses of prospective studies. *Diabetologia* **2022**, *65*, 275–285. [\[CrossRef\]](#) [\[PubMed\]](#)
14. American Diabetes Association Professional Practice Committee. 2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes-2024. *Diabetes Care* **2024**, *47*, S20–S42. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Zhang, J.; Zhang, Z.; Zhang, K.; Ge, X.; Sun, R.; Zhai, X. Early detection of type 2 diabetes risk: Limitations of current diagnostic criteria. *Front. Endocrinol.* **2023**, *14*, 1260623. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Eloranta, S.; Boman, M. Predictive models for clinical decision making: Deep dives in practical machine learning. *J. Intern. Med.* **2022**, *292*, 278–295. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Jahandideh, S.; Ozavci, G.; Sahle, B.W.; Kouzani, A.Z.; Magrabi, F.; Bucknall, T. Evaluation of machine learning-based models for prediction of clinical deterioration: A systematic literature review. *Int. J. Med. Inform.* **2023**, *175*, 105084. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Choi, A.; Lee, K.; Hyun, H.; Kim, K.J.; Ahn, B.; Lee, K.H.; Hahn, S.; Choi, S.Y.; Kim, J.H. A novel deep learning algorithm for real-time prediction of clinical deterioration in the emergency department for a multimodal clinical decision support system. *Sci. Rep.* **2024**, *14*, 30116. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Khalifa, M.; Albadawy, M. Artificial Intelligence for Clinical Prediction: Exploring Key Domains and Essential Functions. *Comput. Methods Programs Biomed. Update* **2024**, *5*, 100148. [\[CrossRef\]](#)
20. Stahl, D. New horizons in prediction modelling using machine learning in older people's healthcare research. *Age Ageing* **2024**, *53*, afae201. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Liu, N.T.; Salinas, J. Machine Learning for Predicting Outcomes in Trauma. *Shock* **2017**, *48*, 504–510. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Rau, C.S.; Kuo, P.J.; Chien, P.C.; Huang, C.Y.; Hsieh, H.Y.; Hsieh, C.H. Mortality prediction in patients with isolated moderate and severe traumatic brain injury using machine learning models. *PLoS ONE* **2018**, *13*, e0207192. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Cardosi, J.D.; Shen, H.; Groner, J.I.; Armstrong, M.; Xiang, H. Machine learning for outcome predictions of patients with trauma during emergency department care. *BMJ Health Care Inform.* **2021**, *28*, e100407. [\[CrossRef\]](#) [\[PubMed\]](#)

24. Ogink, P.T.; Groot, O.Q.; Karhade, A.V.; Bongers, M.E.R.; Oner, F.C.; Verlaan, J.J.; Schwab, J.H. Wide range of applications for machine-learning prediction models in orthopedic surgical outcome: A systematic review. *Acta Orthop.* **2021**, *92*, 526–531. [[CrossRef](#)] [[PubMed](#)]
25. Bergmann, A.; Li, J.; Wang, L.; Schulze, J.; Bornstein, S.R.; Schwarz, P.E. A simplified Finnish diabetes risk score to predict type 2 diabetes risk and disease evolution in a German population. *Horm. Metab. Res.* **2007**, *39*, 677–682. [[CrossRef](#)] [[PubMed](#)]
26. Schiborn, C.; Paprott, R.; Heidemann, C.; Kühn, T.; Fritsche, A.; Kaaks, R.; Schulze, M.B. German Diabetes Risk Score for the Determination of the Individual Type 2 Diabetes Risk. *Dtsch. Arztebl. Int.* **2022**, *119*, 651–657. [[CrossRef](#)] [[PubMed](#)]
27. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
28. Mucherino, A.; Papajorgji, P.J.; Pardalos, P.M. k-Nearest Neighbor Classification. In *Data Mining in Agriculture*; Mucherino, A., Papajorgji, P.J., Pardalos, P.M., Eds.; Springer New York: New York, NY, USA, 2009; pp. 83–106.
29. Costa, L.; Guerreiro, M.; Puchta, E.; de Souza Tadano, Y.; Alves, T.A.; Kaster, M.; Siqueira, H.V. Multilayer perceptron. In *Introduction to Computational Intelligence*; IEEE: New York, NY, USA, 2023; p. 105.
30. Popescu, M.-C.; Balas, V.E.; Perescu-Popescu, L.; Mastorakis, N. Multilayer perceptron and neural networks. *WSEAS Trans. Cir. Sys.* **2009**, *8*, 579–588.
31. Song, Y.Y.; Lu, Y. Decision tree methods: Applications for classification and prediction. *Shanghai Arch. Psychiatry* **2015**, *27*, 130–135. [[CrossRef](#)] [[PubMed](#)]
32. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
33. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [[CrossRef](#)]
34. Wang, R. AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review. *Phys. Procedia* **2012**, *25*, 800–807. [[CrossRef](#)]
35. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
36. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
37. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.
38. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 2623–2631.
39. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc. (NeurIPS): South Lake Tahoe, NV, USA, 2017; Volume 30.
40. Mattina, A.; Raffa, G.M.; Giusti, M.A.; Conoscenti, E.; Morsolini, M.; Mularoni, A.; Fazzina, M.L.; Di Carlo, D.; Cipriani, M.; Musumeci, F.; et al. Impact of systematic diabetes screening on peri-operative infections in patients undergoing cardiac surgery. *Sci. Rep.* **2024**, *14*, 14182. [[CrossRef](#)] [[PubMed](#)]
41. Arokiyanathan, S.; Baalamurugan, H.; Ip, I.; Marcus, L.; Mafullul, D.; Ranchagoda, J. Perioperative Glycaemic Control and Orthopaedic Outcomes: A Narrative Review of Evidence and Future Directions. *Cureus* **2025**, *17*, e96905. [[CrossRef](#)] [[PubMed](#)]
42. Akiboye, F.; Rayman, G. Management of Hyperglycemia and Diabetes in Orthopedic Surgery. *Curr. Diab Rep.* **2017**, *17*, 13. [[CrossRef](#)] [[PubMed](#)]
43. Knowler, W.C.; Barrett-Connor, E.; Fowler, S.E.; Hamman, R.F.; Lachin, J.M.; Walker, E.A.; Nathan, D.M. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N. Engl. J. Med.* **2002**, *346*, 393–403. [[CrossRef](#)] [[PubMed](#)]
44. Lee, C.G.; Heckman-Stoddard, B.; Dabelea, D.; Gadde, K.M.; Ehrmann, D.; Ford, L.; Prorok, P.; Boyko, E.J.; Pi-Sunyer, X.; Wallia, A.; et al. Effect of Metformin and Lifestyle Interventions on Mortality in the Diabetes Prevention Program and Diabetes Prevention Program Outcomes Study. *Diabetes Care* **2021**, *44*, 2775–2782. [[CrossRef](#)] [[PubMed](#)]
45. Bowen, M.E.; Xuan, L.; Lingvay, I.; Halm, E.A. Random Blood Glucose: A Robust Risk Factor For Type 2 Diabetes. *J. Clin. Endocrinol. Metab.* **2015**, *100*, 1503–1510. [[CrossRef](#)] [[PubMed](#)]
46. CDC—Diabetes Risk Factors. Available online: https://www.cdc.gov/diabetes/risk-factors/index.html#cdc_risk_factors_who-prediabetes-and-type-2-diabetes (accessed on 17 July 2025).
47. Feller, S.; Boeing, H.; Pischon, T. Body mass index, waist circumference, and the risk of type 2 diabetes mellitus: Implications for routine clinical practice. *Dtsch. Arztebl. Int.* **2010**, *107*, 470–476. [[CrossRef](#)] [[PubMed](#)]
48. Kato, K.; Otsuka, T.; Saiki, Y.; Kobayashi, N.; Nakamura, T.; Kon, Y.; Kawada, T. Association Between Elevated C-Reactive Protein Levels and Prediabetes in Adults, Particularly Impaired Glucose Tolerance. *Can. J. Diabetes* **2019**, *43*, 40–45.e2. [[CrossRef](#)] [[PubMed](#)]
49. Farooqui, R.; Afsar, N.; Afroze, I.A. Role and Significance of Hematological parameters in Diabetes Mellitus. *Ann. Pathol. Lab. Med.* **2019**, *6*, A158–A162. [[CrossRef](#)]
50. Umeji, L.; Paul, A.; Felix, S.; Umeji, C.; Folake, A.; Chrisitian, O.; Eigege, N.; Okolo, C. Haematological Profile of Diabetes and Non-Diabetes Patients in Abuja, Nigeria. *IJRSI* **2019**, *VI*, 274–277.

