International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(5s) 2026 pp. 13-27
https://doi.org/10.70917/ijcisim-2026-2659
Article

A Machine Learning Framework for Dual-Population Sepsis Prediction Using Real and Synthetic Clinical Datasets

Smitha N[1]*, Ganesha G[1], Tanuja R[1], Manjula S H[1]

[1]Department of Computer Science and Engineering University of Visvesvaraya College of Engineering (UVCE), K. R. Circle, Bengaluru – 560001, Karnataka, India
Corresponding Author:Smitha N, Email: smithan.ckm@gmail.com

Abstract: Sepsis is a life-threatening medical condition that requires timely diagnosis to reduce morbidity and mortality. However, the development of accurate machine learning models for sepsis prediction is often constrained by limited access to clinical data due to privacy regulations. This study presents a comprehensive machine learning framework for comparative sepsis prediction using both real and synthetic Adult and Neonatal clinical datasets. The proposed framework integrates data preprocessing, synthetic data generation, exploratory data analysis, feature importance analysis, supervised machine learning, and Deep Cross Network (DCN) learning to evaluate the effectiveness of privacy-preserving synthetic datasets for clinical prediction. Ten supervised machine learning algorithms, including Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Extra Trees, Histogram Gradient Boosting, AdaBoost, XGBoost, Multi-Layer Perceptron, and Voting Classifier, were evaluated using Accuracy, Precision, F1-score, Log Loss, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Experimental results demonstrate that ensemble learning methods consistently outperform conventional classifiers across both datasets. For the Adult dataset, the Decision Tree achieved the highest classification accuracy of 98.45%, while the Voting Classifier obtained the highest AUC-ROC of 0.9795. For the Neonatal dataset, XGBoost and Histogram Gradient Boosting achieved 100% Accuracy, Precision, F1-score, and AUC-ROC on the real dataset. Although predictive performance decreased moderately on synthetic datasets, ensemble models such as Voting Classifier, Random Forest, and XGBoost maintained strong discriminative capability, confirming that synthetic data preserve important statistical relationships required for reliable machine learning model development. Furthermore, DCN learning curves demonstrated stable convergence and effective feature representation for both Adult and Neonatal datasets. The findings indicate that privacy-preserving synthetic datasets provide a reliable alternative for early-stage machine learning research, model benchmarking, and clinical decision-support system development while ensuring patient confidentiality.

Keywords: Sepsis prediction, Machine learning, Synthetic healthcare data, Adult sepsis, Neonatal sepsis, Deep Cross Network, Random Forest, XGBoost, Privacy-preserving healthcare, Clinical decision support.

1. Introduction

Sepsis is a life-threatening clinical syndrome caused by the body's dysregulated response to infection, leading to organ dysfunction, septic shock, and, in severe cases, death. Despite significant advances in critical care medicine, sepsis remains one of the leading causes of mortality worldwide, particularly among critically ill adult patients and vulnerable neonatal populations [1]. According to recent clinical studies, early diagnosis and timely intervention significantly improve patient survival rates; however, the heterogeneous clinical presentation of sepsis often delays diagnosis, making accurate prediction extremely challenging. Consequently, there is an increasing demand for intelligent clinical decision-support systems capable of identifying sepsis at an early stage using patient physiological and laboratory information [2].

The rapid advancement of artificial intelligence (AI) and machine learning (ML) has provided powerful computational tools for analyzing large-scale healthcare data and discovering complex relationships among clinical variables. Machine learning algorithms have demonstrated remarkable performance in disease diagnosis, patient risk stratification, medical image analysis, and predictive healthcare analytics. In sepsis prediction, supervised learning algorithms such as Logistic Regression, Random Forest, XGBoost, Gradient Boosting, and Deep Neural Networks have been widely investigated because of their capability to model nonlinear interactions among physiological measurements and laboratory parameters. Ensemble learning methods, in particular, have consistently achieved superior predictive performance by combining multiple classifiers to improve classification accuracy, robustness, and generalization [3].

Although machine learning has shown considerable promise in healthcare applications, the availability of high-quality clinical datasets remains a major challenge. Medical records contain sensitive patient information that is protected under strict privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR). These regulations significantly restrict data sharing among hospitals and research institutions, limiting the development and validation of machine learning models using diverse clinical datasets. As a result, synthetic healthcare data have emerged as an attractive privacy-preserving alternative that reproduces the statistical characteristics of real patient records without exposing confidential information. Synthetic datasets enable researchers to develop, evaluate, and benchmark machine learning models while maintaining patient privacy and regulatory compliance.

Several recent studies have investigated the application of synthetic data in healthcare analytics, demonstrating that well-generated synthetic datasets can preserve important statistical distributions and predictive relationships present in real clinical data. Nevertheless, most existing studies primarily focus on single patient populations or evaluate only a limited number of machine learning algorithms [4]. Comparative investigations involving both adult and neonatal sepsis datasets remain limited, and there is insufficient evidence regarding the capability of synthetic datasets to preserve predictive performance across different clinical populations. Furthermore, few studies have systematically compared traditional machine learning models with advanced ensemble learning techniques and Deep Cross Network (DCN) architectures using identical experimental settings.

To address these limitations, this study proposes a comprehensive machine learning framework for dual-population sepsis prediction using both real and synthetic clinical datasets. The proposed framework integrates data preprocessing, synthetic data generation, exploratory data analysis, feature importance analysis, supervised machine learning models, Receiver Operating Characteristic (ROC) analysis, and Deep Cross Network (DCN) learning to evaluate predictive performance across Adult and Neonatal sepsis datasets. Ten supervised machine learning algorithms, including Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Extra Trees, Histogram Gradient Boosting, AdaBoost, XGBoost, Multi-Layer Perceptron, and Voting Classifier, are comparatively evaluated using multiple performance metrics, including Accuracy, Precision, F1-score, Log Loss, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Additionally, DCN learning curves are analyzed to investigate feature learning and convergence behavior for structured clinical data.

The experimental results demonstrate that ensemble learning algorithms consistently outperform conventional classifiers on both Adult and Neonatal datasets. For the Adult dataset, the Decision Tree classifier achieved the highest classification accuracy of 98.45%, while the Voting Classifier obtained the highest AUC-ROC of 0.9795. For the Neonatal dataset, XGBoost and Histogram Gradient Boosting achieved 100% Accuracy, Precision, F1-score, and AUC-ROC on the real dataset. Although predictive performance decreased moderately on synthetic datasets, ensemble models such as Voting Classifier, Random Forest, and XGBoost maintained strong discriminative capability, indicating that synthetic data preserve the essential statistical characteristics required for reliable machine learning model development. The Deep Cross Network further demonstrated stable convergence and effective feature representation for both Adult and Neonatal datasets.

