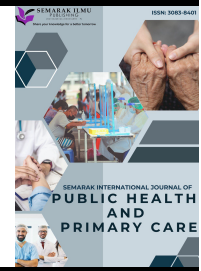




Semarak International Journal of Public Health and Primary Care

Journal homepage:
<https://semarakilmu.my/index.php/sijphpc/index>
ISSN: 3083-8401



SLE-RelapseX: Explainable and Causality-Aware Machine Learning for SLE Relapse Prediction

Muhammad Izzul Islam Faisal^{1,*}, Shahnorbanun Sahran², Syahrul Sazliyana Shaharir³, Nedaa Almansour⁴

¹ Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

² Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

³ Department of Medicine, Faculty of Medicine, Universiti Kebangsaan Malaysia Medical Centre, Kuala Lumpur, Malaysia

⁴ Intelligent Systems Department, Faculty of Artificial Intelligence, Al-Balqa Applied University, Salt, Jordan

ARTICLE INFO

Article history:

Received 15 August 2026

Received in revised form 2 September 2026

Accepted 30 September 2026

Available online 9 October 2026

Keywords:

Systemic lupus erythematosus; relapse prediction; explainable machine learning; leakage-aware modelling; clinical decision support

ABSTRACT

Systemic lupus erythematosus (SLE) relapse is clinically heterogeneous, and small prediction studies are vulnerable to leakage, correlated predictors and causal overinterpretation. This study addressed the lack of a compact relapse model accompanied by an explicit evidence audit. The objective was to develop and critically evaluate a leakage-aware machine-learning framework using routinely recorded variables. A retrospective cohort of 120 patients, including 41 relapses, was divided into 96 training and 24 held-out records. Preprocessing and feature selection were fitted only on training data. Logistic regression, random forest, multilayer perceptron and CatBoost were screened; the selected CatBoost model was evaluated using Shapley additive explanations (SHAP), an Explainable Boosting Machine (EBM), exploratory double machine learning (DML) and a remission-interval sensitivity analysis. The final model used lowest haemoglobin, complement component 3 (C3) status and anti-double-stranded DNA status. It produced precision 0.714, recall 0.625, specificity 0.875, accuracy 0.792 and an F1-score of 0.667. Standard and conditional SHAP preserved the same predictor ranking, while SHAP and EBM converged on a haemoglobin transition of 11.75-11.95 grams per decilitre. Exploratory DML found the strongest evidence compatible with a possible causal relationship for anti-double-stranded DNA positivity (estimate 0.426, 95% confidence interval 0.175-0.677; $p < 0.001$), without establishing causation. Reintroducing remission interval reduced the F1-score to 0.429 and produced discordant thresholds. The framework therefore separates predictive performance, explanation robustness, exploratory causal evidence and leakage sensitivity. Prospective multicentre validation, calibration and clinical-utility assessment are required before deployment.

* Corresponding author.

E-mail address: izzuley97@mail.com

1. Introduction

Systemic lupus erythematosus (SLE) is a heterogeneous autoimmune disease with a relapsing-remitting course. Flares can affect multiple organ systems, require treatment escalation and contribute to irreversible damage. The clinical challenge is not merely to identify patients who have already relapsed, but to recognise an increased-risk state early enough to intensify surveillance while avoiding unnecessary intervention in lower-risk patients.

SLE pathophysiology and clinical heterogeneity complicate relapse surveillance [1]. In a multicentre longitudinal study, McElhone *et al.*, [2] documented frequent early flares despite active management. A focused review found that demographic, disease-related and serological predictors varied substantially between cohorts [3]. A prospective follow-up study likewise identified severe-flare predictors but did not establish a universal marker [4]. Complement component 3 (C3), anti-double-stranded DNA (anti-dsDNA), blood counts, albumin, treatment exposure and prior disease activity therefore require interpretation as a combined and context-dependent signal.

Empirical studies have examined different parts of this risk landscape. Kakati *et al.*, [5] described regional patterns and determinants of flare in a North-East Indian cohort. Ji *et al.*, [6] synthesised relapse risk after glucocorticoid withdrawal, addressing a treatment-specific setting. Ramirez *et al.*, [7] examined infection-associated flares and highlighted the interaction between immune activity, treatment and care processes. Li *et al.*, [8] evaluated significant relapse and maintenance treatment in SLE-associated immune thrombocytopenia. These studies contribute clinically useful but setting-specific evidence, and their differences in populations, outcomes and follow-up limit direct transfer of any single rule.

Machine-learning research offers a way to represent nonlinear relations, missingness and correlated predictors, but it introduces additional sources of bias. Ceccarelli *et al.*, [9] reviewed applications across diagnosis, activity and outcomes and identified both promise and methodological heterogeneity. Zhan *et al.*, [10] emphasised the opportunity for machine learning in lupus medicine while noting the need for interpretable, clinically evaluated systems. Neither predictive accuracy nor a visually persuasive explanation alone establishes transportability, calibration or clinical benefit.

Transparent reporting guidance requires leakage prevention, uncertainty reporting and explicit limits on intended use [11]. Risk-of-bias guidance further requires development and evaluation procedures that match the intended clinical setting [12]. Calibration is essential because discrimination can remain acceptable while predicted risks are systematically wrong [13]. Decision-curve analysis evaluates whether predictions provide net clinical benefit [14]. Broader prediction-model assessment also requires complementary performance measures rather than a single accuracy statistic [15]. An earlier analysis of the same Hospital Canselor Tuanku Muhriz cohort used Shapley additive explanations (SHAP) with CatBoost and reported approximately 79% accuracy, but it retained remission interval and did not test explanation stability under dependent predictors or distinguish predictive importance from exploratory causal evidence [16].

