International Journal of Management and Data Analytics
IJMADA

Vol 6 (1), 283-296
ISSN: 2816-9395
DOI: https://doi.org/10.5281/zenodo.22102794
Journal website: http://ijmada.com

Comparative Evaluation of Multilayer Perceptron, Random Forest, and Support Vector Machine for Sepsis Classification Using Longitudinal Structured Clinical Data

Author(s): Olalekan Akinbosoye Okewale[1], Esther Oduntan[2], Chibuzo Benjamin Onuike[3], Olufemi S. Ojo[1], Sholanke Oluwafunmbi Benedict[1]

1 Department of Computer Science, Ajayi Crowther University, Oyo, Nigeria.
2 Department of Computer Science, Federal University of Technology, Ilaro, Nigeria.
3 Department of Mechanical Engineering, Federal University of Technology, Owerri, Nigeria.

Corresponding Author: okewale.akinbosoye@gmail.com

Received: June, 2026 Published: Aug, 2026

ARTICLE INFO
Keywords:
Clinical decision support ; Machine learning; Sepsis classification; SHAP

ABSTRACT
Sepsis classification from routinely collected clinical data is challenging because positive labels are uncommon, repeated observations from the same patient are correlated, and leakage-prone preprocessing can produce misleadingly high performance. This study re-evaluated multilayer perceptron (MLP), random forest (RF), and support vector machine (SVM) models using a leakage-controlled patient-level design. The source file contained 1,552,210 hourly observations from 40,336 patients; 2,932 patients (7.27%) were ever sepsis-positive. A reproducible stratified sample of 10,000 patients (382,387 hourly observations) was selected, preserving patient-level prevalence, and divided into 8,000 development and 2,000 independent hold-out patients with no patient overlap. Nine routinely available predictors were retained after a development-only missingness audit: heart rate, oxygen saturation, temperature, systolic blood pressure, mean arterial pressure, diastolic blood pressure, respiratory rate, age, and gender. Median imputation, missingness indicators, scaling, imbalance handling, hyperparameter tuning, and threshold selection were confined to development data and patient-grouped cross-validation. On the untouched hold-out cohort, RF achieved the highest ROC-AUC (0.671; 95% CI 0.635–0.706) and PR-AUC (0.041), with sensitivity 0.538 and specificity 0.716. RF significantly exceeded MLP in ROC-AUC (difference 0.0317, p=0.012), whereas RF and SVM did not differ significantly (p=0.054). Calibrated Brier scores were similar (0.0175–0.0176). Permutation importance and SHAP identified heart rate, respiratory rate, temperature, and blood-pressure measures as influential RF predictors. The models show moderate discrimination but low positive predictive value under severe class imbalance; therefore, they should be regarded as feasibility-level decision-support models requiring external, prospective, temporal, and fairness validation before clinical use.

1.0 Introduction

Sepsis is life-threatening organ dysfunction caused by a dysregulated host response to infection and remains a major cause of preventable morbidity and mortality (Singer et al., 2016). Global estimates indicate approximately 48.9 million cases and 11 million sepsis-related deaths in 2017, with a disproportionate burden in low- and middle-income settings where delayed presentation, limited laboratory capacity, and constrained critical-care resources may impede timely escalation (Rudd et al., 2020; World Health Organization [WHO], 2024). Because clinical deterioration can occur rapidly, reliable recognition of sepsis-related physiological patterns remains an important clinical and health-data problem.

Bedside screening commonly draws on organ-dysfunction scores and physiological warning systems such as SOFA, qSOFA, SIRS, NEWS, and MEWS. These tools are useful because they are interpretable and inexpensive, but they apply fixed rules or thresholds and may trade sensitivity against specificity across populations. Machine learning (ML) offers a complementary approach by learning multivariable and non-linear relationships from routinely collected observations. However, published sepsis models have also been criticised for data leakage, inconsistent prediction horizons, incomplete calibration assessment, weak external validation, and limited interpretability (Collins et al., 2024; Do et al., 2026; Yadgarov et al., 2024).

The methodological gap addressed in this study is therefore not simply whether ML can classify sepsis, but whether accessible algorithms retain useful performance when evaluated under a patient-level, leakage-controlled design with severe class imbalance and an untouched hold-out cohort. Three model families were selected because they represent complementary approaches to structured clinical data: MLP for non-linear neural modelling, RF for robust ensemble learning and interpretability, and linear SVM for margin-based classification at large sample size.

The objective was to compare MLP, RF, and SVM using a common longitudinal dataset and a reproducible workflow that prevents patient overlap, confines preprocessing and imbalance handling to development data, reports uncertainty and calibration, and provides model explanations. The supplied hourly SepsisLabel was used as the binary outcome without imposing an additional lead-time shift. Consequently, the contribution is framed as classification of the source sepsis label rather than as proof of a newly defined prospective lead-time prediction system.

2.0 Literature Review

2.1 Clinical context and conventional screening

Sepsis is heterogeneous: infection may trigger inflammatory dysregulation, endothelial injury, coagulation abnormalities, tissue hypoperfusion, and progressive organ dysfunction (Santacroce et al., 2024). Clinical recognition therefore relies on combinations of vital signs, organ-function measures, laboratory results, and clinical judgment. Conventional scores such as SOFA and qSOFA provide transparent bedside summaries, while SIRS, NEWS, and MEWS capture physiological derangement. Their transparency is advantageous, but fixed thresholds may not fully represent complex interactions among changing clinical variables. A direct head-to-head score comparison was not performed in the present dataset because the complete variables required to reconstruct all scores consistently were not available; this is identified as a priority for external validation rather than inferred from incomplete data.

2.2 Machine learning for sepsis

A growing literature has evaluated neural networks, tree ensembles, support-vector methods, gradient boosting, and time-series architectures for sepsis. Goh et al. (2021) demonstrated the value of combining structured and unstructured information, while Tang et al. (2024) used a Transformer-based time-series approach. Zhou et al. (2024) combined imbalance handling and explainability and reported competitive tree-based performance. Lin et al. (2025) emphasised the utility of routinely available blood-count information, and Liu et al. (2025) demonstrated the importance of interpretable models in emergency triage. A systematic review and network meta-analysis by Yadgarov et al. (2024) found promising discrimination across ML approaches but also substantial heterogeneity in study design and validation.

