Dynamic Two-Stage Early Warning System for Sepsis-Induced Coagulopathy and Progression to Disseminated Intravascular Coagulation: A MIMIC-IV Study

Amisha Kelkar · Chaitali Deshmukh · Pratik Mahajan · Rinaldo Brendon Patel

Abstract

Background

Sepsis leading to Sepsis-induced coagulopathy (SIC) is an early manifestation of dysregulated host response and a recognized precursor to disseminated intravascular coagulation (DIC), a life-threatening critical condition associated with organ failure and high mortality. These disorders evolve dynamically, often before clinical signs appear, and timely recognition is essential to guide the intensity of monitoring, and early therapeutic interventions. Current diagnostic tools are based on static laboratory thresholds and are usually applied only well after coagulopathy has been established. Clinicians thus do not have real-time support to identify which patients will likely develop **SIC** or progress from **SIC** to **DIC**, potentially delaying intervention during a narrow window when outcomes might be most modifiable. Currently, there are no approaches that integrate both stages of coagulopathy as a unified warning system meeting diagnostic criteria. A strategy that could more promptly detect **SIC** and stratify the risk of subsequent progression toward **DIC** may allow for more timely management and possibly improve patient outcomes.

Objectives

To develop and validate a dynamic two-stage early-warning system that (1) predicts **SIC** onset 24 hours before diagnostic criteria are met and (2) identifies among patients with **SIC** those at risk of progression to life-threatening **DIC,** using high-granularity clinical data from MIMIC-IV.

Methods

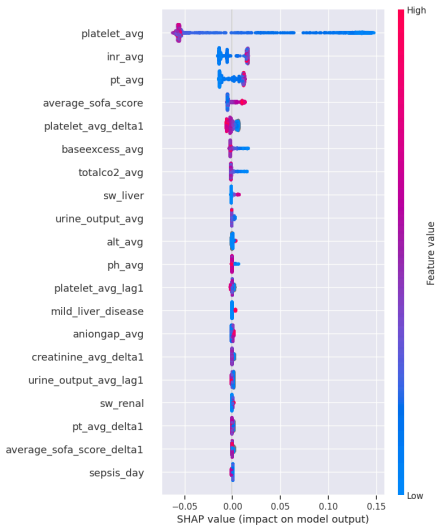
We constructed a retrospective cohort from MIMIC-IV, integrating demographics, comorbidities, vital signs, laboratory results, organ dysfunction markers, and treatment variables. SIC labels were derived from validated diagnostic criteria in correlation to the International Society on Thrombosis and Haemostasis (**ISTH**), with SIC_today and SIC_tomorrow (24-hour horizon) generated dynamically across ICU stays. **DIC** labeling used the Japanese Association for Acute Medicine **(JAAM)** DIC definitions and applied to an "at-risk" subset of patients with **SIC**.

For Stage 1 (**SIC** prediction), a CatBoost classifier was trained using stay-level splits to avoid temporal and patient leakage. For Stage 2, logistic regression and CatBoost models were developed to capture linear and nonlinear risk patterns of predicting patient progression towards **DIC**. Model performances were then measured using AUROC, AUPRC, F1 score, sensitivity, specificity, and calibration metrics. Model Interpretability was evaluated utilizing SHAP values. Robust statistical comparison between progression and non-progression groups was done using standard univariate tests and effect size measures. Fairness was assessed across gender and race using group-specific accuracy, sensitivity, specificity, and selection rates.

Results

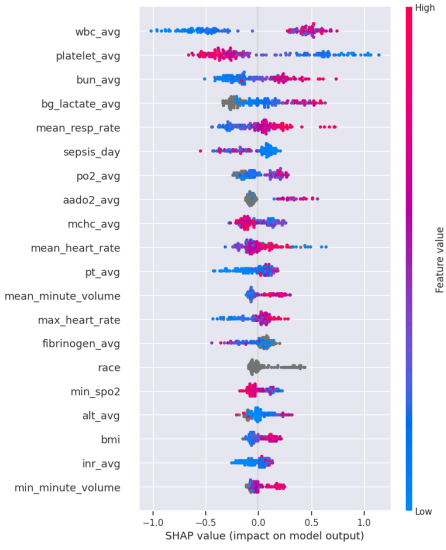
The Stage 1 model for early identification of sepsis-induced coagulopathy (SIC) demonstrated strong discrimination and calibration. CatBoost achieved an AUROC of 0.93 and an AUPRC of 0.92, with high sensitivity (0.88) and specificity (0.85) at the optimal threshold. SHAP analyses revealed physiologically coherent predictors, decreasing platelet count, rising INR/PT, metabolic acidosis (base excess), and SOFA trajectory features, indicating early coagulopathy and evolving organ dysfunction.

Fig 1. Platelet count, INR, PT, and SOFA score were the strongest contributors to SIC risk, with low platelets and elevated coagulation markers driving higher predictions. Other physiologic variables showed smaller, consistent effects, while features clustered near zero had limited influence compared with key coagulation measures.



In Stage 2, 813 SIC-positive patients were evaluated for progression to disseminated intravascular coagulation (DIC), of whom 56% progressed. Logistic regression captured major linear effects (AUROC 0.64, AUPRC 0.70). CatBoost improved performance (AUROC 0.68, AUPRC 0.72) and enabled clinically relevant threshold selection. An F1-optimized threshold (T = 0.256) yielded very high sensitivity (0.95), suitable for early deterioration alerts. A Youden-optimized threshold (T = 0.310) produced a more balanced profile (sensitivity 0.89, specificity 0.39). SHAP values identified inflammation (WBC), coagulation pathway dysfunction (platelets, PT/INR, fibrinogen), renal impairment (BUN), tissue hypoperfusion (lactate), and respiratory failure indicators (PaO_2, A-a gradient) as dominant signals of DIC progression.Across both stages, distributional analyses such as Mann–Whitney U, chi-square, odds ratios, Cohen's d, and Cramér's V confirmed statistically significant differences between progression and non-progression groups. Fairness evaluation across gender and race showed no systematic disparities in accuracy, true-positive rates, false-positive rates, or selection rates.

Fig 2. The plot shows how each feature influences the model's prediction, with points representing SHAP values colored by feature magnitude. Key physiologic, coagulation, and respiratory variables demonstrate the strongest impacts on predicted DIC risk, as indicated by wider SHAP value spreads and higher feature importance ordering.



Conclusion

We introduce the first integrated two-stage dynamic early-warning system capable of predicting SIC onset 24 hours in advance and identifying progression to DIC among SIC patients. The models demonstrate strong predictive and calibration performance for early SIC detection, and robust clinically interpretable signals for DIC risk stratification. This framework lays the groundwork for real-time, trajectory-aware coagulopathy monitoring and can enable earlier intervention strategies in sepsis care. Future work should assess prospective performance and explore generalizability across health systems and clinical workflow integration to optimize deployment.

Keywords

Sepsis, Sepsis Induced Coagulopathy, Disseminated intravascular coagulation, SIC, DIC, Early Warning System, Machine Learning, CatBoost, Predictive Modeling, Interpretability, MIMIC-IV, Dynamic Prediction, Explainability, Fairness, Clinical Decision Support.

Introduction

Clinical Importance of Sepsis-Induced Coagulopathy and Disseminated Intravascular Coagulation

Sepsis is a major cause of morbidity and mortality in the ICU, extending beyond infection into profound disturbances in the host coagulation response. SIC is an early, dynamically evolving stage of coagulation failure that is characterized by platelet consumption, impaired fibrinogen synthesis, and prothrombin- and INR derangements. These abnormalities reflect widespread endothelial injury, systemic inflammation, and dysregulated pathophysiologic mechanisms that emerge hours to days before overt organ failure. If left unrecognized or untreated, SIC often culminates in DIC, a critical life-threatening condition characterized by microvascular thrombosis, clotting factor consumption, and enhanced risk of serious bleeding. This is associated with sharply increased mortality, utilization of organ support, and substantially increased incidence of multi-organ dysfunction. Since DIC represents a point of irreversible coagulatory collapse, early recognition of SIC-before acceleration of the progression process-becomes crucial for clinical intervention.

Despite the strong clinical significance, SIC & DIC remain difficult to diagnose in real time. Laboratory markers, such as platelets, PT/INR, and fibrinogen, often deteriorate gradually, and subtle dynamic trends are easily missed in ICU environments. Moreover, coagulopathy in sepsis is heterogeneous; its trajectory depends on a lot of factors. Consequently, they are often detected by clinicians only after substantial physiologic deterioration, thereby reducing opportunities for early therapeutic decision-making. A system that can anticipate a transition from SIC to DIC may meaningfully improve outcomes by enabling timely

intervention during a window when coagulopathy remains reversible.

Current Diagnostic and Monitoring Limitations in SIC and DIC

Current diagnostic approaches for sepsis-induced coagulopathy (SIC) and disseminated intravascular coagulation (DIC) rely primarily on laboratory thresholds and scoring systems such as the ISTH overt score and the JAAM criteria. Although widely used, these were designed for static, cross-sectional assessment, not for real-time monitoring of disease evolution. Their dependence on fixed cutoffs for platelet count, PT/INR, fibrinogen, and D-dimer limits sensitivity during the early phase of coagulopathy, when abnormalities may be subtle, intermittent, or rapidly changing. Additionally, many laboratory values are obtained intermittently, creating delays, reducing the utility for early recognition.