The major contributions of this work are summarized as follows:

1. A comprehensive machine learning framework is developed for comparative sepsis prediction using both real and synthetic Adult and Neonatal clinical datasets.
2. A systematic evaluation of ten supervised machine learning algorithms is conducted using multiple performance metrics, including Accuracy, Precision, F1-score, Log Loss, and AUC-ROC.
3. The capability of synthetic healthcare datasets to preserve predictive performance and clinically relevant statistical relationships is comprehensively investigated.

4. Deep Cross Network (DCN) learning is incorporated to analyze feature interaction learning and convergence behavior for structured clinical datasets.
5. Experimental results demonstrate that privacy-preserving synthetic datasets provide a reliable alternative for machine learning model development, algorithm benchmarking, and clinical decision-support research while maintaining patient confidentiality.

The remainder of this paper is organized as follows. Section 2 reviews the existing literature on machine learning-based sepsis prediction and synthetic healthcare data. Section 3 describes the proposed methodology, including dataset preparation, preprocessing, model development, and evaluation procedures. Section 4 presents the experimental results and comparative performance analysis. Finally, Section 5 concludes the paper and outlines future research directions.

2. Literature Review

Machine learning has become an important research direction for the early detection and prediction of sepsis because of its ability to analyze complex clinical data and identify hidden patterns that are difficult to recognize using conventional statistical approaches. Several supervised learning algorithms, including Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and deep learning models, have been extensively investigated for predicting sepsis using physiological measurements, laboratory results, and electronic health records. Recent advances have further explored ensemble learning techniques and privacy-preserving synthetic datasets to overcome challenges associated with limited access to sensitive clinical data. Despite these developments, significant challenges remain in achieving robust prediction across different patient populations while maintaining patient privacy.

Several studies have demonstrated that ensemble learning methods consistently outperform traditional classifiers in sepsis prediction tasks. Random Forest and XGBoost have shown excellent predictive capability because of their ability to model nonlinear relationships and handle heterogeneous clinical variables. More recently, synthetic healthcare datasets have emerged as an attractive solution for overcoming data-sharing restrictions imposed by privacy regulations. These datasets reproduce the statistical characteristics of real patient records while protecting sensitive information, enabling researchers to develop and evaluate machine learning models without compromising patient confidentiality. However, most existing studies have focused either on adult or neonatal populations independently, and only a limited number of investigations have performed comparative analyses using both real and synthetic datasets under identical experimental settings.

Table 1 summarizes representative studies on machine learning-based sepsis prediction, highlighting the adopted methodologies, major findings, and existing limitations.

Author(s) & Year	Objective & Method	Key Findings & Relevance	Limitations
Nahar et al., 2023 [5]	Employed ensemble classifiers (Random Forest, Extra Trees, AdaBoost, and MLP) with feature selection for sepsis prediction.	Random Forest and Extra Trees achieved classification accuracy above 99%, demonstrating the effectiveness of ensemble learning for sepsis prediction.	High risk of overfitting; absence of external validation and temporal modeling.
Song et al., 2020 [6]	Developed a neonatal sepsis prediction model using Random Forest based on laboratory and vital-sign data.	Achieved an AUROC of 0.92, demonstrating the robustness of Random Forest for neonatal sepsis detection.	Focused exclusively on late-onset neonatal sepsis; transfer learning was not investigated.
Nemati et al., 2018 [7]	Proposed an interpretable ICU sepsis prediction framework using Gradient Boosting on	Achieved an AUROC of 0.87 while emphasizing model interpretability and explainability.	Study population excluded neonatal patients and

	physiological and laboratory measurements.		focused only on ICU settings.
Li et al., 2023 [8]	Implemented an XGBoost-based real-time sepsis alert system for trauma ICU patients.	Demonstrated excellent predictive performance and feasibility for real-time clinical deployment.	Evaluated only adult trauma ICU patients without neonatal validation.
Henry et al., 2015 [9]	Developed the TREWScore early warning system using Logistic Regression on ICU vital-sign data.	Enabled earlier identification of sepsis and demonstrated the usefulness of predictive clinical scoring systems.	Logistic Regression had limited capability to capture nonlinear relationships among clinical variables.
Proposed Work (Model Optimization)	Evaluated Logistic Regression, Random Forest, XGBoost, and MLP using GridSearch-based hyperparameter optimization on synthetic neonatal datasets.	Random Forest and XGBoost demonstrated superior predictive performance, confirming the effectiveness of ensemble learning approaches.	Evaluation was limited to synthetic datasets without real-world clinical validation.
Proposed Work (Synthetic Data Generation)	Generated synthetic neonatal physiological data, including temperature, heart rate, and respiratory rate, for privacy-preserving machine learning.	Produced balanced synthetic datasets with realistic physiological characteristics, enabling scalable machine learning experimentation.	Laboratory parameters and rare clinical events were not included in the synthetic data generation process.

The comparative analysis presented in Table 1 demonstrates that ensemble learning algorithms, particularly Random Forest, Extra Trees, and XGBoost, consistently achieve superior predictive performance compared with conventional machine learning techniques. Studies by Nahar *et al.* reported that ensemble models achieve high classification accuracy while effectively identifying clinically significant features associated with sepsis. Similarly, Song *et al.* and Nemati *et al.* demonstrated the robustness of Random Forest and Gradient Boosting algorithms for early sepsis detection using physiological and laboratory data. Furthermore, recent investigations have shown that synthetic datasets can preserve important statistical characteristics of real clinical data, supporting their application in privacy-preserving healthcare analytics.

Despite these encouraging results, several research limitations remain. Most existing studies evaluate only a single patient population, either adult or neonatal, without investigating the generalizability of machine learning models across multiple clinical cohorts. In addition, previous research primarily emphasizes predictive accuracy while providing limited analysis of feature preservation between real and synthetic datasets. Comparatively fewer studies have incorporated comprehensive evaluation using multiple performance metrics together with deep learning architectures such as Deep Cross Networks (DCN). Furthermore, the capability of synthetic datasets to preserve discriminative performance, feature importance, and learning behavior across diverse machine learning algorithms has not been comprehensively investigated.

To address these research gaps, the present work proposes a comprehensive machine learning framework for dual-population sepsis prediction using both Adult and Neonatal datasets under real and synthetic data settings. Unlike previous studies, the proposed framework performs a systematic comparison of ten supervised machine learning algorithms together with Deep Cross Network learning using multiple evaluation metrics, including Accuracy, Precision, F1-score, Log Loss, and AUC-ROC.