2.3 Methodological validity, calibration, and explainability

High discrimination is not sufficient for clinical decision support. Predictions must be evaluated using an analysis design that matches the intended use, prevents leakage

between patients or time points, quantifies uncertainty, and assesses whether predicted probabilities correspond to observed risk. Do et al. (2026) showed that evaluation strategy can materially alter apparent sepsis-model performance. TRIPOD+AI similarly emphasises transparent reporting of data provenance, outcome definition, missing-data handling, validation, calibration, and intended use (Collins et al., 2024). Explainability is also important because clinicians need to understand which observations drive alerts; SHAP and permutation importance provide complementary global and local views of model behaviour (Bomrah et al., 2024; Liu et al., 2025).

Table 1: Selected related studies and relevance to the present work

Study	Approach	Main contribution	Relevance/limitation
Goh et al. (2021)	AI using structured/unstructured data	Demonstrated multimodal value for sepsis risk	Different data modalities and deployment context
Tang et al. (2024)	Transformer time-series model	Dynamic sequential modelling	Not directly comparable with the present compact tabular models
Zhou et al. (2024)	ML + imbalance processing + SHAP	Tree-based performance and explainability	Supports explicit imbalance and interpretability analysis
Yadgarov et al. (2024)	Systematic review/meta-analysis	Promising ML discrimination; reporting heterogeneity	Motivates cautious interpretation and stronger validation
Do et al. (2026)	Evaluation-strategy study	Metrics vary with evaluation design	Supports patient/time-aware validation and clear outcome framing
Present study	MLP, RF, linear SVM	Leakage-controlled patient split, calibration, CIs, SHAP	Single public source; external validation still required

3.0 Methodology

3.1 Study design and data provenance

A retrospective quantitative machine-learning design was used. The source dataset was obtained from the public Kaggle "Prediction of Sepsis" repository (Hussaini, n.d.). The file was loaded programmatically rather than through spreadsheet software to address the reviewer concern that a value near the Excel row limit could indicate truncation. The audited source contained 1,552,210 hourly observations from 40,336 unique patients, including 27,916 sepsis-positive hourly labels (1.80%). At patient level, 2,932 patients (7.27%) were ever sepsis-positive and 37,404 (92.73%) were never positive.

To make repeated experimentation computationally feasible while preserving patient structure, a reproducible stratified sample of 10,000 patients was selected according to whether a patient ever had a positive SepsisLabel. The resulting cohort contained 9,273 non-sepsis and 727 ever-sepsis-positive patients (92.73% and 7.27%, respectively) and 382,387 hourly observations. Patients contributed a median of 38 hourly observations (mean 38.24, range 8–336).

3.2 Outcome definition and patient-level partitioning

The binary outcome was the hourly SepsisLabel supplied by the source dataset (0 = negative; 1 = positive). The analysis did not derive a new clinical sepsis definition from downstream outcomes and did not apply a second temporal shift to the source label. This avoids creating an unsupported lead-time claim. For stratification, each patient was classified as ever positive if any hourly SepsisLabel equalled 1. The 10,000 patients were split 80:20 at patient level into 8,000 development and 2,000 independent hold-out patients. The

development cohort contained 582 ever-positive patients and 7,418 never-positive patients; the hold-out cohort contained 145 ever-positive and 1,855 never-positive patients. No patient appeared in both cohorts.

3.3 Predictor audit, missing data, and leakage control

Feature decisions were made using the development cohort only. Identifiers, the outcome, explicit time/length-of-stay variables (Hour, ICULOS, HospAdmTime), and care-unit administrative indicators were excluded. Variables with more than 80% missingness in development data were excluded from the primary model. This removed sparsely measured laboratory variables such as lactate, glucose, WBC, creatinine, platelets, and related tests from the primary predictor set rather than imputing overwhelmingly absent measurements. Nine routinely available variables were retained: HR, O2Sat, Temp, SBP, MAP, DBP, Resp, Age, and Gender.

Table 2: Retained predictors and development-cohort missingness

Predictor	Clinical interpretation	Missing (%)
Heart rate (HR)	Cardiovascular stress/tachycardia	9.60
Oxygen saturation (O2Sat)	Oxygenation status	12.84
Temperature (Temp)	Fever/hypothermia	66.19
Systolic BP (SBP)	Arterial pressure/circulatory status	14.35
Mean arterial pressure (MAP)	Perfusion pressure	12.35
Diastolic BP (DBP)	Arterial pressure	31.51
Respiratory rate (Resp)	Respiratory/metabolic stress	14.99
Age	Demographic risk factor	0.00
Gender	Demographic variable	0.00

Within every training fold, missing values were replaced using median imputation and missingness indicators were added. StandardScaler was fitted inside training folds for MLP and SVM; RF was not scaled. No post-outcome variables such as mortality, dialysis, acute kidney injury, or intubation were used. Outlier values were audited but were not removed solely because they were extreme, since critical illness can produce physiologically unusual observations and indiscriminate trimming could remove genuine high-risk patterns.

3.4 Cross-validation, tuning, and imbalance strategy

Hyperparameters were tuned on the 8,000-patient development cohort using three-fold StratifiedGroupKFold grouped by Patient_ID and RandomizedSearchCV with eight candidates per model. The primary tuning metric was ROC-AUC; PR-AUC, recall, and F1 were also recorded. All preprocessing steps were fitted within each training fold. The initial tuning pipeline used SMOTE with sampling_strategy=0.20 inside training folds only. The hold-out cohort was not used for imputation, scaling, SMOTE, tuning, threshold selection, or calibration fitting.

The best MLP configuration used one hidden layer with 32 units, ReLU activation, Adam optimization, learning rate 0.0005, batch size 512, alpha 0.0001, early stopping with a 10% internal validation fraction, and a maximum of 150 iterations. The best RF used 200 trees, maximum depth 12, min_samples_split=10, min_samples_leaf=5, and max_features="log2". The computationally scalable SVM was LinearSVC with C=0.01, squared-hinge loss, no class weighting, dual="auto", and maximum 5,000 iterations.

