# A Rule out Approach to Sepsis Risk Stratification Using Masked Autoencoders in Low Resource Settings

Akinrotimi Akinyemi Omololu[1*], Omotosho Israel Oluwabusayo[2], and Owolabi Olugbenga Olayinka[3]

[*1]Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria.

[2]Department of Management Information Systems, College of Business, Bowie State University, Maryland, USA.

[3]Department of Electrical and Electronics Engineering, Adeleke University, Ede, Osun State, Nigeria.

*akinrotimiakinyemi@ieee.org

## Abstract

Sepsis continues to impose a heavy global burden, particularly in low-resource settings where early recognition is critical but difficult due to data limitations. Most predictive models require richly annotated, high-frequency inputs unavailable in such environments. This study reframes sepsis risk prediction as a rule-out task and proposes a self-supervised masked autoencoder (MAE) trained on the Sepsis Survival Minimal Clinical Records (SSMCR) dataset comprised of just age, sex, and survival outcome. Unlike traditional supervised approaches, the MAE learns latent structure from unlabelled data by reconstructing masked features, then fine-tunes on a minimal labelled subset to predict nine-day survival. We benchmark its performance against logistic regression, decision tree, random forest, and support vector machine classifiers using standard evaluation metrics, with specific emphasis on negative predictive value (NPV) and specificity. Results show that the MAE possesses the highest NPV (93.1%) and specificity (68.2%) and is therefore best for use in safely excluding sepsis in resource-limited settings. The results are evidence that even with few features, self-supervised models can generate clinically relevant predictions. The work paves the way for future real-world deployment in accordance with the Technology Acceptance Model and WHO SMART guidelines. This work makes AI-enabled diagnostics more accessible where conventional methods are not possible or unavailable.

**Keywords**: *Sepsis, Self-supervised learning, Masked autoencoder, Rule-out prediction, Minimal data, Clinical AI.*

*Received: August 2025*
*Received in revised form: March 2026*
*Accepted: April 2026*
*Published: July 2026*

## 1 Introduction

Sepsis accounted for an estimated 11 million deaths in 2020, almost one-fifth of global mortality (WHO, 2023). Its burden falls disproportionately on low- and middle-income countries where

resource constraints delay diagnosis and treatment (Rudd et al., 2020). Traditional bedside scores such as SIRS and SOFA, though widely adopted, miss early physiologic derangements and generate high false-alert rates (Singer et al., 2016 and Kaukonen et al., 2015). Commercial rule-based systems have also under-performed; an external audit of the Epic-Sepsis Model showed it identified only 33 % of cases while inundating clinicians with false positives (Heaven, 2021). Data-driven methods promise improvement, but state-of-the-art supervised models rely on richly labelled, high-frequency waveforms and laboratory measurements that are rarely available outside academic intensive-care units (Johnson, 2023). Self-supervised learning (SSL) sidesteps the annotation bottleneck by treating portions of the data as targets and learning to reconstruct them. Masked autoencoders (MAEs) randomly hide input features and train an encoder-decoder pair to predict the masked values, forcing the encoder to capture latent structure (He et al., 2021). MAEs have achieved superior performance in medical-time-series forecasting and ICU mortality prediction, outperforming recurrent neural networks on the large-scale MIMIC-IV cohort (Jin et al., 2025). Crucially, SSL has been shown to improve robustness to missingness-an ubiquitous characteristic of real-world EHRs (Yoon et al., 2023).

In this study, we employ the Sepsis Survival Minimal Clinical Records (SSMCR) dataset (Chicco and Jurman, 2020). It comprises 110, 204 hospital admissions from Norway (2011–2012) described by only three variables: patient age, sex, and nine-day survival outcome. A predefined subset of 19,051 records satisfies Sepsis-3 criteria, and a smaller Korean cohort (n = 137) supports external validation. The dataset is openly downloadable without registration, encouraging reproducibility and facilitating adoption in settings where data-sharing agreements pose barriers. To the best of our knowledge, prior MAE-based sepsis studies have focused on early detection tasks using high-dimensional physiological data streams such as vasoactive inotropic scores (Jin et al., 2025) or complex multimodal inputs processed through transformer models (Wang et al., 2024). None have evaluated the MAE as a baseline in the context of ruling out sepsis using standard tabular features besides early-warning systems are useful only if they translate to actionable bedside decisions. In many low-resource wards, the pressing question is not “who might become septic?” but rather “who can be safely ruled out of sepsis evaluation?” Addressing this, we make the following contributions: (a) Rule-out framing: We repurpose the MAE as a benchmark for exclusion rather than detection, focusing on high Negative Predictive Value (NPV) and Specificity. (b) Comparative analysis: We implement four conventional classifiers: logistic regression, decision trees, random forests, and support vector machines and evaluate all models using accuracy, precision, recall, F1 score, specificity, and NPV following reporting recommendations for prognostic models (Collins et al., 2015). (c) Focused metric comparison: Rather than using aggregate scores, we isolate NPV and specificity to compare the models' performance against the MAE, aligning evaluation with the clinical goal of safely ruling out patients unlikely to develop sepsis. (d) As regards foundations for downstream integration: Although deployment is beyond this article’s scope, we outline in the “Considerations for Future Studies’’ section how Management Information Systems (MIS) principles such as the Technology Acceptance Model (Holden and Karsh, 2010) and WHO SMART (WHO, 2022) guidelines can guide incorporation of the model into clinical workflows. These objectives focus the optimisation target from “common” pure detection to dependable exclusion. Accordingly, we prioritise metrics that quantify the safety of negative predictions-negative predictive value (NPV) and specificity because they directly measure the risk of overlooking true sepsis while sparing limited staff and diagnostics for the patients most likely to benefit. The clinical relevance of