The research gap is therefore the absence of a compact SLE relapse model accompanied by an explicit audit of leakage, explanation dependence, threshold reproducibility and causal overinterpretation. The primary objective was to develop a relapse model from routinely recorded variables using training-only preprocessing and feature selection. Secondary objectives were to compare explanation methods, identify clinically readable transition regions, explore adjusted relationships without claiming causal proof and challenge remission interval as a potentially unsafe predictor. The novelty is the integration of these checks into one evidence chain. The study is significant because this audit makes the model's limitations visible and provides a reusable framework for safer clinical prediction research in small, heterogeneous datasets.

2. Methodology

2.1 Study Design, Cohort and Outcome

This retrospective study used 120 medical records of patients with SLE managed by the Rheumatology and Nephrology services at Hospital Canselor Tuanku Muhriz (HCTM), Kuala Lumpur [16]. Patients fulfilled the 2019 European League Against Rheumatism/American College of Rheumatology (EULAR/ACR) classification criteria [17]. They also fulfilled the 2012 Systemic Lupus International Collaborating Clinics (SLICC) classification criteria [18]. For patients with multiple relapses within 24 months, the most recent episode was selected. Forty-one patients (34.2%) were classified as relapse and 79 (65.8%) as non-relapse. Table 1 summarises the cohort structure and design safeguards.

Each record used a group-specific reference date: the selected relapse visit for patients classified as relapse and the last recorded clinic visit for those classified as non-relapse. The parent documentation states that demographic and laboratory data were extracted 12 months before relapse for relapse records and at the last recorded clinic visit for non-relapse records [16]. Thus, for relapse records, the fields labelled 12-month C3 status and 12-month anti-dsDNA status refer to status 12 months before the selected relapse date; for non-relapse records, the corresponding status was abstracted at the last recorded visit. The parent report described flare as a measurable increase in disease activity in one or more organ systems with new or worsening clinical signs, symptoms and/or laboratory measurements, but did not document one activity-index threshold for assigning the binary outcome [16]. The source study did not specify a separate future outcome window; this is a retrospective case-referenced design rather than a uniform prospective prediction time point.

Table 1

Cohort structure and design safeguards

Element	Specification	Why it matters
Cohort	120 patients; 41 relapses (34.2%)	Defines event prevalence and limits model complexity
Primary split	Stratified 80:20; random state 42	Preserves relapse prevalence across development and held-out sets
Training set	96 patients	Only source for imputation, transformation, feature selection and tuning
Held-out set	24 patients; 8 relapses and 16 non-relapses	Used for provisional performance and explanation analyses
Leakage exclusion	Remission interval and downstream age/timing fields removed	Prevents outcome-timing information from entering the primary model
Explicit third categories	Category 2 denotes the relevant test/procedure not performed or treatment/medication not received	Preserves a variable-specific care-process state distinct from physiological categories

2.2 Leakage-Aware Data Pipeline

The dataset was split before preprocessing. Numeric missing values were replaced using medians estimated from the training set. For binary clinical variables, unrecorded source cells were encoded as an explicit third category (2): the relevant test or procedure was not performed, or the treatment or medication was not received, depending on the variable. Category 0 and 1 meanings were retained from each variable's data definition rather than assigned one universal interpretation; non-binary categorical gaps used the training mode. Continuous variables with substantial skew and influential extreme observations were transformed according to their distributions: lowest haemoglobin and lowest albumin underwent Box-Cox transformation, while highest C-reactive protein, lowest white-

cell count and lowest neutrophil count were log-transformed. All transformation parameters were learned from the training set and applied unchanged to the held-out set. Extreme observations were retained because they may contain clinically meaningful disease information; transformation reduced their disproportionate influence on model fitting rather than deleting or winsorising them. A one-hot encoded unscaled branch supported tree and EBM models, and a scaled branch supported logistic regression and the multilayer perceptron. Remission interval, age at relapse and other downstream fields were excluded from the primary feature matrix. The complete workflow is shown in Figure 1.

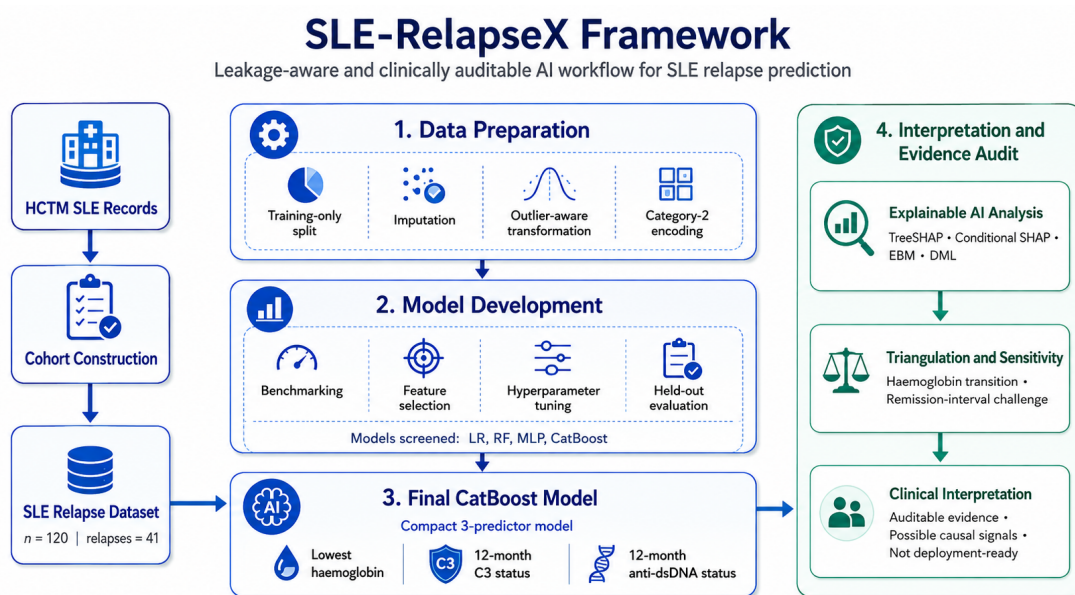


Fig. 1. SLE-RelapseX converts model development into a sequence of explicit evidence challenges