51. Getu, F.; Aynalem, M.; Bizuneh, S.; Enawgaw, B. The Prevalence of Coagulopathy and Associated Factors Among Adult Type II Diabetes Mellitus Patients Attending the University of Gondar Comprehensive Specialized Hospital, Northwest Ethiopia. *Diabetes Metab. Syndr. Obes.* **2022**, *15*, 579–590. [[CrossRef](#)] [[PubMed](#)]
52. Li, X.; Weber, N.C.; Cohn, D.M.; Hollmann, M.W.; DeVries, J.H.; Hermanides, J.; Preckel, B. Effects of Hyperglycemia and Diabetes Mellitus on Coagulation and Hemostasis. *J. Clin. Med.* **2021**, *10*, 2419. [[CrossRef](#)] [[PubMed](#)]
53. Huang, Y.; Li, W.; Macheret, F.; Gabriel, R.A.; Ohno-Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 621–633. [[CrossRef](#)] [[PubMed](#)]
54. Kiran, M.; Xie, Y.; Anjum, N.; Ball, G.; Pierscionek, B.; Russell, D. Machine learning and artificial intelligence in type 2 diabetes prediction: A comprehensive 33-year bibliometric and literature analysis. *Front. Digit. Health* **2025**, *7*, 1557467. [[CrossRef](#)] [[PubMed](#)]
55. Garcia-Carretero, R.; Vigil-Medina, L.; Barquero-Perez, O. The Use of Machine Learning Techniques to Determine the Predictive Value of Inflammatory Biomarkers in the Development of Type 2 Diabetes Mellitus. *Metab. Syndr. Relat. Disord.* **2021**, *19*, 240–248. [[CrossRef](#)] [[PubMed](#)]
56. Alghamdi, M.; Al-Mallah, M.; Keteyian, S.; Brawner, C.; Ehrman, J.; Sakr, S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project. *PLoS ONE* **2017**, *12*, e0179805. [[CrossRef](#)] [[PubMed](#)]
57. Farran, B.; AlWotayan, R.; Alkandari, H.; Al-Abdulrazzaq, D.; Channanath, A.; Thanaraj, T.A. Use of Non-invasive Parameters and Machine-Learning Algorithms for Predicting Future Risk of Type 2 Diabetes: A Retrospective Cohort Study of Health Data From Kuwait. *Front. Endocrinol.* **2019**, *10*, 624. [[CrossRef](#)] [[PubMed](#)]
58. Casanova, R.; Saldana, S.; Simpson, S.L.; Lacy, M.E.; Subauste, A.R.; Blackshear, C.; Wagenknecht, L.; Bertoni, A.G. Prediction of Incident Diabetes in the Jackson Heart Study Using High-Dimensional Machine Learning. *PLoS ONE* **2016**, *11*, e0163942. [[CrossRef](#)] [[PubMed](#)]
59. Islam, M.S.; Qaraqe, M.K.; Belhaouari, S.B.; Abdul-Ghani, M.A. Advanced Techniques for Predicting the Future Progression of Type 2 Diabetes. *IEEE Access* **2020**, *8*, 120537–120547. [[CrossRef](#)]
60. Cho, A.R.; Lee, S.B.; Hong, K.-W.; Jung, D.H. C-reactive protein-to-albumin ratio and 8-year incidence of type 2 diabetes: The Korean genome and epidemiology study. *Acta Diabetol.* **2021**, *58*, 1525–1532. [[CrossRef](#)] [[PubMed](#)]
61. Nusinovi, S.; Tham, Y.C.; Chak Yan, M.Y.; Wei Ting, D.S.; Li, J.; Sabanayagam, C.; Wong, T.Y.; Cheng, C.-Y. Logistic regression was as good as machine learning for predicting major chronic diseases. *J. Clin. Epidemiol.* **2020**, *122*, 56–69. [[CrossRef](#)] [[PubMed](#)]
62. Razavian, N.; Blecker, S.; Schmidt, A.M.; Smith-McLallen, A.; Nigam, S.; Sontag, D. Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors. *Big Data* **2015**, *3*, 277–287. [[CrossRef](#)] [[PubMed](#)]
63. Deberneh, H.M.; Kim, I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. *Int. J. Environ. Res. Public Health* **2021**, *18*, 3317. [[CrossRef](#)] [[PubMed](#)]
64. Liu, Q.; Zhou, Q.; He, Y.; Zou, J.; Guo, Y.; Yan, Y. Predicting the 2-Year Risk of Progression from Prediabetes to Diabetes Using Machine Learning among Chinese Elderly Adults. *J. Pers. Med.* **2022**, *12*, 1055. [[CrossRef](#)] [[PubMed](#)]
65. Idris, N.A.; Yusof, N.A.; Ismail, M.Z.H.; Juhari, S.N.; Hassan, N.M.; Daud, N.; Yunus, N.I.; Pauzi, M.F. Predicting postpartum glucose intolerance in women with gestational diabetes mellitus in primary care: A machine learning approach using XGBoost and SHAP values. *Diabetes Res. Clin. Pract.* **2026**, *232*, 113098. [[CrossRef](#)] [[PubMed](#)]
66. Yoo, D.; Maggiore, U.; Joliet, O. Enhancing prediabetes and diabetes detection through a machine learning-enabled self-assessment approach. *J. Clin. Epidemiol.* **2026**, *195*, 112266. [[CrossRef](#)] [[PubMed](#)]
67. Wu, J.; Zhang, M.; Cui, S.; Yang, G.; Wang, L.; Duan, H.; Xue, F. Comparison of multiple machine learning methods for predicting postoperative hyperglycemia in patients without diabetes undergoing cardiac surgery. *Front. Cardiovasc. Med.* **2025**, *12*, 1699809. [[CrossRef](#)] [[PubMed](#)]
68. Devnath, H.; Devnath, L.; Hossain, M.M. Explainable artificial intelligence for automated flatfoot detection in foot x-ray images. *Discov. Artif. Intell.* **2026**, *6*, 782. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.