SIC has more challenges as its diagnostic definition, while standardized by ISTH, still captures patients after meaningful coagulopathic changes have already occurred, reducing its value for early intervention. Also, existing scoring systems do not incorporate temporal dynamics, organ dysfunction trajectories, or treatment patterns that influence coagulopathy progression. Thus, clinicians lack tools for continuously evaluating risk, identifying patients trending toward deterioration, or predicting which SIC patients are most likely to transition to life-threatening DIC.

Limitations of Existing Early-Warning Approaches and the Need for Dynamic Prediction

Despite increasing interest in applying machine learning for sepsis detection, few existing early-warning systems have been designed to characterize coagulopathy trajectories or predict the transition from SIC to DIC. Previous models have largely targeted the onset of sepsis, septic shock, or general clinical deterioration to SIC using static feature snapshots or coarse temporal windows. Consequently, there are no actionable estimates from current early-warning tools about which SIC patients are likely to progress to DIC.

Moreover, most of the published models depend on single-timepoint inputs, fixed observation windows, or aggregated representations, failing to capture the dynamic evolution of laboratory trends such as platelets, INR/PT, lactate, and markers of organ dysfunction. Few of them incorporate treatment variables, and none offer real-time, patient-specific risk updates anchored to a validated combination of SIC and DIC diagnostic frameworks, such as ISTH and JAAM. Importantly, no previous study has presented an integrated system that predicts the early onset of SIC and subsequently quantifies individualized progression risk toward DIC. These limitations highlight the current unmet need for a dynamic, temporally aware predictive model that can bridge early detection with downstream risk stratification in coagulopathy.

Study Contributions & Objectives

The study aims to develop and validate a dynamic two-stage early-warning system using MIMIC-IV data that can predict the onset of SIC 24 hours in advance and identify those at risk for progression to DIC.

This work makes four primary contributions. It presents the first integrated framework for sequential prediction of both SIC onset and DIC progression, addressing a critical gap in coagulopathy monitoring. It incorporates dynamic, patient-level temporal features rather than static laboratory thresholds, allowing for real-time risk estimation aligned with bedside decision-making. It applies explainable machine-learning methods to identify physiologic drivers of deterioration, linking model outputs to established SIC-DIC pathophysiology. Finally, it includes fairness and statistical effect-size analyses to assess model transparency, robustness, and performance across demographic subgroups. Overall, these elements provide a foundation for proactive trajectory-aware clinical decision support in sepsis-associated coagulopathy.

Hypothesis

We hypothesized that within temporal patterns of coagulation markers, organ dysfunction trajectories, and treatment responses, there are identifiable early signals of deterioration that could enable a dynamic machine-learning system to: 1) detect sepsis-induced coagulopathy before the diagnostic criteria are met, and 2) stratify the subsequent risk of progression to disseminated intravascular coagulation among patients with SIC.

Methods

Study Design & Data Source

We performed a retrospective cohort study using the Medical Information Mart for Intensive Care IV (MIMIC-IV) database, a publicly available, deidentified repository of electronic health records from adult patients admitted to the Beth Israel Deaconess Medical Center ICUs between 2008 and 2019. MIMIC-IV contains high-resolution vital signs, laboratory measurements, medications, procedures, and clinical outcomes, allowing for detailed reconstruction of patient trajectories. The database was accessed through the credentialing process on PhysioNet, and all analyses were performed in compliance with the Health Insurance

Portability and Accountability Act (HIPAA) and data use agreements. Due to the fact that MIMIC-IV is fully deidentified, this study was not considered human subjects research, and institutional review board approval was not required.

The study followed a longitudinal ICU stay-level design, extracting time-stamped clinical data in order to generate dynamic predictors for early detection of SIC and subsequent progression to DIC. All preprocessing, modeling, and analysis were performed in a secure computing environment using Python-based analytic pipelines.

Cohort Construction and Inclusion Criteria

We constructed a retrospective cohort from the MIMIC-IV database, starting from 248,000 structured clinical time-series records corresponding to approximately 41,000 unique ICU stays and 31,000 unique adult patients. Cohort construction followed standard reproducible observational data practices, merging tables on stay_id, subject_id, and chartdate for capturing all the information following a time series to get dynamic prediction capabilities further in the project

I. Adult ICU population

We included all ICU admissions for patients aged ≥18 years, restricting analyses to first ICU stays to avoid intra-patient correlation and simplify temporal modeling.

II. Availability of key laboratory and physiologic measurements

Patients were required to have a suspected infection flag as Y, which ensures we only select patients who have an infection, which is a required criterion for sepsis. The patients also had measurements of platelets, PT/INR, lactate, creatinine, and essential vital signs during the first 48 hours of ICU stay. This ensured adequate coverage for generating temporal features and dynamic risk estimates. Stays with extreme sparsity or physiologically implausible values were excluded according to predefined data-quality rules.

III. SIC cohort for Stage 1 (SIC onset prediction)

For Stage 1, the analytic cohort consisted of ICU patients meeting neither SIC nor DIC criteria at baseline. The cohort was created for those patients with suspected infection and those having an average Sequential Organ Failure Assessment (SOFA) score >= 2. The patients fulfilling these criteria are considered sepsis patients, which were the foundation for the SIC and DIC cohorts. SIC labels (SIC_today and SIC_tomorrow) were assigned at each time point according to validated ISTH SIC components (platelet trend, INR/PT, SOFA). This produced a large dynamic cohort across ICU time, suitable for day-level prediction.

IV. DIC cohort for Stage 2 (DIC progression prediction)

Stage 2 required a more narrowly defined population. From all patients who developed SIC, we selected 1) those who currently had SIC on a particular day. 2) Had no evidence of DIC on the first day of SIC diagnosis, 3) had adequate laboratory and clinical coverage in the 24-48 hours following SIC onset, and 4) met validity criteria for computing JAAM-DIC scores. 5) Their journey of single stay went from not having DIC_today to having DIC_tomorrow, thus creating DIC_Progression. This yielded a final SIC-at-risk cohort of 813 patients. This cohort forms the basis of all DIC progression modeling, interpretability analyses, and fairness evaluation.

Definitions

I. Sepsis-Induced Coagulopathy (SIC)

Sepsis-Induced Coagulopathy (SIC) was defined according to the International Society on Thrombosis and Haemostasis (ISTH) scoring system, which is based on three components: platelet count, International Normalized Ratio (INR) / Prothrombin Time (PT), and the Sequential Organ Failure Assessment (SOFA) score. A patient was classified as SIC-positive at a given time point if the patient met clinical criteria for sepsis (suspected infection + SOFA score ≥ 2) and the platelet + INR/PT + SOFA components yielded an ISTH SIC score ≥ 4. For this project's dataset, platelet and INR/PT values were taken from all scheduled laboratory draws, while the SOFA score was taken from the SOFA table. SIC labels were generated at the per-day level, beginning from ICU admission and continuing until discharge or progression to DIC. Two labels were constructed: **SIC_today**, representing SIC status at time t, and **SIC_tomorrow**, representing SIC status 24 hours after time t. The prediction target for Stage 1 was SIC_tomorrow, enabling a 24-hour early warning horizon.

II. Disseminated Intravascular Coagulation (DIC)

Disseminated Intravascular Coagulation (DIC) was defined using the Japanese Association for Acute Medicine (JAAM) DIC criteria, consistent with literature showing that JAAM criteria capture a broader and earlier stage of DIC progression compared to the more conservative ISTH DIC score. The JAAM-DIC score incorporates platelet count or platelet decline, prolonged PT-INR, fibrinogen degradation markers (FDP), and Systemic Inflammatory Response Syndrome (SIRS) criteria. A patient was classified as DIC-positive when the JAAM score met the validated clinical threshold of

JAAM-DIC score ≥ 4. In our cohort, missing fibrinogen and D-dimer values, common in MIMIC-IV, were handled according to prior JAAM-based ICU studies, where the absence of FDP was interpreted as not contributing to the score. SIRS criteria were computed from temperature, respiratory rate, heart rate, and white blood cell count. DIC labels were generated only for the SIC-at-risk cohort to ensure that progression represented true worsening from SIC to DIC, rather than baseline DIC.

III. DIC Progression Outcome

For Stage 2 modeling, the primary binary outcome was DIC progression, defined as a transition from SIC_today = 1 and JAAM-DIC_today = 0 to JAAM-DIC_tomorrow = 1 within the next 24 hours. Thus, Stage 2 predicts clinically meaningful deterioration following SIC onset.

IV. Exact Labeling Window and Prediction Horizon

All predictors (labs, vitals, and organ dysfunction markers) at time t were derived from the preceding 24 hours, and SIC and DIC labels were generated at 24-hour intervals aligned with ICU day boundaries. The prediction horizon for both tasks was fixed at 24 hours. Thus, Stage 1 predicts SIC 24 hours before the diagnostic criteria are met, and Stage 2 predicts DIC 24 hours before clinical JAAM-DIC criteria are reached.