After tuning, imbalance strategies were compared using patient-grouped out-of-fold development predictions: no resampling, SMOTE ratios 0.20, 0.50, and 1.00, plus balanced class weighting where applicable. The final strategy was chosen by discrimination among

configurations meeting the recall requirement. MLP used SMOTE 0.20, RF used no SMOTE, and SVM used SMOTE 0.50. Operating thresholds were fixed from development out-of-fold predictions by maximising F2 among thresholds with recall at least 0.50: 0.159914 for MLP, 0.017969 for RF, and -0.276308 for SVM decision scores.

3.5 Evaluation, uncertainty, calibration, and explainability

Final evaluation was performed once on the untouched hold-out cohort. Metrics included accuracy, precision, sensitivity/recall, specificity, F1, ROC-AUC, PR-AUC, and confusion matrices. Ninety-five percent confidence intervals were estimated using 1,000 patient-cluster bootstrap resamples so that repeated hourly observations from the same patient remained clustered. Paired McNemar tests compared thresholded classification errors. Pairwise ROC-AUC differences were evaluated using paired patient-cluster bootstrap intervals and two-sided bootstrap p-values.

For probability calibration, Platt-style logistic calibration models were fitted to development out-of-fold scores only and then frozen before application to the hold-out scores. Brier score, calibration intercept, and calibration slope were reported. Decision-curve analysis was conducted over low threshold probabilities appropriate to the rare positive outcome. For RF explainability, permutation importance quantified the decrease in hold-out ROC-AUC after shuffling each predictor, and SHAP was used for global feature attribution and a patient-level true-positive explanation.

3.6 Ethical consideration and reproducibility

The study used anonymised, publicly accessible secondary data and involved no direct participant contact or intervention. The analysis was implemented in Python/Google Colab with a fixed random state of 42. Six sequential notebooks were used for source audit, patient-level splitting, feature audit, tuning, imbalance/threshold selection, hold-out evaluation, calibration, statistical comparison, and explainability. The notebooks and derived analysis files are retained by the authors and can be supplied to the journal or readers on reasonable request; a permanent public repository should be created before final publication if required by the journal.

4.0 Analysis and Discussions

4.1 Cohort characteristics and cross-validation performance

The sampled cohort preserved the source patient-level prevalence: 727 of 10,000 patients (7.27%) were ever sepsis-positive. The development cohort contained 305,719 hourly observations and the hold-out cohort 76,668. Row-level positive-label prevalence was 1.816% in development and 1.797% in hold-out, demonstrating severe class imbalance despite patient-level stratification.

Table 3: Best three-fold patient-grouped cross-validation performance during tuning

Model	ROC-AUC mean±SD	PR-AUC mean±SD	Recall mean±SD	F1 mean±SD
MLP	0.6102 ± 0.0112	0.0329 ± 0.0005	0.0500 ± 0.0062	0.0577 ± 0.0027
RF	0.6231 ± 0.0123	0.0321 ± 0.0026	0.0000 ± 0.0000	0.0000 ± 0.0000
SVM	0.6495 ± 0.0076	0.0378 ± 0.0031	0.0020 ± 0.0020	0.0038 ± 0.0038

These default-threshold cross-validation recall values were poor, illustrating why threshold choice is consequential in rare-event screening. Development-only threshold optimisation increased sensitivity to approximately 0.50 while preserving moderate specificity. Importantly, imbalance handling was not uniformly beneficial: RF performed best without SMOTE, whereas MLP selected SMOTE 0.20 and SVM selected SMOTE 0.50. This finding argues against assuming that synthetic oversampling necessarily improves every model.

4.2 Independent hold-out performance

Table 4 presents the definitive hold-out results using the fixed development strategies and thresholds. RF achieved the highest ROC-AUC (0.6707) and PR-AUC (0.0409), as well as the highest precision and F1. MLP achieved the highest sensitivity (0.5581), whereas SVM achieved marginally the highest specificity (0.7161). The ROC curves (Fig. 1) show modest discrimination for all three models, with RF consistently providing the strongest overall ranking performance. The precision-recall curves (Fig. 2) are particularly informative under the severe row-level class imbalance: all three models remained above the prevalence baseline over useful portions of the recall range, but absolute precision was low. The low precision of all models (2.86–3.35%) reflects the 1.80% row-level positive prevalence and indicates a substantial false-alert burden at operating points designed to detect approximately half of positive observations.