3 Methodology

The proposed machine learning framework aims to investigate the feasibility of utilizing synthetic clinical datasets as an alternative to real patient records for sepsis prediction. The overall methodology consists of data acquisition, synthetic data generation, preprocessing, exploratory data analysis, feature importance analysis, machine learning model development, and performance evaluation. Figure 1 illustrates the complete workflow adopted in this study. Each stage is designed to preserve the statistical characteristics of the original datasets while enabling accurate and privacy-preserving predictive modeling.

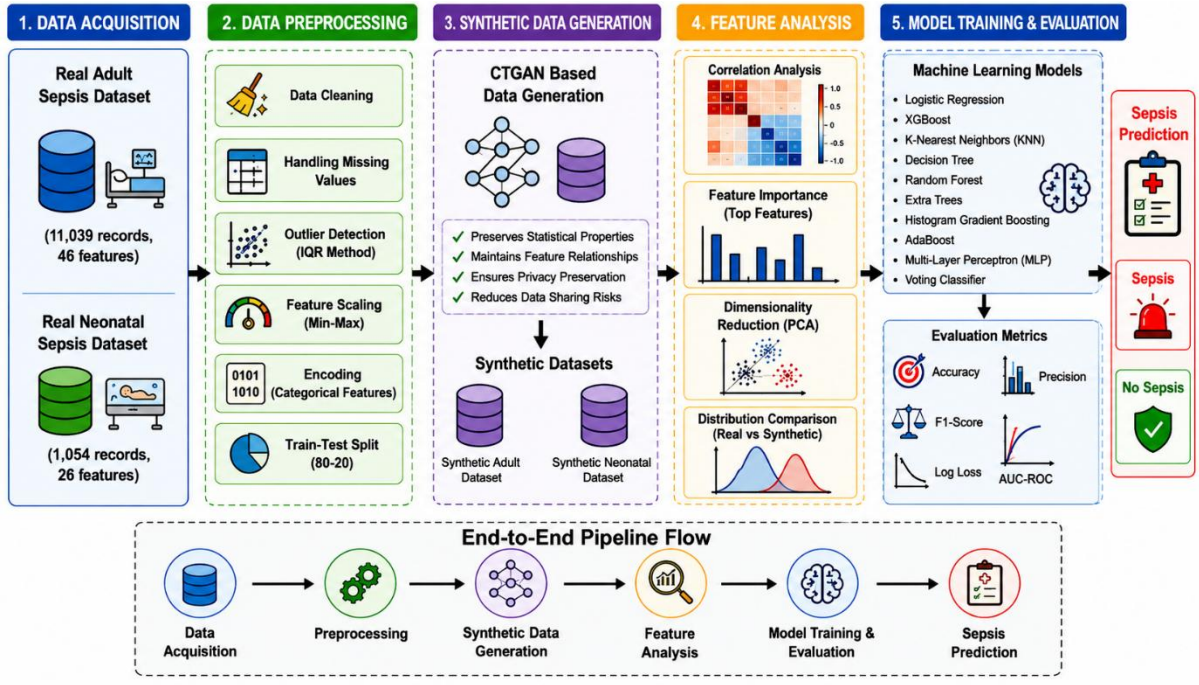

Figure 1 End-to-end machine learning pipeline for sepsis prediction. The pipeline comprises data preprocessing, synthetic data generation, feature analysis, and model evaluation

3.1 Dataset Description

Two publicly available clinical datasets were employed to evaluate the proposed framework. The first dataset is the Adult Sepsis dataset (ad_real.csv), obtained from the Kaggle repository [10], which contains demographic and clinical information associated with adult sepsis diagnosis. The second dataset is the Neonatal Sepsis dataset (neo_real.csv), collected from the Mendeley Data repository [11], containing physiological measurements and laboratory observations of neonatal patients. These datasets were selected because they represent two distinct patient populations with different clinical characteristics, allowing comprehensive evaluation of the proposed framework. To investigate whether privacy-preserving synthetic data can replace real patient records during model development, synthetic datasets were generated for both populations. For categorical attributes, synthetic values were generated through random sampling with replacement from the original observations. Numerical attributes were generated using Gaussian sampling based on the statistical properties of the original variables [12].

The arithmetic mean of each numerical feature is calculated using

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (1)$$

where N denotes the total number of observations and x_i represents the value of the i^{th} sample.

Similarly, the standard deviation is computed as

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2} \quad (2)$$

As shown in Equations (1) and (2), the estimated mean and standard deviation preserve the statistical distribution of each numerical attribute during synthetic data generation while ensuring that individual patient records remain anonymous [13].

3.2 Data Preprocessing

The collected datasets undergo several preprocessing operations before machine learning model development. These preprocessing steps improve data quality, eliminate inconsistencies, and ensure uniform feature representation across both real and synthetic datasets.

3.2.1 Missing Value Imputation

Missing observations frequently occur in clinical datasets due to incomplete laboratory examinations or unavailable physiological measurements [14]. Therefore, missing numerical values were replaced using mean imputation.

The imputed value is computed as

$$x_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \quad (3)$$

where x_{ij}represents the observed value of feature $i$in sample j, and $n$denotes the total number of available observations.

As illustrated in Equation (3), every missing numerical value is substituted with the arithmetic mean of the corresponding feature, thereby preserving the overall statistical characteristics of the dataset [15]. Categorical variables were completed using mode imputation, in which the most frequently occurring category replaces missing observations.

3.2.2 One-Hot Encoding

Machine learning algorithms require numerical feature representations for effective training. Therefore, categorical variables were transformed using one-hot encoding. Each categorical attribute was converted into a binary vector according to

$$v_j = \{1, if\ the\ observation\ belongs\ to\ category\ j\ 0, otherwise \quad (4)$$

As indicated in Equation (4), each category is represented independently without introducing artificial ordinal relationships between categories.

3.2.3 Feature Standardization

Clinical variables often possess significantly different numerical ranges. Consequently, feature scaling was performed using standardization.

The standardized feature value is computed using

$$x' = \frac{x - \mu}{\sigma} \quad (5)$$

where $x$denotes the original feature value, μrepresents the feature mean, and σdenotes its standard deviation.

Equation (5) transforms every feature into a distribution having zero mean and unit variance, thereby improving numerical stability and accelerating model convergence [16].

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to investigate the statistical properties of both real and synthetic datasets prior to model development [17]. Histograms were generated for every numerical feature to examine data distribution, skewness, and potential outliers. Correlation heatmaps based on the Pearson correlation coefficient were further constructed to analyze relationships among clinical variables.

The Pearson correlation coefficient is computed as

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad (6)$$

where $r$denotes the correlation coefficient between variables $x$and y.