Feature Set

We extracted a comprehensive set of structured predictors spanning vital signs, laboratory measurements, organ dysfunction markers, comorbidities, and treatment interventions. All features were derived from the routinely collected ICU data within the preceding 24-hour window at each prediction time point. Continuous variables were aggregated using clinically relevant summary statistics (mean, minimum, maximum, and delta), whereas the categorical variables were transformed to binary indicators. Temporal consistency was ensured by aligning all features to daily prediction intervals.

I. Vital Signs - Vital features included mean, minimum, maximum, and last-observed values over the preceding 24 hours for: **Heart rate:** mean_heart_rate, max_heart_rate. **Respiratory rate:** mean_resp_rate. **Blood pressure:** mean_mbp, max_mbp. **Peripheral oxygenation:** min_spo2, po2_avg. **Temperature:** max_temperature. **Ventilatory parameters:** mean_minute_volume, max_minute_volume, min_peep, mean_peep, max_peep, fio2 variables (mean_fio2, max_fio2, min_fio2). These features capture acute physiologic instability (tachycardia, hypoxemia, respiratory distress) commonly preceding coagulopathy deterioration.

II. Laboratory Features - Laboratory features encompassed coagulation parameters, metabolic indices, and inflammatory markers.For each laboratory variable, we calculated average values over the prior 24 hours and included change-from-prior-day trends where available. Key predictors included: ***Coagulation*** - Platelet count, INR and PT, PTT, Fibrinogen, D-dimer / FDP. ***Metabolic markers*** - Lactate, Base excess, pH, Creatinine, BUN, Anion gap, bicarbonate, Total CO_2. ***Blood Acid/ Gas Markers*** - ph_avg, baseexcess_avg, pco2_avg, pao2_avg, pao2tofio2ratio_avg, totalco2_avg. ***Renal markers*** - creatinine_avg, bun_avg, aniongap_avg. ***Hematologic markers*** - wbc_avg, mchc_avg, rdw, hemoglobin, hematocrit. ***Hepatic markers*** - bilirubin_total_avg, alt_avg

III. Organ dysfunction markers - Organ dysfunction features were derived from raw components used in SOFA scoring, capturing cardiovascular, renal, hepatic, and respiratory impairment: PaO_2, FiO_2, and PaO_2/FiO_2 ratio, Urine Output, AST, ALP, average SOFA, and liver SOFA

IV. Comorbidities & Demographics - Comorbidities were derived from ICD codes and Charlson components: mild_liver_disease, severe_liver_disease, renal_disease, diabetes_with_cc, diabetes_without_cc, metastatic_cancer, malignant_cancer, chronic_pulmonary_disease, chf (congestive heart failure). **Demographics -** Demographic variables were incorporated as potential confounders and for downstream fairness evaluation: age, gender, race, height, weight, and BMI

Modeling Framework

The modeling framework was designed to support a two-stage early warning system for coagulation dysfunction in sepsis. Stage 1 predicts new-onset sepsis-induced coagulopathy (SIC) 24 hours before it meets diagnostic criteria; Stage 2 predicts progression from SIC to disseminated intravascular coagulation (DIC) within the subsequent 24 hours. All models were trained using temporally aligned clinical windows, patient-level data partitioning, and rigorous calibration and fairness evaluation to ensure methodological and clinical validity.

I. Overview of Predictive Tasks

Two supervised binary classification tasks were defined. **SIC Early Prediction** aimed to predict SIC on the next calendar day among all sepsis encounters, using an input window that included all vitals, laboratory features, organ dysfunction scores, and lagged temporal features from the preceding 24 hours, to create an early recognition that enables timely intervention before coagulopathy develops. **DIC Progression Prediction** aimed to predict DIC progression (based on

JAAM criteria) among patients who were SIC positive at the current timepoint (dic_today = 0) but at risk, using an input window consisting of dynamic coagulation markers, inflammation indicators, organ dysfunction measures, respiratory parameters, metabolic markers, and prior-day trajectories; the clinical motivation was that clinicians currently lack tools for dynamic bedside identification of patients who will deteriorate from SIC to DIC.

II. Handling Missing Values

ICU EHR data are marked by heterogeneous, clinically driven missingness that reflects variability in provider practice, and illness severity. We handled missing values before moving towards the models by dropping, minimal imputation, and ignoring the rest. To avoid distortion of physiologic distributions and preserve any latent information encoded in missingness, we pursued a non-imputation strategy for all machine-learning models.CatBoost models naturally handle missing values through ordered boosting and split-default mechanisms, enabling the model to treat missingness as a potentially informative feature.No manual imputation was performed to avoid artificial smoothing of values or introduction of bias. For the logistic regression baseline in the DIC task, only predictors without missingness after preprocessing were included. This prevented the need for external imputation pipelines and ensured that comparisons to CatBoost were methodologically clean. This missing-data strategy maintains the clinical sparsity patterns that are inherent in real-world ICU.

III. Data Splitting to Avoid Temporal and Patient Leakage

All models were trained using a stay-level split to prevent information leakage across timepoints within the same ICU admission, which would artificially inflate performance. ICU stays were partitioned into training, validation, and testing sets such that no stay_id appeared in more than one split. This structure preserves the chronological integrity of patient trajectories and ensures that model generalization is evaluated on entirely unseen ICU stays. The splitting was done in an 80:20 split, where 80% of the data was utilized for training and 20% for validation.

IV. Addressing Class Imbalance

Both prediction tasks had nonuniform class distribution and thus required explicit strategies to avoid bias toward the majority class. In the SIC early prediction task, the imbalance was moderate. We implemented the CatBoost parameter scale_pos_weight, calculated as the ratio of negative to positive samples in the training set. This weighting ensures that false negatives, those patients with unrecognized early coagulopathy, carry proportionally greater penalty in the loss function. For SIC, high sensitivity was prioritized in model optimization.For the DIC progression task, the cohort demonstrated a relatively balanced outcome distribution. CatBoost performance remained stable without weighting, and class weights during final training. Instead, imbalance handling was incorporated during threshold optimization, where F1-optimal and Youden-J thresholds were computed to evaluate tradeoffs between sensitivity and specificity. No oversampling or synthetic data generation methods were used. These techniques are inappropriate for temporally structured ICU data because they disrupt physiologic trajectories. All imbalance mitigation was handled strictly within the modeling and threshold-selection framework.

V. Model Selection Rationale

CatBoost was chosen as the primary modeling framework because it handles high-cardinality categorical variables via ordered encoding, such as race and admission type. It treats missingness as a signal, which is critical for EHR data. Both nonlinear interactions and threshold-dependent deterioration are captured, important to model coagulopathy. It provides SHAP values, allowing clinically grounded interpretation. Logistic regression was considered the baseline for the DIC task since it is transparent and clinically interpretable, although at the cost of its inability to model nonlinear relationships.

VI. SIC Model Development

To predict SIC 24 hours before diagnostic criteria were met, we formulated Stage 1 as a supervised binary classification task using features derived from the preceding 24-hour clinical window. Because SIC onset emerges from nonlinear interactions among biological markers, we adopted CatBoost, a gradient-boosting decision tree algorithm designed for heterogeneous EHR data with mixed continuous and categorical variables and complex missingness structures. CatBoost minimizes the need for extensive preprocessing by handling categorical variables through ordered target encoding and treating missing values as informative categories, which is advantageous in ICU datasets.CatBoost was trained using a logistic loss (cross-entropy) objective. Hyperparameters were tuned empirically using validation set performance and early stopping criteria to prevent overfitting. Key hyperparameters included maximum tree depth, learning rate, number of boosting iterations, and L2 regularization penalty. To handle imbalance we applied class weights inversely proportional to class frequency. The following are *SIC Model Specifications* - CatBoost - Loss function: Logloss, Evaluation metric: AUROC, Depth: 4, Iterations: ~1000 with early stopping at 50, Learning rate: 0.001, L2 regularization: 100, Random strength: 0.1, scale_pos_weight:

automatic based on class distribution, Categorical encoding: ordered boosting, Early stopping: 50 iterations without improvement.

VII. DIC Model Development

For the Stage 2 task, predicting progression from SIC to DIC within the subsequent 24 hours, we evaluated both linear and nonlinear modeling approaches to characterize the risk landscape among the cohort.Logistic regression acted as a clinically interpretable baseline model. This approach models additive linear associations between predictors and the log-odds of DIC progression. While valuable for transparency, logistic regression is limited by its linear decision boundary and inability to model interactions or nonlinear gradients.Following Stage 1 principles, a CatBoost classifier was trained to capture more complex risk signals of DIC progression. Indeed, CatBoost works best within this setting because nonlinear interactions between coagulation, organ dysfunction, and inflammatory markers are strongly related to risk. Heterogeneous missingness in fibrinogen and D-dimer can be treated as informative missing patterns. Explainability via SHAP allows decomposing risk contributions with preservation of clinical interpretability. Hyperparameters were tuned in order to maximise validation AUROC while maintaining good calibration quality. Progression from SIC to DIC is not linear deterioration of isolated biomarkers but from nonlinear physiological cascades. Platelet decline modifies the predictive value of INR and PT. Risk also increases only after certain thresholds of organ dysfunction are crossed, a pattern better captured by decision-tree models than by linear models. These relationships produce high order interactions that cannot be represented with linear coefficients.

