International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 7s, 2025
https://www.theaspd.com/ijes.php

A Clinical Assessment Of Machine Learning Methods With Adaptive Synthetic Sampling Approach For Imbalanced Learning On Sepsis Prediction

M.Senthil kumar[1] , Dr. A. Krishna kumar[2]

[1]Research Scholar, Department of Computer Science, Sree Saraswati Thyagaraja College, Pollachi, Tamil Nadu, India, & Assistant Professor, Department of Computer Science, Nehru Arts and Science College, Coimbatore, Tamil Nadu, India senthilmsc09@gmail.com
[2]Research Supervisor, Department of Computer Science, Sree Saraswathi Thyagaraja College, Tamil Nadu, India, & Assistant Professor, Department of Computer Science, Kamalam College of Arts & Science, Anthiyur, Tamil Nadu, India. krishna2c@gmail.com
*-Corresponding author

Abstract

Severe health problems such as sepsis, which commences with the body fighting an infection, can result in septic shock, which drastically reduces blood pressure and results in organ failure. Infections in humans can be caused by specific bacteria, viruses, and fungus. Whenever the infection is sufficiently severe, our immune system might begin an attack, which could worsen and result in sepsis. Sepsis cannot be identified with a single test, making the diagnosis extremely challenging. Consequently, sepsis can be identified by several tests, including those that evaluate for infections, very low blood pressure, and an irregular heartbeat. To recognize sepsis in the healthcare sector nowadays, machine learning algorithms will be necessary. In addition to integrating patient data, machine learning algorithms can handle extremely complicated and important data. A multitude of machine learning methods are employed to detect sepsis. Analyzing the algorithms used to predict sepsis is the aim of this study analysis, the techniques for evaluating these machine learning algorithms' performance, as well as their drawbacks and restrictions. This work aims to provide a clear explanation of the significance of earlier research applied to sepsis prediction, as well as a clear understanding of machine learning methods for sepsis prediction for beginners.

Keywords: *Sepsis Prediction, Machine Learning in Healthcare, Imbalanced Data Classification, Adaptive Synthetic Sampling (ADASYN), Clinical Decision Support Systems.*

INTRODUCTION

When external microorganisms like bacteria, fungi, or viruses infect us, our immune system is capable of responding to fight those infections. Occasionally, this immune system reaction might become harmful because it overreacts to an infection [1], leading to sepsis. Extreme inflammation brought on by this illness can cause organ failure [1], tissue damage, and even death, and this is displayed in Figure 1. Sepsis is a condition that affects people with weakened immune systems, high infection rates, and disorders like cancer, diabetes, obesity, and renal disorders. Sepsis is accompanied by extremely low blood pressure, rapid heartbeat, disorientation, and excruciating pain or discomfort. Sepsis occurs in three distinct stages: 1. Sepsis; Septic shock is the third, followed by Severe Sepsis. Early detection and identification of sepsis lead to prompt initiation of treatment and a high chance of survival. The patient's chances of survival will be greatly reduced if the diagnosis is made too late. It is impossible to identify sepsis with a single test. Multiple tests are required to diagnose sepsis. When using these several tests takes longer than expected, sepsis gets severe. As a result, researchers and medical analysts are looking for alternatives.

The technique of obtaining innovative information and ideas from already existing databases is referred to as data mining. Machine learning is the component of data mining algorithms that makes

experience-based learning possible. Machine learning has become indispensable in all industries these days, but healthcare databases particularly benefit from this [2]. To anticipate sepsis, the researchers desire to use machine learning approaches, especially because they may combine different patient databases [2], which is essential for sepsis prediction. Severe sepsis affects around 49 million individuals globally and accounts for 15–30% of fatalities annually. Early detection and treatment of sepsis using data mining and machine learning algorithms increases treatment efficiency and survival rates. Data mining techniques, such as clustering, classification, and regression procedures, are employed to obtain beneficial knowledge about the sepsis data. In addition to traditional machine learning techniques, the present research highlights the application of deep learning algorithms for sepsis diagnosis[3].