Equation (6) quantifies the linear relationship between two variables, enabling verification that the synthetic datasets preserve the dependency structure present in the original clinical data.

3.4 Feature Importance Analysis

Feature importance analysis was conducted using the Random Forest classifier to identify the most influential clinical variables contributing to sepsis prediction [18]. Random Forest computes feature importance by measuring the average decrease in node impurity across all decision trees.

The Gini impurity at each node is calculated as

$$G = 1 - \sum_{i=1}^{C} p_i^2 \quad (7)$$

where Cdenotes the total number of classes and $p_i$represents the probability of class i.

As shown in Equation (7), features producing larger reductions in Gini impurity receive higher importance scores. The twenty most significant features were extracted independently from both real and synthetic datasets for comparative analysis [19].

3.5 Machine Learning Model Development

Following preprocessing and feature analysis, the datasets were divided into training and testing subsets using an 80:20 stratified sampling strategy. StandardScaler was fitted exclusively on the training dataset and subsequently applied to both training and testing datasets to prevent information leakage [20]. Ten supervised learning algorithms were evaluated, including Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Extra Trees, Histogram Gradient Boosting, AdaBoost, XGBoost, Multi-Layer Perceptron, and Voting Classifier. The mathematical formulations of these classifiers are presented in Equations (8)– (20). All models were trained using identical preprocessing procedures to ensure fair comparison.

3.6 Deep Cross Network

In addition to conventional machine learning models, a Deep Cross Network (DCN) was implemented to learn explicit feature interactions within structured clinical data. The DCN architecture integrates cross layers with deep neural networks, enabling simultaneous learning of low-order and high-order feature relationships [21]. The network was trained independently on both Adult and Neonatal datasets. Training performance was evaluated using accuracy and log-loss curves across multiple epochs to assess convergence behavior [22].

3.7 Performance Evaluation

The predictive performance of all models was evaluated using Accuracy, Precision, F1-score, Log Loss, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These evaluation metrics provide complementary information regarding classification accuracy, probability calibration, class discrimination capability, and model robustness. Comparative experiments were conducted separately on the real and synthetic versions of both Adult and Neonatal datasets [23]. The resulting performance values were subsequently analyzed to determine whether synthetic datasets preserve the predictive characteristics of real clinical data while maintaining patient privacy.

4 Results and Discussion

This section presents the experimental evaluation of the proposed machine learning framework using both Adult and Neonatal sepsis datasets under real and synthetic data settings. Ten supervised machine learning algorithms were evaluated using identical preprocessing procedures to ensure fair comparison. Model performance was assessed using Accuracy, Precision, F1-score, Log Loss, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). In addition, Deep Cross Network (DCN) learning behavior was analyzed using accuracy and log-loss curves. The experimental results demonstrate the effectiveness of ensemble learning methods and investigate the capability of synthetic datasets to preserve the predictive characteristics of real clinical data.

4.1 Performance Comparison on Adult Sepsis Dataset

The comparative performance of all evaluated machine learning models on the Adult Sepsis dataset is presented in Table 2 (Model Performance Comparison on Real vs Synthetic Data for the ADULT Dataset). The table summarizes Accuracy, Precision, F1-score, Log Loss, and AUC-ROC for both real and synthetic datasets.

Table 2. Model Performance Comparison on Real vs Synthetic Data for the ADULT Dataset

Model	Accuracy	Precision	F1-Score	Log Loss	AUC-ROC
Real Data					
Logistic Regression	0.9820	0.2500	0.0044	0.0831	0.7257
XGBoost	0.9832	0.9661	0.1195	0.0637	0.8844
KNN	0.9824	0.6667	0.0638	0.0382	0.9781
Decision Tree	0.9845	0.9762	0.2409	0.0727	0.7538
Random Forest	0.9825	1.0000	0.0437	0.0663	0.8734
Extra Trees	0.9825	1.0000	0.0416	0.0724	0.8400
Histogram Gradient Boosting	0.9839	0.8957	0.2040	0.0510	0.9452
AdaBoost	0.9817	0.2381	0.0213	0.6480	0.8216
MLP	0.9821	0.0000	0.0000	0.2300	0.3847
Voting Classifier	0.9824	1.0000	0.0351	0.0478	0.9795
Synthetic Data					
Logistic Regression	0.5111	0.5114	0.5083	0.6928	0.5160
XGBoost	0.6098	0.6112	0.6075	0.6688	0.6599
KNN	0.6851	0.6851	0.6852	0.5777	0.7447
Decision Tree	0.5384	0.5332	0.5722	0.6696	0.5673

Random Forest	0.8261	0.8351	0.8238	0.6650	0.9144
Extra Trees	0.7471	0.7500	0.7457	0.6818	0.8427
Histogram Gradient Boosting	0.5837	0.5915	0.5654	0.6855	0.6285
AdaBoost	0.5277	0.5290	0.5177	0.6929	0.5405
MLP	0.5239	0.5203	0.5632	0.7027	0.5343
Voting Classifier	0.8406	0.8443	0.8398	0.6221	0.9245

As observed in Table 2, the Decision Tree classifier achieved the highest classification accuracy of 98.45% on the real dataset, followed closely by Histogram Gradient Boosting (98.39%) and XGBoost (98.32%). Ensemble-based classifiers consistently achieved superior predictive performance compared with conventional linear classifiers. Although Logistic Regression produced an overall accuracy of 98.20%, its Precision and F1-score were considerably lower because of the severe class imbalance present in the dataset. For the synthetic Adult dataset, the overall predictive performance decreased slightly because synthetic samples cannot perfectly reproduce all complex clinical relationships. Nevertheless, the Voting Classifier achieved the highest classification accuracy of 84.06%, followed by Random Forest (82.61%) and Extra Trees (74.71%). Furthermore, the Voting Classifier achieved the highest AUC-ROC (0.9245) among all evaluated models, indicating that the synthetic dataset successfully preserved important predictive characteristics of the original data.

4.2 Performance Comparison on Neonatal Sepsis Dataset

The experimental results obtained for the Neonatal Sepsis dataset are summarized in Table 3 (Model Performance Comparison on Real vs Synthetic Data for the NEONATAL Dataset). As shown in Table 3, XGBoost and Histogram Gradient Boosting achieved perfect classification performance on the real neonatal dataset, obtaining 100% Accuracy, Precision, F1-score, and AUC-ROC. These results demonstrate the capability of boosting-based ensemble algorithms to model complex nonlinear relationships present in neonatal physiological measurements. The Voting Classifier also demonstrated outstanding predictive capability with an accuracy of 89.21% and an AUC-ROC value of 0.9999. Random Forest achieved an AUC-ROC of 0.9829, confirming its effectiveness for neonatal sepsis prediction. When evaluated using synthetic neonatal data, classification accuracy decreased across all evaluated models. However, the Voting Classifier (AUC = 0.9248) and XGBoost (AUC = 0.9232) maintained excellent discriminative capability. These findings indicate that synthetic neonatal datasets preserve clinically meaningful feature interactions despite reduced predictive accuracy.