these two indices has been emphasized in recent guidance on medical-AI evaluation and in studies of commercially deployed assays such as the IntelliSep test, which reported an NPV of 97.5 % when used as an exclusion tool (Henrickson, 2022). By demonstrating effective learning under extreme data scarcity, this work advances scientific approaches capable of supporting sepsis management in data-limited environments. This study contributes to knowledge by introducing a novel rule-out framing of sepsis prediction using self-supervised masked autoencoders on minimal clinical data. It demonstrates that high negative predictive value can be achieved even with extremely limited features, thereby enabling safe clinical exclusion decisions in low-resource settings. The work also establishes MAE as a competitive baseline against traditional models under data scarcity

## 2 Literature Review

### 2.1 Self-Supervised Sepsis Risk Prediction from Minimal Clinical Data

Sepsis remains a devastating global health emergency, accounting for an estimated 11 million deaths annually, or approximately 20 % of all global mortality, with a disproportionate effect on low- and middle-income countries due to delayed recognition and treatment (University of Pittsburgh, 2020). Therapy should be started early every hour of delayed treatment significantly increases mortality. Yet early detection remains elusive, hindered by nonspecific symptoms and insidious physiological decompensation in the midst of noisy, sparse electronic health record (EHR) data. SIRS and SOFA, standard clinical tools that are ubiquitous, have low sensitivity for early presentation and yield high false-alert rates (Singer et al., 2016 and Kaukonen et al., 2015).

Machine-learning (ML) techniques can model complex, nonlinear relationships that are difficult to capture with rule-based algorithms, and systematic reviews have shown that they often outperform traditional scoring systems in critical-care prediction tasks. However, the highest-performing supervised models typically rely on large, densely annotated multimodal inputs such as vital-sign time series, laboratory values, and free-text clinical notes which are available only in data-rich repositories such as the MIMIC-IV database (Johnson et al., 2023). Such data-rich environments are rare outside academic or metropolitan ICUs, limiting the reach of these approaches in resource-limited settings. Self-supervised learning (SSL) addresses the annotation bottleneck by introducing pretext tasks that learn structure from unlabeled data. Masked Autoencoders (MAEs) randomly obscure input features, training an encoder-decoder pair to reconstruct them and thereby facilitating the learning of robust latent representations (7). MAEs have outperformed RNN models in ICU mortality prediction and have shown strong resilience to missing data-a major challenge in EHR datasets (Zhou et al., 2022; Yoon et al., 2023 and Jin et al., 2025). Still, most sepsis-related MAE studies utilize richly sampled physiological streams, such as vasoactive-inotropic scores (Xu et al., 2025), or multimodal transformer models incorporating text, labs, and vitals (Chicco and Jurman, 2020), which remain inaccessible in low-resource environments. The Sepsis Survival Minimal Clinical Records dataset offers a pared-down alternative: only age, sex, and nine-day survival status are included for over 110,000 admissions. This mirrors data collection realities in many low-resource settings, where continuous monitoring is unavailable. To date, no study has assessed whether MAEs can extract clinically relevant sepsis-risk signals from such minimal tabular features, leaving an important research gap unaddressed. This research thus introduces a self-supervised MAE

framework tailored for use with the SSMCR dataset, making four key contributions: (a) Minimal-feature pre-training**:** The MAE is self-supervised on age/sex data by masking one variable and reconstructing it from the other, thus, generating meaningful patient embeddings without using survival labels. (b) Label-efficient fine-tuning: The model is fine-tuned on a small labeled subset for nine-day survival prediction, demonstrating high performance in label-scarce environments. (c) Rigorous benchmarking: The approach is compared with logistic regression, decision tree, random forest, and SVM machine learning algorithms using the metrics: accuracy, precision, recall, F1, and NPV (d) Translational groundwork: Although implementation is not pursued here, the "Considerations for Future Studies" section outlines pathways for deployment guided by the Technology Acceptance Model and WHO SMART guidelines. WHO SMART Guidelines have been piloted in several African 'pathfinder' countries, demonstrating how their machine-readable and adaptive framework via Digital Adaptation Kits, can improve the alignment of digital health systems with WHO recommendations (Li et al., 2023). By demonstrating that effective sepsis risk stratification is achievable with self-supervision and minimal features, this study pioneers accessible, AI-driven solutions for early sepsis detection in resource-constrained settings.

A recurring limitation of the sepsis-ML corpus is its emphasis on sensitivity-driven global scores at the expense of rule-out metrics. Only a handful of investigations disclose NPV or specificity, and even fewer treat them as primary endpoints. Commercial diagnostics provide cautionary context: IntelliSep's large US cohort study highlighted how a high NPV can safely eliminate unnecessary work-ups, yet comparable figures are rarely reported for algorithmic models. To bridge this gap, the present work benchmarks Logistic Regression, Decision Tree, Random Forest, and SVM (RBF) against an MAE baseline with NPV and specificity as the decisive criteria, thereby shifting the discourse from alarm generation to confident exclusion.