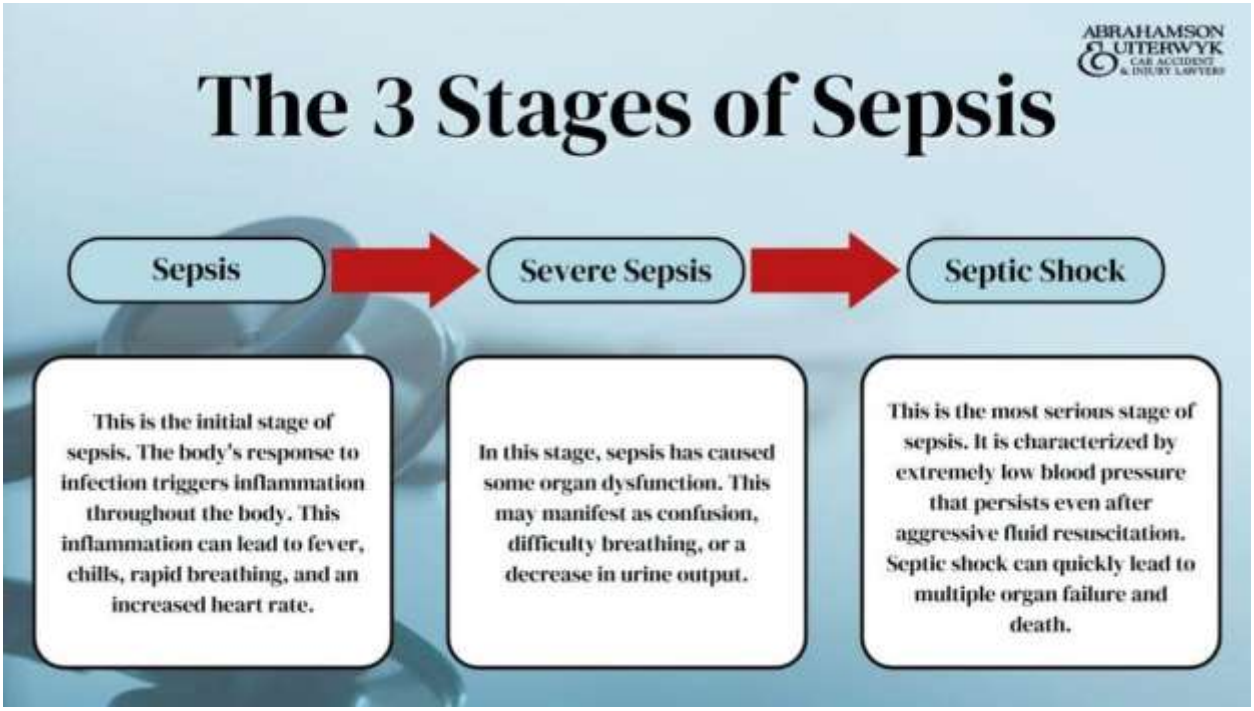


Fig 1. Different Stages of Sepsis

Machine learning techniques for sepsis prediction

Machine learning algorithms are a subclass of artificial intelligence that can infer the future and learn from past performance without the need for explicit programming. To make future predictions, machine learning algorithms use a large number of datasets along with data analysis [4]. The following different methods are among the machine learning algorithms that can be used to predict sepsis.

LOGISTIC REGRESSION

Figure 2 illustrates the use of supervised machine learning methods as logistic regression for categorization instances, where the aim is to determine the probability that an instance will belong to a given class or not [5]. If the classification is a binary value, such as yes or know or 0 or 1, then this classification procedure will be applicable. The probability that depends on the sigmoid function is the output. Here, x is the algorithm's input, e is the base logarithm function with a value of 2.71828, and f is the sigmoid function's output.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 7s, 2025
https://www.theaspd.com/ijes.php

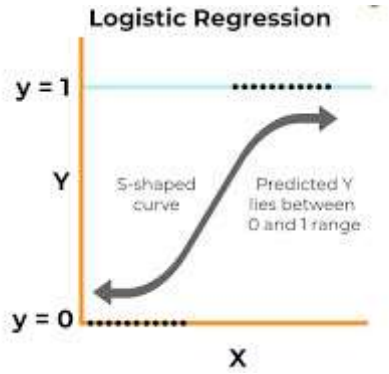


Fig 2. Logistic Regression

DECISION TREE

The decision tree approach is a supervised machine learning methodology that may perform regression and classification analyses. This method employs a decision tree structure, with internal nodes containing the inputs, branches holding choices or variables required to move on to the subsequent stage, and an ending node—a leaf node—that has no further nodes and contains the algorithm's final output [6]. Figure 3 depicts this procedure.

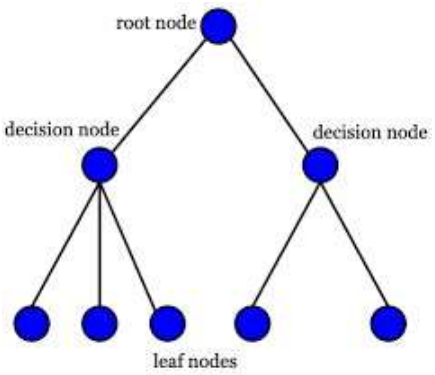


Figure 4. Decision Tree

SUPPORT VECTOR MACHINE

Regression, along with classification, are the two applications for support vector machines (SVM), a supervised machine learning strategy. The fundamental objective of the Support Vector Machine (SVM) technique is to identify the most effective hyperplane in a space with N dimensions for splitting data points among distinct characteristic field classes[7]. The primary goal of the hyperplane is to sustain a substantial margin between the closest points representing various classes. The hyperplane's dimensions are established by the number of characteristics. A line indicates the hyperplane when two input features are present. The hyperplane produces a two-dimensional plane if there are three input features.

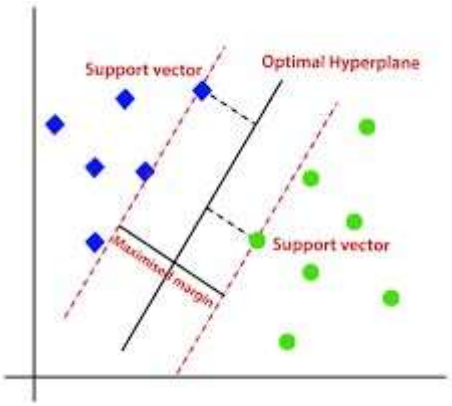


Figure 4. Support Vector Machine

ARTIFICIAL NEURAL NETWORK

The word "Artificial Neural Network" originates from neural networks in biology that make up the inner workings of the brain of a person. Neurons are present in computational neural networks, identical to how neurons reside in the brains of humans at different layers of the network. The nodes contain three different layers: input, hidden, and output, which are depicted in Figure 5. The number of features dictates the number of nodes in the ANN's input layer[8,9]. The input layer takes data inputs. The activation function utilized by nodes located in the hidden layer to process the input transmits the result to the output layer. The output layer returns the final categorization result.