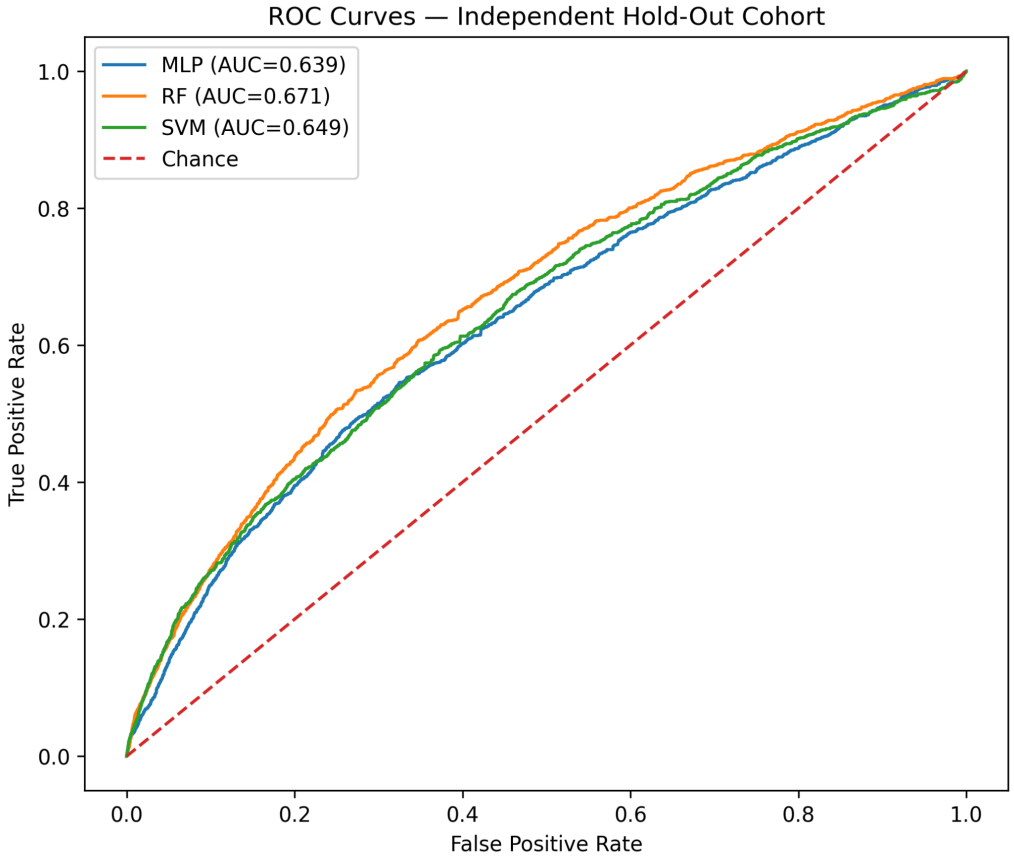

Fig. 1 Receiver operating characteristic curves for MLP, RF, and SVM on the independent hold-out cohort.

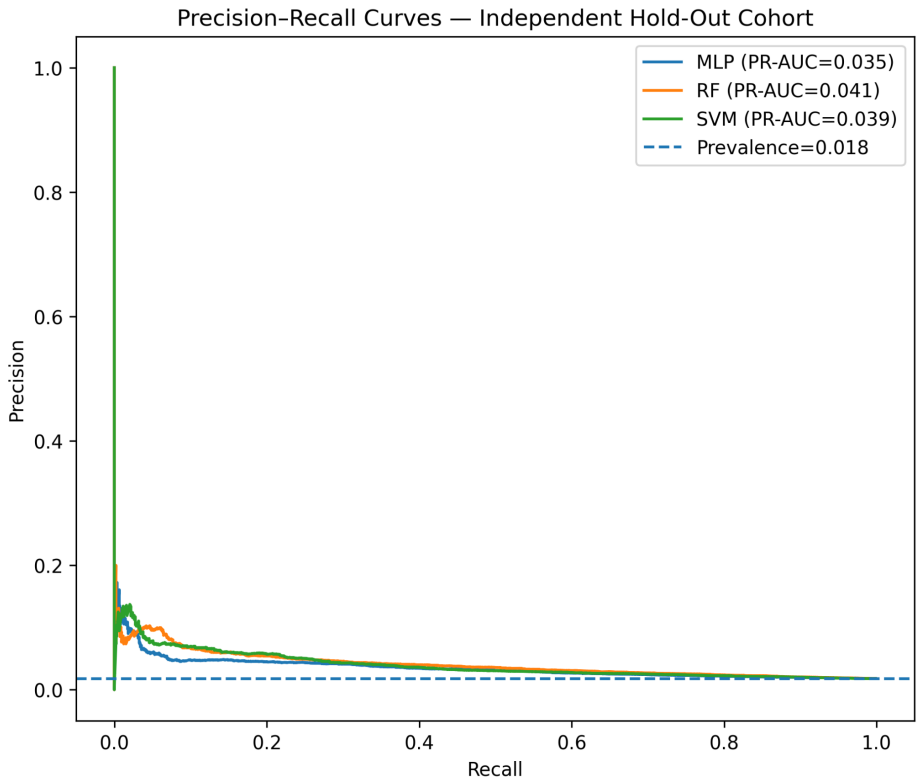

Fig. 2 Precision-recall curves for MLP, RF, and SVM on the independent hold-out cohort; the dashed line indicates the row-level positive prevalence (0.018).

Table 4: Independent hold-out performance with patient-cluster bootstrap 95% confidence intervals

Metric	MLP	RF	SVM
Accuracy	0.6520 (0.6382–0.6638)	0.7125 (0.6987–0.7256)	0.7120 (0.6978–0.7263)
Precision	0.0286 (0.0237–0.0336)	0.0335 (0.0277–0.0395)	0.0306 (0.0254–0.0361)
Sensitivity	0.5581 (0.5023–0.6100)	0.5377 (0.4776–0.5940)	0.4898 (0.4293–0.5474)
Specificity	0.6537 (0.6394–0.6659)	0.7157 (0.7016–0.7293)	0.7161 (0.7011–0.7308)
F1	0.0545 (0.0454–0.0634)	0.0630 (0.0525–0.0739)	0.0576 (0.0479–0.0676)
ROC-AUC	0.6390 (0.6053–0.6711)	0.6707 (0.6347–0.7063)	0.6491 (0.6095–0.6861)
PR-AUC	0.0350 (0.0272–0.0457)	0.0409 (0.0314–0.0530)	0.0393 (0.0299–0.0542)

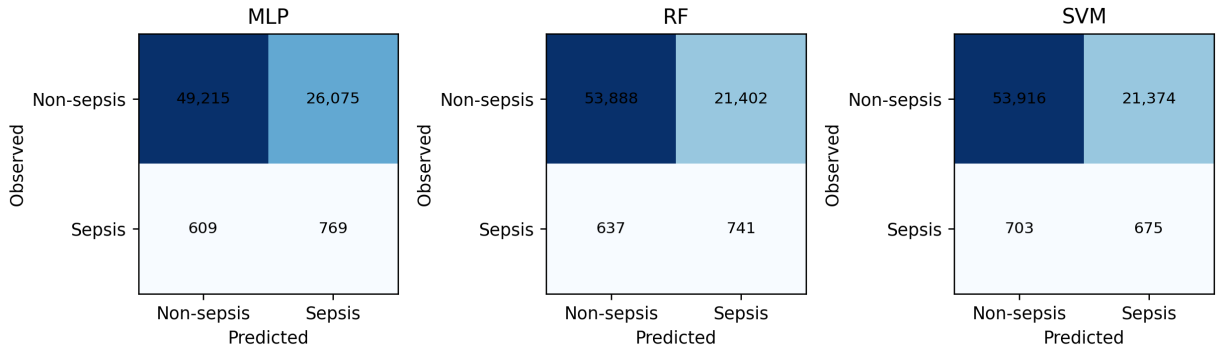

Fig. 3 Independent hold-out confusion matrices for MLP, RF, and SVM.