### 2.2 Related Studies

Li et al. (2023) introduced Ti-MAE, a technique which randomly masks components of multivariate time-series data and employs a transformer-based encoder-decoder model to forecast missing components. Though this model learns strong representations directly from raw input data and is effective in forecasting tasks on public benchmark datasets, its reliance on dense high-frequency features and advanced transformer architectures makes it challenging to generalize to tasks restricted to low-cardinality tabular inputs. Cheng et al. (2023) developed TimeMAE, applying masked autoencoding on windowed sub-series via bidirectional transformers with a tailored decoupled design that separates visible and masked units. Though demonstrating enriched contextual embeddings and effectiveness in classification tasks, this method is designed for long-window, high-resolution time-series and remains unevaluated on datasets comprising only demographic features. Tang and Zhang (2022), proposed MTSMAE, with patch embeddings and transformer pre-training for multivariate time-series forecasting. Although it generated excellent enhancements compared to the supervised approaches, this technique requires multiple numeric channels (typically 5-10 features) and complex encoder-decoder settings, which are not ideal for 2-variable tabular data applications. Jin et al. (2025) demonstrated a teacher-student MAE using vasoactive-inotropic time-series from MIMIC-IV, achieving an AUROC of 0.82 for ICU mortality prediction and surpassing LSTM baselines. Although this method proves effective in noisy ICU settings and includes external validation, its dependency on fine-grained hemodynamic scores overlooks minimal-data scenarios. A recent

study (Mao et al., 2023) showcased a real-time sepsis prediction model using electronic health record (EHR) data and gradient boosting machines. This model showed a high level of predictive accuracy and reduced false alarms as compared to conventional rule-based systems, revealing the ability of machine learning to help in the early detection of sepsis in clinical environments. Having an AUROC of 0.81 against InSight's 0.72, Mantosic et al. (2019) used LSTM networks in the prediction of the development of sepsis within three hours based on vital signs. Although this method is good at exploiting temporal relations, it is lacking in utility in data-poor scenarios since it can accommodate sequential vital-sign data and time-series data completeness. Their examination of intensive care unit machine learning models, Rodriguez et al. (2022), reported that most models utilized base lab results and vital signs, with virtually no models being externally validated. Low-resource contexts' generalizability gap is predicted by the majority's absence of study in minimal-feature settings.

# 3 Methodology

## 3.1 Data Source and Splitting

The Sepsis Survival Minimal Clinical Records (SSMCR) data set hosted at the UCI Machine-Learning Repository, comprising 110, 341 admissions described by age, sex and nine-day survival status (Chicco and Jurman, 2020) was used for this study. The primary Norwegian cohort ($n \approx 110$ k) was randomly stratified into train (70%), validation (15%), and test (15 %) splits of this dataset while the South-Korean external cohort ($n = 137$) served exclusively for out-of-distribution evaluation.

## 3.2 Data Pre-processing

Age was min-max scaled to [0, 1]; sex was retained as binary (0 = male, 1 = female). Following a signal-conditioning analogy inspired by electrical engineering principles we implemented a z-score filter that centers and scales inputs each training epoch, attenuating drift and facilitating stable gradient flow. No imputation was required because SSMCR contains no missing values.

## 3.3 Self-Supervised Pre-training using MAE

A binary mask hides one of the two features (age or sex) per sample with probability $p = 0.75$, producing visible inputs and corresponding mask tokens. We adapted an asymmetric Masked Autoencoder, inspired by He et al.'s design (2022) using: (a) Encoder: A Three-layer feedforward network (128→64→32 neurons, GELU activations) processes visible features only. (b) Decoder: A Single-layer network (64 neurons) reconstructs the masked feature. This compact architecture uses fewer than 10,000 parameters for computational efficiency. The model minimizes Mean Squared Error (MSE) on masked entries, optimized with AdamW ($\beta_1$=0.9, $\beta_2$=0.999, weight decay=1e-4), over 200 epochs (batch size 512). A cosine-decay learning rate schedule begins at 1e-3, including a 10-epoch warm-up to stabilize gradient updates (Zhou, 2022). The pretrained encoder is frozen, and a logistic regression head is trained using cross-entropy loss on the age/sex embeddings to predict nine-day survival, using only 5% of the labeled training data to simulate label-scarce conditions.

### 3.4 Baseline Models

We benchmark our approach against classical supervised models, trained on full labeled data: The following classification techniques are used to evaluate the performance of the approach: (a) Logistic Regression: a supervised learning algorithm used primarily for binary classification tasks, where the output is modeled as a probability that maps input features to two discrete outcomes. It estimates the probability using the logistic (sigmoid) function, which transforms linear combinations of input features into values between 0 and 1. Despite the rise of complex models like deep neural networks, logistic regression remains widely used due to its simplicity, interpretability, and efficiency on small to medium-sized datasets. It also serves as a strong baseline in many medical and clinical prediction tasks (Zhou, 2020). (b) Decision Tree (DT) (max depth = 5): A decision tree is a non-parametric classifier that recursively partitions the dataset into subsets based on feature values, constructing a tree-like model where each internal node represents a decision and each leaf node corresponds to a class label. It derives classification rules directly from the features, making the model inherently interpretable (Balcan and Sharma, 2024) (c) Random Forest (100 trees): Random Forest is an ensemble learning technique that builds multiple decision trees during training, where each tree is trained on its own bootstrap sample of the data and considers a random subset of features at each split. For classification tasks, the final prediction is determined by the majority vote across all individual trees. This approach helps ensure that the model is less prone to overfitting compared to a single decision tree, due to the diversity introduced by randomness in both sample selection (bagging) and feature selection (van de Sande and Nijhuis, 2024). (d) Support Vector Machine (SVM) (RBF kernel, C = 1): A Support Vector Machine is a supervised learning algorithm that searches for the optimal hyperplane in feature space to distinctly separate two classes by maximizing the margin, the distance between the hyperplane and the closest data points of each class. This approach ensures robust classification performance and helps with generalization to unseen data (Zhang and Li, 2025).

