IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 September 2022, accepted 24 September 2022, date of publication 28 September 2022,
date of current version 10 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3210575

RESEARCH ARTICLE

A Comprehensive Machine Learning Based Pipeline for an Accurate Early Prediction of Sepsis in ICU

B. C. SRIMEDHA, RASHMI NAVEEN RAJ, AND VEENA MAYYA, (Senior Member, IEEE)
Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka 576104, India

Corresponding authors: Rashmi Naveen Raj (rashmi.naveen@manipal.edu) and Veena Mayya (veena.mayya@manipal.edu)

ABSTRACT Sepsis is a lethal infection-related illness that has an extremely high fatality rate, especially among intensive care unit patients. Early and precise recognition of sepsis is critical as delayed treatment increases the mortality rate dramatically. System inflammatory response syndrome, quick sequential organ failure assessment, and modified early warning score are the traditional clinical score systems in practice to detect sepsis. But the scoring systems fail in the early prediction of sepsis, a stage in which if a patient is treated immediately, the mortality rate will reduce significantly. The proposed classifier can accurately predict sepsis up to six hours before the disease is clinically diagnosed. The patient's electronic medical records, demographics, and vital signs are used to achieve this. The study uses data set adaptive data preprocessing strategies. The proposed method adds value to existing literature by introducing a novel outlier-based mean-median data imputation technique that enhances the prediction's overall accuracy. The primary factors that influence the classifier's predictions have been outlined, making the model easier to understand for medical professionals. For the classification of patients as sepsis positive or negative, four algorithms were investigated: Random Forest, Logistic Regression, Gradient Boosting, and Decision Tree. Of all the prediction algorithms, Random Forest gives the best results with an accuracy of 99.01%, F1-score of 99%, and an area under the receiver operator characteristic curve of 99.99%. Even for a 24-hour early prediction of sepsis, the random forest method is proven to provide greater prediction accuracy while logistic regression provides the least prediction accuracy. We attribute this to the fact that, unlike regression models, random forests do not require that the model have a linear relationship between the dependent and independent variables. The evaluation measures produced are useful and can be tremendously valuable in predicting sepsis in a timely and accurate manner.

INDEX TERMS Sepsis, prediction, machine learning, mortality, imputation, classifier, early diagnosis, artificial intelligence.

I. INTRODUCTION

Sepsis is certainly a life-threatening medical illness that occurs when the body's reaction to infection harms the body's tissues. It is a bloodstream infection that causes a slew of symptoms, including low blood pressure, increased heart rate, temperature, confusion or disorientation, acute pain or discomfort, as well as shortness of breath, etc. Reference [1] Sepsis can progress to septic shock if not handled properly. Septic shock is caused by a sudden drop in blood pressure, and it can induce organ malfunction and even organ failure [2]. Patients with cancer or a weak immune system are more likely to develop septic shock. In the adult Intensive Care Unit (ICU), sepsis and septic shock with consequent multi-organ failure, are currently the leading causes of death [1]. According to the survey, one in five deaths [3], half of all hospital deaths [4], and 19.7% of the whole world's deaths were reported due to sepsis in [5]. Even though the

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan.

surgical and pharmacological approaches to sepsis treatment are improving all the time, epidemiological data reveals that the frequency of sepsis has increased over the previous 20 years.

Most of the symptoms caused by sepsis can also be caused by a variety of other medical diseases, making early detection challenging. Some of the symptoms are caused by general weakness and are thus overlooked until they worsen. The morbidity rate significantly increases with a delay in every minute of the treatment. To avoid severe consequences, there is a critical need to develop an early prediction algorithm as a decision support system that will surely alarm the clinicians if a patient is predicted to have sepsis so that the physicians can closely monitor the patient and treatment procedures can be initiated beforehand. Timely initiation of sepsis management through fluid resuscitation or intravenous microbial results in a better quality of life for the sepsis survivors and mitigates hospital costs [6].

Common reasons for the delay in starting the sepsis management are: operational formalities at the hospital like the opinion of the physicians from multiple departments based on patient history to decide the treatment strategies, admission or registration procedures, the financial status of the patient, timely availability of the infrastructure and medicines, etc. Researchers in the healthcare service management system can propose a personalized scheduling system that can reduce the waiting time of the patient and thereby reduce the harmful delays in getting treatment.

The clinical decision is still based on the Systemic Inflammatory Response Syndrome (SIRS) score, i.e., the presence of two or more SIRS criteria paired with a suspicion of infection. The four SIRS criteria are summarized in Table 1. From the Sepsis-3 definition [2], the Sequential Organ Failure Assessment (SOFA) score is used for assessing organ dysfunction. The score is made up of six factors, one for each of the respiratory, neurological, cardiovascular, kidney, coagulation, and liver systems. Each parameter is given a number between 0 and 4, with zero signifying the lowest risk of organ failure as shown in Table 2. A SOFA score greater than 2 is considered to be associated with a 10% increase in in-hospital mortality. Quick SOFA (qSOFA) and Modified Early Warning Score (MEWS) in which the scores are manually generated based on lab-test results and vital signs are still the existing methods used to decide the pre-hospitalization or ICU admission [7].

TABLE 1. SIRS parameters [8].

Sl.No.	SIRS parameter	Range
Tachycardia	Heart rate	> 90 beats/min
Hypothermia	Body temperature	$> 38^{O}$C or $< 36^{O}$C
Tachypnea	Respiratory rate	$>$20 breaths/min
Leukopenia	White cell count	$> 12 \times 10^9$ cells/L or $< 4 \times 10^9$ cells/L

Even though there have been many attempts to use machine learning (ML) to find sepsis early, it remains a significant concern for healthcare stakeholders worldwide. ML accelerates data processing and analysis, which can greatly aid in early prediction. With minor changes in deployment, predictive analytic approaches that use machine learning could be trained on increasingly larger data sets and provide deeper analysis on a variety of aspects. Early prediction aims at identifying the onset of sepsis well before a physician can do it. An accurate early predictive model would help the physicians to have a closer look at the patients well before the onset of sepsis and could reduce the morbidity rate as well as the financial cost of treating the patient. The four major categories of ML algorithms are supervised, unsupervised, semi-supervised, and reinforcement learning, as shown in Figure 1.