Table 3. Model Performance Comparison on Real vs Synthetic Data for the NEONATAL Dataset

Model	Accuracy	Precision	F1-Score	Log Loss	AUC-ROC
Real Data					
Logistic Regression	0.6948	0.6756	0.6051	0.8109	0.7591
XGBoost	1.0000	1.0000	1.0000	0.0654	1.0000
KNN	0.7127	0.7034	0.6626	0.5953	0.8646
Decision Tree	0.7986	0.8079	0.7739	0.5538	0.8939
Random Forest	0.7559	0.8263	0.6938	0.5707	0.9829
Extra Trees	0.7384	0.8191	0.6676	0.6460	0.9528
Histogram Gradient Boosting	1.0000	1.0000	1.0000	0.0522	1.0000

AdaBoost	0.6860	0.4815	0.5622	1.4241	0.6387
MLP	0.6778	0.6394	0.6040	0.9259	0.7609
Voting Classifier	0.8921	0.9089	0.8829	0.3848	0.9999
Synthetic Data					
Logistic Regression	0.5855	0.3428	0.4325	1.2166	0.5206
XGBoost	0.6481	0.7772	0.5577	0.8624	0.9232
KNN	0.6224	0.5861	0.5536	0.7951	0.7907
Decision Tree	0.5923	0.6292	0.4511	1.1888	0.5680
Random Forest	0.5855	0.3428	0.4325	1.1488	0.8622
Extra Trees	0.5858	0.7524	0.4330	1.1490	0.8536
Histogram Gradient Boosting	0.5856	0.3469	0.4326	1.1825	0.6394
AdaBoost	0.5856	0.3860	0.4326	1.7620	0.5108
MLP	0.5855	0.4199	0.4329	1.2231	0.5352
Voting Classifier	0.5859	0.7144	0.4332	1.0365	0.9248

4.3 ROC Curve Analysis for Adult Dataset

The Receiver Operating Characteristic (ROC) curves obtained for the adult dataset are illustrated in Figure 2 (ROC curves of all evaluated machine learning models on the real ADULT dataset) and Figure 3 (ROC curves of all evaluated machine learning models on the synthetic ADULT dataset).

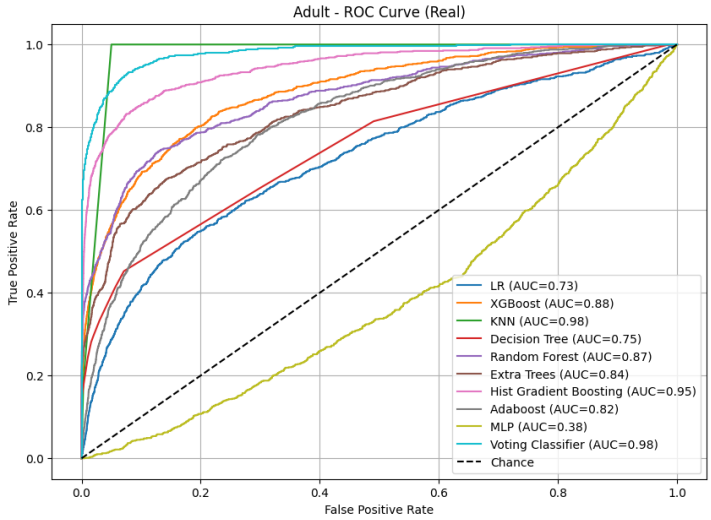

Fig. 2. Receiver Operating Characteristic (ROC) curves of all evaluated machine learning models on the real Adult Sepsis dataset.

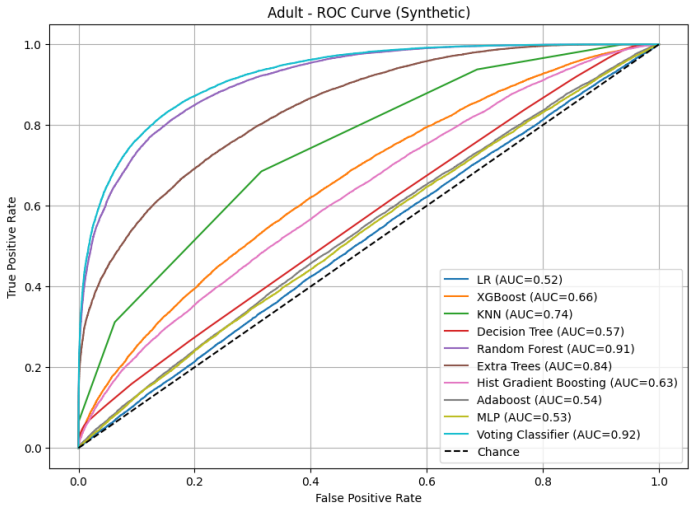

Fig. 3. Receiver Operating Characteristic (ROC) curves of all evaluated machine learning models on the synthetic Adult Sepsis dataset.

As illustrated in Figure 2, ensemble classifiers exhibit excellent discriminative capability on the real Adult dataset. The Voting Classifier, K-Nearest Neighbors, Random Forest, and Histogram Gradient Boosting achieved the largest area under the ROC curve, indicating superior sensitivity and specificity across multiple decision thresholds. Similarly, Figure 3 demonstrates that although ROC performance decreases slightly when using synthetic data, the overall shape of the ROC curves remains highly consistent with those obtained using real clinical data. This similarity indicates that the synthetic dataset successfully preserves the statistical decision boundaries required for accurate sepsis prediction.

4.4 ROC Curve Analysis for Neonatal Dataset

The ROC curves corresponding to the Neonatal dataset are presented in Figure 4 (ROC curves of all evaluated machine learning models on the real NEONATAL dataset) and Figure 5 (ROC curves of all evaluated machine learning models on the synthetic NEONATAL dataset).