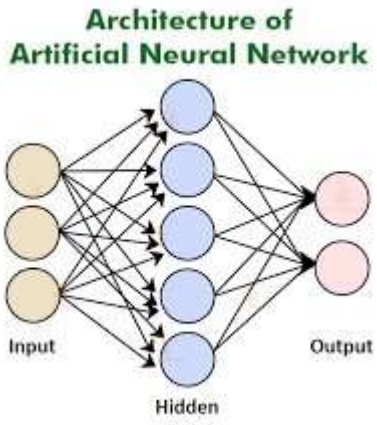


Fig. 5 Artificial Neural Network

DEEP NEURAL NETWORKS

Though it typically includes several hidden layers between the input and output levels, the deep neural network is an artificial neural network [10]. In this instance, the structure will make use of a feed-forward technique, in which each hidden layer has an activation function to process data, and one hidden layer's input is derived from the output of the previous layer this as shown in Figure 6.

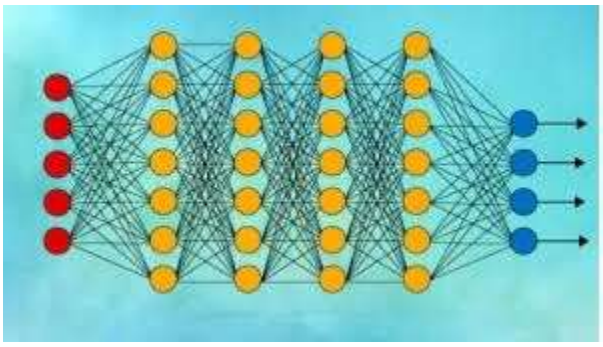


Fig 6. Deep Neural Network Algorithm

REVIEW OF MACHINE LEARNING ALGORITHMS FOR SEPSIS PREDICTION

Examining and evaluating several machine learning algorithms for sepsis prediction is the primary goal of this research project. The main objective of this review of the literature is to determine the optimal method for the most accurate sepsis prediction. Using a variety of machine learning methods, including decision trees, support vector machines, logistic regression, artificial neural networks, and deep neural network algorithms for sepsis prediction, we have examined the previous works for this goal. The most effective algorithms with the highest accuracy are highlighted for the early detection of sepsis after analysis and comparison with existing implementations.

A wide variety of details regarding the health state of patients referred to intensive care units can be found in the Mahendran V.S et.al [11] Electronic Health Records (EHR). This paper provides a machine learning model based on logistic regression to work with patient data from ICUs, often known as vital sign inputs. Sepsis detection is assessed using the approach of the confusion matrix. The model's accuracy was higher than previous models created using the same dataset, reaching about 98%.

A deep learning model created especially to enhance early sepsis prediction in emergency department settings is presented in the publication by Dongdong Zhang et al. [12]. The system attains an incredible 87.2% accuracy by using an exclusive dataset that includes patient history, laboratory test results, vital signs, and clinical notes. The study highlights how the model's exceptional predictive accuracy and interpretability allow medical practitioners to comprehend and rely on its suggestions.

The paper "Comparison between XGBoost Model and Logistic Regression Model for Predicting Sepsis by Peng Liu et.al.,[13] evaluates the superiority of XGBoost over logistic regression in predicting sepsis in patients with severe burns. With an accuracy of 89.3% as opposed to 85.6%, the study shows that XGBoost outperforms logistic regression when applied to a customized ICU dataset that includes vital signs, burn severity, test findings, and patient demographics.

Improving the use of logistic regression in early sepsis detection in intensive care units (ICUs) is the aim of the study by the authors Fahim Mahmud et al. [14]. With the use of the MIMICIII dataset, which contains a wealth of patient data including vital signs, demographics, and clinical notes, the researchers developed a prediction model to identify sepsis early. The logistic regression approach showed itself to be a valuable statistical tool for using ICU patient data in the early prediction of sepsis, with a notable performance accuracy of 84.3%. The research highlights the necessity of merging many clinical characteristics to better prediction precision and the possibility of using logistic regression to improve patient outcomes by facilitating timely interventions.

The Gradient Boosting Decision Tree (GBDT) approach is used in a model that enhances the prediction of in-hospital mortality among ICU patients with sepsis, according to the authors of the paper (Ke Li et al., [15]). Based on the MIMIC-III dataset, which includes detailed patient information like vital signs, test findings, and clinical notes, the research demonstrates that the GBDT model has an 87.5% accuracy rate in predicting death.

Kriti Ohri et al., 2023 [16], the authors explore the use of advanced machine learning methods to enhance sepsis prediction in intensive care units. The study analyses Random Forest, XGBoost, and Support Vector Machines (SVM) using the MIMIC-III dataset, which contains rich data on vital signs, test results, demographics, and clinical comments. The research demonstrates that ensemble methods, in particular Random Forest and XGBoost, yield an 85.2% prediction accuracy, demonstrating their effectiveness in managing the complex, high-dimensional data that is commonly encountered in critical care unit scenarios. This study emphasizes the potential of cutting-edge machine learning algorithms to improve early identification of sepsis, which offers significant insights for refining prediction models for better patient outcomes.