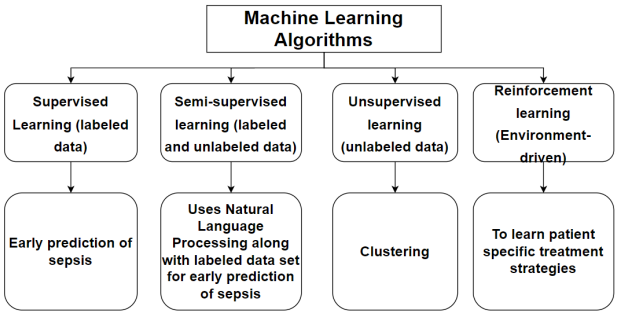

FIGURE 1. Different machine learning algorithms.

- Supervised algorithms use labeled data sets to train the predictive or classification model and then the trained model is tested on unlabeled data sets to measure the effectiveness of the model. Supervised learning is used for sepsis prediction
- An unsupervised algorithm uses unlabeled data set and is especially used for clustering applications.
- Semi-supervised learning is the hybrid of supervised and unsupervised learning algorithms that use both labeled and unlabeled data sets. Using additional information from unlabeled data will surely improve the prediction outcome and is attempted by a few authors in early sepsis prediction [4].
- Reinforcement Learning (RL) does not need any data set and the learning agent continuously interacts with the environment to derive an optimum strategy for sequential-decision problems. RL algorithms are used by researchers for finding patient-specific sepsis treatment strategies [9].

The performance of the ML-based prediction algorithm is highly dependent on the data set. The sampling frequency of the data depends on the source of data, i.e., the pathological/biomedical/physiological data captured through various laboratory tests or sensors in different departments. Hence, there is a high chance of variation in the time at which data recording is started and also the periodicity of the data. This indirectly accounts for one of the reasons for missing data in the data set. The proposed work uses unique data

TABLE 2. SOFA score calculation from parameters of six different organs.

Parameter	0	1	2	3	4
Respiratory, $PO_2/FiO_2, mmHg(kPa)$	$\geq$ 400 (53.3)	$<$ 400 (53.3)	$<$ 300 (40)	$<$ 200 (26.7) with respiratory support	$<$ 100 (13.3) with respiratory
Coagulation, Platelets, $\times 10^3/mm^3$	$\geq$ 150	$<$ 150	$<$ 100	$<$ 50	$<$ 20
Liver, Bilirubin, mg/dL	$<$ 1.2	1.2-1.9	2.0 - 5.9	6.0 - 11.9	$>$ 12
Cardiovascular	MAP $\geq 70mmHg$	MAP $<$ 70 mmHg	Dopamine $<$ 5 or dobutamine (any dose)	Dopamine 5.1 - 15 or epinephrine $\leq$ 0.1 or norepinephrine $\leq$ 0.1	Dopamine $>$ 15 or epinephrine $>$ 0.1 or norepinephrine $>$ 0.1
The central nervous system, Glasgow Coma Scale	15	13-14	10-12	6-9	$<$ 6
Renal, Creatinine, mg/dL. Urine output, mL/d	$<$ 1.2	1.2-1.9	2.0-3.4	3.5-4.9, $<$ 500	$>$ 5.0, $<$ 200

processing techniques with supervised learning for the early prediction of sepsis using the Challenge 2019 data set, available here [10]. The main contributions of the paper are:

- Development of an outlier-based adaptive data preprocessing technique to improve the prediction accuracy.
- Development of an improved and more accurate machine learning-based early prediction model.
- Extensive benchmarking results demonstrate the superior performance of the proposed model compared to the prior studies in terms of various performance metrics.

Section II presents a comprehensive literature survey of early detection of sepsis including all the research papers up to April 2022. Section III discusses the methodology used and Section IV presents the analysis of the various prediction models used in Section III. The proposed work is concluded in Section V with a scope for future work.

II. LITERATURE SURVEY

Researchers from the medical field are focusing on the complications, management, and treatment strategies of sepsis [11], [12]. At the same time, the researchers, academicians, and scientists from the technical side are attempting to develop a model for the early detection of sepsis. This section discusses the state-of-the-art research in applying ML models for sepsis prediction.

Few authors [13], [14], [15] utilised a time frame onset duration of 12 hours, and few authors employed a duration of 6 hours [16], [17], [18], [19]. Early detection of sepsis facilitates disease prevention and treatment planning. Liu *et al.* [16] have used clinical data to suggest an attention-based sequential representation algorithm for the early prediction of sepsis. The two main elements of the proposed model are clinical event interaction extraction using heterogeneous event aggregation and temporal interaction captured using LSTM. The proposed split storage approach of aggregation representation with distinct heads preserves the temporal interactions of events and can decrease clinical event sequences for improved temporal dependence modeling.

Lauritsen *et al.* [20] have provided a novel deep learning technique for early sepsis detection in a heterogeneous data set outside of intensive care units. The system learns representations of crucial components and relationships from the unprocessed event sequence data without having to resort to a time-consuming feature extraction technique. This study demonstrates that sequential deep learning models are capable of early sepsis detection. The clinical effectiveness of the model is also recommended to be assessed using a novel retrospective evaluation method that takes both blood culture and intravenous antibiotic requisitions into consideration.

Shankar *et al.* [21] have investigated four imputation methodologies for the early identification of sepsis in patients. Each technique was evaluated using six models. It is evident from a comparison of the different imputation strategies that the ''mixed filling'' algorithm that is suggested yields the best results since it chooses and combines elements from the best filling strategy. Out of all the models this study looked at, the LGBM classifier generates the best metrics.

Authors in [22] combined a Sepsis risk assessment using dynamic data and a downstream deep learning model with probabilistic continuous function estimation. The MGP-RNN has a C-statistic of 0.88 for diagnosing sepsis within 4 hours of onset in comparison to the C-statistics of RF, CR, PLR, SIRS, NEWS, and qSOFA.

Nakhashi *et al.* [13] have used two training sets from two different hospitals and have carried out a 10-fold cross-validation to see if the random forest is learning. On training set A, this model had a ''test accuracy of 63.20 %'', a ''test AUROC of 0.621'', and a ''test F-score of 0.067''. On the presented challenge 2019 data set [10], the suggested technique has performed decently. Selcuk *et al.* used Box-cox followed by Min-max transformations and have observed an improvement in the efficiency of the ensemble algorithms [23] which were tested on a dataset of 200 patients. Few independent parameters like vitamin D and patient history which is usually available as a clinical notes are also linked with higher possibility of sepsis [24], [25]. Zhang *et al.* have proposed ensemble based machine learning model and have demonstrated the performance using the dataset from multiple hospitals: eICU from US and clinical care database from China. Liu *et al.* [26]

proposed an objective function for the XGBoost framework along with the first-order and second-order gradients of the objective function, which are used to train the sepsis prediction models. The authors found that lower ranking features may have a negligible impact on the outcomes and might therefore be eliminated for feature reduction. Additionally, the authors emphasised that finding a mechanism to impute the missing values could further enhance performance.