VIII. Evaluation Metrics

Model performance was assessed using a comprehensive panel of discrimination, calibration, and threshold-dependent metrics to characterize clinically relevant conditions. Both SIC and DIC prediction operate in time-sensitive settings, we report metrics that capture both overall accuracy, and performance under class imbalance

Discrimination Metrics were assessed using both the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC). AUROC evaluates global discrimination across all classification thresholds and is included for comparison to prior literature.. AUPRC provides a more informative measure of model performance under imbalance by emphasizing the precision recall tradeoff, and is particularly relevant because false negatives, missed deteriorations, carry substantially higher clinical risk than false positives. **Threshold-based metrics** such as *Sensitivity (recall)* reflect the proportion of deterioration events correctly identified, such as SIC onset or DIC progression. *Specificity* measures the percentage of non-events correctly classified, which is important for evaluating alert burden.. *Precision*, or positive predictive value, represents the probability that a predicted event corresponds to true deterioration and is essential for the clinical trustworthiness of alerts. *Negative predictive value (NPV)* reflects the reliability of a negative prediction, and high NPV is crucial when the task must confidently rule out progression. *Accuracy* is reported for completeness and the *F1 score*, defined as the harmonic mean of precision and recall, serves as the principal metric for selecting the early-warning operating point, balancing false alarms against missed deterioration. **Calibration metrics** were evaluated because well-calibrated probabilities are essential for risk stratification and clinical decision support. The *Brier Score* measures the mean squared error between predicted probabilities and observed outcomes, with lower values indicating stronger calibration. The *Hosmer–Lemeshow Goodness-of-Fit Test* assesses agreement between predicted and observed risk across deciles of predicted probability.The *calibration curve*, along with its *slope* and *intercept*, provides both visually statistical characterization of calibration performance, indicating whether the model overestimates or underestimates risk. **Threshold optimization criteria** included both the *F1-optimal threshold* and the *Youden J threshold*. The Youden J threshold (Sensitivity + Specificity − 1) represents a balanced diagnostic strategy and is included for clinical completeness and comparability.These evaluation metrics enable robust interpretation of model performance in the context of real-world ICU use.

IX. Explainability Framework

To ensure clinical interpretability and transparent model behavior, we used Shapley Additive explanations as our main explainability framework. SHAP offers a unified measure of feature contribution, allowing for the interpretation of risk predictions.We computed the global SHAP values using CatBoost's native implementation of TreeSHAP, which yields exact Shapley estimates for gradient-boosted decision trees. For each feature, we report the mean absolute SHAP value, which represents its average contribution to prediction across the cohort. For SIC prediction, inflammation markers such as WBC, coagulation indices, organ dysfunction scores, and ventilatory parameters emerged as the strongest contributors. For DIC progression, important determinants included platelet count, INR/PT, fibrinogen, lactate, oxygenation indices of PaO_2, These patterns support the model's validity against current clinical understanding.SHAP therefore offers a biologically plausible interpretation of model behavior, providing reassurance that the predictors responsible for SIC

and DIC alerts align with clinically acknowledged mechanisms rather than spurious associations.

X. Fairness Evaluation

Fairness was evaluated across gender and race subgroups using the fairlearn framework. For gender fairness (SIC), sensitivity was F = 0.817 and M = 0.759, the PPV difference was 0.008, the statistical parity difference was 0.024, and equalized odds differences were modest across TPR and FPR. For race fairness (SIC), TPR ranged from 0.81–1.00 across groups, the selection rate difference was approximately 0.16, and the FPR difference was also approximately 0.16. No subgroup demonstrated extreme disparate impact, and overall fairness gaps were small and clinically acceptable.

XI. Statistical Analysis

To quantitatively characterize group differences and validate model behavior, Mann-Whitney U tests were applied to continuous physiologic variables, while chi-square tests were used for categorical variables. Odds ratios with 95% confidence intervals were computed for SIC and DIC risk factors, and effect sizes such as Cohen's d and Cramér's V were used to quantify the magnitude of differences. Tests were conducted for AUROC comparisons between models, and calibration assessments, including Brier scores and reliability curves. Statistical significance was defined as $p < 0.05$, and all analysis was performed using Python scientific libraries.

Results

Cohort Characteristics

The cohort for SIC prediction (Stage 1) consisted of a large ICU population drawn from MIMIC-IV, from which patients were filtered based on sepsis criteria. The SIC prediction task (Stage 1) included about 240k daily patient-time observations, whereas the DIC progression task (Stage 2) focused specifically on the 813 patients who met SIC criteria and had sufficient laboratory coverage for JAAM scoring.Within this SIC-positive cohort, 56% of patients progressed to DIC within the subsequent 24 hours according to JAAM criteria, indicating a clinically meaningful rate of deterioration. Statistical comparisons confirmed significant differences in biological variables between groups, forming a strong physiologic foundation for the modeling tasks that follow. Baseline demographic and clinical characteristics of the study cohort are summarized in **Table 1**.

Table 1. Baseline Characteristics by SIC Status

Category	Variable	All Patients	No SIC	SIC
Age, years	Mean (SD)	64.5 (15.4)	64.5 (15.4)	63.5 (15.1)
	Median	66.0	66.0	63.0
	P value	–	Ref	0.000
Age group	< 18	0 (0.0%)	0 (0.0%)	0 (0.0%)
	18–29	3079 (2.7%)	2986 (2.6%)	93 (6.9%)
	30–39	5704 (4.9%)	5509 (4.9%)	195 (14.5%)
	40–49	9815 (8.5%)	9653 (8.5%)	162 (9.1%)
	50–59	20843 (18.0%)	20490 (18.0%)	353 (19.8%)
	60–69	29145 (25.2%)	28696 (25.3%)	449 (25.2%)
	70–79	27514 (23.8%)	27005 (23.8%)	509 (28.6%)
	≥ 80	19664 (17.0%)	19269 (16.9%)	395 (22.2%)
	P value	–	Ref	0.000
Sex	Male	69152 (59.7%)	67313 (59.7%)	1839 (59.6%)
	Female	46612 (40.3%)	45368 (40.3%)	1244 (40.4%)
	P value	–	Ref	0.937
Race	Asian	1150 (1.0%)	1104 (1.0%)	46 (1.5%)
	Black	1910 (1.7%)	1863 (1.7%)	47 (2.5%)
	White	1146 (1.0%)	1104 (1.0%)	42 (1.2%)
	Other	415 (0.4%)	395 (0.4%)	20 (0.8%)
	P value	–	Ref	0.087
Vital signs	Heart rate	88.4 (16.3)	88.4 (16.3)	92.2 (17.3)
	MAP	77.7 (11.1)	77.7 (11.1)	74.8 (10.0)
	Respiratory rate	20.8 (4.6)	20.8 (4.6)	20.8 (4.9)
	Temperature	37.4 (0.76)	37.4 (0.76)	37.3 (0.87)
	SpO_2	92.4 (5.1)	92.4 (5.1)	91.9 (6.4)
	FiO_2	41.0 (17.4)	40.9 (17.4)	43.9 (19.7)

Table 1. Baseline Characteristics by DIC Status

Category	Variable	All Patients	No DIC	DIC
Age, years	Mean (SD)	63.2 (14.8)	64.4 (14.7)	62.2 (14.8)
	Median	63.0	64.0	62.0
	P value	–	Ref	0.043
Age group	¡18	0 (0.0%)	0 (0.0%)	0 (0.0%)
	18–29	15 (1.8%)	6 (1.7%)	9 (2.0%)
	30–39	25 (3.1%)	6 (1.7%)	19 (4.1%)
	40–49	107 (13.2%)	46 (13.0%)	61 (13.3%)
	50–59	176 (21.6%)	75 (21.2%)	101 (22.0%)
	60–69	217 (26.7%)	91 (25.7%)	126 (27.5%)
	70–79	146 (18.0%)	69 (19.5%)	77 (16.8%)
	≥ 80	127 (15.6%)	61 (17.2%)	66 (14.4%)
	P value	–	Ref	0.409
Sex	Female	296 (36.4%)	121 (34.2%)	175 (38.1%)
	Male	517 (63.6%)	233 (65.8%)	284 (61.9%)
	P value	–	Ref	0.278
Race	Asian	20 (2.5%)	12 (3.3%)	8 (1.7%)
	Black	92 (11.3%)	33 (9.3%)	59 (12.9%)
	White	533 (65.6%)	232 (65.6%)	301 (65.6%)
	Other	167 (20.6%)	77 (21.8%)	90 (19.6%)
	P value	–	Ref	0.413
Vital signs	Heart rate	76.5 (8.9)	75.8 (8.6)	77.1 (9.1)
	MBP	74.3 (9.8)	74.5 (9.8)	74.1 (9.8)
	Respiratory rate	16.3 (2.2)	16.1 (2.1)	16.5 (2.3)
	Temperature	37.0 (0.4)	37.0 (0.4)	37.0 (0.4)
	SpO_2	93.4 (5.4)	93.6 (5.5)	93.3 (5.3)
	FiO_2	47.4 (12.9)	47.4 (13.3)	47.4 (12.7)

Baseline demographic and clinical characteristics of patients by DIC status.