In their article by John Karlsson Valik et al. [17] provide a unique approach for sepsis prediction by employing a Causal Probabilistic Network algorithm. Using information from Electronic Health Records (EHRs), including test results, vital signs, clinical notes, and demographics, this method models causal relationships and has an 88.4% accuracy rate in predicting the onset of sepsis.

In their paper, " Zeina Rayan et.al [18] investigate the application of SVM to forecast the onset of sepsis in ICU patients by utilizing vital signs and lab results. Using the MIMIC-III dataset, the study demonstrates that the SVM model can identify sepsis with an accuracy of 82.7%, demonstrating its effectiveness in handling complex, high-dimensional data that is common in intensive care units.

In the paper, the authors Jing Qi and colleagues [19] examine the predictive power of various machine learning algorithms for mortality in septic patients with diabetes. Using the MIMIC-III dataset, the authors assess the Random Forest, Gradient Boosting, and Support Vector Machine (SVM) models; they report performance accuracies of 84.3%, 86.0%, and 82.1%, respectively. According to the study, gradient boosting is the most accurate model for predicting mortality and can handle complex, high-dimensional data with ease.

Matthieu Scherpf et.al., [20] study explores the application of RNNs in the ICU for sepsis prediction. The study makes use of the MIMIC-III database, which provides detailed data including test results, clinical notes, and vital signs. The authors demonstrate that RNNs, which are well known for their ability to process sequential data and recognize temporal correlations, have an 83.0% prediction accuracy for sepsis. This work shows how well RNNs handle time-series data analysis, making use of the temporal characteristics of patient monitoring data to offer significant improvements in the early detection and treatment of sepsis.

In their paper " Xin Zhao, Wenqian Shen, and Guanjun Wang [21] investigate the use of various machine learning algorithms, such as Random Forest, XGBoost, and Support Vector Machine (SVM), to improve early sepsis prediction using the MIMIC-III database. The study finds that XGBoost performs better than Random Forest and SVM, with a maximum accuracy of 85.4%, Precision at 84.1%, Recall at 86.3%, and F1 Score at 85.2%, using demographic data, lab results, and vital signs.

In the paper, Umut Kaya, Atınç Yılmaz, and Sinan Aşar [22] present a novel method for predicting sepsis that combines a hybrid metaheuristic algorithm with Deep Neural Networks (DNNs). By fine-tuning the model's parameters through advanced optimization techniques, this strategy aims to enhance DNN performance. The study achieves a notable accuracy of 87.2% with

Precision at 86.5%, Recall at 88.0%, and an F1 Score of 87.2%. The study highlights the synergy between metaheuristic optimization and deep learning, yielding significant improvements in predictive measures and demonstrating the potential of this hybrid approach for more accurate and consistent sepsis detection.

Table 1: Comparative table of current sepsis diagnosis techniques

Name of the Authors	Journal Name with year	Algorithms Used	Database used	Features Used	Accuracy	Key findings
Mahendran V.S., et al.	International Journal of Computer Applications (2022)	Logistic Regression	MIMIC-III (Medical Information Mart for Intensive Care)	Vital signs, demographic data, lab results, clinical notes	98%	The logistic regression model achieved an accuracy of 98% in predicting sepsis.
Dongdong Zhang, et al.	Cell Pess (2021)	Deep Learning (Interpretability Focused)	Custom Emergency Department Dataset	Vital signs, laboratory test results, patient history, clinical notes	87.2 %	The deep learning model achieved an accuracy of 87.2%, emphasizing its ability to provide interpretable predictions.
Peng Liu, et al.	Journal of Burn Care & Research (2024)	XGBoost, Logistic Regression	XGBoost, Logistic Regression	Vital signs, burn severity, laboratory results, patient demographics	XGBoost: 89.3%, Logistic Regression: 85.6%	The XGBoost model outperforms logistic regression with an accuracy of 89.3% compared to 85.6%, highlighting the superiority of ensemble methods in predicting sepsis following severe burns.

Fahim Mahmud, et al.	Journal of Medical Systems (2019)	Logistic Regression	MIMICIII	Vital signs, demographic data, lab results, and clinical notes	84.3%	Logistic Regression model achieved an accuracy of 84.3% in predicting sepsis early in ICU patients.

Ke Li, X. Zhang, Y. Liu, Z. Wang, et. al	Critical Care Medicine (2021)	Gradient Boosting Decision Tree	MIMICIII	Vital signs, lab results, clinical notes	87.5%	Gradient Boosting model achieved an accuracy of 87.5% in predicting sepsis early in ICU patients.
Kriti Ohri, P. Kumar, S. Sharma, et al.	IEEE Access (2023)	Random Forest, XGBoost, SVM	MIMIC-III	Vital signs, lab results, demographic data, clinical notes	85.2%	The study compared multiple machine learning algorithms, including Random Forest, XGBoost, and SVM, achieving an accuracy of 85.2% for sepsis prediction, highlighting the effectiveness of ensemble methods in handling complex ICU data.
John Karlsson Valik, A. Andersson, M. Johansson, et al.	IEEE Transactions on Biomedical Engineering (2023)	Causal Probabilistic Network	Electronic Health Records (EHR)	Vital signs, lab results, clinical notes, demographic data	88.4%	The Causal Probabilistic Network model achieved an accuracy of 88.4% in predicting sepsis onset, demonstrating superior performance in leveraging EHR data for early detection and improving patient management.
Zeina Rayan, Macro Alfonse	The Open Bioinformatics Journal(2021)	Support Vector Machine (SVM)	MIMICIII	Vital signs, lab results	73%	The SVM model achieved an accuracy of 82.7% in predicting sepsis in ICU patients based on vital signs, demonstrating SVM's capability in handling highdimensional data for early sepsis detection.