Using a variety of techniques, research is now being done on early sepsis predictions. A few of these are artificial intelligence-based sepsis predictions like machine learning and deep learning algorithms. Having a good data set is the greatest obstacle to producing an accurate prediction. The quality of the data might not be particularly good due to defective machinery, mistakes made by humans, etc. As a result, finding reliable data is difficult. The goal of this research is to build effective data preparation techniques along with implementing a machine learning model as a means of resolving this problem. Beyond gathering consistent and error-free data, work can be done in the future on applying reinforcement learning and advanced deep learning [25] to detect and to derive an optimal personalized treatment strategies for sepsis [27]. The deep learning models [28], [29], [30] are computationally intensive. It is challenging to apply deep learning models in a real-time clinical situation since they are difficult to comprehend and take a great deal of processing resources. Consequently, the proposed work focuses on machine learning-based methodologies with novel data processing paradigms that can be easily deployed in a clinical setup.

III. METHODOLOGY

The overall methodology of the proposed work is shown in Figure 2.

A. DATA SOURCE

This investigation makes use of PhysioNet Challenge 2019 data [10] which is one of the most often utilized data sets for the classification of Sepsis patients [10]. Each row in this time series data set represents an hour's worth of patient data. The data set was labeled using the Sepsis-3 criteria, according to [2], with a 36-hour observation period for each patient. Additionally, data up to 6 hours before the onset has been deemed favorable for early disease prediction if a patient is diagnosed with sepsis.

B. DATA SPLITTING

In the public release of the data collection, there is information on 40,336 patients from two hospitals with 41 patient attributes. All the data is stored in the form of PSV files. A PSV file is a pipe-separated value file. It is similar to a CSV file, i.e., a file in which the values are separated by commas, whereas in a PSV file the values are separated by the pipe symbol. This data is split into 3 sets, namely: train, test, and validation data. The data used for training the model is approximately 75 percent of the total data i.e., information of 30,000 patients. The data used for both testing and validation is around 12.5 percent, which sums up to an approximation of 5,000 patient data.

C. DATA PREPROCESSING

The quality of the output in any algorithm is highly dependent on the quality of the input data. One such stage in producing excellent input data from raw data is data preprocessing [32]. Data preprocessing turns the raw data into a format that is both practical and effective. Data preprocessing is generally divided into 3 categories: data cleaning, data transformation, and data reduction. Data cleaning deals with missing and noisy data. Noisy data can be handled by various methods such as binning, regression, clustering, etc. The data transformation step is used to turn the data into a form that is suitable for data mining. Different types of data transformation techniques are normalization, attribute selection, etc. Data reduction is used to avoid difficulties while handling large amounts of data. This can be used to speed up the process and also for efficient storage of data. Some very common data reduction techniques used are dimensionality reduction, aggregation, etc.

D. EXPLORATORY DATA ANALYSIS

After Data Preprocessing, Exploratory Data Analysis (EDA) is performed. EDA is a technique for assessing or understanding data to draw conclusions or identify important features. Graphical analysis and non-graphical analysis are the two subtypes of EDA. EDA is a crucial step as it helps in understanding the problem statement and the various relationships between the data attributes. This study used a variety of graphical techniques to analyze the distribution of the data, the number of null values, and outliers. Figure 3 depicts the percentage of missing values in each feature.

1) DATA BALANCING

The data needs to be balanced before we proceed with the preprocessing techniques. The data of all the patients were concatenated to create a single data set. One substantial data set was produced as a result, and each model's cross-validation was then carried out using this data set. To train and test the models, each row was considered a separate sample. The data is very unbalanced with a huge partiality towards negative samples. As there are insufficient instances of the minority class (positive sample), imbalanced classification has the drawback that a model cannot efficiently learn the decision boundary. To solve this issue, the minority class can be over-sampled or the majority class can be under-sampled. It is in general preferred to over-sample the data instead of under-sampling as the latter can result in the loss of some data. Although keeping in mind the amount of time taken and the computational expense, if the data is too large, under-sampling is preferred. In this study, both these balancing techniques are combined to form a balanced data set. The minority class has 22560 samples whereas the majority class has

TABLE 3. Literature survey overview.

Sl.no	Research work (Year)	Data set used	Methodology used	Improvements made
1	Shankar et al. (2021) [21]	PhysioNet challenge 2019 data set	Four imputation techniques using 6 models: RF, LR, NN, LSTM, LightGBM and XGBoost	A new filling algorithm called Mixed filling.
2	Liu et al. (2019) [16]	PhysioNet Challenge 2019 data set	Traditional LSTM, hierarchical neural networks to model long sequence	Attentional events Aggregation. Aggregates heterogenous clinical events and captures temporal interactions of the aggregated representations with LSTM
3	Lauritsen et al. (2020) [20]	Retrospective data taken from various hospitals(Danish) over 7 years	Combination of convolutional neural network and LSTM network	Richer data set and Deep learning model that can learn key factors from the raw event.
4	Ackerman et al. (2022) [6]	Review of existing work	This scoping review's main objective is to provide a summary of the literature on current approaches to early diagnosis of sepsis using AI.	The findings will provide information on the state of current science and developments for applying AI systems to diagnose sepsis in the initial stages of the health care system.
5	Nakashi et al. (2019) [13]	PhysioNet Challenge 2019 dataset	RF-based ensemble ML technique.	Combined classifier and an early predictor approach. A utility metric score is used to evaluate the early predictor.
6	Bedoya et al. (2020) [22]	Data from an academic quaternary hospital with 43,000 inpatient beds and 1 million aftercare visits annually	Three clinical ratings used to diagnose sepsis are random forest (RF), cox regression (CR), and penalized logistic regression (PLR).	A deep learning model that has been internally constructed and dynamically verified network
7	Selcuk et al. (2022) [23]	Dataset is taken from Acibadem Kadikoy Hospital in Istanbul, Turkey	Eight different ML methods besides a generated ensemble model along with the traditional prognostic scores are employed.	To improve the effectiveness of algorithms, Box-Cox and Min-Max transformations are used repeatedly to data and parameter adjustment.
8	Zhang et al. (2022) [31]	Dataset is taken from multiple hospitals : eICU from US and clinical care database from China	Support Vector Machine, Random Forest, Neural Network, and Extreme Gradient Boost are used as four first-level learners via stacking algorithm.	An ensemble model for early prediction of S-AKI onset was developed and it demonstrated good performance in multicenter external datasets.
9	Jiu et al. (2022) [26]	Utilised MIMIC-III dataset and PhysioNet Challenge 2019 dataset.	The statistical counting method is used to create count-dependent characteristics for the clinical intervention data.	Proposed an objective function for the XGBoost framework along with the first-order and second-order gradients.