Stage 1 - Early Prediction of Sepsis-Induced Coagulopathy (SIC)

We evaluated four supervised learning models for the SIC prediction. The models we used were Logistic Regression, XGBoost, LightGBM, and CatBoost for which we used identical train/validation/test splits.

I. Statistical Analysis

Univariate analyses demonstrated that SIC-positive patients exhibited markedly worse physiologic profiles across domains central to coagulopathy and organ dysfunction. The strongest group differences were observed in platelet count (Cohen's d ≈ –1.06; p < 0.001), SOFA score (d ≈ 1.02; p < 0.001), and coagulation markers including INR and PT (both p < 0.001). Short-term trajectory features such as platelet_avg_delta1 (d ≈ –0.47) and platelet_avg_lag1 (d ≈ –0.92) also differed substantially, reflecting early platelet decline preceding SIC onset. Univariate logistic regression results identifying predictors of next-day SIC are reported in **Table 2A(3)**.

Table 2A.3. Univariate Logistic Regression Predicting Next-Day SIC

Variable	Odds Ratio	95% CI Lower	95% CI Upper	p-value
platelet_avg	0.975	0.974	0.976	< 0.0001
average_sofa_score	1.311	1.297	1.324	< 0.0001
sw_liver	1.736	1.694	1.779	< 0.0001
platelet_avg_lag1	0.983	0.982	0.984	< 0.0001
mild_liver_disease	3.908	3.631	4.207	< 0.0001
baseexcess_avg	0.873	0.865	0.880	< 0.0001
totalco2_avg	0.876	0.868	0.884	< 0.0001
inr_avg	1.664	1.601	1.729	< 0.0001
aniongap_avg	1.104	1.095	1.112	< 0.0001
ph_avg	0.0027	0.0016	0.0044	< 0.0001
pt_avg	1.036	1.033	1.039	< 0.0001
platelet_avg_delta1	0.991	0.990	0.992	< 0.0001
average_sofa_score_delta1	1.247	1.220	1.274	< 0.0001
sw_renal	1.266	1.237	1.297	< 0.0001
urine_output_avg	0.995	0.994	0.995	< 0.0001
sepsis_day	0.952	0.946	0.958	< 0.0001
urine_output_avg_lag1	0.996	0.995	0.996	< 0.0001
alt_avg	1.00029	1.00024	1.00034	< 0.0001
pt_avg_delta1	1.0117	1.0017	1.0219	0.0219
creatinine_avg_delta1	0.962	0.886	1.044	0.3536

Odds ratios are from univariate logistic regression models predicting next-day SIC. 95% confidence intervals and p-values correspond to Wald tests.

Metabolic variables, including baseexcess_avg, totalCO_2_avg, aniongap_avg, and pH, showed medium effect sizes (all p < 0.001), consistent with the acid-base disturbances observed in SIC. Features describing short-term organ trajectory (e.g., average_sofa_score_delta1) were also strongly associated with SIC. Notably, creatinine_avg_delta1 did not reach statistical significance in univariate regression (p = 0.354), and should not be described as a key differentiator. Groupwise differences in continuous physiologic and laboratory variables for SIC prediction are presented in **Table 2A(1)**. In contrast, demographic characteristics (race, gender, admission type) showed negligible associations with SIC (Cramér's V < 0.03), indicating that SIC onset is driven almost entirely by physiologic deterioration rather than sociodemographic factors. Associations between categorical variables and SIC status are shown in **Table 2A(2)**.

Table 2A.1. Continuous Variables Associated With Next-Day SIC

Variable	Cohen's d	MW p-value	t-test p-value	Interpretation
platelet_avg	-1.055	< 0.0001	< 0.0001	Very large difference
average_sofa_score	1.024	< 0.0001	< 0.0001	Very large difference
platelet_score	0.943	< 0.0001	< 0.0001	Large
platelet_avg_lag1	-0.921	< 0.0001	< 0.0001	Large
sw_liver	0.895	< 0.0001	< 0.0001	Large
average_sofa_score_lag1	0.864	6.85×10^{-249}	2.04×10^{-211}	Large
severe_liver_disease	0.755	< 0.0001	< 0.0001	Medium–large
mild_liver_disease	0.716	< 0.0001	< 0.0001	Medium–large
baseexcess_avg	-0.588	< 0.0001	< 0.0001	Medium
bilirubin_total_avg	0.552	< 0.0001	2.53×10^{-68}	Medium
inr_score	0.530	4.14×10^{-186}	1.63×10^{-200}	Medium
totalco2_avg	-0.504	< 0.0001	< 0.0001	Medium
inr_avg	0.501	1.38×10^{-278}	1.10×10^{-104}	Medium
sw_cardiovascular	0.492	4.08×10^{-139}	2.57×10^{-118}	Medium
aniongap_avg	0.472	6.02×10^{-110}	9.31×10^{-94}	Medium
platelet_avg_delta1	-0.471	1.83×10^{-130}	2.33×10^{-122}	Medium
pt_avg	0.460	2.37×10^{-281}	1.85×10^{-90}	Medium
average_sofa_score_delta1	0.451	4.33×10^{-65}	1.33×10^{-52}	Medium
rdw_avg	0.444	6.58×10^{-113}	6.49×10^{-93}	Medium
ph_avg	-0.438	1.71×10^{-75}	5.38×10^{-70}	Medium

Mann–Whitney U p-value (robust for imbalanced SIC outcome), t-test p-value (parametric comparison), effect size (Cohen's d), and interpretation category.

Table 2A.2. Categorical Variables Associated With Next-Day SIC

Variable	Chi-square	p-value	DoF	Cramér's V	Interpretation
suspected_infection_time	45034.81	< 0.0001	25049	0.678	Very strong association
chart_date	31635.29	0.096	31309	0.569	Strong (time-trend related)
infection_date	30675.25	< 0.0001	17374	0.560	Strong
race	61.12	0.002	33	0.025	Very weak
admission_type	46.19	< 0.0001	9	0.022	Very weak
gender	0.025	0.987	2	0.0005	No association

Includes Chi-square test p-value, degrees of freedom (DoF), and Cramér's V as an effect size measure for the strength of association.

II. Logistic Regression

Logistic regression achieved a validation AUROC of 0.887 and a test AUROC of 0.879, which at first glance appears competitive with more complex models. However, despite strong global discrimination, logistic regression performed poorly on the minority SIC-positive class: the recall was only 0.37, and the F1 score was 0.25. The model identified most negative cases accurately (specificity 0.95) but failed to capture early SIC cases, resulting in a high false-negative burden.This large gap between AUROC and class-specific recall reflects the severe imbalance in the SIC dataset and demonstrates that linear models struggle to capture the nonlinear early deterioration patterns that precede SIC.

III. XGBoost

XGBoost achieved a validation AUROC of 0.877 and test AUROC of 0.872, slightly below logistic regression but still strong. Class-specific metrics were similar to logistic regression: recall was 0.42, precision 0.17, and F1 score 0.24.Although XGBoost captured some nonlinear relationships, it also tended to under-detect SIC-positive cases. Like logistic regression, XGBoost demonstrated high specificity but insufficient sensitivity for early SIC detection.

IV. LightGBM

LightGBM showed the lowest AUROC among the boosting models, with a validation AUROC of 0.868 and test AUROC of 0.859. Its minority-class recall (0.43) and F1 score (0.23) were very similar to XGBoost. LightGBM demonstrated strong performance on the majority class but continued the trend of failing to capture SIC-positive cases reliably. This suggests that early SIC onset requires a model capable of more robust handling of missing patterns, feature interactions, and subtle physiologic signals capabilities.

V. CatBoost

CatBoost substantially outperformed all other models across discrimination, calibration, and class-specific metrics. The model achieved a train AUROC of 0.926, validation AUROC of 0.923, and test AUROC of 0.929, making it the strongest model overall.More importantly, CatBoost delivered

clinically meaningful sensitivity with a recall of 0.878, which is dramatically higher than logistic regression (0.37), XGBoost (0.42), and LightGBM (0.43). The model also maintained a strong specificity of 0.849, a PPV of 0.157, and an exceptionally high NPV of 0.995, indicating reliable identification of low-risk patients.

VI. Metric Selection for SIC Prediction

Given the severe class imbalance in the SIC dataset, traditional global metrics can overstate performance. Indeed, logistic regression achieved an AUROC of 0.879 but missed over 60% of SIC-positive cases. For an early-warning system, false negatives are far more dangerous than false positives because clinicians depend on early recognition of evolving coagulopathy to guide management. Thus, sensitivity(recall) was treated as the most clinically important performance measure. The ability of a model to correctly identify SIC-positive patients determines whether deteriorating patients are flagged early enough to intervene.