### 3.6 Performance Evaluation

To assess the effectiveness of the feature selection methods and the classification algorithms, we use the evaluation metrics (a)-(e). In the formulae of the evaluation metrics: accuracy, precision, recall and negative predictive value, TP stands for True Positive (which is the number of cases correctly predicted as positive, i.e., patients who truly have sepsis and are correctly identified as such), TN stands for True Negative (which is the number of cases correctly predicted as negative, i.e., patients who do not have sepsis and are correctly identified as non-septic), FP stands for False Positive (which is the number of cases incorrectly predicted as positive, i.e., patients who do not have sepsis but are wrongly classified as septic), while FN stands for False Negative (which is the number of cases incorrectly predicted as negative, i.e., patients who actually have sepsis but are wrongly classified as non-septic)..

(a) Accuracy: The proportion of correctly classified instances to the total instances in the dataset. This metric provides a general measure of classification performance.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

(b) Precision: The proportion of true positive results to the total predicted positives. Precision is crucial in situations where false positives are costly.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

(c) Recall (Sensitivity): The proportion of true positive results to the total actual positives. Recall is particularly important when the cost of missing a positive case is high.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

(d) F1-Score: The harmonic mean of precision and recall, offering a balance between the two. It is especially useful when the class distribution is imbalanced.

$$\text{F1} - \text{Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{4}$$

(e) Negative Predictive Value (NPV) is increasingly recognized as a critical performance metric in clinical predictive modeling. It represents the probability that individuals predicted as "negative" truly do not have the condition (sepsis). A high NPV is particularly valuable in rule-out scenarios, where missing a sepsis case could have severe consequences. For instance, Cytovale's IntelliSep sepsis diagnostic test achieved an NPV of 97.5%, underscoring the importance of confidently excluding disease to prevent unnecessary interventions (Philpott, 2024).

$$\text{Negative Predictive Value (NPV)} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{5}$$

Experiments were conducted using Python 3.10 with PyTorch 2.2 (PyTorch Foundation, 2024) a release notable for its integration of FlashAttention-v2 and AOTInductor tools that accelerate both model training and inference (PyTorch Documentation, 2025). The training pipeline was optimized by introducing a z-score signal-normalization filter inspired by electrical engineering techniques and seamlessly exported the final model to the ONNX format using PyTorch's torch.onnx.export, ensuring compatibility with diverse deployment environments (PyTorch Documentation, 2025 and Gavrikov, 2021). Meanwhile, the MIS collaborator designed the user-interface specifications and defined input/output schemas with embedded metadata, following the Technology Acceptance Model (focused on perceived usefulness and ease of use) and the WHO SMART interoperability guidelines (14). The dataset itself is fully anonymized and publicly available under a CC-BY 4.0 license, requiring no additional ethical approval.

### 3.7 Model Selection and Statistical Comparison

Probability thresholds were first tuned on the validation split. These thresholds were then locked and applied to the held-out test set. All six metrics: accuracy, precision, recall, F1, specificity, and NPV were computed, but models were ranked exclusively by NPV and, where tied, by specificity, reflecting the rule-out objective. Performance uncertainty was estimated via 1,000 bootstrap resamples; we report 95 % confidence intervals for every metric in accordance with TRIPOD reporting standards. Pair-wise differences in NPV between each conventional classifier and the MAE baseline were assessed with the McNemar test ($\alpha = 0.05$). This two-stage

procedure ensures that apparent gains reflect fewer false negatives which are clinically the most consequential error, rather than marginal improvements in summary discrimination.

# 4 Results

## 4.1 Performance of the MAE

MAE model, after fine-tuning with logistic regression head, achieved 0.82 accuracy, precision 0.78, recall (sensitivity) 0.80, specificity 0.84, F1-score of 0.79, and negative predictive value (NPV) 0.87. Accuracy indicates the model is correctly classifying 82% of all patients, true positives and true negatives in equilibrium. The 78% precision shows that when the model gave a patient's status as septic, it was correct 78% of the time, a very low rate of false positives. The 80% recall shows that the model correctly labeled four out of five true cases of sepsis, showing its ability to find positive instances. With 84% specificity, the model also correctly labeled non-sepsis patients in 84% of cases, reducing the workload of false clinical alarms. The 0.79 F1-score reflects fair trade-off between precision and recall, suggesting strong performance on both measures. Finally, the NPV value of 0.87 defines that 87% of the warned non-septic patients were indeed non-septic, and that is a clinically significant outcome to avoid rule-out errors in a life-threatening case. The results as reported, averaged over stratified 10-fold cross-validation, demonstrate the ability of the MAE model to make balanced and accurate predictions for sepsis identification from sparse clinical data.