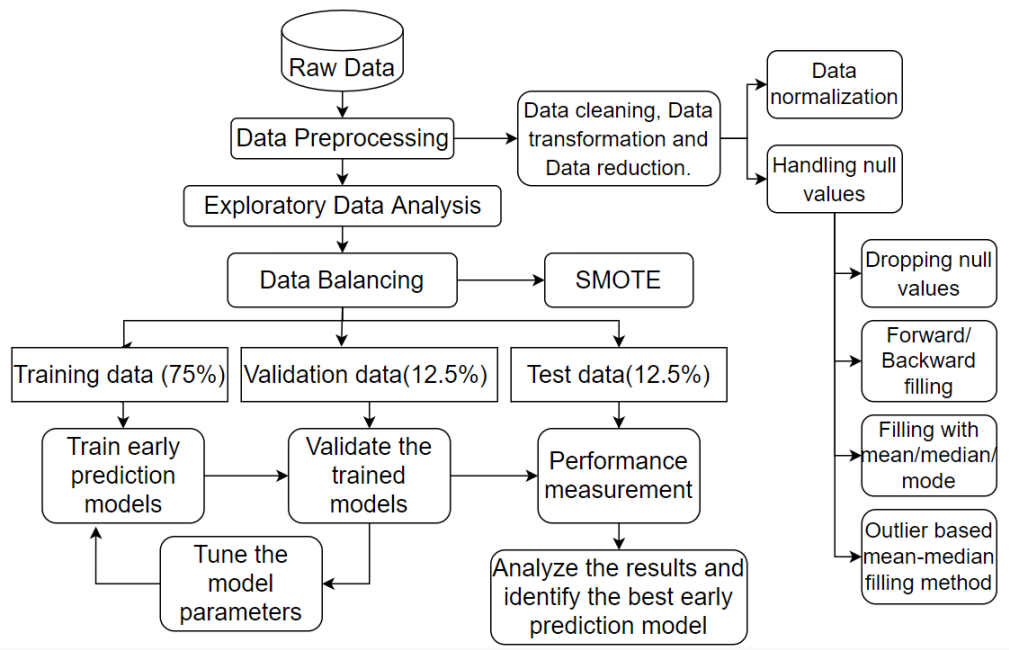

FIGURE 2. Flowchart depicting the methodology of the proposed model.

1148835 samples. Firstly, random under sampling is applied in the ratio of 4:1. This yields us approximately 90000 majority class and 22560 minority class samples. Then Synthetic Minority Over-sampling Technique (SMOTE) technique has been used to oversample the minority class [33]. SMOTE selects examples in the feature space that are close to one another, draws a line between the examples, and then creates a new sample at a location along the line. After the whole process, we have 90000 samples of each minority and majority class. A good improvement in the overall results can be observed due to balancing. These results are presented in section IV.

2) FEATURE IMPORTANCE

Feature importance has been implemented in this study to create a concrete baseline. The term feature importance refers to the methods that assign each input feature in a given model a score, which simply indicates how important that feature

TABLE 4. A Summary of the physionet challenge data set [10].

Category	Parameters
Demographics	HR: Heart Rate (beats per minute), Temp: Temperature (Deg C), O2Sat: Pulse Oximetry (%), MAP - Mean arterial pressure (mm Hg), DBP: Diastolic BP (mm Hg), SBP: Systolic BP (mm Hg), EtCO2: End-tidal carbon dioxide (mm Hg), Resp: Respiration rate (breaths per minute)
Vital signs	BaseExcess: The measure of the amount of excess bicarbonate (mmol/L), FiO2: Fraction of inspired oxygen (%), HCO3: Bicarbonate (mmol/L),PaCO2: CO2's partial pressure from arterial blood (mm Hg), AST: Aspartate Transaminase (In/L), Phosphate: (mg/dL), BUN: Blood Urea Nitrogen (mg/dL), SaO2: Saturation of 02 from arterial blood(%), Calcium: (mg/dL), Chloride: (mmol/L), Creatinine: (mg/dL), Alkalinephos: Alkaline Phosphatase (IU/L), Glucose: Serum glucose (mg/dL), Bilirubin direct: (mg/dL), Lactate: Lactic acid(mg/dL), Magnesium: (mmol/dL), Potassium: (mmol/L), Bilirubin total: Total bilirubin(mg/dL), Troponinl: (ng/mL), Hct: Hematocrit (%), Hgb: Hemoglobin (g/dL), PTT: Partial thromboplastin time (seconds), Fibrinogen: (mg/dL), WBC: Leukocyte count (count"10"3/ML), Platelets: (count*10"3/ML)
Laboratory Values	Age - Years (100 for patients 90 or above) Gender: Female(0) and Male(1) Unit1: Administrative identifier for ICU unit (MICU), Unit2: Administrative identifier for ICU unit (SICU), HospAdm-Time: Number of hours between hospital admit and ICU admit, ICULOS: length-of-stay in ICU (hours since ICU admit)

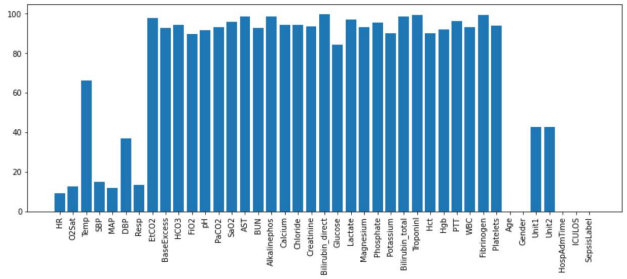

FIGURE 3. Percentage of missing values in each feature.

is. A higher score indicates that the particular characteristic will have more of an impact on the model. This also helps in understanding what features are irrelevant and can be ignored. This will not only help in improving the performance of the model but also helps with the execution time. Since the unnecessary features are excluded, the execution time will reduce by a decent amount. The top 28 features are taken into consideration in this study for the models to operate at their best. This decision is taken after careful monitoring of how many features are considered together to produce the best result. Both accuracy and fit time are used to gauge this performance. ICULOS, the length of stay in the ICU, had the highest score as the most significant feature followed by Unit 1, the administrative identifier for the ICU unit (MICU), and Unit 2, the administrative identifier for the ICU unit (SICU). These top 28 features are ICULOS, Unit 1, Unit 2, BUN, HR, Resp, Gender, Bilirubin_direct, Bilirubin_total, Creatinine, EtCO2, Hct, Hgb, WBC, Fibrinogen, Temp, MAP, DBP, Calcium, PTT, Glucose, SBP, BaseExcess, Platelets, Magnesium, HCO3, AST, HospAdmTime. Each of these features is briefly defined in Table 4. The top 15 features plotted against their feature importance scores are shown in Figure 4.