2.3 Model Development and Feature Selection

Four model families were evaluated because they represent different bias-variance and interpretability profiles: least absolute shrinkage and selection operator (L1)-regularised logistic regression, random forest, multilayer perceptron (MLP) and base CatBoost. Feature selection was matched to each model rather than imposed as a universal filter. L1 coefficients selected 10 encoded predictors for logistic regression; Boruta selected lowest haemoglobin for random forest; a two-of-three consensus retained five encoded features for the MLP; and BorutaSHAP retained three source predictors for CatBoost. Initial held-out screening did not identify CatBoost as universally superior: MLP achieved the highest screening F1-score (0.533) and recall (1.000) but specificity was 0.125, whereas base CatBoost achieved precision 0.600, recall 0.375, specificity 0.875 and an F1-score of 0.462. CatBoost was advanced because it offered high precision and specificity with a compact predictor set and natively handled categorical variables [19]. The general SHAP framework supported a coherent downstream explanation audit [20], while its tree-specific implementation enabled efficient CatBoost explanations [21]. This was a methodological selection rather than an equal-budget algorithm victory.

CatBoost hyperparameters were selected using five-fold cross-validation within the 96-patient training set. The final settings were depth 3, learning rate 0.03, L2 leaf regularisation 3 and 300 iterations. The 24-patient held-out set was not used to fit preprocessing parameters or model weights. Because several models and analysis stages were nevertheless compared against the same

small held-out set, all test estimates are described as provisional internal performance rather than definitive validation.

2.4 Multi-Layer Explanation and Threshold Triangulation

Standard TreeSHAP quantified each predictor's contribution to the final CatBoost output using the general SHAP additive-attribution framework [20]. Its tree-ensemble algorithm provided efficient, consistent feature attribution [21]. Because standard feature attribution can redistribute importance when predictors are dependent, the same 24 records were also explained with the shap package's conditional inference tree (CTREE) approach [22]. Conditional inference trees estimate contributions while respecting dependencies between mixed predictors [23]. All eight coalitions for the three-predictor model were evaluated with 1,000 Monte Carlo samples. Additivity was checked by reconstructing the CatBoost raw margin for every held-out record; all 24 residuals were within numerical tolerance.

An Explainable Boosting Machine (EBM) was then trained with the same source predictor groups. EBM is a glass-box generalised additive model that exposes the learned shape of each feature contribution [24]. Its purpose was not external validation because it used the same cohort and held-out records, but independent algorithmic triangulation: a transition was treated as more credible when SHAP and EBM identified a similar region and direction.

2.5 Exploratory Causal Analysis and Leakage Sensitivity

Double/debiased machine learning (DML) was used to explore whether each selected predictor retained evidence compatible with a possible causal relationship with relapse after adjustment for the other measured selected variables. The analysis used five-fold cross-fitted residualisation within the 96-patient training set: CatBoost nuisance models estimated the relapse outcome and each candidate predictor from the remaining candidate covariates in out-of-fold observations. The relationship between outcome and predictor residuals was then estimated using heteroskedasticity-consistent type 3 (HC3) robust standard errors [25]. For binary indicators, estimates approximate adjusted risk-difference contrasts; continuous haemoglobin was restored to its clinical scale before DML. Because the data are observational and the analysis was post-selection, unmeasured confounding, weak overlap, model misspecification and ambiguous timing cannot be excluded. DML results are therefore exploratory evidence of possible causal relationships, not proof of causation.

A secondary sensitivity analysis reintroduced remission interval using the same split, model settings and explanatory tools. This variable was deliberately kept outside the primary model because the duration of remission can be entangled with when the relapse outcome is observed. The sensitivity analysis was designed as a challenge: a useful baseline predictor should improve or preserve performance and produce stable transitions across explanation methods.

3. Results

3.1 Held-Out Predictive Performance

The tuned CatBoost classified 19 of 24 held-out patients correctly: 5 true positives, 14 true negatives, 2 false positives and 3 false negatives. Precision was 0.714, recall 0.625, specificity 0.875, accuracy 0.792, balanced accuracy 0.750 and the F1-score was 0.667. Wilson 95% confidence intervals (CIs) were 0.359-0.918 for precision, 0.306-0.863 for recall, 0.640-0.965 for specificity and 0.595-0.907 for accuracy. These wide intervals are a direct consequence of the small test set and

should accompany, rather than be hidden behind, the point estimates. Table 2 summarises these results, and Figure 2 visualises the trade-offs.

Table 2

Held-out model comparison

Model	Selected predictors*	Precision	Recall	Specificity	F1
CatBoost (tuned)	3	0.714	0.625	0.875	0.667
CatBoost (base)	3	0.600	0.375	0.875	0.462
MLP	5 encoded	0.364	1.000	0.125	0.533
Random forest	1	0.300	0.375	0.563	0.333
Logistic regression	10 encoded	0.167	0.125	0.688	0.143

*Predictor counts reflect model-specific feature-selection spaces. Logistic regression, random forest, MLP and base CatBoost are initial screening benchmarks; tuned CatBoost followed training-only cross-validation. These rows should not be interpreted as an equal-budget hyperparameter contest.