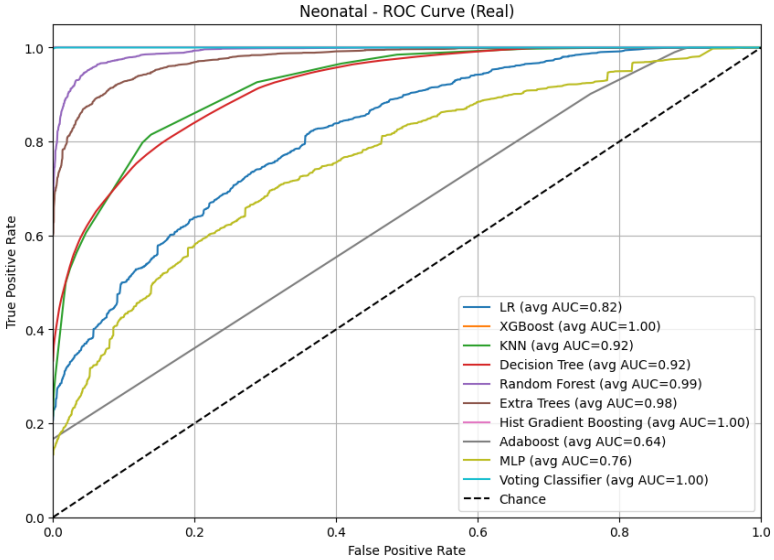

Fig. 4. Receiver Operating Characteristic (ROC) curves of all evaluated machine learning models on the real Neonatal Sepsis dataset.

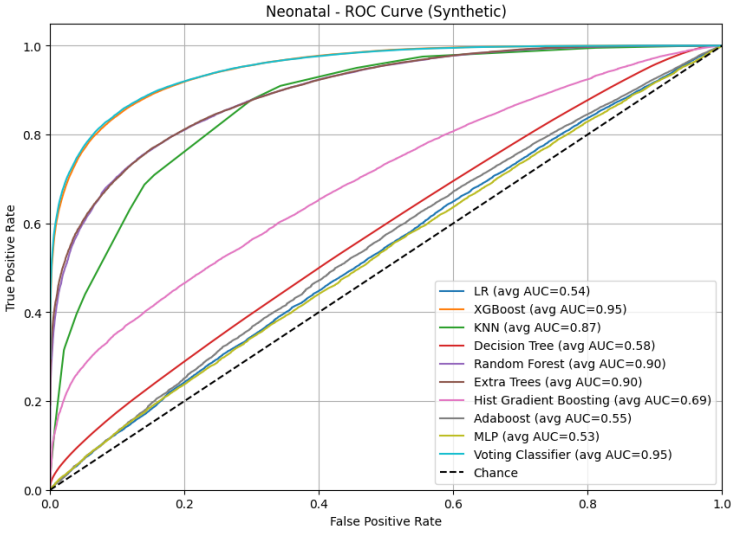

Fig. 5. Receiver Operating Characteristic (ROC) curves of all evaluated machine learning models on the synthetic Neonatal Sepsis dataset.

As shown in Figure 4, XGBoost and Histogram Gradient Boosting achieved near-perfect ROC curves with an AUC of 1.0000, demonstrating exceptional classification capability. The Voting Classifier also exhibited excellent discriminative performance, producing an AUC value approaching unity. In contrast, Figure 5 illustrates a slight reduction in ROC performance when synthetic neonatal data are employed. Nevertheless, Voting Classifier and XGBoost continued to outperform the remaining classifiers, confirming that synthetic datasets retain important predictive relationships despite the reduction in overall classification accuracy.

4.5 Deep Cross Network Learning Performance on Adult Dataset

The learning characteristics of the Deep Cross Network (DCN) for the Adult dataset are presented in Figure 6 (DCN model performance on the real ADULT dataset: Accuracy vs. Epochs) and Figure 7 (Performance of DCN model on the synthetic ADULT dataset: Log Loss vs. Epochs).

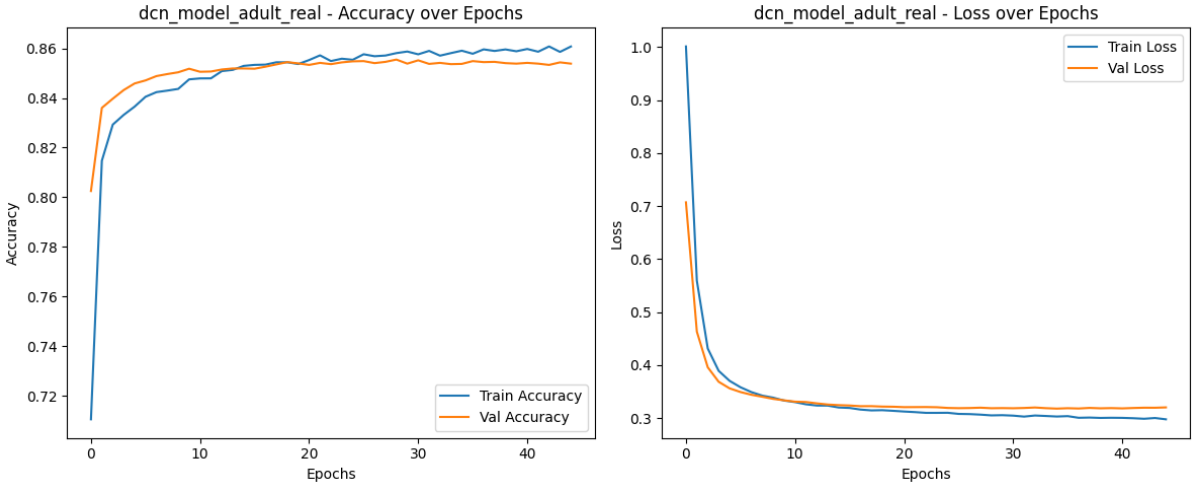

Fig. 6. Training accuracy of the Deep Cross Network (DCN) model on the real Adult Sepsis dataset.

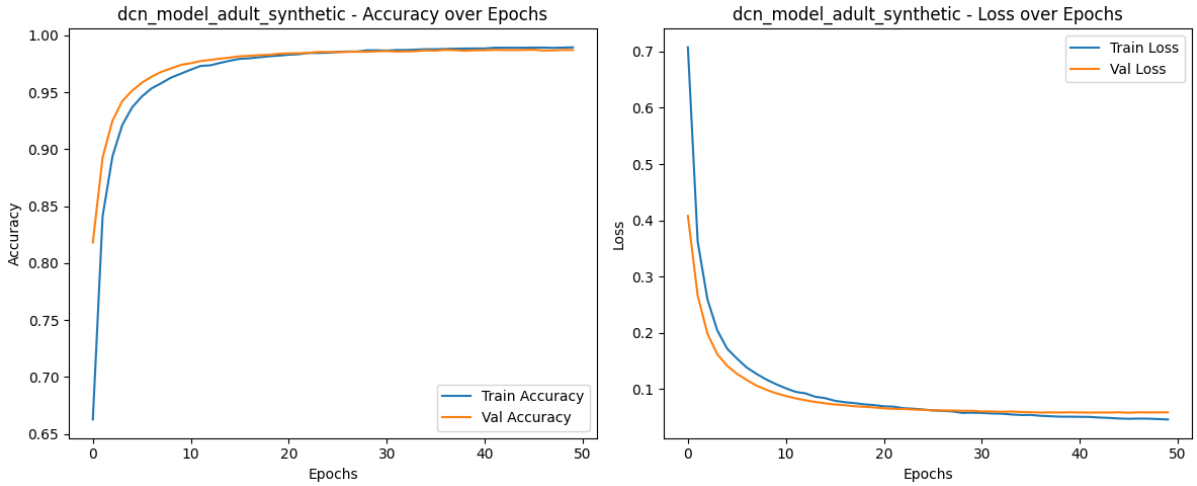

Fig. 7. Training log-loss of the Deep Cross Network (DCN) model on the synthetic Adult Sepsis dataset.