The features after ICULOS in Figure 4 are not visible due to the high importance score of ICULOS. Figure 5 will show a clear comparison of the features ranking 2 to 15.

3) IMPUTATION METHODS

Each model is trained using various imputation approaches. This is done on each PSV file individually. The imputation methods investigated for each algorithm are: dropping null values, front and back filling, filling with mean, filling with median, filling with mode, and mean-median filling. For all the models, several metrics are used for performance measurement and are discussed in Section IV. Each of these methods is explained here in detail.

(i) Dropping null values: The dropna() function in the python data frame is used to remove null values from data sets. This strategy is typically employed when the data set is too vast and the number of null values is small. All the null-valued rows can be removed from the data set using the dropna() method. Dropping of null values isn't very useful for small data sets. In this study, a row was dropped if any of the continuous variables present in that particular row were identified as null.

(ii) Front and back filling: The 'ffill' and 'bfill' methods can be utilized instead of filling the null values with another type of data. The first method, forward fill, replaces null values with past data, whereas the second method, backward fill, fills null values with the next genuine value in the data set. For ffill, suppose a value in the n^{th} row of a feature is null, that null value is assigned the $(n–1)^{th}$ row value. Similarly, for bfill, if a value in the n^{th} row of a feature is null, then that null value is assigned the $(n + 1)^{th}$ row value. For all features in this study, ffill was applied primarily. If the value of the first row is null for a feature, then ffill will not work. In such cases, bfill has been employed.

(iii) Filling with mean, median, and mode: The mean and median are filled in for the null values of numerical data, whereas the mode is filled in for the null values of categorical data. Filling in null entries with mean values may modify the data's mean and standard deviation if the data collection contains outliers. Therefore,

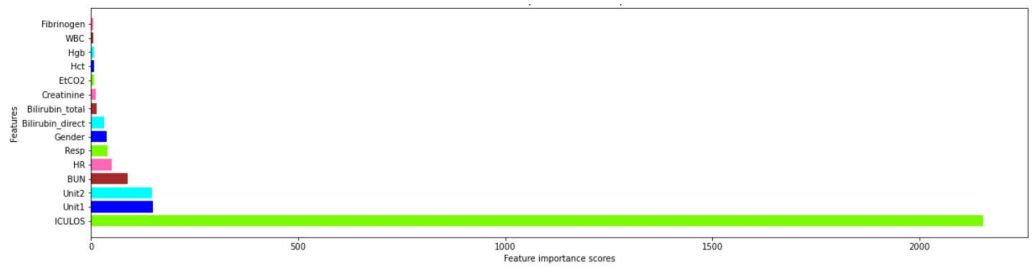

FIGURE 4. Top 15 features based on the feature importance scores.

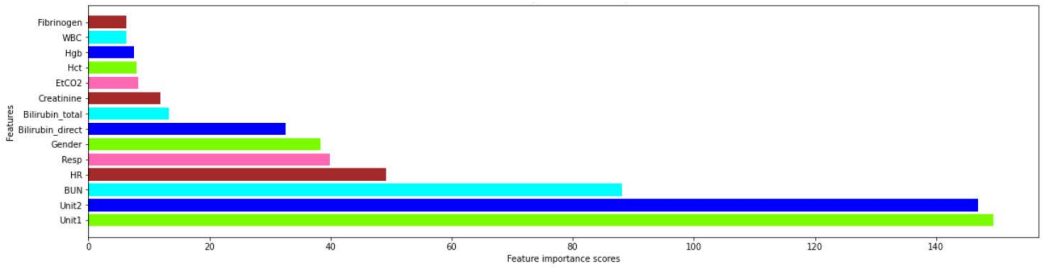

FIGURE 5. 2-15 features based on the feature importance scores.

in such cases, compared to mean, replacing null data with median values can be a very successful tactic.

All three above-mentioned methods (i.e., dropping null values, forward/backward fill, mean/median/mode fill) have been used to check how well the performance of the model is with each of them. Out of these three, forward/backward fill gave the best results. Hence, these results have been used for comparison with the novel outlier-based mean-median algorithm that has been introduced in this study

(iv) Outlier-based mean-median filling: Depending on the number of outliers present in a data collection, filling the null values with the mean and median will have a varying impact. If there are several outliers, the mean value will be greatly impacted, which will produce an inaccurate result. In such cases, it is preferred to fill the null values with the median. This imputation algorithm decides whether to fill the null values with mean or median accordingly. The Interquartile range is used to decide the number of outliers present in a particular feature. An outlier is an observation on a sample dataset that deviates from the general pattern. By dividing a data set into quartiles, the IQR is used to measure variability. The information is divided into 4 equal portions and arranged in ascending order. The values that divide the four equal halves are known as the first, second, and third quartiles, or Q1, Q2, and Q3, respectively.

The interquartile range, or IQR, is the space between the first and third quartiles, or Q1 and Q3: IQR = Q3–Q1. Outliers are data points that are either below or above the median (Q1–1.5 IQR or Q3 + 1.5 IQR).

TABLE 5. Outlier percentage of different features present in the data set.

Catergory	Feature name	Outlier %
Vital signs	HR	1.094
	Temperature	1.835
	SBP	1.103
	MAP	1.661
	DBP	2.296
	Resp	1.743
	EtCO2	27.431
Laboratory values	BaseExcess	5.419
	HCO3	4.195
	AST	13.051
	BUN	8.757
	Bilirubin_direct	11.956
	Bilirubin_total	11.408
	Hct	1.089
	Hgb	1.120
	Calcium	5.430
	Creatinine	11.940
	Glucose	5.150
	Magnesium	2.640
	Fibrinogen	3.899
	PTT	10.294
	WBC	3.462
	Platelets	3.450
Demographics	HospAdmTime	14.008
	ICULOS	4.282

The outlier percentage of different features is listed in Table 5 and the threshold considered in this study is 5%. If for a particular feature, it is greater than "5", then the null values of that particular feature are filled with median and otherwise with mean.