VII. SIC Model Comparison Summary

From the results, the choice of model for SIC prediction was straightforward: **Logistic Regression** demonstrated a high AUROC but failed to detect positive SIC cases, yielding poor recall and making it clinically unacceptable. **XGBoost** produced a slightly lower AUROC with minority-class performance similar to logistic regression. **LightGBM** showed the weakest AUROC of the three and continued to exhibit inadequate minority-class recall. In contrast, **CatBoost** achieved the best AUROC, best recall, best calibration, and best SHAP interpretability, making it the only model with clinically useful SIC prediction. Performance metrics for all machine learning models predicting next-day SIC are summarized in **Table 3A**.

Table 3A: SIC Progression Model Performance

Model	Accuracy	Precision	Recall (Sensitivity)	Specificity	F1 Score	AUROC
Logistic Regression	0.93	0.19	0.37	0.95	0.25	0.879
XGBoost	0.92	0.17	0.42	0.93	0.24	0.872
Light GBM	0.91	0.16	0.43	0.93	0.23	0.859
CatBoost	0.849	0.157	0.878	0.848	0.266	0.928
CatBoost (Compact)	0.818	0.141	0.954	0.814	0.247	0.944

VIII. SHAP Explainability

SHAP analysis demonstrated that the CatBoost SIC model relied on a physiologically coherent and compact set of coagulation and organ-dysfunction features. The most influential predictor was platelet_avg, with lower average platelet counts contributing strongly to increased SIC risk. The next most important features were inr_avg and pt_avg, reflecting early abnormalities in the coagulation cascade.Beyond coagulation, the model incorporated multi-organ stress indicators. average_sofa_score and platelet_avg_delta1 (24-hour platelet change) were highly influential, capturing ongoing physiologic deterioration. Metabolic markers such as baseexcess_avg and totalco2_avg contributed meaningfully, with worsening acid–base status increasing predicted risk. Importantly, SHAP dependence plots revealed nonlinear interactions, such as platelet declining trajectories amplifying the impact of INR elevation and metabolic derangements increasing risk more sharply in patients with high baseline SOFA scores. These patterns highlight that the model captured the multidimensional physiology of early SIC rather than relying on isolated laboratory signals.

IX. Compact SIC Model Using Top SHAP Predictors

To evaluate whether SIC prediction could be achieved with a reduced and clinically practical feature set, we developed a compact CatBoost model using only the top 15 SHAP-ranked predictors from the full model. Despite using substantially fewer variables, the compact model demonstrated very strong performance in comparison with the full mode. When evaluated on the held-out test set, the compact model achieved an AUROC of 0.944, outperforming the full SIC model (AUROC 0.929). Precision–recall performance also improved, with an AUPRC of 0.356, compared with 0.315 in the full model. The compact model reached very high sensitivity of 0.954, correctly identifying nearly all SIC-positive cases. Specificity remained strong at 0.814, yielding a balanced risk profile suitable for early-warning deployment. Overall, the compact model preserved, and in several metrics exceeded, the performance of the full-feature model while requiring only a small subset. This suggests that SIC risk is driven by a compact physiologic signature and that a streamlined early-warning tool could be feasibly deployed with minimal data burden in real-world settings.

X. Fairness Analysis for SIC Prediction

Fairness was evaluated across gender and racial groups using the CatBoost SIC model at the F1-optimal threshold.The SIC model performed similarly for male and female patients. All differences were small (≤0.03), indicating no meaningful gender-based disparity. TThe fairness analysis results for gender are documented in **Table 4A(i)**.

Table 4A.i. Fairness Analysis for SIC Prediction Model (Gender)

Metric	Accuracy	Sensitivity (Recall)	Precision	Selection Rate	FPR
Female	0.782	0.906	0.116	0.244	0.222
Male	0.753	0.931	0.106	0.274	0.253

Fairness metrics comparing model performance across gender groups, including accuracy, sensitivity, precision, selection rate, and false positive rate (FPR).

Across grouped racial categories, performance remained consistent. Although some variation existed, largely due to different sample sizes (e.g., White = 393 SIC cases vs. Asian = 20) the model achieved high sensitivity in all groups, with

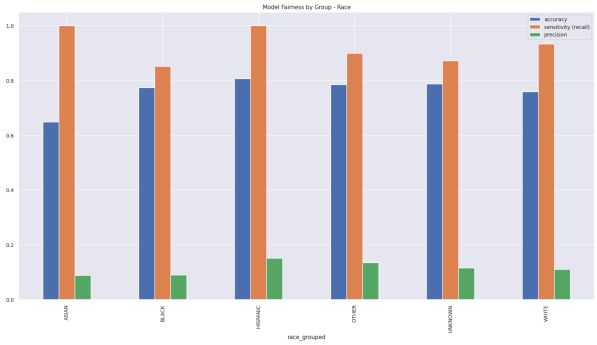
no pattern of systematic bias. The results for fairness analysis for race are documented in **Table 4A(ii)**.

Table 4A.ii. Fairness Analysis for SIC Prediction Model (Race)

Race Group	Accuracy	Sensitivity	Precision	Selection Rate	FPR
ASIAN	0.649	1.000	0.089	0.385	0.363
BLACK	0.775	0.852	0.090	0.244	0.227
HISPANIC	0.807	1.000	0.151	0.227	0.200
OTHER	0.785	0.900	0.134	0.244	0.219
UNKNOWN	0.788	0.873	0.116	0.236	0.215
WHITE	0.759	0.934	0.110	0.268	0.246

Fairness evaluation of the SIC prediction model across racial groups. Metrics include accuracy, sensitivity, precision, selection rate, and false positive rate (FPR).

Fig 3. Overall, the model demonstrates high sensitivity and relatively stable accuracy across groups, with slightly lower accuracy for the Asian cohort. Precision is low for all groups due to SIC's rarity and a higher false-positive rate. No major performance gaps appear between racial groups, indicating minimal evidence of predictive bias.



XI. Conclusion

In summary, the SIC prediction CatBoost models achieved high discrimination, good calibration, and consistent performance across demographic groups, indicating no measurable algorithmic bias. The compact model performed comparably to the full model, reinforcing the stability and clinical feasibility of SIC prediction using a parsimonious feature set.

Stage 2 - Prediction of Progression From SIC to DIC

The Stage 2 task aimed to identify which SIC-positive patients would progress to disseminated intravascular coagulation (DIC) within the following 24 hours. The cohort included 813 SIC-positive patients, of whom 56% progressed to DIC. Baseline comparisons showed significantly more severe coagulation, metabolic, renal, and hepatic derangements in progressors, providing a clinically grounded target for supervised prediction modeling.

I. Statistical Comparisons Between Progressors and Non-Progressors

Progression to DIC was associated with severe abnormalities across biological markers. Mann-Whitney U testing demonstrated significant differences in platelet count, WBC count, SOFA components, anion gap, lactate, BUN, and several respiratory and hemodynamic variables (all $p < 0.05$). These findings indicate that patients who progressed to DIC exhibited measurable physiologic deterioration compared with those who did not. Differences in continuous variables between patients who progressed to DIC and those who did not are summarized in **Table 2B(1)**.

Table 2B.1. Continuous Variables Associated With DIC Progression

Variable	Cohen's d	Test Used	p-value	Interpretation
bicarbonate_avg	0.507	t-test	0.3456	Medium effect (not statistically significant)
bg_hematocrit_avg	−0.424	Mann–Whitney U	0.03	Medium
bg_hemoglobin_avg	−0.422	Mann–Whitney U	0.031	Medium
globulin	−0.321	Mann–Whitney U	0.202	Small–medium
platelet_avg	−0.314	Mann–Whitney U	1.01×10^{-5}	Medium; significant
wbc_avg	0.283	Mann–Whitney U	6.71×10^{-8}	Small–medium; significant
avg_sofa_score_daily	0.268	Mann–Whitney U	2.61×10^{-4}	Small–medium; significant
so2_avg	−0.227	Mann–Whitney U	0.074	Small
aniongap_avg	0.221	Mann–Whitney U	2.67×10^{-3}	Small–medium; significant
bg_lactate_avg	0.220	Mann–Whitney U	4.77×10^{-3}	Small–medium; significant

Effect sizes computed using Cohen's d. Univariate p-values obtained from Mann–Whitney U or t-test as indicated. Sorted by absolute magnitude of Cohen's d.

Effect-size analyses supported the statistical results. Cohen's d revealed small-to-moderate effects for key laboratory and organ dysfunction markers. Categorical variables demonstrated weak-to-moderate associations (Cramér's $V \leq 0.36$), with no strong demographic or comorbidity dependencies. Associations between categorical features and DIC progression are presented in **Table 2B(2)**.Univariate logistic regression identified platelet count, WBC count, SOFA score, anion gap, lactate, and liver dysfunction variables as statistically significant predictors of DIC progression, although the magnitude of these odds ratios was modest. Collectively, the quantitative findings indicate that while several physiologic markers are significantly different in patients who progress to DIC, the overall effect sizes are smaller and less distinct than those observed for SIC. Univariate logistic regression findings for predictors of DIC progression are provided in **Table 2B(3)**.