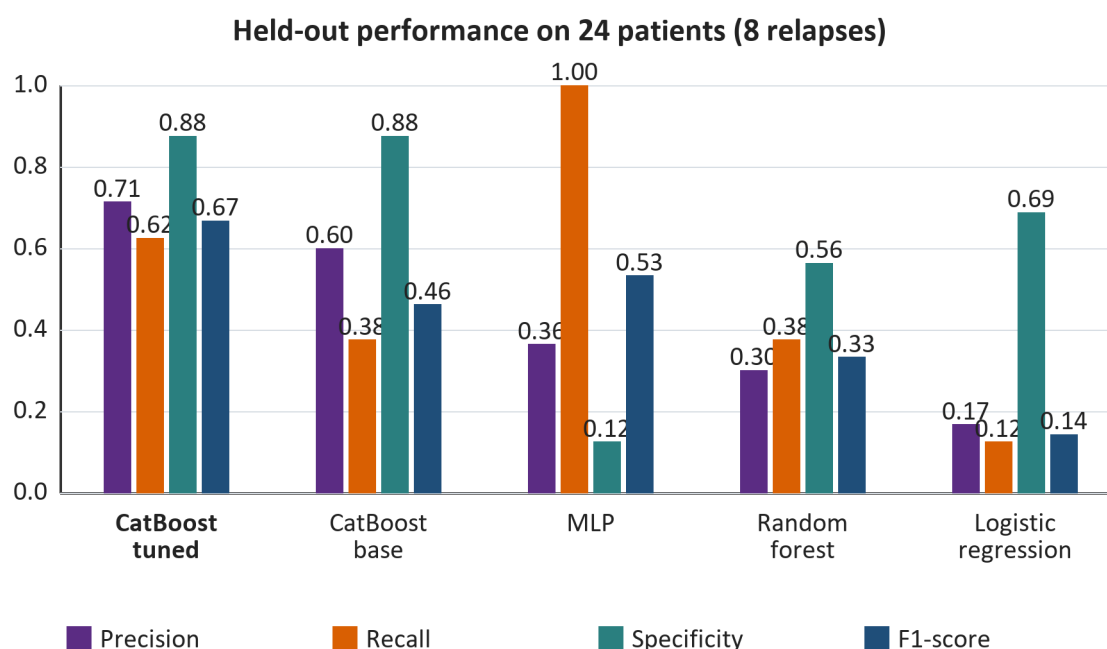


Fig. 2. Initial screening showed complementary trade-offs: MLP maximised recall but produced 14 false positives, while base CatBoost combined precision 0.600 with specificity 0.875. Training-only tuning then improved CatBoost to precision 0.714, recall 0.625 and F1 0.667

3.2 Compact Predictor Set and Explanation Stability

BorutaSHAP retained only three source predictors: lowest haemoglobin, 12-month C3 status and 12-month anti-dsDNA status. This compact set is operationally attractive because all three variables arise from routine lupus monitoring. Standard and Conditional SHAP produced the same order: lowest haemoglobin first, C3 status second and anti-dsDNA status third. Spearman and Kendall rank correlations were both 1.00. Conditional attribution increased the mean absolute contribution of C3 from 0.635 to 0.732 but did not alter feature order, showing rank stability but not magnitude equivalence. Figure 3 compares standard and dependence-aware attribution magnitudes.

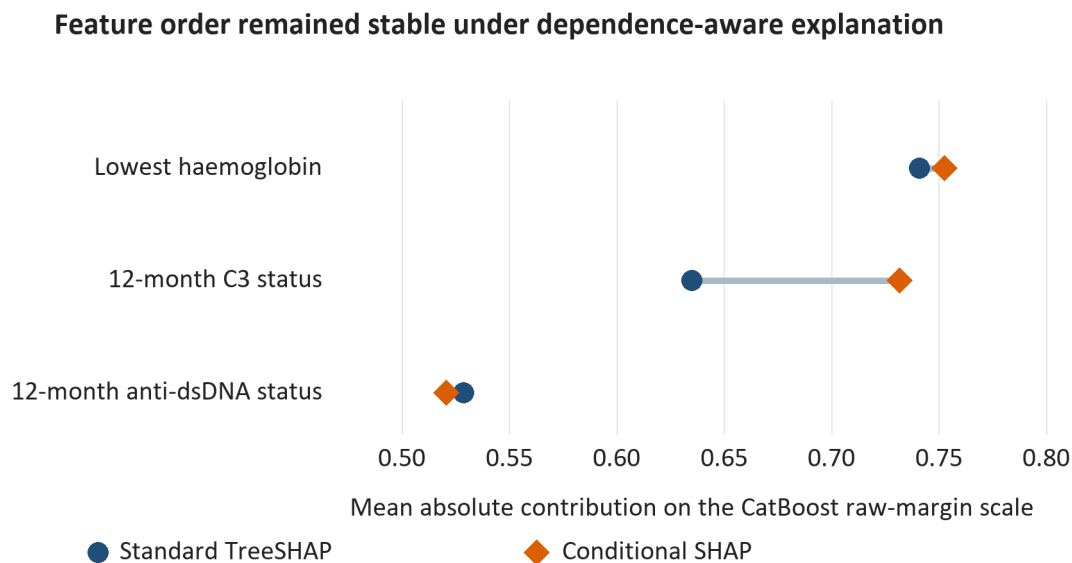


Fig. 3. Dependence-aware SHAP preserved the three-feature order while changing attribution magnitude, particularly for C3 status

3.3 Haemoglobin Transition

Lowest haemoglobin was selected by all four model-specific feature-selection pipelines. Standard SHAP identified an observed sign-change interval of 11.5-12.0 g/dL with a midpoint of 11.75 g/dL. EBM identified its largest interpretable transition at 11.95 g/dL. Both methods assigned higher predicted relapse risk below the transition and lower risk above it; the relative threshold difference was 1.7%. The 11.75-11.95 g/dL band is therefore the strongest internally triangulated threshold-like finding, but it remains model-derived and must not be presented as a treatment cutoff. Figure 4 shows the convergent threshold evidence.