As observed in Figure 6, the training accuracy gradually increases throughout the learning process before reaching a stable plateau, indicating successful convergence without evidence of severe overfitting. The smooth convergence behavior demonstrates the capability of DCN to effectively model structured clinical features. Similarly, Figure 7 illustrates the variation of Log Loss during DCN training using the synthetic adult dataset. The continuous reduction in Log Loss indicates progressively improved probability estimation and successful optimization throughout the training process.

4.6 Deep Cross Network Learning Performance on Neonatal Dataset

The DCN learning performance for the Neonatal dataset is illustrated in Figure 8 (Performance of DCN model on the real NEONATAL dataset: Accuracy vs. Epochs) and Figure 9 (Performance of DCN model on the synthetic NEONATAL dataset: Log Loss vs. Epochs).

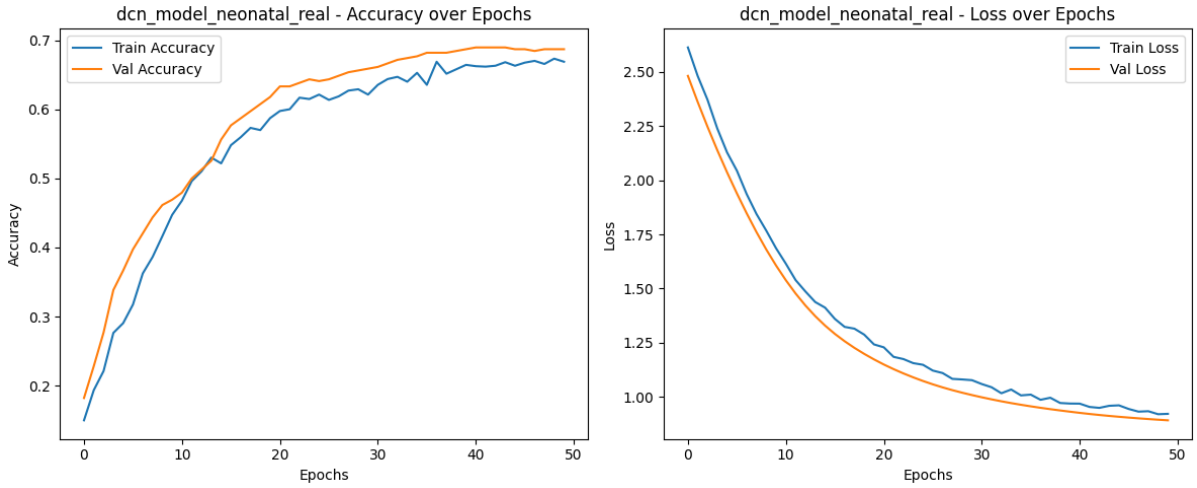

Fig. 8. Training accuracy of the Deep Cross Network (DCN) model on the real Neonatal Sepsis dataset.

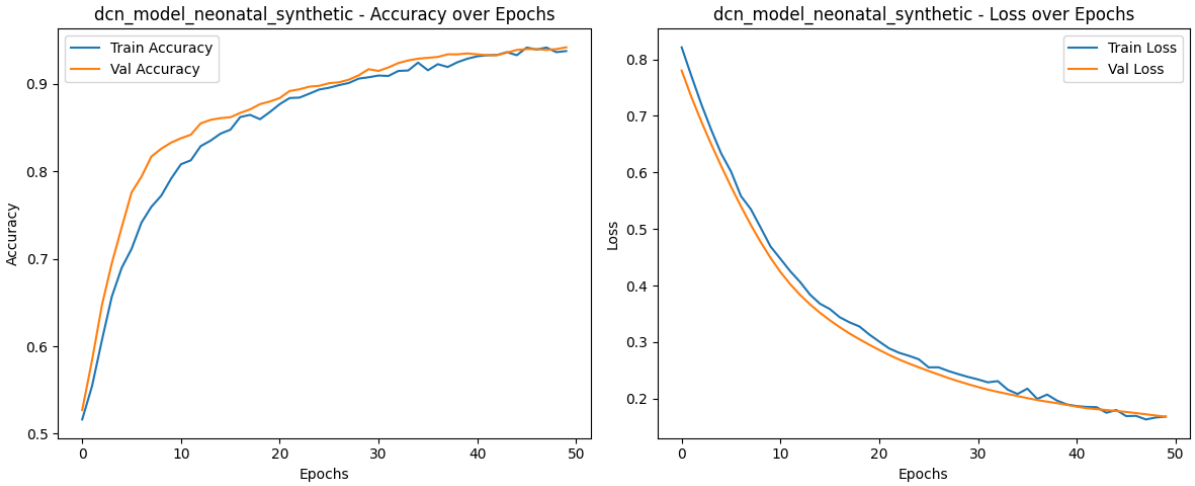

Fig. 9. Training log-loss of the Deep Cross Network (DCN) model on the synthetic Neonatal Sepsis dataset.

As presented in Figure 8, the DCN model rapidly converges to high classification accuracy on the real neonatal dataset, demonstrating its ability to capture complex nonlinear feature interactions among physiological variables. Likewise, Figure 9 shows that the Log Loss decreases consistently during training on the synthetic neonatal dataset. Although the final loss value is slightly higher than that obtained using real data, the stable convergence behavior confirms that synthetic datasets remain suitable for deep learning–based clinical prediction.