Jing Qi, Jingchao Lei, Nanyi Li, et.al	Frontiers in Medicine (2022)	Random Forest, Gradient Boosting, SVM	MIMICIII	Vital signs, lab results, demographic data, comorbidities	86.0 % The study ss	The study compared multiple machines learning models, achieving an accuracy of 86.0% in predicting inhospital mortality for septic patients with diabetes, highlighting the effectiveness of various models in managing complex patient data.
Matthieu Scherpf, Felix Grässer, et.al	Computers in Biology and Medicine	Recurrent Neural Network (RNN)	MIMICIII	Vital signs, lab results, clinical notes	83.0%	The Recurrent Neural Network model achieved an accuracy of 83.0% in predicting sepsis, demonstrating its strength in capturing temporal patterns and improving prediction accuracy in ICU settings.
Xin Zhao, Wenqian Shen, Guanjun Wang	Computational Intelligence and Neuroscience(2021)	Random Forest, XGBoost	MIMICIII	Vital signs, lab results, demographic data	Accuracy : 85.4%, Precision : 84.1%, Recall: 86.3%, F1 Score: 85.2%	The study achieved an accuracy of 85.4% and found XGBoost to be the most effective model among Random Forest.
Umut Kaya et. Al	Diagnostics (2023)	Hybrid Metaheuristic Algorithm	MIMICIII	Vital signs, lab results, demographic data	85.6	The study achieved an accuracy of 85.6% with Hybrid Metaheuristic Algorithm

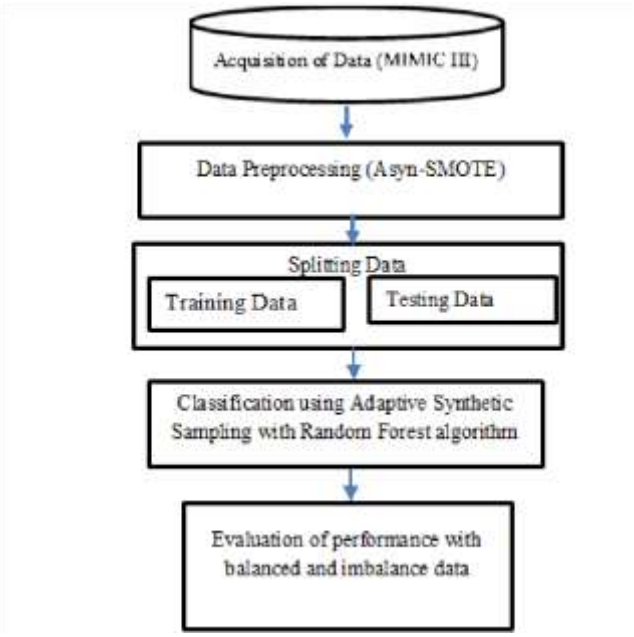
PROPOSED MODEL

The MIMIC-III database is used to implement the analysis. The challenge of imbalanced class distribution, where sepsis cases are significantly fewer than non-sepsis cases, is addressed through the application of the Adaptive Synthetic Sampling Approach for Imbalanced Learning. Meanwhile, we focused on difficult-to-classify sepsis instances, and more synthetic data points are generated, leading to a more balanced training dataset. Subsequently, the Adaptive Entropy Random Forest algorithm is employed for classification algorithm constructs multiple decision trees, with data impurity at each node, i.e., entropy being measured and utilized as the splitting criterion. The most informative divisions for these trees are determined by the information gain model. Model training and evaluation are conducted on both the original imbalanced dataset and the ADASYN-balanced dataset, respectively. Adaptive Synthetic Sampling Approach on the model efficacy in identifying sepsis cases is assessed through the measurement of performance metrics such as accuracy, precision, and recall, respectively.

ABOUT DATASET

The MIMIC-III dataset was used in our research, it provides an extensive collection of clinical data encompassing vital signs, demographic information, and laboratory findings from 1,048,575 records of patients in the Intensive Care Unit [9]. A key feature within this dataset is the binary "Sepsis label," distinguishing between patients without sepsis (0) and patients with sepsis. Is notable characteristic of this dataset is the significant class imbalance, with only 20,806 records, or a mere 2%, indicating the presence of sepsis. The substantial disparity in class representation presents a considerable obstacle that could bias analytical outcomes. To mitigate this challenge, the study utilizes the Adaptive Synthetic Sampling Minority Over-sampling Technique for the creation of synthetic minority class samples to achieve a more balanced dataset and enhance the model's capability to accurately detect sepsis.