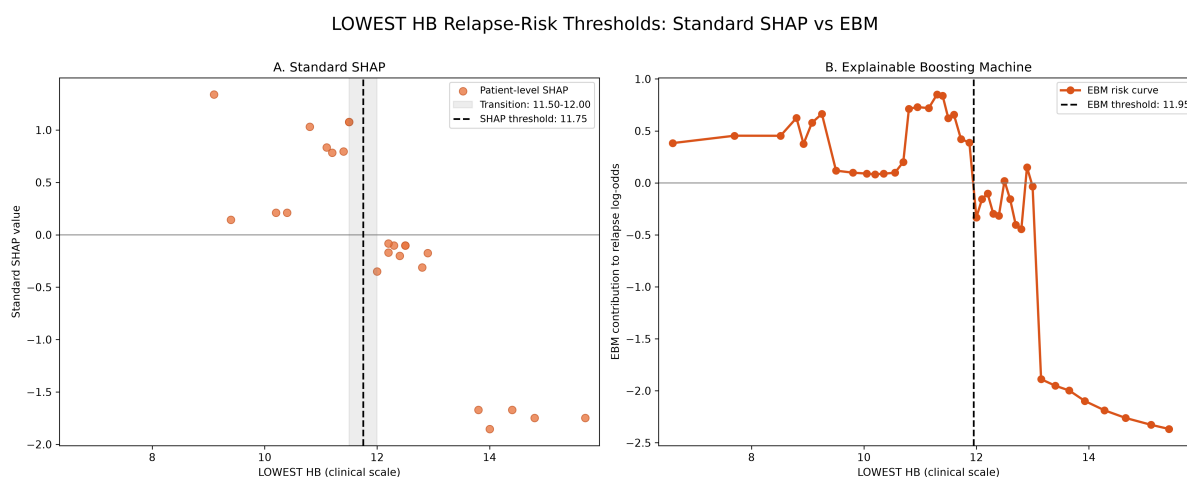


Fig. 4. SHAP and EBM converged on a narrow haemoglobin transition band with the same risk direction

3.4 Predictor-Level Evidence Triangulation

Table 3

Cross-method interpretation of the final predictors

Predictor	Predictive evidence	Exploratory causal evidence	Defensible interpretation
Lowest haemoglobin	Rank 1 in standard and conditional SHAP; SHAP/EBM transition at 11.75-11.95 g/dL	Effect -0.055; 95% CI - 0.122 to 0.012; $p=0.110$	Strong nonlinear predictive marker; no clear DML evidence of an independent causal relationship
Anti-dsDNA positive	Rank 3; positive category contribution	Effect 0.426; 95% CI 0.175 to 0.677; $p<0.001$	Strongest evidence in this study consistent with a possible causal relationship; not causal proof
Low C3	C3 source feature ranked 2	Effect 0.063; 95% CI - 0.183 to 0.308; $p=0.617$	No clear DML evidence of an independent causal relationship
C3 not tested	High EBM importance within the C3 group	Effect -0.238; 95% CI - 0.435 to -0.041; $p=0.018$	Statistically significant testing/care-process relationship; not protective C3 biology

The distinction between low C3 and C3 not tested is critical. In this dataset, category 2 means that the C3 test was not performed; it is an explicit care-process category, not a physiological value. Only 61 of 120 patients underwent 12-month C3 testing in the analysis dataset underlying the earlier cohort manuscript [16]. Category 2 is therefore not a physiological complement state. Its negative DML coefficient must not be described as protective C3 biology; it may instead reflect testing patterns, clinical workflow, patient selection or other care processes.

3.5 Exploratory Causal Findings

Anti-dsDNA positivity produced a positive statistically significant DML estimate (0.426, 95% CI 0.175-0.677; $p<0.001$), providing the strongest evidence in this study consistent with a possible causal relationship with relapse after measured-variable adjustment. This remains exploratory and is not causal proof. Lowest haemoglobin did not show clear DML evidence of an independent causal relationship despite being the strongest predictor. This is not a contradiction: prediction can exploit nonlinear risk patterns, proxy information and interactions even when a selected variable does not retain a stable residual relationship after adjustment. Low C3 likewise showed no clear DML evidence of an independent causal relationship. The C3-not-tested category had a statistically significant negative coefficient ($p=0.018$), but because it denotes that testing was not performed and was not corrected for multiple comparisons, it is interpreted as a possible testing, workflow, selection or care-process relationship rather than biological protection. Figure 5 summarises these exploratory DML estimates and confidence intervals.