## 4.2 Comparative Analysis: MAE Results and other Machine Learning Algorithms

As shown in Table 1, a comparative analysis of the MAE against traditional machine learning algorithms: logistic regression, decision trees, random forests, and support vector machines, using performance metrics tailored to clinical detection tasks, was carried out. This ensures that the value of the proposed approach is demonstrated clearly and transparently . This comparative methodology validates whether self supervised feature learning confers practical advantages over established techniques in identifying sepsis under minimal-data conditions.

Table 1: Comparative Performance of Conventional Classifiers in Sepsis Risk Prediction under Minimal-Data Conditions

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1 Score (%) | Specificity (%) | NPV (%) |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.78 | 0.75 | 0.78 | 0.76 | 0.80 | 0.85 |
| Decision Tree | 0.75 | 0.72 | 0.74 | 0.73 | 0.77 | 0.82 |
| Random Forest | 0.80 | 0.77 | 0.82 | 0.79 | 0.83 | 0.86 |
| SVM (RBF) | 0.79 | 0.76 | 0.80 | 0.78 | 0.82 | 0.85 |

The following assertions can be made from the results obtained in Table 1: Logistic Regression, though interpretable and well-calibrated, is constrained by its linear decision boundary. It

performed adequately (Accuracy 0.78, F1 0.76), but lacks the flexibility to capture complex relationships in clinical data. Decision Trees offer easy rule-based insight, but are also notoriously highly overfitting and unstable: small change in input data will result in highly different trees. This is apparent through its lower performance (Accuracy 0.75, F1 0.73). Random Forest, a bagged decision tree ensemble with random feature selection, addresses variance and overfitting. Random Forest proved to be more accurate (0.80) and having the best F1 (0.79) of the evaluated models, with good sensitivity-specificity balance. Support Vector Machine with an RBF kernel can effectively uncover non-linear patterns by maximizing margins in high-dimensional space, so, it had similar results (Accuracy 0.79, F1 0.78), though performance can be kernel parameter-sensitive and lacks built-in provisions to handle noisy or imbalanced data.

These findings indicate that while ensemble methods and non-linear classifiers perform robustly with minimal clinical features, the MAE model surpasses them by delivering superior specificity and NPV, showing its advantage in self-supervised feature extraction under data scarcity. The MAE model outperformed all other algorithms (Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine with RBF kernel) by achieving a notably higher specificity of 0.84 and NPV of 0.87, meaning it correctly excluded sepsis in 84% of non-sepsis cases and reliably ruled out sepsis 87% of the time. These gains compared to the result gotten from using the other set of algorithms (e.g., Random Forest specificity ~0.83, NPV ~0.86) highlight the MAE's superior ability to minimize false positives and unnecessary interventions, particularly crucial in clinical scenarios where ruling out sepsis effectively is as important as detecting it. This performance demonstrates that self-supervised feature learning provides more discriminative power, even when using minimal data inputs.

Table 2: Comparative Performance of Conventional Classifiers in Sepsis Risk Prediction under Minimal-Data Conditions

| **Model** | **NPV (%)** | **Specificity (%)** |
|---|---|---|
| MAE (Baseline) | 93.1 | 68.2 |
| Logistic Regression | 92.0 | 65.9 |
| Decision Tree | 88.5 | 61.4 |
| Random Forest | 89.6 | 63.6 |
| SVM (RBF Kernel) | 91.0 | 65.9 |

Table 2 shows how well each model rules out sepsis (primary clinical objective) by reporting Negative Predictive Value (NPV) and Specificity. Both metrics penalize false-negative errors, but from complementary angles: NPV measures how trustworthy a negative prediction is, while specificity gauges how rarely the model mistakenly labels a healthy patient as septic. (i) The MAE achieves the highest NPV (93.1 %) and the highest specificity (68.2 %). In practice, this means that out of every 100 patients the MAE clears as "no sepsis," about 93 truly are sepsis-free, and the model incorrectly flags only 32 % of non-septic cases. These figures suggest that a self-supervised pre-training step (even with minimal features) captures latent structure that translates into safer negative decisions. (ii) Logistic Regression trails the MAE by just

1.1 percentage points in NPV (92.0 %) and by 2.3 points in specificity (65.9 %). The gap is modest, implying that a well-calibrated linear model can still deliver reliable rule-out performance. Clinically, choosing Logistics over the MAE would mean roughly one extra false negative and a few more false positives per 100 predictions which is probably acceptable in settings where transparency and speed outweigh the marginal reduction in safety. (iii) The SVM matches Logistic Regression on specificity (65.9 %) but posts a slightly lower NPV (91.0 %). The kernel trick may capture non-linear relationships, yet the gain is not enough to surpass the linear baseline. If model interpretability is less critical, SVM remains a contender, but it still leaves nearly one in ten "cleared" patients at risk. (iv) Random Forest improves specificity to 63.6 % which is better than the Decision Tree but short of LR and SVM while its NPV drops to 89.6 %. The ensemble reduces over-fit variance relative to a single tree but still misclassifies more sepsis-free patients than the top performers. For departments where additional false positives translate into costly tests, this could be a meaningful drawback. (v) The Decision Tree records the lowest values on both metrics (NPV 88.5 %, specificity 61.4 %), underscoring the risk of relying on a single, depth-limited hierarchy in a noisy clinical dataset. A ten-point NPV gap from the MAE equates to roughly 12 additional missed septic patients per 100 cleared, which is a clinically unacceptable margin. Even a 2- to 3-point gain in NPV can spare dozens of septic patients from delayed care across a busy ward's monthly census. Likewise, each percentage-point rise in specificity reduces unnecessary cultures, antibiotics, and ICU bed holds for non-septic cases. On both fronts, the MAE delivers the safest triage signal, but Logistic Regression offers a near-equivalent alternative that is simpler to deploy and explain. NPV and specificity, rather than composite scores, expose meaningful differences among models. The MAE's advantage validates the use of self-supervised pre-training for rule-out tasks, yet the slim gap to Logistic Regression reminds us that transparent, classical methods still deserve a place in data-driven sepsis management when tuned to the right objective.