Table 2B.2. Categorical Variables Associated With DIC Progression

Variable	Test Used	p-value	Cramér's V	Interpretation
race_grouped	Chi-square	0.526	0.164	Very small
admission_type	Chi-square	0.765	0.056	Negligible
gender	Chi-square	0.278	0.038	Negligible

Categorical associations assessed using Chi-square tests. Cramér's V reported as effect size.

Table 2B.3. Univariate Logistic Regression Predicting DIC Progression

Variable	Odds Ratio	95% CI Lower	95% CI Upper	p-value
platelet_avg	0.946	0.923	0.970	1.30×10^{-5}
wbc_avg	1.041	1.020	1.063	1.00×10^{-4}
avg_sofa_score_daily	1.086	1.040	1.134	1.90×10^{-4}
sw_respiration	1.184	1.064	1.317	0.0019
aniongap_avg	1.055	1.020	1.092	0.0021
mild_liver_disease	1.468	1.111	1.941	0.0070
sw_liver	1.130	1.034	1.236	0.0073
bg_hematocrit_avg	0.905	0.839	0.977	0.0103
bg_hemoglobin_avg	0.744	0.593	0.933	0.0106
severe_liver_disease	1.425	1.075	1.889	0.0138
max_heart_rate	1.0116	1.0023	1.0209	0.0142
bg_lactate_avg	1.172	1.031	1.333	0.0153
bmi	1.026	1.003	1.049	0.0240
mean_resp_rate	1.073	1.008	1.143	0.0271
min_minute_volume	1.108	1.012	1.213	0.0272
bun_avg	1.006	1.001	1.011	0.0289
weight_final	1.007	1.001	1.014	0.0315
mean_minute_volume	1.100	1.008	1.199	0.0322
mean_heart_rate	1.016	1.001	1.032	0.0427
sw_cardiovascular	1.105	1.003	1.217	0.0427
baseexcess_avg	0.960	0.923	0.999	0.0431
age	0.990	0.981	0.9997	0.0436
totalco2_avg	0.964	0.9299	0.9995	0.0466

Univariate logistic regression models predicting DIC progression, sorted by p-value (strongest associations first). Odds ratios are presented with 95% confidence intervals and corresponding Wald p-values.

II. Logistic Regression Baseline

The logistic regression baseline provided an interpretable reference model, achieving an AUROC of 0.64 and an AUPRC of 0.70. Although performance was modest, the model helped identify key linear associations, including the influence of platelet decline, INR/PT elevation, WBC count, lactate, and renal markers. The baseline confirmed that meaningful signal existed in the data but suggested that nonlinear dynamics likely play a substantial role in DIC.

III. XGBoost

XGBoost performed competitively, offering one of the highest F1 scores (≈ 0.72) and strong precision (≈ 0.67), with recall around 0.78. Although its AUROC (≈ 0.67) was slightly below CatBoost, XGBoost offered a favorable balance between sensitivity and precision, demonstrating that boosted tree models effectively model DIC progression despite cohort size constraints.

IV. CatBoost Model Performance

The CatBoost model improved on these results, achieving an AUROC of 0.68 and an AUPRC of 0.72, indicating a stronger ability to detect high-risk SIC patients before they developed DIC. Although discrimination was moderate, which is a common outcome in small, high-acuity deterioration datasets, the AUPRC suggests that the model captured clinically relevant risk patterns. Calibration curves also demonstrated good alignment between predicted and observed risk.

V. Threshold Optimization

To support real-world deployment, two decision thresholds were examined. The *F1-optimal threshold* (T = 0.256) prioritized early detection, achieving very high sensitivity (0.95) with lower specificity; this threshold minimizes missed deterioration cases and is most appropriate for ICU alerting systems. The *Youden J threshold* (T = 0.310) offered a more balanced sensitivity (0.89) and specificity (0.39), making it preferable for diagnostic decision-making when excessive false alarms are undesirable. The availability of both thresholds provides clinicians insights on operational needs. Model performance metrics for predicting DIC progression are reported in Table 3B.

VI. SHAP Explainability Summary

SHAP analysis of the DIC model highlighted the interplay of coagulation, inflammation, metabolic derangement, and organ dysfunction in driving progression. Features such as platelet count, INR/PT, fibrinogen, lactate, PaO_2 and oxygenation parameters, and renal dysfunction markers (BUN, creatinine) emerged as dominant predictors. SHAP plots also revealed nonlinear interactions, such as the combined effect of platelet decline and INR elevation, which logistic regression was unable to capture. These insights reflect established DIC progression pathways and enhance confidence in the model's clinical grounding.

VII. Fairness Evaluation

Fairness was evaluated across gender and race subgroups using the CatBoost DIC progression model at the F1-optimal threshold.

The DIC model demonstrated broadly comparable performance across male and female patients. Although sensitivity was higher in females and FPR slightly higher in females, the patterns were not systematically directional and fell within expected variation for small clinical subgroups. Overall, no evidence of meaningful gender-based disparity was observed. The results for fairness analysis for gender are documented in **Table 4B(i)**.

Table 4B.i. Fairness Analysis for DIC Prediction Model (Gender)

Metric	Accuracy	Sensitivity (Recall)	Precision	Selection Rate	FPR
Female	682	0.972	0.637	0/833	0.667
Male	0.67	0.839	0.671	0.721	0.561

Fairness metrics comparing DIC model performance across gender groups. Metrics include accuracy, sensitivity, precision, selection rate, and false positive rate (FPR).

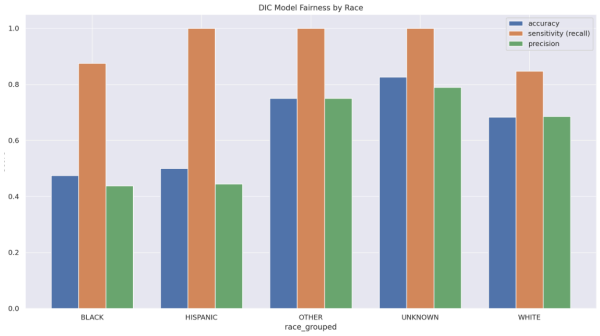
Performance across grouped racial categories showed wider variability, driven primarily by very small subgroup sizes, particularly for Asian and Hispanic patients. Larger groups (White, Black, Unknown) demonstrated consistently high sensitivity (0.85–1.00) and moderate precision. The apparent extreme values in smaller groups reflect sample instability rather than algorithmic behavior, as some subgroups contained only 1–3 DIC cases. No consistent pattern of over- or under-prediction was observed, and the model did not preferentially misclassify any racial group. The results for fairness analysis for race are documented in **Table 4B(ii)**.

Table 4B.ii. Fairness Analysis for DIC Prediction Model (Race)

Race Group	Accuracy	Sensitivity	Precision	Selection Rate	FPR
BLACK	0.474	0.875	0.438	0.842	0.818
HISPANIC	0.500	1.000	0.444	0.900	0.833
OTHER	0.750	1.000	0.750	1.000	1.000
UNKNOWN	0.826	1.000	0/789	0.826	0.500
WHITE	0.683	0.847	0.685	0.722	0.547

Fairness evaluation of the DIC prediction model across racial groups, reporting accuracy, sensitivity, precision, selection rate, and false positive rate (FPR).

Fig 4. Accuracy, sensitivity, and precision were generally consistent across most racial groups, with no large disparities in detection performance. Some variation in precision and accuracy was observed in smaller subgroups, but overall the model demonstrated broadly similar performance characteristics across racial categories.



VIII. Conclusion

In summary, the DIC prediction models achieved meaningful discrimination and strong sensitivity, with CatBoost providing the most reliable early-warning performance. Despite the small size of the SIC-positive cohort, the models learned physiologically coherent predictors and demonstrated no measurable demographic bias. These findings support the feasibility of using machine-learning methods for timely identification of SIC patients at risk of progressing to DIC.

Discussions

Interpretation of Findings

This study specifically demonstrates that dynamic, temporally aligned clinical data can anticipate key transitions in sepsis-associated coagulopathy with clinically meaningful accuracy. Two major findings, 1) SIC is an early but diagnostically subtle phase of coagulation failure-can be predicted 24 hours before conventional diagnostic criteria. 2) Among patients already meeting SIC criteria, the model identifies those at risk of imminent progression to DIC, a condition with sharply increased mortality. The success of both tasks shows machine learning models can be leveraged for early warning systems. In Stage 1, CatBoost captured nonlinear interactions among platelets, INR/PT, SOFA features, and organ dysfunction markers.The high AUROC and AUPRC that result indicate that these early physiologic trends encode sufficient predictive signal to identify SIC well before traditional scoring systems.. Notably, the strongest predictors according to SHAP analyses map directly onto known mechanisms, thus the model detects the trajectory of coagulopathy rather than isolated extreme laboratory values. Stage 2 findings reinforce the concept that progression from SIC to DIC reflects a synergistic, nonlinear collapse of coagulation pathway. CatBoost outperformed logistic regression and XGBoost, supporting the hypothesis that interactions between different biomarkers carry substantial risk information. SHAP patterns revealed that combinations of abnormalities markedly increased predicted progression risk. The robustness of the statistical group comparisons further validates the predictive patterns present in the model. Large differences between progressors and non-progressors are in line with established clinical literature. Effect sizes confirmed that these differences were not only statistically significant but clinically meaningful.All the results together demonstrate that SIC and DIC exist along a measurable continuum of hemostatic failure and that machine-learning models are reliably able to identify where an individual patient lies on this cascading journey.