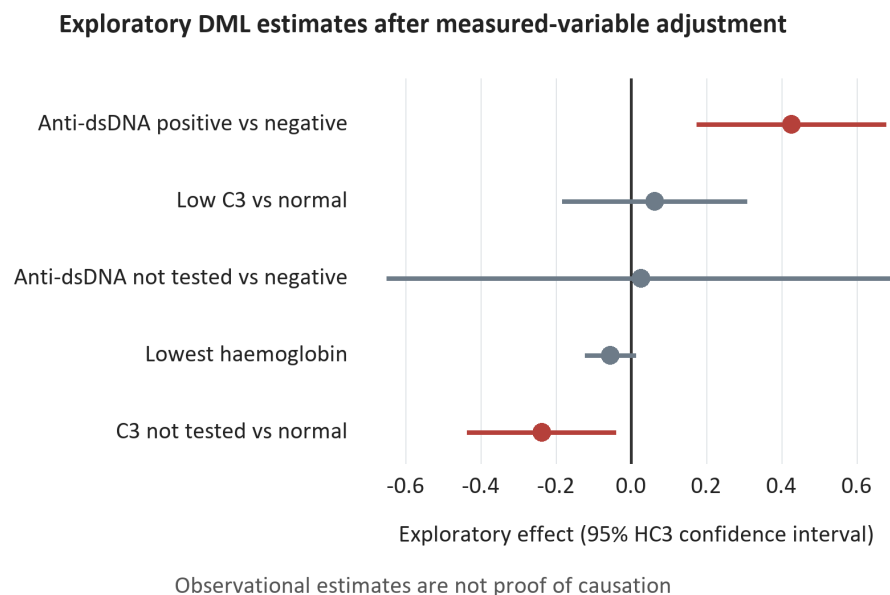


Fig. 5. Exploratory DML evidence was strongest for a possible anti-dsDNA relationship, while the C3-not-tested coefficient may reflect testing or care processes rather than C3 biology

3.6 Remission-Interval Sensitivity Analysis

Reintroducing remission interval did not strengthen the model. Precision fell to 0.500, recall to 0.375 and F1-score to 0.429; the confusion matrix contained 13 true negatives, 3 false positives, 5 false negatives and 3 true positives. Both explanatory models associated shorter remission with higher predicted relapse risk, but SHAP identified a transition near 24 months (observed interval 20-28 months) whereas EBM identified 16 months. The direction agreed but the threshold did not. Performance deterioration and threshold discordance support the original decision to exclude remission interval from the primary model rather than convert it into a clinical rule.

4. Discussion

4.1 Principal Contribution

SLE-RelapseX advances the earlier cohort analysis by converting a compact relapse model into an auditable clinical-AI framework. Its principal contribution is an explicit evidence chain that tests leakage, explanation dependence, threshold reproducibility, exploratory causal evidence and temporal sensitivity within a single workflow. RelapseX prevents major preprocessing leakage, tests explanations under predictor dependence, asks an interpretable model to reproduce the principal threshold, uses DML to explore possible causal relationships under explicit assumptions, and rejects a clinically attractive variable when its timing and sensitivity behaviour are unsafe.

This design responds directly to the concern that black-box performance is insufficient in high-stakes healthcare [26]. It also demonstrates that interpretability is not a single plot. SHAP answered which variables drove CatBoost predictions; Conditional SHAP tested whether their order depended on unrealistic independence; EBM asked whether a transparent model learned a similar risk transition; DML explored whether predictors retained evidence compatible with possible causal relationships after measured-variable adjustment, without treating those estimates as proof; and the remission sensitivity analysis attempted to falsify, rather than merely confirm, the preferred model.

4.2 Clinical Interpretation

The three final predictors are clinically plausible. A focused review found repeated associations between anti-dsDNA and disease activity or flare risk [3]. A prospective follow-up study separately identified predictors of severe flare [4]. Haematological abnormalities are common in active SLE, and lower haemoglobin may integrate inflammation, renal involvement, haemolysis, marrow suppression, treatment burden or chronic disease. Complement consumption is biologically relevant, but the present study does not show a clear independent low-C3 signal. Biomarkers can therefore be informative at the population level while remaining inconsistent for individual relapse prediction.

The haemoglobin transition deserves prospective testing because two different explanatory models converged within 0.20 g/dL. It should nevertheless be treated as a hypothesis-generating monitoring region. A clinician could reasonably view a haemoglobin value below approximately 12 g/dL as a prompt to inspect trajectory, renal status, haemolysis, inflammation and treatment context, but RelapseX does not establish that correcting haemoglobin will prevent relapse.

4.3 Evolution Beyond the Predecessor Model

Table 4

What RelapseX changes relative to the earlier cohort analysis

Dimension	Earlier analysis [16]	SLE-RelapseX	Why the change matters
Primary predictors	Remission interval, haemoglobin, immunosuppressant use, disease duration	Haemoglobin, 12-month C3 status, 12-month anti-dsDNA status	Removes an outcome-timing variable and reduces the model to routine surveillance data
Held-out profile	Accuracy 0.79; precision 0.64; recall 0.88; F1 0.74	Accuracy 0.792; precision 0.714; recall 0.625; specificity 0.875; F1 0.667	Shows a shift toward fewer false positives, not universal performance superiority
Explanation	Standard SHAP	Standard plus Conditional SHAP; complete additivity audit	Tests attribution robustness under feature dependence
Threshold evidence	SHAP patterns	SHAP plus EBM convergence	Separates an isolated plot from cross-model triangulation
Causal assessment	Conventional association and model importance	Exploratory DML for possible causal relationships with explicit assumption limits	Distinguishes prediction from evidence compatible with causation without claiming proof
Leakage challenge	Remission interval retained	Remission interval analysed separately and rejected	Makes temporal deployability part of the model-selection decision

4.4 Innovation and Translational Value

The competition value of RelapseX rests on four characteristics. First, it is feasible: the final model uses three routinely available variables rather than expensive omics or imaging. Second, it is transparent: every headline predictor is accompanied by an explanation-stability result and an evidence grade. Third, it is self-critical: remission interval and the C3-not-tested care-process category are reported as threats to validity rather than converted into promotional claims. Fourth, it is reusable: the same audit sequence can be transferred to other small clinical datasets in which leakage, incomplete measurement and correlation are more dangerous than algorithmic complexity.