The confusion matrices show the practical consequence of the selected operating points. RF correctly identified 741 positive hourly observations but generated 21,402 false positives and missed 637 positives. MLP detected more positives (769) but at the cost of 26,075 false positives. SVM produced 675 true positives and 21,374 false positives. Therefore, accuracy alone would be misleading; the clinically relevant trade-off is between sensitivity and alert burden.

4.3 Statistical comparison

Paired patient-cluster bootstrap comparison showed that RF exceeded MLP in ROC-AUC by 0.0317 (95% CI 0.0068–0.0578, p=0.012). RF exceeded SVM by 0.0216, but the interval included zero (95% CI -0.0002–0.0456, p=0.054). MLP and SVM were also not significantly different (difference -0.0101, 95% CI -0.0365–0.0166, p=0.436). Thus, RF had the highest discrimination, but the evidence does not support a claim that it was statistically superior to SVM at the conventional 0.05 level.

Table 5: Paired McNemar comparison of thresholded hold-out classifications

Comparison	A correct/B wrong	A wrong/B correct	p-value
MLP vs RF	6,451	11,096	<0.001
MLP vs SVM	6,581	11,188	<0.001
RF vs SVM	6,686	6,648	0.749

McNemar testing similarly found no meaningful difference between RF and SVM thresholded error patterns (p=0.749), while both differed strongly from MLP. The statistical results reinforce a restrained interpretation: RF provided the strongest overall balance in this dataset, but model ranking depends on the metric and operating threshold.

4.4 Calibration and clinical utility

Table 6: Hold-out probability calibration after development-only Platt calibration

Model	Brier score	Calibration intercept	Calibration slope
MLP	0.017552	0.406	1.112
RF	0.017580	1.571	1.397
SVM	0.017523	-0.226	0.942

All three calibrated Brier scores were similar. SVM had calibration parameters closest to the ideal intercept of 0 and slope of 1, whereas RF showed greater calibration-in-the-large deviation despite having the best discrimination. The reliability plot (Fig. 4) shows that calibrated predictions were concentrated in the low-probability range expected for this rare outcome; therefore, the full 0–1 display should be interpreted together with the Brier scores and calibration slope/intercept rather than by visual proximity alone. Decision-curve analysis (Fig. 5) showed small positive net benefit for the model-based strategies over low threshold probabilities and clear improvement over the treat-all strategy once its net benefit became negative. Because the absolute net-benefit gains were small and the analysis was retrospective, these findings support further prospective utility assessment rather than immediate clinical deployment.

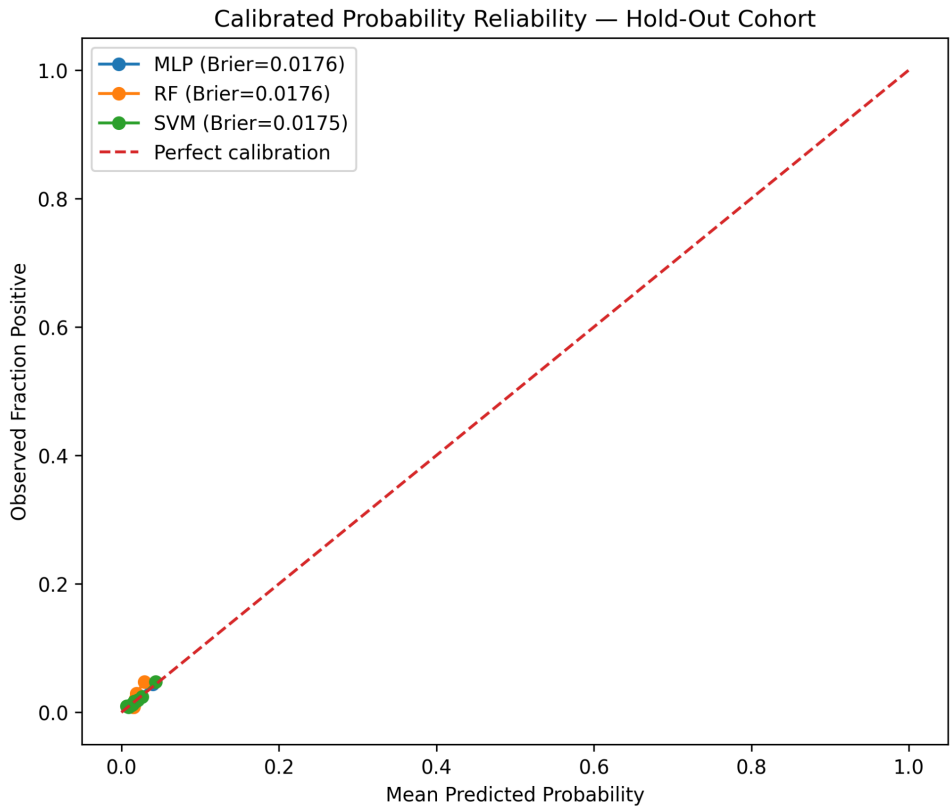

Fig. 4 Calibrated probability reliability curves on the independent hold-out cohort after development-only Platt calibration.