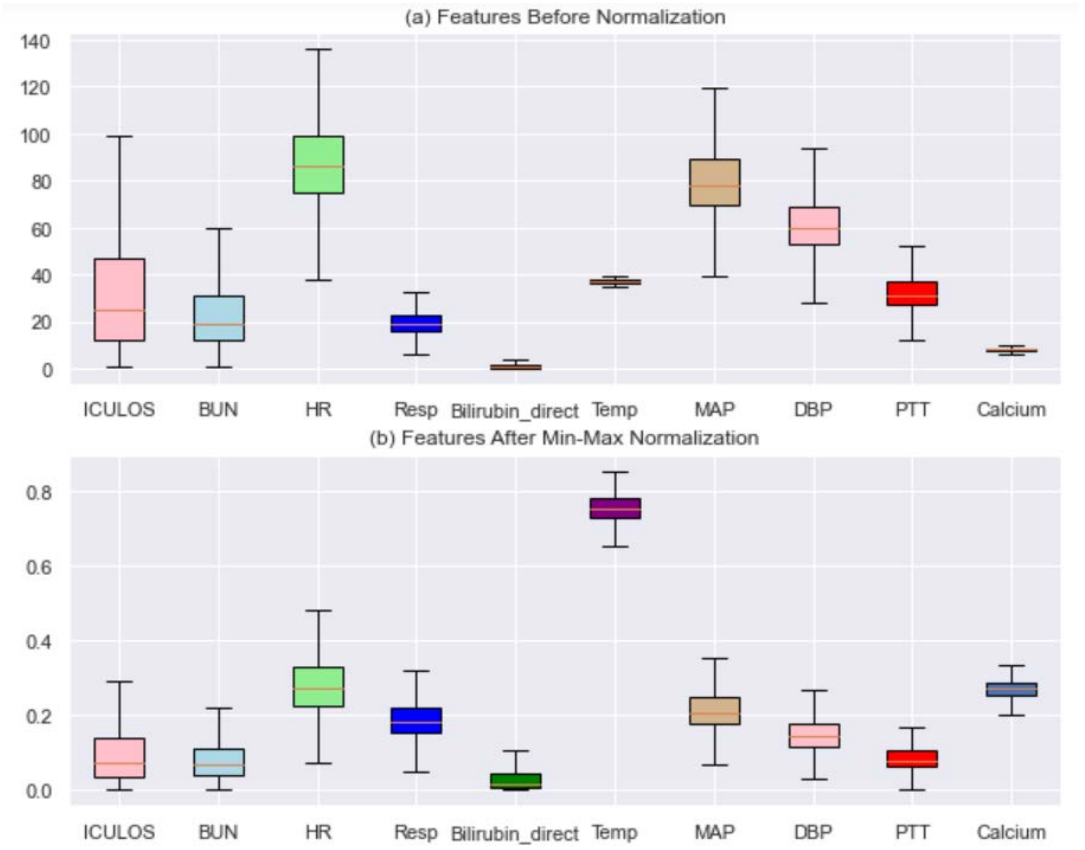

FIGURE 6. Sample box plot of 10 features before and after min-max normalization.

4) NORMALIZATION

To the data set obtained after applying the above filling methods, min-max normalization is applied. Re-scaling a set of data is what min-max normalization does. The information is changed to fit a target range of [0, 1]. The smallest value from the initial set would be changed to 0. The largest value from the initial set would be moved to 1. Any additional values would be translated into something between those ranges. After applying min-max normalization the results have shown significant improvement over the results of the unscaled data set. A sample plot of the data set before and after min-max normalization is shown in Figure 6.

5) STRATIFIED KFOLD CROSS-VALIDATION

After the above-mentioned processes are investigated, and the data is normalized, stratified KFold cross-validation is applied. The cross-validation technique is used to gauge the proficiency of machine learning models. A single parameter, k, which determines how many groups a given data sample should be divided into, is used in the method. As a result, K-fold cross-validation is a common name for this procedure. One-fold is considered a test set and the rest are used for training. This is done for every unique fold. Stratified kfold cross-validation is a refinement of ordinary kfold cross-validation for classification issues where the desired class ratio is the same in each fold as it is in the overall dataset, rather than the splitting being fully random. When we have unbalanced data, we usually utilize stratified kfold cross-validation. The k value, in this case, is taken as 3, i.e., the data set is split into 3 folds.

E. CLASSIFIERS FOR SEPSIS DETECTION

A typical supervised learning task is classification. It is employed whenever there is a need to make a categorical type prediction, such as whether or not a given example falls into a particular category. Some of the popular classifiers that can be used for early sepsis detection are Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes Classifier, Support Vector Machine (SVM), Random Forest Classifier, etc. While logistic regression and SVM are linear classification models, the rest come under non-linear classification models. There are different ways of assessing the performance of a classification model. A successful binary classification model should have a log loss value that is close to 0. If the anticipated value differs from the actual value, the value of the log loss rises. The higher accuracy of the model is shown by the lower log loss. The confusion matrix describes the performance of the model and gives us a matrix or table as an output. The matrix is made up of the results of the forecasts in a simplified way, together with the

total number of right and wrong predictions. The following four algorithms were implemented and used for comparison:

1) Logistic Regression:
The classification algorithm logistic regression is useful in determining the probability of event success and failure. In the case of a binary dependent variable (0/1, True/False, Yes/No), LR is utilized. After being trained on a labeled data set, this is used to categorize data into these distinct classes. It is based on the sigmoid function, with probability as the output and input ranging through the set of real numbers. Logistic regression is quite easy to implement, and it is also very efficient to train. When it comes to the classification of unknown records, it works quickly. When the data provided is linearly separable, this algorithm works very efficiently. Since it has a linear decision surface, it cannot perform well when it comes to non-linear data. In real-world scenarios, linear data is very rare. One of the fundamental drawbacks of logistic regression is the requirement for linearity between the dependent and independent variables.
2) Decision Tree:
A decision tree algorithm is a machine learning technique that is used for making predictions. The training data is repeatedly split into smaller data samples. This algorithm works based on conditions. At every node, we check the condition and split the data accordingly. This method can be used to build both classification and regression models. Normalization and scaling of data are not required for this algorithm. This requires a lot less data preprocessing work as compared to the other ML algorithms. To a reasonable extent, missing values in the data set will not affect the process of building a tree. One major disadvantage of decision trees is that even a minor change in the data might cause a significant change in the tree structure. This might lead to instability. It also usually requires more time to train the model.
3) Gradient Boosting:
Gradient boosting iteratively learns from weak learners and builds a strong model. There are several reasons why gradient boosting tree algorithms are used: In comparison with other methods, they are generally more accurate, and train faster, especially on bigger data sets and most of them allow categorical features. Gradient boosting is prone to overfitting. Models trained on this algorithm can be computationally expensive and training them takes a long time. These are some issues with gradient boosting.
4) Random Forest:
The bagging method is the foundation of the Random Forest. It develops as many trees as it can on different input portions before combining the results [34]. As a result, the accuracy of decision trees is increased while the over-fitting issue and variation are both minimized. Both categorical and continuous variables respond well to this. Missing values can be effectively handled using this approach. Given that it creates several trees (instead of just one as in decision trees) and bases judgments on the majority of votes, Random Forest requires substantially more training time than decision trees. The algorithms are implemented in Python using the Scikit-learn libraries.