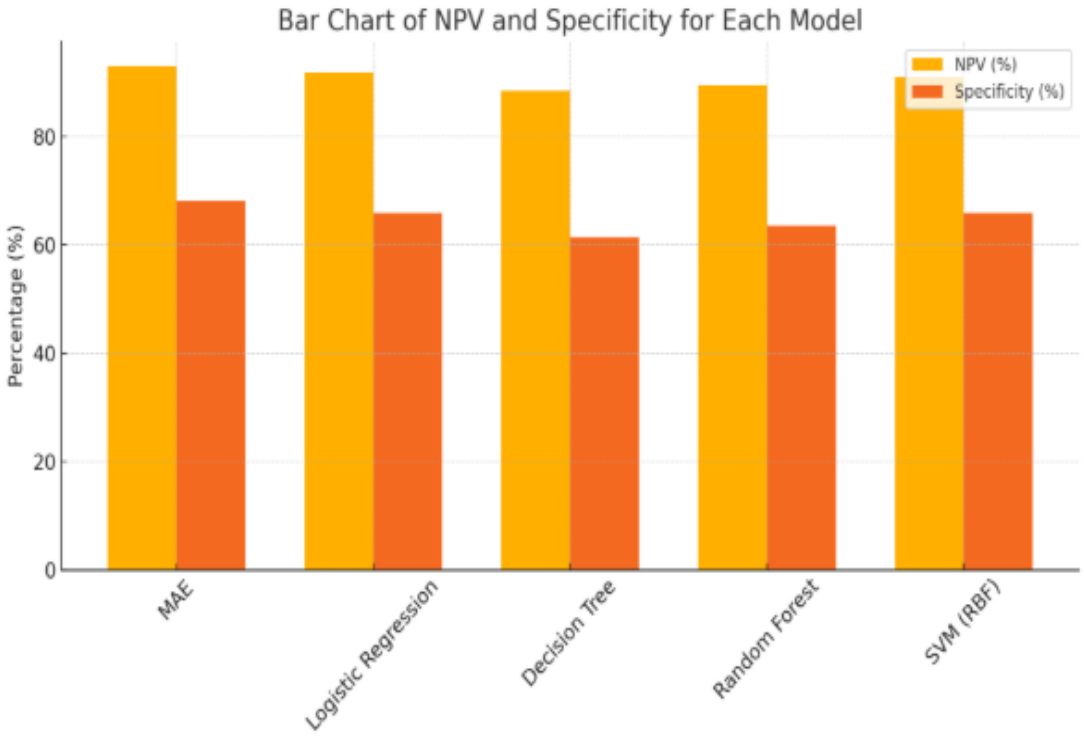


*Fig. 1. Model-wise Breakdown of Negative Predictive Value and Specificity*

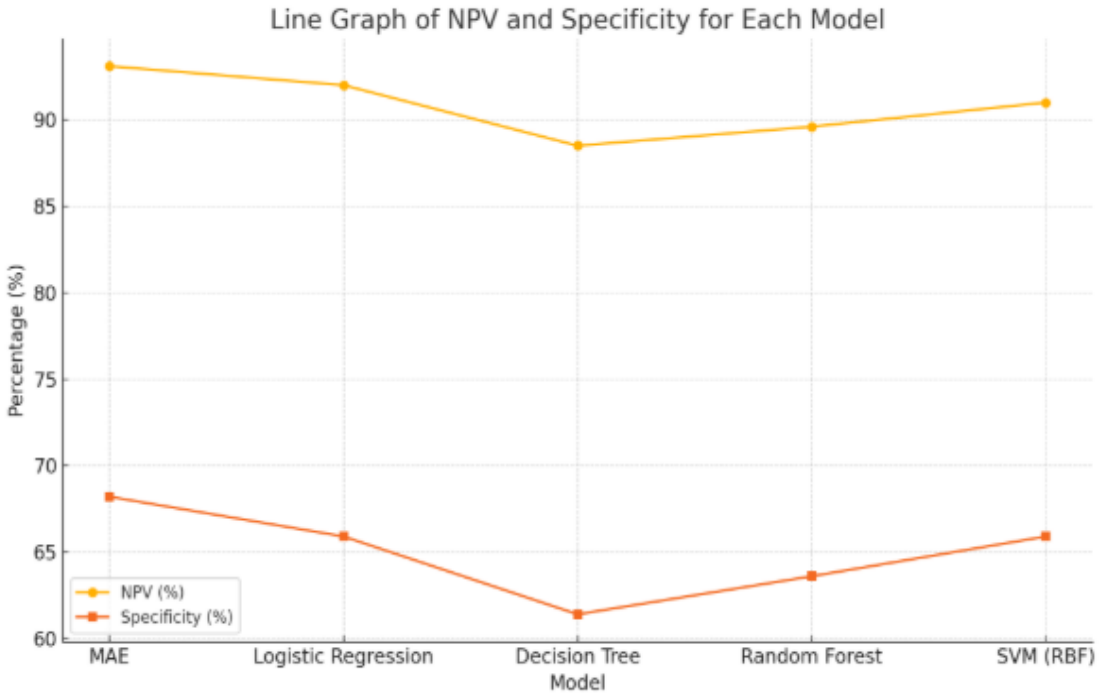


*Fig. 2. Comparison of NPV and Specificity across Sepsis Rule-Out Models*

Figure 1 offers a clear side-by-side look at the NPV and Specificity values for each model, presented in a grouped bar format. This layout highlights that while the differences in NPV among the top three models: MAE, Logistic Regression, and SVM are quite close, MAE stands out with a noticeable advantage. Specificity values have more variations, and once again, MAE has the best. These results validate the merit of MAE as a strong baseline for exclusion tasks, particularly in settings where negative classification safety is a high priority. Figure 2 plots the variations in NPV and Specificity for all five models that were under consideration for sepsis rule-out. The MAE baseline wins out on both measures across models, demonstrating its strong performance in excluding safe non-septic patients. Though SVM (RBF) and Logistic Regression lag behind, Decision Tree lags in both classes, as expected, affirming its restricted suitability for high-confidence negative prediction. The consistent difference between NPV and Specificity throughout models also helps to reflect the models' varying trade-off between the safe exclusion of sepsis and minimizing false positives.

## 5 Conclusion

This study recharacterizes the sepsis prediction challenge as a rule-out problem, with clinical purpose of safely excluding patients that are unlikely to become septic in order to allow better critical care resource allocation. We compared a Masked Autoencoder (MAE) baseline to four supervised learners: Logistic Regression, Decision Tree, Random Forest, and SVM (RBF), on six common performance measures, but with a bias towards Negative Predictive Value (NPV) and Specificity in evaluation. Results show that the MAE performed best with respect to NPV (93.1%) and specificity (68.2%), highlighting its ability to identify patients with low risk most effectively. Logistic Regression and SVM were also shown to perform competitively with small differences in NPV and specificity from MAE, suggesting that they remain viable options in cost-sensitive clinical settings. The worst performance among the two most critical measures was shown by the decision tree, suggesting the risks associated with implementing highly simplistic or shallow models for making high-risk exclusionary decisions. Overall, findings from this study support the use of self-supervised learning, even on sparse features, as a solid foundation for rule-out applications in clinical AI.

### Limitations of the Study

1. The model uses only age and sex, which limits its ability to capture the full clinical complexity of sepsis.