A safe translational pathway has four gates. Gate 1 is a locked, clinically defined prediction time point with prospective data collection. Gate 2 is nested internal validation with calibration, threshold

uncertainty and decision-curve analysis. Gate 3 is external multicentre validation with subgroup assessment across sex, ethnicity, organ involvement and treatment strata. Gate 4 is human-in-the-loop implementation in which preprocessing, category semantics and model versioning are packaged together and predictions support, rather than replace, specialist judgment. This pathway aligns the project with United Nations Sustainable Development Goal 3 (SDG 3): Good Health and Well-Being while avoiding premature automation.

4.5 Limitations

The study is retrospective, single-centre and small. Predictor anchoring differed between relapse and non-relapse records, and the source study did not specify a uniform future outcome window; performance therefore represents retrospective discrimination under a case-referenced design, not validated real-time forecasting. Only 24 patients, including 8 relapses, were available for held-out evaluation; performance intervals are consequently wide. The held-out set was used across several model and explanation comparisons, which can create optimistic selection if repeated iteration responds to test results. No external validation, calibration curve, area under the receiver operating characteristic curve (ROC-AUC), decision-curve analysis or net-benefit evaluation was available in the consolidated outputs. The cohort is predominantly female and drawn from one Malaysian tertiary centre, limiting transportability.

Testing and care processes are not neutral. The importance of the C3 source feature is partly driven by the explicit C3-not-tested category, which may encode clinical workflow, patient selection or care processes rather than disease biology. Standard and Conditional SHAP rank agreement is based on only three final predictors and does not quantify resampling uncertainty. EBM is an independent algorithmic check, not an independent dataset. DML was performed after feature selection on 96 observations, without a prespecified causal graph, a broad confounder set, overlap diagnostics or multiplicity correction; unmeasured confounding and model misspecification remain possible, so its estimates are exploratory evidence rather than proof of causation. Finally, the current saved CatBoost artefact does not encapsulate the haemoglobin transformation and category dictionary, so direct entry of raw clinical values would be unsafe until preprocessing is packaged and verified.

5. Conclusions

SLE-RelapseX produced a compact three-predictor CatBoost model with provisional held-out precision 0.714, recall 0.625, specificity 0.875 and F1-score 0.667. Lowest haemoglobin was the most stable predictive marker, with SHAP and EBM converging on a model-derived transition around 11.75-11.95 g/dL, but DML did not provide clear evidence of an independent causal relationship. Anti-dsDNA positivity provided the strongest exploratory evidence in this study consistent with a possible causal relationship after measured-variable adjustment, whereas low C3 did not. The significant negative C3-not-tested coefficient was interpreted as a testing or care-process relationship rather than protective biology. Remission interval failed the leakage sensitivity challenge and remained excluded. The principal innovation is therefore not a claim of clinical readiness or causal proof, but a reusable audit framework that converts prediction into transparent, contestable evidence. Prospective multicentre validation, calibration, clinical-utility testing and end-to-end preprocessing are required before deployment.

Acknowledgement

The authors acknowledge the Department of Medicine and the Rheumatology and Nephrology Clinics of Hospital Canselor Tuanku Muhriz for supporting the parent clinical record study, and the Center for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia, for analytical collaboration. This research was not funded by any grant.

Ethics and Data Governance

The parent retrospective study received ethical approval from the Research Ethics Committee of Universiti Kebangsaan Malaysia (UKM; JEP-2025-190), with institutional approval from the Faculty of Medicine UKM and Hospital Canselor Tuanku Muhriz under project code FF-2025-157. This secondary analysis used the resulting de-identified dataset, and no attempt was made to re-identify participants.

References

- [1] Tsokos, George C. 2011. "Systemic Lupus Erythematosus." *New England Journal of Medicine* 365: 2110-2121. <https://doi.org/10.1056/NEJMr1100359>
- [2] McElhone, Kathleen, Janice Abbott, Margaret Hurley, Jane Burnell, Peter Lanyon, Anisur Rahman, Chee-Seng Yee et al. "Flares in patients with systemic lupus erythematosus." *Rheumatology* 60, no. 7 (2021): 3262-3267. <https://doi.org/10.1093/rheumatology/keaa777>
- [3] Andrade, Samuel de Oliveira, Paulo Rogerio Julio, Diego Nunes de Paula Ferreira, and Simone Appenzeller. "Predicting lupus flares: epidemiological and disease related risk factors." *Expert Review of Clinical Immunology* 17, no. 2 (2021): 143-153. <https://doi.org/10.1080/1744666X.2020.1865156>
- [4] Fatemi, Alimohammad, Elaheh Keivani-Boroujeni, and Abbas Smiley. "Predictors of severe lupus flare: a prospective follow-up study." *Bmc Rheumatology* 7, no. 1 (2023): 10. <https://doi.org/10.1186/s41927-023-00333-y>
- [5] Kakati, Sanjeeb, Runjun Teronpi, and Bhupen Barman. "Frequency, pattern and determinants of flare in systemic lupus erythematosus: A study from North East India." *The Egyptian Rheumatologist* 37, no. 4 (2015): S55-S59. <https://doi.org/10.1016/j.ejr.2015.08.002>
- [6] Ji, Lanlan, Wenhui Xie, Serena Fasano, and Zhuoli Zhang. "Risk factors of flare in patients with systemic lupus erythematosus after glucocorticoids withdrawal. A systematic review and meta-analysis." *Lupus science & medicine* 9, no. 1 (2022): e000603. <https://doi.org/10.1136/lupus-2021-000603>
- [7] Ramirez, Giuseppe A., Chiara Calabrese, Marta Secci, Luca Moroni, Gabriele D. Gallina, Giovanni Benanti, Enrica P. Bozzolo, Marco Matucci-Cerinic, and Lorenzo Dagna. "Infection-associated flares in systemic lupus erythematosus." *Pathogens* 13, no. 11 (2024): 934. <https://doi.org/10.3390/pathogens13110934>
- [8] Li, He-Jun, Yi-Qing Zheng, Ling Chen, Shun-Ping Lin, and Xiang-Xiong Zheng. "Risk factors of significant relapse and appropriate maintenance therapy strategy in SLE-associated immune thrombocytopenia." *Therapeutic Advances in Chronic Disease* 14 (2023): 20406223231160688. <https://doi.org/10.1177/20406223231160688>
- [9] Ceccarelli, Fulvia, Francesco Natalucci, Licia Picciariello, Claudia Ciancarella, Giulio Dolcini, Angelica Gattamelata, Cristiano Alessandri, and Fabrizio Conti. "Application of machine learning models in systemic lupus erythematosus." *International Journal of Molecular Sciences* 24, no. 5 (2023): 4514. <https://doi.org/10.3390/ijms24054514>
- [10] Zhan, Kevin, Katherine A. Buhler, Irene Y. Chen, Marvin J. Fritzler, and May Y. Choi. "Systemic lupus in the era of machine learning medicine." *Lupus science & medicine* 11, no. 1 (2024): e001140. <https://doi.org/10.1136/lupus-2023-001140>
- [11] Collins, Gary S., Karel GM Moons, Paula Dhiman, Richard D. Riley, Andrew L. Beam, Ben Van Calster, Marzyeh Ghassemi et al. "TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods." *bmj* 385 (2024). <https://doi.org/10.1136/bmj-2023-078378>
- [12] Moons, Karel GM, Johanna AA Damen, Tabea Kaul, Lotty Hooft, Constanza Andaur Navarro, Paula Dhiman, Andrew L. Beam et al. "PROBAST+ AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods." *bmj* 388 (2025). <https://doi.org/10.1136/bmj-2024-082505>
- [13] Van Calster, Ben, David J. McLernon, Maarten Van Smeden, Laure Wynants, Ewout W. Steyerberg, and Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative Bossuyt Patrick Collins Gary S.