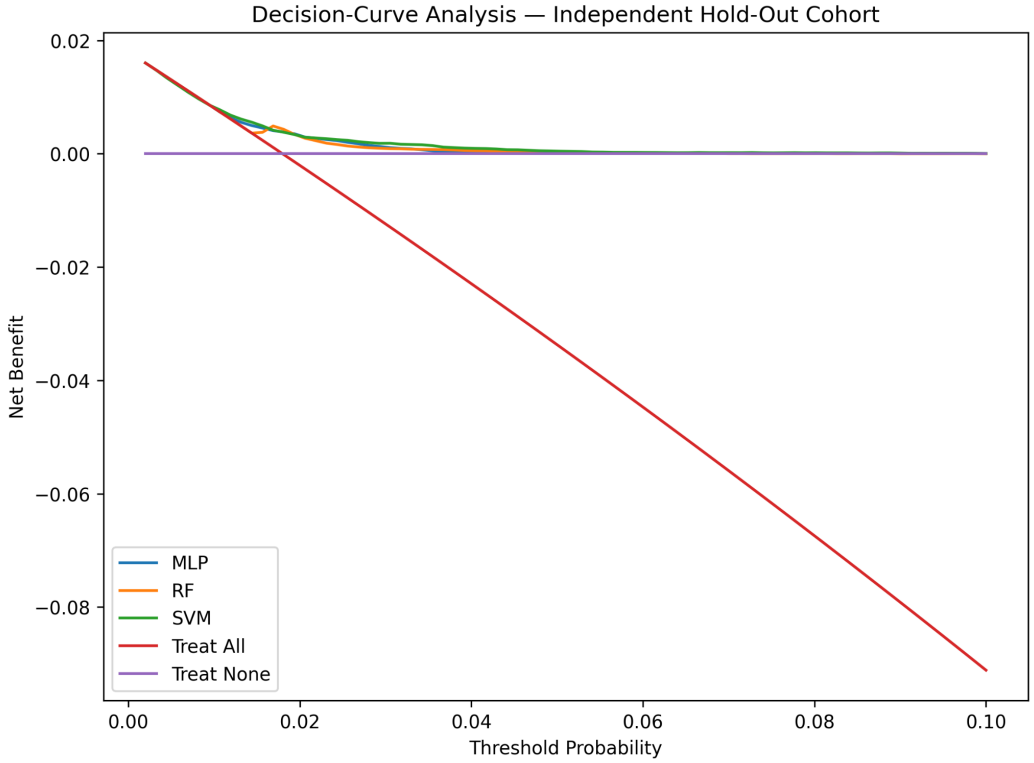

Fig. 5 Decision-curve analysis on the independent hold-out cohort across low threshold probabilities, including treat-all and treat-none reference strategies.

4.5 Explainability

RF permutation importance identified HR as the most influential predictor, followed by respiratory rate, temperature, SBP, and MAP. The mean ROC-AUC decreases after permutation were 0.0575, 0.0403, 0.0288, 0.0184, and 0.0174, respectively. Gender, DBP, age, and O2Sat had smaller contributions. These findings are clinically plausible because cardiovascular, respiratory, thermal, and perfusion abnormalities are common manifestations of systemic deterioration, although importance does not imply causality.

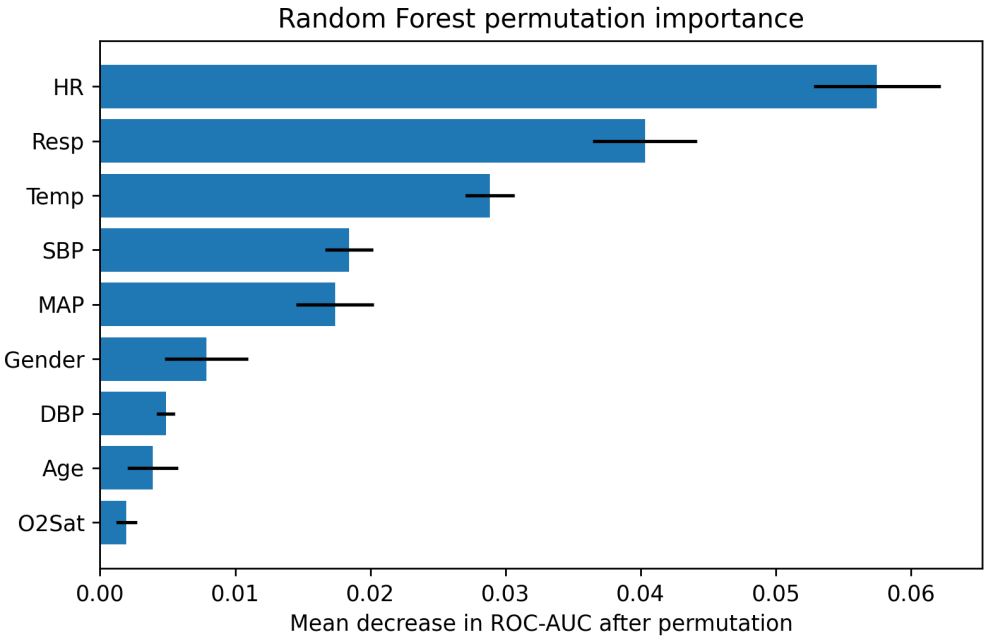

Fig. 6 Random Forest permutation importance on the independent hold-out cohort.

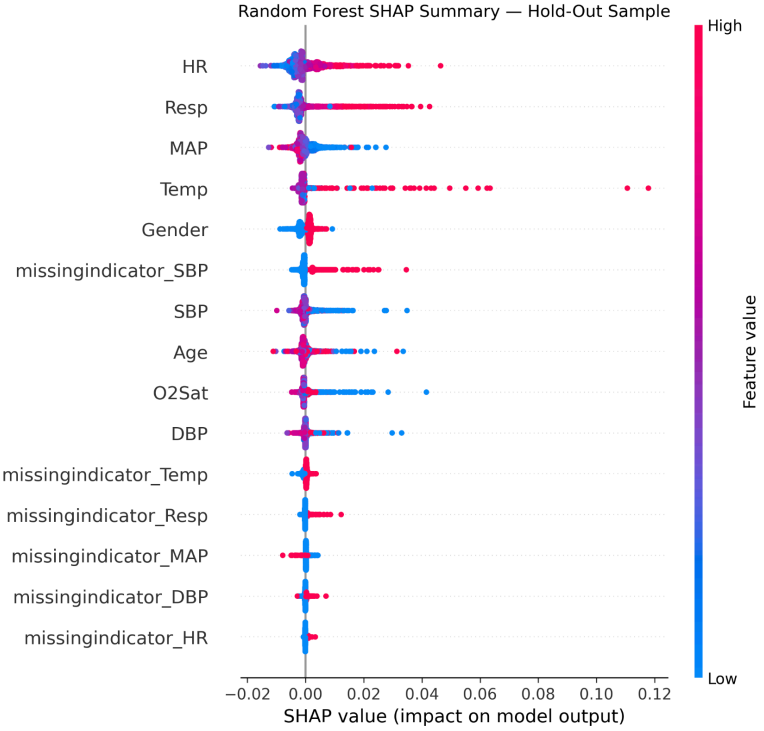

Fig. 7 SHAP summary for the Random Forest model on a reproducible hold-out sample.