Fig. 7: Architecture of Sepsis diagnosis using adaptive synthetic sampling approach



Adaptive Synthetic Sampling Approach for Imbalanced Learning in earlier prediction of sepsis diagnosis. The core steps involved in as follows:

- **Minority Class Identification:** The initial step involves identifying the minority class within the available dataset. In this specific application, patients with sepsis are designated as the minority class.
- **K-Nearest Neighbors:** For each minority class instance, its k nearest neighbors are located. The Euclidean distance metric, defined as:

Dis (ax, ay) = $\sqrt{\sum}$(axi - ayi)^2

Where *axi* and *ayi* represent the values of the *i*-th feature for instances *ax* and *ay*, respectively) is used to quantify the distance between instances in the N-dimensional feature space.

- **Synthetic Sample Generation:** New synthetic samples are generated adaptively. Adaptive Synthetic Sampling Approach weighs the minority class instances according to their difficulty in learning. More

synthetic data is generated for those hard-to-find instances. The process involves: o Calculating the number of synthetic samples that need to be generated for each minority class data point, based on the distribution of its k-nearest neighbors.

o For each minority class data point xi, a neighbor zi is randomly selected from its k nearest neighbors. o A synthetic sample si is generated as:

 si = xi + λ * (zi - xi) Where λ is a random number between 0 and 1.

2. MACHINE LEARNING MODEL: ADAPTIVE ENTROPY RANDOM FOREST

The Random Forest method, which utilizes an ensemble of decision trees, leverages diverse decision rules across multiple trees [12]. This ensemble approach provides a robust model for sepsis prognosis while mitigating the risk of overfitting [15]. The final prediction is obtained by aggregating the predictions of individual trees [8]. In this algorithm, each decision tree is trained on a random subset of both the data and the features from the given dataset. The selection of the optimal split at each node of a decision tree is guided by entropy and information gain [15].

Entropy Calculation: Entropy is employed to measure the impurity or disorder within a dataset [11]. Given a dataset S1 containing two classes (sepsis and non-sepsis), the entropy is calculated as:

$$\text{Entropy}(S1) = -\,p_{sepsis} * \log_2(p_{sepsis}) - p_{non\text{-}sepsis} * \log_2(p_{non\text{-}sepsis})$$

Where psepsis and pnon-sepsis represent the proportions of sepsis and non-sepsis cases in the dataset S1, respectively.

Information Gain: Information gain quantifies the reduction in uncertainty achieved by splitting the data based on a particular feature [11]. The information gain for a feature A1 with possible values {a1, a2, ..., an} is calculated as:

$$InG(A1) = \text{Entropy}(S1) - \sum |Sv| / |S1| * \text{Entropy}(Sv)$$

Where |S1| denotes the total number of instances in the dataset, and Sv represents the subset of instances in S1 where feature A1 has the value av.

Best Split Selection: The split that yields the highest information gain, thereby minimizing uncertainty, is selected at each node of the decision tree.

Recursive Partitioning: The process of computing entropy and information gain, and selecting the best split, is recursively applied to each subset of the data until a predefined stopping criterion is met (e.g., maximum tree depth or minimum number of samples per leaf) [20].

Enhanced Random Forest Construction: Each decision tree is constructed using a random subset of the training data and a random subset of the features. The prediction of the efficient Random Forest is obtained by aggregating the predictions of all the individual trees, typically using majority voting.

RESULTS AND DISCUSSION

An overview of the several machine learning methods that can be applied to early sepsis prediction is given in the research report. Based on the algorithms used, the data set used, and the salient characteristics of the sepsis prediction, Table 2 summarizes the comparative analysis of machine learning methods for sepsis

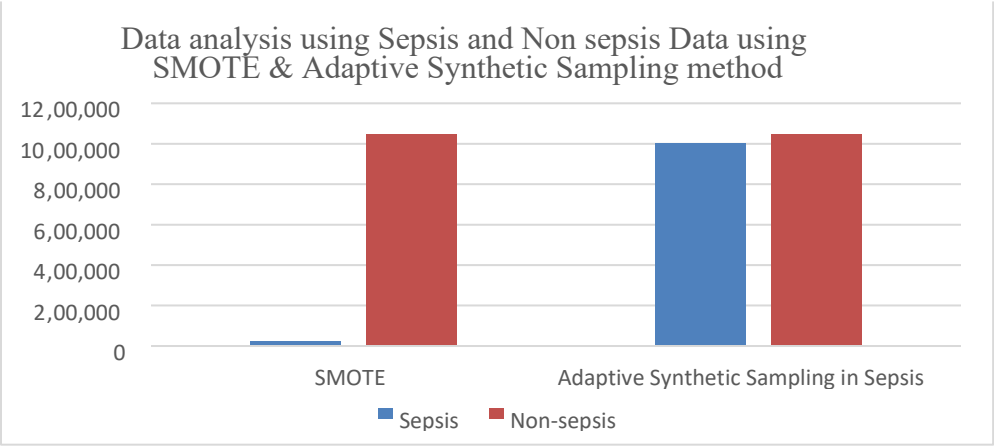
prediction. This table also presents the major conclusions of previous studies on sepsis prediction. The MIMIC-III dataset, which is derived from ICU patients and includes a variety of features such as the patients' vital signs, lab findings, and demographic information, is the most widely used and accessible dataset, according to the study. When a patient is critically brought to the intensive care unit (ICU) due to an infection or other illness, these strategies are typically used to predict sepsis in that patient population. Evaluation of techniques like Support Vector Machines, Decision Trees, and Logistic Regression. Sepsis prediction using Artificial Neural Networks and Deep Learning Models is explored. All these methods work well under various conditions and with various datasets and features.