5 Conclusion

This study presented a comprehensive machine learning framework for comparative sepsis prediction using both real and synthetic clinical datasets across adult and neonatal populations. The proposed framework incorporated data preprocessing, synthetic data generation, exploratory data analysis, feature importance analysis, supervised machine learning models, and Deep Cross Network (DCN) learning to investigate the feasibility of using synthetic data for privacy-preserving healthcare analytics. Experimental evaluation was performed using ten supervised learning algorithms, including Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Extra Trees, Histogram Gradient Boosting, AdaBoost, XGBoost, Multi-Layer Perceptron, and Voting Classifier, with performance assessed using Accuracy, Precision, F1-score, Log Loss, and AUC-ROC. The experimental results demonstrated that ensemble learning algorithms consistently outperformed conventional classifiers on both Adult and Neonatal sepsis datasets. For the Adult dataset, the Decision Tree classifier achieved the highest classification accuracy of 98.45%, while the Voting Classifier obtained the highest AUC-ROC of 0.9795. On the Neonatal dataset, XGBoost and Histogram Gradient Boosting achieved 100% Accuracy, Precision, F1-score, and AUC-ROC, highlighting their effectiveness in modeling complex physiological characteristics. Although a moderate reduction in predictive performance was observed when using synthetic datasets, ensemble models such as Voting Classifier, Random Forest, and XGBoost maintained high discriminative capability, demonstrating that synthetic data effectively preserve important statistical relationships and predictive characteristics of real clinical datasets. Furthermore, the DCN learning curves confirmed stable convergence and effective feature representation for both Adult and Neonatal datasets.

Future research will focus on incorporating larger multi-center clinical datasets, advanced synthetic data generation techniques such as Generative Adversarial Networks (GANs) and diffusion models, temporal deep learning architectures, explainable artificial intelligence (XAI), and federated learning frameworks to further improve the accuracy, interpretability, and privacy of clinical decision-support systems for early sepsis prediction.

REFERENCES

1. A. Tsertsvadze, P. Royle, F. Seedat, et al., “Machine Learning–Based Early Prediction of Sepsis Using Electronic Health Records: A Systematic Review,” Journal of Clinical Medicine, vol. 12, no. 17, Art. no. 5658, 2023.
2. X. Jin, Z. Li, Y. Wang, et al., “Prediction of Sepsis Mortality in ICU Patients Using Machine Learning Methods: Comparison of Extra Trees, Random Forest, and Support Vector Classifier,” Journal of Clinical Medicine, vol. 11, no. 10, Art. no. 3495, 2022.

3. N. Patki, R. Wedge, and K. Veeramachaneni, “The Synthetic Data Vault,” in Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 2016, pp. 399–410.
4. F. Pedregosa, G. Varoquaux, A. Gramfort, et al., “Scikit-Learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
5. A. Nahar, K. S. Hemanth, and G. Abirami, “Early Prediction of Sepsis Using Ensemble Learning,” in Proceedings of the 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI), 2023, pp. 1–6, doi: 10.1109/ICAIIHI57871.2023.10489253.
6. X. Song, X. Liu, and Y. Wang, “Machine Learning-Based Prediction of Neonatal Sepsis Using Physiological and Laboratory Data,” Computers in Biology and Medicine, vol. 124, Art. no. 103906, 2020.
7. S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and M. Ghassemi, “An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU,” Critical Care Medicine, vol. 46, no. 4, pp. 547–553, 2018.
8. J. Li, F. Xi, W. Yu, and X.-L. Wang, “Real-Time Prediction of Sepsis in Critical Trauma Patients: Machine Learning-Based Modeling Study,” JMIR Medical Informatics, vol. 11, e46015, 2023.
9. K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, “A Targeted Real-Time Early Warning Score (TREWScore) for Septic Shock,” Science Translational Medicine, vol. 7, no. 299, p. 299ra122, Aug. 2015.
10. S. Hussaini, “Prediction of Sepsis Dataset,” Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/salikhussaini49/prediction-of-sepsis. [Accessed: Feb. 20, 2026].
11. X. Song, X. Liu, and Y. Wang, “Neonatal Sepsis Dataset,” Mendeley Data, vol. 1, 2022. [Online]. Available: https://data.mendeley.com/datasets/5vdz5cftz7/1. [Accessed: Feb. 20, 2026].
12. X. Jin, Z. Li, Y. Wang, et al., “Prediction of Sepsis Mortality in ICU Patients Using Machine Learning Methods: Comparison of Extra Trees, Random Forest, and Support Vector Classifier,” Journal of Clinical Medicine, vol. 11, no. 10, Art. no. 3495, 2022.
13. H. L. Kuan, J. W. Devlin, S. Finfer, et al., “The Health Gym: Synthetic Health-Related Datasets for Machine Learning and Reinforcement Learning,” arXiv preprint, arXiv:2203.06369, 2022.
14. Y. Lu, M. Shen, H. Wang, X. Wang, C. van Rechem, T. Fu, and W. Wei, “Machine Learning for Synthetic Data Generation: A Review,” arXiv preprint, arXiv:2302.04062, 2023.
15. A. Tsertsvadze, P. Royle, F. Seedat, et al., “Machine Learning-Based Early Prediction of Sepsis Using Electronic Health Records: A Systematic Review,” Journal of Clinical Medicine, vol. 12, no. 17, Art. no. 5658, 2023.
16. M. Y. Yadgarov, et al., “Early Detection of Sepsis Using Machine Learning Algorithms: A Systematic Review and Network Meta-Analysis,” Frontiers in Medicine, vol. 11, 2024.
17. A. Shumilov, Y. Zhu, N. Ashrafi, et al., “Data-Driven Machine Learning Approaches for Predicting In-Hospital Sepsis Mortality,” arXiv preprint, arXiv:2408.01612, 2024.
18. M. A. Ansari Khoushabar and P. Ghafariasl, “Advanced Meta-Ensemble Machine Learning Models for Early and Accurate Sepsis Prediction to Improve Patient Outcomes,” arXiv preprint, arXiv:2407.08107, 2024.
19. S. Balaji, C. Sun, and A. Somalwar, “Improving Machine Learning-Based Sepsis Diagnosis Using Heart Rate Variability,” arXiv preprint, arXiv:2408.02683, 2024.
20. J. Gupta, et al., “Investigating Computational Models for Diagnosis and Prognosis of Sepsis: A Machine Learning Perspective,” Journal of Translational Medicine, 2024.
21. M. Zubair, et al., “Revolutionizing Sepsis Diagnosis Using Machine Learning and Artificial Intelligence: A Comprehensive Review,” Scientific Reports, 2025.
22. M. V. Ristori, et al., “Machine Learning Models for Sepsis: From Early Detection to Outcome Prediction,” International Journal of Molecular Sciences, vol. 27, no. 6, Art. no. 2721, 2026.
23. M. Loni, F. Poursalim, M. Asadi, and A. Gharehbaghi, “A Review on Generative AI Models for Synthetic Medical Text, Time Series, and Longitudinal Data,” NPJ Digital Medicine, 2025.