The SHAP analysis broadly supported the permutation findings, with HR, respiratory rate, MAP, and temperature among the dominant contributors. SHAP and permutation importance need not produce identical rankings because they quantify different aspects of model behaviour. A local true-positive explanation further showed that a temperature of 38.06°C and respiratory rate of 24.5 breaths/min increased the RF output for the selected observation, demonstrating how patient-level explanations can accompany an alert.

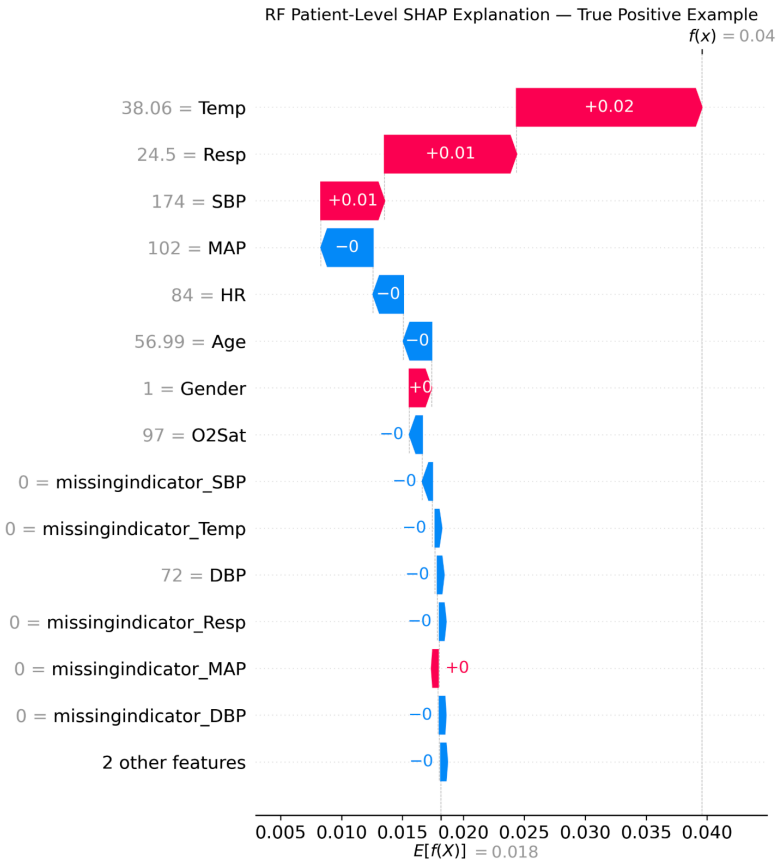

Fig. 8 Patient-level RF SHAP explanation for a correctly classified positive observation.

4.6 Clinical interpretation and resource-limited settings

The revised results are substantially more conservative than the originally reported near-perfect performance, which disappeared after patient-level separation and leakage-controlled preprocessing. This is scientifically informative: it demonstrates how evaluation design can determine apparent model quality. RF appears the most promising of the three models for further investigation because it combined the highest discrimination with moderate sensitivity and specificity, but its precision remained low. In a real clinical workflow, such a model could create alert fatigue if each positive hourly classification triggered an interruptive alert.

A practical deployment would therefore require workflow-level aggregation, refractory periods between alerts, clinician-adjustable thresholds, and prospective evaluation of the number of alerts per patient-day. Infrastructure requirements for the present nine-variable RF model are modest compared with deep time-series systems: routine vital-sign capture, a basic electronic data interface, a lightweight inference service, and an audit log would be sufficient technically. The larger barriers in low-resource settings are likely to be data completeness, interoperability, staffing, governance, and response capacity rather than computation itself. Cost-effectiveness cannot be inferred from retrospective classification performance and should be assessed prospectively against clinical outcomes, workload, and resource use.

Traditional tools such as SOFA, qSOFA, SIRS, NEWS, and MEWS remain important benchmarks because they are transparent and widely understood. A direct comparative claim is not made here because the source file did not provide a consistently complete set of variables required to reconstruct all scores. Future validation should calculate these scores prospectively from the same patients and compare discrimination, calibration, alert burden, and net benefit under identical prediction times.

4.7 Limitations and external validity

This study has several limitations. First, it used a single public data source and therefore does not establish transportability to other hospitals, countries, or care settings. Second, the supplied SepsisLabel was treated as the outcome without independently reconstructing a Sepsis-3 onset time; accordingly, the study should not be interpreted as demonstrating a new fixed-hour early-warning horizon. Third, laboratory variables with very high missingness were excluded from the primary model, which improves operational simplicity but may remove informative biomarkers. Fourth, the analysis evaluated hourly rows while accounting for patient clustering in splitting and uncertainty; a future patient-episode or time-to-event design would more directly answer prospective warning questions. Fifth, no subgroup fairness analysis by age, sex, ethnicity, hospital, or socioeconomic context was possible beyond the variables available in the source. Finally, calibration and decision utility require confirmation after transport to a new clinical population.

5.0 Conclusion

This study provides a leakage-controlled comparative evaluation of MLP, RF, and SVM for classification of a supplied hourly sepsis label using longitudinal structured clinical data. A reproducible 10,000-patient cohort was partitioned at patient level, preprocessing and imbalance handling were confined to development folds, thresholds were selected without using the hold-out cohort, and final performance was reported with patient-cluster confidence intervals, calibration, statistical comparisons, and explainability analyses. RF achieved the highest hold-out ROC-AUC (0.671) and PR-AUC (0.041) and significantly exceeded MLP in ROC-AUC, although its advantage over SVM was not statistically significant. MLP achieved the highest sensitivity, while SVM showed calibration parameters closest to ideal after development-only calibration. All models had low precision because of severe class imbalance, indicating that false-alert burden would be a major implementation concern. The findings support continued investigation of compact ML decision support but do not justify autonomous diagnosis or immediate deployment. The next research stage should include multicentre external validation, prospective and temporal validation with an explicit prediction horizon, direct comparison with SOFA/qSOFA and other warning scores, subgroup fairness assessment, workflow-based alert-fatigue analysis, and health-economic evaluation. Clinical translation will require close collaboration among clinicians, data scientists, health informaticians, and health-system managers.