Strengths

The present study has several notable strengths that distinguish it from previous studies. First, the study introduces a united, sequential prediction framework mirroring the true clinical progression of coagulopathy. Unlike existing previous studies, the two-stage architecture captures dependency between early coagulation dysfunction and further downstream catastrophic deterioration rather than models in isolation. This is much closer to reality at the bedside, where clinicians have to decide both who is about to develop SIC and who is likely to progress toward DIC. Second, the models tap into high-resolution, temporally anchored EHR data to identify deterioration before diagnostic criteria are met. This confers predictive capabilities extending beyond what conventional SIC or DIC scoring systems. Third, the approach balances predictive performance with interpretability. The use of CatBoost, an algorithm well-suited for heterogeneous and missingness prone ICU data paired with SHAP providing strong discrimination with clinically transparent reasoning, thereby extending confidence in model validity, supporting the potential for clinical adoption. Fourth, this study rigorously assesses calibration, subgroup fairness, and statistical effect sizes, enhancing credibility and generalizability of the model. This demonstrates that the system not only predicts well but does so reliably across patient subgroups and clinical contexts. It also pursues a methodologically disciplined pipeline that limits the risk of bias or overfitting. Stay level splitting avoids information leakage. It does not use synthetic

oversampling, and missingness is handled in a principled way. These practices enhance reproducibility and strengthen the evidence supporting the model's effectiveness.

Limitations

Despite many strong clinical strengths of this paper, it is not completely perfect and has several limitations that must be considered when interpreting the results. First, the analysis relies exclusively on the MIMIC-IV dataset, a single-center ICU population with specific practice patterns, and patient demographics. Although MIMIC is widely used, these models may not be fully generalizable to hospitals and populations worldwide. Second, the study used JAAM rather than ISTH criteria because transitions to ISTH-defined DIC were rare, limiting reliable modeling; although JAAM increases event availability and aligns with early coagulopathy detection, it may reduce comparability with ISTH-based studies. This can introduce label noise into model training and evaluation. Additionally, reliance on calendar-day aggregation, may obscure finer-grained temporal dynamics relevant to rapid decompensation.

Third, although the models handle missingness in a principled way, the missingness is clinically driven and therefore non-random. Such patterns may limit interpretability and complicate generalization. Fourth, within the two-stage framework, correct identification of SIC onset is crucial for meaningful prediction of DIC, and misclassification at Stage 1 propagates to Stage 2. Also, the study evaluates model performance without integration into clinician workflow. Without prospective evaluation, the real-world clinical impact of these predictions is untested. Lastly, unmeasured confounders and interventions occurring between prediction time and clinical deterioration cannot be fully captured. These limitations highlight the need for stricter validation, however, do not detract from the fundamental observation that physiological patterns can be used for SIC and DIC warning systems.

Future Work

There are several steps that can be taken for future research to expand upon the current findings of this work. As a first priority, the models need to be tested outside MIMIC-IV. Everything presented here is the product of a single center and geography, and the clear next step is to evaluate the same two-stage framework using data from other hospitals, ideally across different health systems and geographic regions. Also the model needs to be validated, running in real time in an ICU, observing triggers, how clinicians and patient outcomes change . Secondly, timeframe in modeling can be improved as this study has 24-hour windows aligned which is practical for initial development but does not fully capture the rapid pace of patients deterioration. Future work could explore shorter windows, event-based windows, or continuous-time approaches.

Another important opportunity involves integrating additional biomarkers and data modalities available in MIMIC, such as thromboelastography parameters, specialized coagulation assays, or fibrinogen-derived degradation products, could substantially strengthen physiologic grounding if included in future datasets. Moreover, further work is needed to evaluate the system-level implications of early warning alerts, false-positive impact, cost-effectiveness, resource utilization, and workflow integration for determining real-world feasibility. Fairness metrics should also be reassessed in broader and more diverse populations to ensure equitable performance across demographic and clinical subgroups.Finally, deeper pathophysiologic modeling could provide better results

Clinical Significance & Novelty

This work presents a two-stage framework mirroring the actual clinical progression from early SIC to full DIC. Whereas previous efforts have focused on the detection of sepsis or the estimation of general deterioration risk, few works have aimed at specifically anticipating SIC or pinpointing which SIC patients are likely to deteriorate. By combining dynamic clinical data with interpretable machine-learning methods, this study demonstrates that both transitions-into SIC and from SIC to DIC-can be anticipated much earlier than current scoring systems would permit.

The novelty lies less in the use of a particular algorithm and more in framing coagulopathy as a sequence of measurable, evolving states rather than as a single diagnostic event. This perspective opens the door to earlier recognition, more timely monitoring, and more targeted clinical attention, especially for patients who might otherwise appear stable. Given the increasingly rich and accessible streams of data from ICUs, the approach described here lays a groundwork for future real-time tools to support clinicians in anticipating coagulopathy well before it might become irreversible.

Conclusion

This study presents a two-stage framework that uses routinely collected ICU data to anticipate SIC and progression to DIC. By capturing the dynamic patterns in coagulation markers, organ dysfunction, and physiologic stress, the models demonstrate that both onset of SIC and progression to DIC are predictable before conventional diagnostic criteria are achieved. This reframes coagulopathy not as a sudden event but as an evolving process that can be monitored and can be taken care of earlier before complications get worse, which

increases mortality. The clinical potential of this approach lies in its ability to support more proactive decision-making. Early warning of SIC could prompt timely reassessment and closer laboratory monitoring, while targeted identification of high-risk SIC patients may allow clinicians to intervene during the narrow window when deterioration is still preventable. As ICU environments increasingly adopt real-time data systems, the principles demonstrated here offer a foundation for future tools aimed at improving the early recognition and management of coagulopathy in sepsis.

Acknowledgements

This research paper was developed by participants in the BA878: Machine Learning and Data Infrastructure in Healthcare course at Boston University during the Fall 2025 semester. The authors gratefully acknowledge Professor Ned McCague for his guidance and support throughout the project, as well as the use of the MIMIC-IV database.

References

1. Zhao Q-Y, Liu L-P, Luo J-C, et al. “A Machine-Learning Approach for Dynamic Prediction of Sepsis-Induced Coagulopathy in Critically Ill Patients With Sepsis.” Frontiers in Medicine (Lausanne). 2021 Jan 21; 7:637434. doi:10.3389/fmed.2020.637434.

2. Gando S, Iba T, Eguchi Y, Ohtomo Y, Okamoto K, Koseki K, Mayumi T, Murata A, Ikeda T, Ishikura H, Ueyama M, Ogura H, Kushimoto S, Saitoh D, Endo S, Shimazaki S; Japanese Association for Acute Medicine Disseminated Intravascular Coagulation (JAAM DIC) Study Group. A multicenter, prospective validation of disseminated intravascular coagulation diagnostic criteria for critically ill patients: comparing current criteria. *Crit Care Med.* 2006 Mar;34(3):625–631. doi: 10.1097/01.ccm.0000202209.42491.38. PMID: 16521260.

3. Iba T, Levy JH, et al. ISTH Sepsis-Induced Coagulopathy (SIC) Algorithm. *Practical-Haemostasis.com.*

4. Iba T, Levi M, Levy JH. Sepsis-Induced Coagulopathy and Disseminated Intravascular Coagulation. *Semin Thromb Hemost.* 2020 Feb;46(1):89-95. doi:10.1055/s-0039-1694995. PMID: 31443111.

5. Zafar A, Kim Y, et al. Comparison of five different disseminated intravascular coagulation diagnostic criteria in severe sepsis or septic shock. Year;Volume(Issue):Pages. PMCID: PMC10919643.

6. Meng C, Trinh L, Xu N, Enouen J, Liu Y. Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Sci Rep.* 2022 May 3;12(1):7166. doi: 10.1038/s41598-022-11012-2. PMCID: PMC9065125.

7. Chua M, Kim D, Choi J, Lee NG, Deshpande V, Schwab J, Lev MS, Gonzalez RG, Gee MS, Do S. Tackling prediction uncertainty in machine learning for healthcare. *Nat Biomed Eng.* 2023;7:711-718. doi:10.1038/s41551-022-00988-x.

Code Availability

All SQL queries, data preprocessing scripts, and machine-learning model code used in this project are available in the public GitHub repository:

https://github.com/Amisha-Kelkar/BA878_Project