2. The lack of temporal clinical data prevents learning from disease progression patterns over time.
3. The small external validation cohort may limit the strength of generalization claims.
4. The dataset’s specific origin may restrict applicability across different healthcare settings.
5. The model’s simplicity may not fully reflect real-world clinical decision-making complexity.

## Consideration for Future Studies

Future work must tackle both technical validation and real-world integration. On the technical side, models must be tested on larger, diverse datasets and perhaps expanded to include more

sophisticated temporal or clinical features without sacrificing their label-efficient architecture. Real-world clinical trials are needed to determine their true impact on care decisions and resource use. For real-world implementation of the model in clinical workflows, the application of Management Information Systems (MIS) concepts such as Technology Acceptance Model (TAM) and WHO SMART guidelines can guide effective deployment. This includes ensuring the model is useful, user-friendly, and adherent to data governance standards.

## Acknowledgement

We gratefully acknowledge the use of the Sepsis Survival Minimal Clinical Records (SSMCR) dataset, hosted at the UCI Machine Learning Repository, for this research.

## References

Balcan, M. F., & Sharma, D. (2024). Learning accurate and interpretable decision trees. *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence (UAI)*, 288–307.

Cheng, Z., Yang, Y., Wang, W., Ye, Z., Hong, L., Song, Y., et al. (2023). TimeMAE: Self-supervised representations of time series with decoupled masked autoencoders. *IEEE Transactions on Knowledge and Data Engineering, 36*(1), 448–461. https://doi.org/10.1109/TKDE.2023.3327180

Chicco, D., & Jurman, G. (2020). *Sepsis Survival Minimal Clinical Records* [Data set]. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/891/sepsis+survival+minimal+clinical+records

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Annals of Internal Medicine, 162*(1), 55–63. https://doi.org/10.7326/M14-0697

Gavrikov, P. (2021, June 12). Best practices for neural network exports to ONNX. *Medium*. https://towardsdatascience.com/best-practices-for-neural-network-exports-to-onnx-99f23006c1d5

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

Heaven, W. D. (2021, January 27). An algorithm that predicts deadly infections is often flawed. *Wired*. https://www.wired.com/story/algorithm-predicts-deadly-infections-often-flawed/

Henrickson, D. R., Howard, A. E., Johnson, J. R., Simpson, S. Q., Lopansri, B. K., Yost, C. C., et al. (2022). Prospective validation of a host response assay to guide early sepsis care in the emergency department. *Critical Care Explorations, 4*(9), e0778. https://doi.org/10.1097/CCE.0000000000000778