- Macaskill Petra McLernon David J. Moons Karel GM Steyerberg Ewout W. Van Calster Ben van Smeden Maarten Vickers Andrew J. "Calibration: the Achilles heel of predictive analytics." *BMC medicine* 17, no. 1 (2019): 230. <https://doi.org/10.1186/s12916-019-1466-7>
- [14] Vickers, Andrew J., and Elena B. Elkin. "Decision curve analysis: a novel method for evaluating prediction models." *Medical Decision Making* 26, no. 6 (2006): 565-574. <https://doi.org/10.1177/0272989X06295361>
- [15] Steyerberg, Ewout W., Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J. Pencina, and Michael W. Kattan. "Assessing the performance of prediction models: a framework for some traditional and novel measures." *Epidemiology (Cambridge, Mass.)* 21, no. 1 (2010): 128. <https://doi.org/10.1097/EDE.0b013e3181c30fb2>
- [16] Norazilee, Nur Ayesah Ameerah, Nurthani Nayli Mohd Lizam, Regina Swaaty Ramis, Thipashini Mahendran, Jonathan Xin Rong Lau, Afiq Nur Aiman Abdul Mustafa Kamal, Muhammad Izzul Islam Faisal, Shahnorbanun Sahran, Lim Kuan Yee, Azmawati Mohammed Nawawi, and Syahrul Sazliyana Shahrir. 2026. "Machine Learning Model of Factors Associated with Systemic Lupus Erythematosus Relapse." *Medicine & Health* 21 (2 Suppl): 9. <https://doi.org/10.17576/MH.2026.s2102>.
- [17] Aringer, Martin, Karen Costenbader, David Daikh, Ralph Brinks, Marta Mosca, Rosalind Ramsey-Goldman, Josef S. Smolen et al. "2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus." *Arthritis & rheumatology* 71, no. 9 (2019): 1400-1412. <https://doi.org/10.1002/art.40930>
- [18] Petri, Michelle, Ana-Maria Orbai, Graciela S. Alarcón, Caroline Gordon, Joan T. Merrill, Paul R. Fortin, Ian N. Bruce et al. "Derivation and validation of the Systemic Lupus International Collaborating Clinics classification criteria for systemic lupus erythematosus." *Arthritis & Rheumatism* 64, no. 8 (2012): 2677-2686. <https://doi.org/10.1002/art.34473>
- [19] Prokhorenkova, Liudmila, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. "CatBoost: Unbiased Boosting with Categorical Features." *Advances in Neural Information Processing Systems* 31: 6638-6648. <https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html>
- [20] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).
- [21] Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." *arXiv preprint arXiv:1802.03888* (2018).
- [22] Sellereite, Nikolai, and Martin Jullum. "shapr: An R-package for explaining machine learning models with dependence-aware Shapley values." *Journal of Open Source Software* 5, no. 46 (2020): 2027. <https://doi.org/10.21105/joss.02027>
- [23] Redelmeier, Annabelle, Martin Jullum, and Kjersti Aas. 2020. "Explaining Predictive Models with Mixed Features Using Shapley Values and Conditional Inference Trees." *arXiv:2007.01027*. <https://arxiv.org/abs/2007.01027>
- [24] Nori, Harsha, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. "InterpretML: A Unified Framework for Machine Learning Interpretability." *arXiv:1909.09223*. <https://arxiv.org/abs/1909.09223>
- [25] Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. "Double/debiased machine learning for treatment and structural parameters." (2018): C1-C68. <https://doi.org/10.1111/ectj.12097>
- [26] Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature machine intelligence* 1, no. 5 (2019): 206-215. <https://doi.org/10.1038/s42256-019-0048-x>