Acknowledgment

The authors acknowledge the public availability of the sepsis dataset used for this secondary analysis and the open-source Python ecosystem used for reproducible modelling. No external funding was reported for this study.

Declarations

Funding: No external funding was reported for this manuscript.
Conflict of interest: The authors declare no conflict of interest.
Ethical consideration: The study used publicly accessible, anonymised secondary data and involved no direct contact with human participants.
Data availability: The source dataset is publicly available from the Kaggle repository cited below. The analysis notebooks and derived reproducibility records are available from the corresponding author on reasonable request.
Author contribution statement: All listed authors contributed to the conceptual direction, manuscript development, review, and approval of the final article.
Data and Code Availability: The dataset used in this study is publicly available from the Kaggle "Prediction of Sepsis" repository:
https://www.kaggle.com/datasets/salikhussaini49/prediction-of-sepsis.

The Python/Google Colab implementation used for dataset auditing, patient-level cohort construction, preprocessing, feature selection, model development and tuning, class-imbalance analysis, independent hold-out evaluation, calibration, statistical comparison, and explainability is publicly available at the project GitHub repository:
https://github.com/olalekanokewale/sepsis_detection

References

[1] Abd El-Aziz, R. M., & Alanazi, R. (2025). Early detection of sepsis using machine learning algorithms. Alexandria Engineering Journal, 111, 47–56. https://doi.org/10.1016/j.aej.2024.10.005

[2] Bignami, E. G., Berdini, M., Panizzi, M., Domenichetti, T., Bezzi, F., Allai, S., Damiano, T., & Bellini, V. (2025). Artificial intelligence in sepsis management: An overview for clinicians. Journal of Clinical Medicine, 14(1), 286. https://doi.org/10.3390/jcm14010286

[3] Bomrah, S., Uddin, M., Upadhyay, U., Komorowski, M., Priya, J., Dhar, E., Hsu, S. C., & Syed-Abdul, S. (2024). A scoping review of machine learning for sepsis prediction—Feature engineering strategies and model performance: A step towards explainability. Critical Care, 28, 180. https://doi.org/10.1186/s13054-024-04948-6

[4] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[6] Collins, G. S., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. https://doi.org/10.1136/bmj-2023-078378

[7] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273–297. https://doi.org/10.1007/BF00994018

[8] Do, D.-K., Rockenschaub, P., Boie, S. D., Kumpf, O., Volk, H.-D., Balzer, F., von Dincklage, F., & Lichtner, G. (2026). The impact of evaluation strategy on sepsis prediction model performance metrics in intensive care data: Retrospective cohort study. Journal of Medical Internet Research, 28, e72083. https://doi.org/10.2196/72083

[9] Goh, K. H., Wang, L., Yeow, A. Y. K., Poh, H., Li, K., Yeow, J. J. L., & Tan, G. Y. H. (2021). Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nature Communications, 12, 711. https://doi.org/10.1038/s41467-021-20910-4

[10] Hussaini, S. (n.d.). Prediction of sepsis [Data set]. Kaggle. https://www.kaggle.com/datasets/salikhussaini49/prediction-of-sepsis

[11] Lin, T. H., Chung, H. Y., Jian, M. J., Chang, C. K., Lin, H. H., Yen, C. T., Tang, S. H., Pan, P. C., Perng, C. L., Chang, F. Y., Chen, C. W., & Shang, H. S. (2025). AI-driven innovations for early sepsis detection by combining predictive accuracy with blood count analysis in an emergency setting: Retrospective study. Journal of Medical Internet Research, 27, e56155. https://doi.org/10.2196/56155

[12] Liu, Z., Shu, W., Li, T., Zhang, X., & Chong, W. (2025). Interpretable machine learning for predicting sepsis risk in emergency triage patients. Scientific Reports, 15, 887. https://doi.org/10.1038/s41598-025-85121-z

[13] Prescott, H. C., et al. (2026). Surviving Sepsis Campaign: International guidelines for management of sepsis and septic shock 2026. Critical Care Medicine, 54(4), 725–812. https://doi.org/10.1097/CCM.0000000000007075

[14] Rudd, K. E., et al. (2020). Global, regional, and national sepsis incidence and mortality, 1990–2017: Analysis for the Global Burden of Disease Study. The Lancet, 395(10219), 200–211. https://doi.org/10.1016/S0140-6736(19)32989-7

[15] Santacroce, E., et al. (2024). Advances and challenges in sepsis management: Modern tools and future directions. Cells, 13(5), 439. https://doi.org/10.3390/cells13050439

[16] Singer, M., et al. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA, 315(8), 801–810. https://doi.org/10.1001/jama.2016.0287

[17] Tang, Y., Zhang, Y., & Li, J. (2024). A time series driven model for early sepsis prediction based on transformer module. BMC Medical Research Methodology, 24, 23. https://doi.org/10.1186/s12874-023-02138-6

[18] World Health Organization. (2024, May 3). Sepsis. https://www.who.int/news-room/fact-sheets/detail/sepsis

[19] Yadgarov, M. Y., et al. (2024). Early detection of sepsis using machine learning algorithms: A systematic review and network meta-analysis. Frontiers in Medicine, 11, 1491358. https://doi.org/10.3389/fmed.2024.1491358

[20] Zhou, L., Shao, M., Wang, C., & Wang, Y. (2024). An early sepsis prediction model utilizing machine learning and unbalanced data processing in a clinical context. Preventive Medicine Reports, 45, 102841. https://doi.org/10.1016/j.pmedr.2024.102841