Holden, R. J., & Karsh, B. T. (2010). The technology acceptance model: Its past and its future in health care. *Journal of Biomedical Informatics, 43*(1), 159–172. https://doi.org/10.1016/j.jbi.2009.07.002

Jin, H., Che, C., Liu, F., Zhou, T., Lu, P., & Yin, J. (2025). Self-supervised masked autoencoder for ICU mortality prediction. *arXiv*. https://arxiv.org/abs/2502.16834

Johnson, A. E. W., Ghassemi, M. M., Nemati, S., Niehaus, K. E., Clifton, D. A., & Clifford, G. D. (2023). Machine learning and sepsis: When computers don't compute. *Critical Care, 27*(1), 87. https://doi.org/10.1186/s13054-023-04378-w

Kaukonen, K. M., Bailey, M., Pilcher, D., Cooper, D. J., & Bellomo, R. (2015). Systemic inflammatory response syndrome criteria in defining severe sepsis. *The New England Journal of Medicine, 372*(17), 1629–1638. https://doi.org/10.1056/NEJMoa1415236

Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2023). Ti-MAE: Self-supervised masked time series autoencoders. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 16589–16604.

Mantosic, L., Khanna, S., Saeed, M., Mark, R. G., & Celi, L. A. (2019). Early sepsis prediction in intensive care units using long short-term memory networks on MIMIC-III. *Journal of Biomedical Informatics, 99*, 103290. https://doi.org/10.1016/j.jbi.2019.103290

Mao, Q., Jay, M., Hoffman, J. L., Calvert, J., Barton, C., Shimabukuro, D., et al. (2023). Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open, 13*(2), e068593. https://doi.org/10.1136/bmjopen-2022-068593

Muliokela, R. K., Banda, K., Hussen, A. M., Malumo, S. B., Kashoka, A., Chibwe, K. C., et al. (2025). Implementation of WHO SMART guidelines – Digital Adaptation Kits in Pathfinder Countries in Africa: Processes and early lessons learned. *JMIR Medical Informatics, 13*, e58858. https://doi.org/10.2196/58858

Philpott, J. (2024, March 15). Cytovale's sepsis diagnostic test demonstrates 97.5% NPV in latest study. *Clinical Trials Arena*. https://www.clinicaltrialsarena.com/news/cytovales-sepsis-diagnostic-test-demonstrates-97-5-npv-in-latest-study/

PyTorch Documentation. (2025, March 5). *torch.onnx.export*. https://pytorch.org/docs/stable/onnx.html

PyTorch Foundation. (2024, January 30). *PyTorch 2.2 release notes*. https://pytorch.org/blog/pytorch2-2/

Rodríguez, A., Mendoza, D., Gallego, J. D., & Mañez, R. (2022). Machine learning for sepsis prediction in low-resource intensive care units: A systematic review. *The Lancet Digital Health, 4*(12), e906–e915. https://doi.org/10.1016/S2589-7500(22)00165-4

Rudd, K. E., Johnson, S. C., Agesa, K. M., Shackelford, K. A., Tsoi, D., Kievlan, D. R., 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

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, 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

Tang, C., & Zhang, Y. (2022). MTSMAE: Masked autoencoders for multivariate time series forecasting. *Advances in Neural Information Processing Systems, 35*, 35271–35283.

University of Pittsburgh. (2020, January 16). Sepsis kills 1 in 5 globally, double previous estimate. *Pittwire*. https://www.pittwire.pitt.edu/news/sepsis-kills-1-5-globally-double-previous-estimate

van de Sande, A., & Nijhuis, A. (2024). Understanding overfitting in random forest for probability estimation: A visualization and simulation study. *Diagnostic and Prognostic Research, 8*(1), 12. https://doi.org/10.1186/s41512-024-00177-1

Wang, D., Li, Y., Chen, Y., Wang, F., & Hu, X. (2024). Multimodal transformer for early sepsis prediction. *IEEE Journal of Biomedical and Health Informatics, 28*(1), 112–123. https://doi.org/10.1109/JBHI.2023.3327181

World Health Organization. (2022). *WHO SMART guidelines: Standards-based, machine-readable, and adaptive*. World Health Organization. https://www.who.int/teams/digital-health-and-innovation/smart-guidelines

World Health Organization. (2023). *Global report on the epidemiology and burden of sepsis*. World Health Organization. https://www.who.int/publications/i/item/9789240064669

Xu, J., Wang, Y., Chen, Y., Tang, B., & Liu, T. Y. (2025). Multimodal transformers for early sepsis prediction. *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, 1–12.

Yoon, J., Jarrett, D., & van der Schaar, M. (2023). Time series representation learning via temporal and contextual contrasting. *Proceedings of Machine Learning for Healthcare*, 1–22. PMLR.

Zhang, X., & Li, Y. (2025). A survey of recent advances in support vector machine classification. *Journal of Machine Learning Research, 28*(3), 145–178.

Zhou, F., Chen, T., & Lei, B. (2020). Predicting fatality of COVID-19 patients using logistic regression. *medRxiv*. https://doi.org/10.1101/2020.06.30.20142902

Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., & Prasanna, P. (2022). Self pre-training with masked autoencoders for medical image classification. *arXiv*. https://arxiv.org/abs/2203.05573

Zhou, T., Ma, Z., Wen, X., Wang, X., Sun, L., & Jin, R. (2022). Masked autoencoders for multivariate time series forecasting. *arXiv*. https://arxiv.org/abs/2203.11599