IV. RESULTS

The purpose of this research study is to facilitate the detection of sepsis in patients before it is diagnosed by clinicians. This is extremely important as detecting sepsis as early as possible is crucial in treating sepsis as early treatments prove to be effective. To achieve this, better data preprocessing techniques were used to enhance the quality of the data. Once the data preprocessing had been performed, several models were trained using Cross Validation (CV). The "fit time" for each cv split indicates how long it will take to install the estimator on the train set. The amount of time spent scoring the estimator on the test set is indicated by "score time" for each cv split.

The performance of the various models employed in this study was assessed using the following metrics: precision, recall, F1-score, accuracy, and AUC-ROC. The mathematical formulae for these metrics are given by (1)-(4).

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

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F_1 - score = 2 \times \frac{precision \cdot recall}{precision + recall} \tag{4}$$

The True Positive (TP) suggested that the classification of a positive sample was accurate. For instance, a person with sepsis has been identified as such. A True Negative (TN) is when a negative sample is correctly identified, which means that a healthy individual (i.e., one who does not have sepsis) is identified as such. False Negatives (FN) happen when a positive sample is incorrectly classified as negative, and False Positives (FP) happen when a negative sample is wrongly classified as positive (for example, an incorrect cancer diagnosis) (for example, when you do not have an illness but the test says you do).

Precision is the percentage of relevant instances found among the recovered instances, whereas recall (sensitivity) is the fraction of relevant examples detected. As a result, relevance forms the cornerstone of both precision and memory. Average precision for ranked retrieval results is a metric that combines recall and accuracy. After every pertinent document has been obtained, the precision scores are averaged to determine the average precision for a specific information requirement.

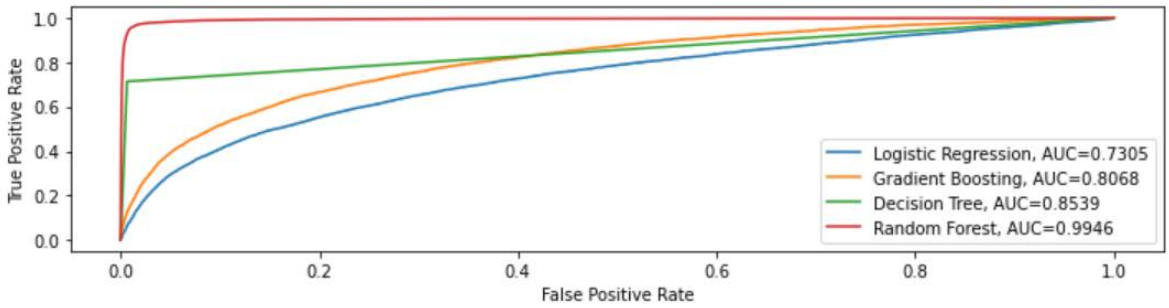

FIGURE 7. ROC curves for the models implemented.

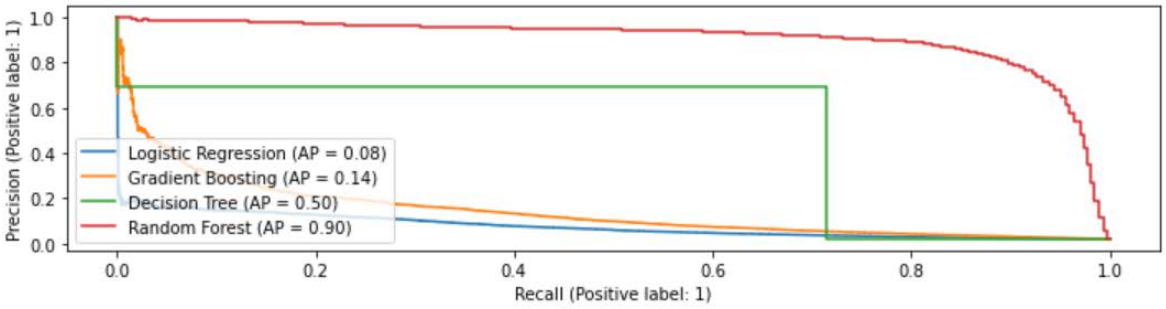

FIGURE 8. Precision-Recall curves for the models implemented.

One of the most commonly used methods for evaluating the effectiveness of ML algorithms is the ''AUC-ROC'' curve. The ROC curves summarize the trade-off between the true positive and the false positive rate for the predictive models. The evaluation metric for a binary classification problem is the Receiver Operator Characteristic (ROC) curve. It is a probability curve that contrasts the TPR and FPR at various threshold values to distinguish the ''signal'' from the ''noise''. The ROC curve's summary, the Area Under the Curve (AUC), gauges a classifier's capacity to differentiate between classes. The model's ability to differentiate between positive and negative classes is shown by the AUC. The AUC should be as high as possible. The ROC curve can be used to determine the optimal operating point.

A. PERFORMANCE OF SEPSIS DETECTION

This section compares the performances of the classifiers. The effect of each filing method on the classification models is depicted in terms of ROC curves and Precision-Recall curves in Figure 7 and Figure 8 respectively. When trained on the proposed novel mean-median fill data set, the Random Forest model stand out among the other models. The metrics achieved by the mean-median filling algorithm are higher since it combines the best elements of the other individual techniques.

In contrast to the decision tree and random forest, which show a considerable improvement, logistic regression and gradient boosting have fared better in forward fill than the novel outlier-based mean-median algorithm. The reasons listed below may be the cause of this. More so than decision trees and random forests, the outliers have an impact on the methods for logistic regression and gradient boosting. This occurs because outliers can alter the classification boundary in logistic regression, making it less accurate at predicting new data. As the boosting bases each tree on the residuals and errors of prior trees, outliers can be detrimental to the process. Gradient boosting will pay disproportionately more attention to outliers because they will have higher residuals than non-outliers. It should be noted that because the partitioning is based on the percentage of samples falling inside the split ranges rather than on absolute values, decision trees are not sensitive to outliers. The same rules apply to random forest.