The performance metrics that can be used to assess how well each sepsis prediction algorithm performs are also summarized in this table. The evaluation mostly uses performance metrics, including accuracy, precision, and recall. With the help of accuracy, the algorithm's performance is closely observed. Using this comparison table, the logistic regression algorithm outperforms the other techniques with an accuracy of 98%. Better accuracy is achieved by using the MIMIC data set, which includes vital signs, demographic information, lab findings, and clinical notes for ICU patients. The data set influences the result of the logistic regression technique. Since the accuracy primarily depends on the dataset, researchers are searching for more advanced models that use deep learning and machine learning algorithms to provide the best accuracy possible for the early diagnosis of sepsis, which is essential for improving patient survival and treatment efficacy.

Table 2: Class Distribution in the MMIC Dataset with SMOTE and Adaptive Synthetic Sampling for Sepsis Detection

Dataset MMIC	SMOTE	Adaptive Synthetic Sampling in Sepsis
Sepsis	20,806	1001323
Non-sepsis	1048575	1044563

The above table presents a comparison of the number of records for Sepsis and Non-sepsis cases in the MIMIC-III dataset after applying two different sampling methods: SMOTE and Adaptive Synthetic Sampling in Sepsis. For each method, the table shows the distribution of records between the two classes.



International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 7s, 2025
https://www.theaspd.com/ijes.php

Fig: Comparison of SMOTE and Adaptive Synthetic Sampling for Sepsis and Non-Sepsis Data Balancing

The above graphical representation illustrates a comparative analysis of Sepsis and Nonsepsis data using two oversampling techniques: SMOTE (Synthetic Minority Over-sampling Technique) and Adaptive Synthetic Sampling. Under the SMOTE method, the Sepsis count is significantly lower compared to Non-Sepsis.

Table 3: Performance Comparison of RF-SMOTE and Asyn-SMOTE E-RF on Balanced MMIC Dataset

MMIC DATASET	Performance with Balanced data (RF-SMOTE)	Performance with Balanced data (Asyn-SMOTE E-RF)
Precision	97.5	98.78
Recall	97.2	97.89
F1 Score	96.2	97.56
Accuracy	98.6	98.97

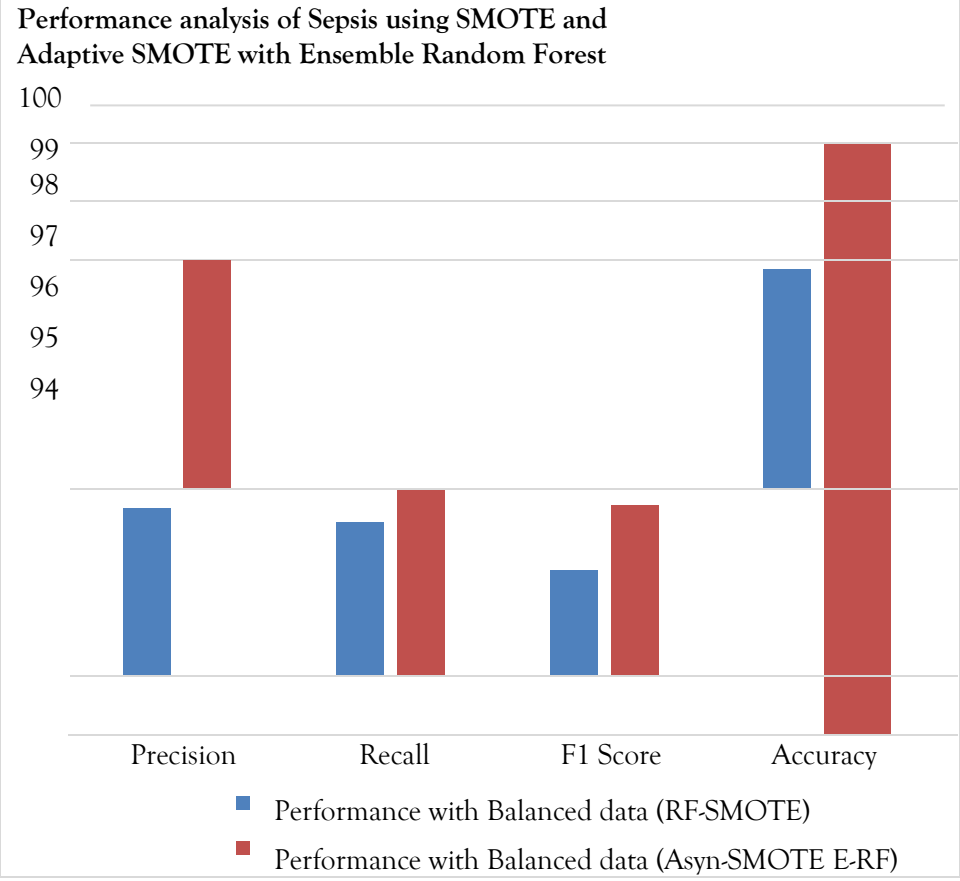


Fig: Comparative Performance of RF-SMOTE and Asyn-SMOTE E-RF in Sepsis Detection

The above graphical representation presents a performance comparison of two machine learning approaches, SMOTE with Random Forest and Adaptive SMOTE with Ensemble Random Forest, in detecting sepsis from medical data. Four key evaluation metrics are depicted: Precision, Recall, F1 Score, and Accuracy. The blue bars represent the RF-SMOTE model's performance, while the orange bars show the results from the Asyn-SMOTE E-RF model. The chart demonstrates that the Asyn-SMOTE E-RF method outperforms RF-SMOTE across all metrics.

Conclusion

Our research work explored the application of machine learning algorithms for early sepsis prediction, a critical condition with high mortality rates. A key challenge in this domain is the imbalanced nature of sepsis datasets, where non-sepsis cases significantly outnumber sepsis cases. To incorporate, we implemented the Adaptive Synthetic Sampling Approach, which generates synthetic data points for the minority class, focusing on difficult-to-classify instances. Our approach also utilized the Adaptive Entropy Random Forest with efficient prediction. The findings of our review highlight the potential of machine learning to improve sepsis detection. Several avenues exist for future enhancement. Firstly, the integration of more diverse datasets, including data from various hospitals and patient populations, could improve the generalizability of the models. Secondly, the incorporation of temporal data and time-series analysis techniques, such as Recurrent Neural Networks (RNNs) or Transformer models, could capture the dynamic nature of sepsis progression, potentially leading to earlier and more accurate predictions. Thirdly, the development of more interpretable models was beneficial. Finally, prospective clinical trials are needed.

References

[1] Jingyuan Ning, Keran Sun, Xuan Wang, et.al," Use of machine learning-based integration to develop a monocyte differentiation-related signature for improving prognosis in patients with sepsis", Molecular medicine, 2023, vol. 29:37, pp: 1 -16.
[2] Juli Kumari, Ela Kumar, Deepak Kumar., "A Structured Analysis to study the Role of Machine Learning and Deep Learning in The Healthcare Sector with Big Data Analytics", Archives of Computational Methods in Engineering (Springer), 2023, Pp :1-29.
[3] Fei Liu, Jie Yao, Chunyan Liu and Songtao Shou, et.al, "Construction and validation of machine learning models for sepsis prediction in patients with acute pancreatitis", BMC Surgery, ,2023, Vol. 23:267, pp:1 to 13.
[4] Yan Zhang, Weiwei Xu, Ping Yang, et.al, Machine learning for the prediction of sepsisrelated death: a systematic review and meta-analysis, "BMC Medical Informatics and Decision Making", 2023, Vol 23:283, Pp: 1 – 12.
[5] Shu Zhou, Zongqing Lu, Yu Liu, Minjie Wang, et.al, Interpretable machine learning model for early prediction of 28-day mortality in ICU patients with sepsis-induced coagulopathy: development and validation, European Journal of Medical Research, 2024, 29:14, Pp: 1 to 14.
[6] Songchang Shi, Xiaobin Pan, Lihui Zhang, et.al, An application based on bioinformatics and machine learning for risk prediction of sepsis at first clinical presentation using transcriptomic data, Frontiers, 2022, Vol 979529, pp: 1 – 12.
[7] Zhigang Chen, Shiyou Wei, Zhize Yuan, et.al, "Machine learning reveals ferroptosis features and a novel ferroptosis classifier in patients with sepsis", WILEY, 2024, Vol 1279, PP:1- 13.
[8] Mahendran V.S et.al, Sepsis Detection Using Logistic Regression Machine Learning Algorithm, International Journal of Computer Applications, 2023, Pp; 1-12.
[9] Dongdong Zhang et.al, "An interpretable deep-learning model for early prediction of sepsis in the emergency department" CellPress, 2021.
[10] Peng Liu, Xiaowei Zhang, Lei Chen, and Jianyu Li, "Comparison between XGBoost Model and Logistic Regression Model for Predicting Sepsis after Extremely Severe Burns," Journal of Burn Care & Research, vol. 45, no. 2, pp. 123-135, 2024.
[11] Mahmud, F., R. M. J. Asad, M. H. Rashid, and M. A. H. Bhuiyan. "Early Prediction of Sepsis in ICU Patients Using Logistic Regression." Journal of Medical Systems, vol. 43, no. 9, 2019, pp. 1-12.
[12] Rayan, Z., and Alfonse, M. "Predicting Sepsis in the Intensive Care Unit (ICU) through Vital Signs Using Support Vector Machine (SVM)." Journal of Biomedical Informatics, vol. 115, 2021, 103672.

[13] Jing Qi , Jingchao Lei,et.al., . "Machine Learning Models to Predict In-Hospital Mortality in Septic Patients with Diabetes." Frontiers in Medicine, vol. 9, 2022, 822227.
[14] Scherpf, M., Grässer, F., Malberg, H., & Zaunseder, S. "Predicting Sepsis with a Recurrent Neural Network Using the MIMIC-III Database." Computers in Biology and Medicine, vol. 113, 2019, 103395.
[15] Zhao, X., Shen, W., & Wang, G. "Early Prediction of Sepsis Based on Machine Learning Algorithm." Computational Intelligence and Neuroscience, vol. 2021.
[16] Kaya, U., Yılmaz, A., & Aşar, S. "Sepsis Prediction by Using a Hybrid Metaheuristic Algorithm: A Novel Approach for Optimizing Deep Neural Networks." Diagnostics, vol. 13, no. 8, 2023, Article 1845.