The features with the highest importance scores, as determined by the findings of the univariate feature importance analysis, are ICULOS, HR, and Resp. The outlier threshold is taken into account for filling the null values with the mean as 5, meaning that any feature with an outlier % under 5 will have all of its null values replaced by the mean, while the null values for the other features will be replaced by the median. ICULOS: 4.28, HR: 1.09, and Resp: 1.74 are the outlier percentages for the top characteristics with high significance scores. Therefore, the mean is used to replace the null values for these features.

Logistic regression and gradient boosting techniques may suffer if the mean value is significantly affected by the outliers and remains a little abnormal compared to the other values around it. The threshold value can be lowered to avoid this.

TABLE 6. Average metrics of forward fill data set before data balancing.

Model	Logistic Regression			Decision Tree			Gradient Boosting			Random Forest		
K-Value	0	1	2	0	1	2	0	1	2	0	1	2
Accuracy	0.727834	0.728356	0.745188	0.616744	0.611909	0.620726	0.814659	0.809380	0.807474	0.928017	0.930468	0.927766
Precision	0.166667	0.184397	0.168000	0.302633	0.293777	0.310382	0.479452	0.745763	0.650000	0.902062	0.904505	0.897084
Recall	0.002660	0.003457	0.002793	0.343883	0.337101	0.348271	0.004654	0.005851	0.003457	0.069814	0.066755	0.069548
ROC AUC Score	0.723637	0.727071	0.723641	0.664209	0.660766	0.666656	0.794525	0.789572	0.795020	0.938833	0.934993	0.938820
Fit time	3.016576	2.044996	2.134997	9.554984	9.286411	9.139908	127.95399	136.55676	134.06907	106.640954	114.629819	116.544809

TABLE 7. Average metrics of outlier-based mean-median fill dataset before data balancing.

Model	Logistic Regression			Decision Tree			Gradient Boosting			Random Forest		
K-Value	0	1	2	0	1	2	0	1	2	0	1	2
Accuracy	0.667024	0.671638	0.670376	0.789688	0.795144	0.780854	0.718959	0.716677	0.717152	0.932686	0.932655	0.923391
Precision	0.116667	0.170940	0.157895	0.505191	0.514167	0.514372	0.941176	0.714286	0.882353	0.948293	0.952489	0.939345
Recall	0.001862	0.002660	0.001995	0.556516	0.557447	0.568750	0.004255	0.003989	0.001995	0.387766	0.391888	0.389229
ROC AUC Score	0.706373	0.720832	0.707721	0.772906	0.773552	0.779103	0.767020	0.759582	0.760623	0.983284	0.982841	0.983922
Fit time	1.264168	1.329116	1.379081	5.340294	5.185083	5.173229	89.948875	94.357645	96.409582	65.167174	73.530906	76.148572

TABLE 8. Average metrics of mean-median fill dataset after feature selection and before data balancing.

Model	Logistic Regression			Decision Tree			Gradient Boosting			Random Forest		
K-Value	0	1	2	0	1	2	0	1	2	0	1	2
Accuracy	0.709420	0.703622	0.709236	0.808487	0.816000	0.819946	0.740720	0.737537	0.743880	0.974377	0.977731	0.974074
Precision	0.215827	0.173913	0.230769	0.640704	0.625413	0.648074	0.666667	0.727273	0.786667	0.919132	0.918091	0.918250
Recall	0.003989	0.003191	0.003191	0.658511	0.654521	0.669016	0.006915	0.007447	0.007846	0.557713	0.550000	0.552660
ROC AUC Score	0.735273	0.726825	0.734545	0.825629	0.823412	0.830941	0.801483	0.799451	0.793210	0.991663	0.993091	0.991883
Fit time	3.127841	3.027978	3.952363	22.303683	22.684189	25.216650	259.282505	259.042776	260.196455	202.675841	204.129152	206.366575

TABLE 9. Average metrics of mean-median fill dataset after data balancing and feature selection.

Model	Logistic Regression			Decision Tree			Gradient Boosting			Random Forest		
K-Value	0	1	2	0	1	2	0	1	2	0	1	2
Accuracy	0.692613	0.694835	0.694384	0.949115	0.946344	0.946222	0.803693	0.808020	0.808856	0.990157	0.989625	0.989072
Precision	0.723998	0.725598	0.724935	0.935276	0.934070	0.932841	0.836511	0.839055	0.840787	0.992909	0.992741	0.992759
Recall	0.622555	0.626648	0.626478	0.965011	0.960481	0.961681	0.754931	0.762250	0.762010	0.987366	0.986463	0.985330
ROC AUC Score	0.752912	0.754034	0.752721	0.949115	0.946344	0.946222	0.883734	0.888003	0.888154	0.999311	0.999241	0.999293
Fit time	1.958690	2.163484	2.213525	67.200356	72.696535	71.772631	126.002376	127.116729	126.961694	429.478352	436.859091	414.422554

Both logistic regression and gradient boosting show improvement when the threshold value is set at 2.5. ICULOS, the feature with the highest relevance score, has its null values filled by the median for a threshold value of 2.5. Further progress can be noticed if the threshold value is lowered to 1.

Tables 6 to 9 display the findings of this research. These tables provide a detailed illustration of the results from several model testing phases. Table 6 displays the results of forward/backward filling the null values, whereas Table 7 displays the outcomes of the ''mean-median filling'' technique used to fill the null values. Table 8 displays the improvement after doing univariate feature selection, and Table 9 displays the results after balancing the number of patients with and without sepsis. Every stage shows improvement, and the final result is yielding a commendable outcome.

TABLE 10. Results after considering 2 hours of data.

Model	Decision Tree	Gradient Boosting	Random Forest
Accuracy	0.819946	0.743880	0.934832
Precision	0.529255	0.586957	0.861386
Recall	0.554318	0.075419	0.243017
ROC AUC Score	0.773851	0.805027	0.949202
Fit time	1.478256	22.832855	19.585103

TABLE 11. Results after considering 8 hours of data.

Model	Decision Tree	Gradient Boosting	Random Forest
Accuracy	0.834565	0.812343	0.957446
Precision	0.641717	0.862745	0.943820
Recall	0.698913	0.048913	0.478261
ROC AUC Score	0.846089	0.819370	0.983861
Fit time	2.048635	24.792092	19.289202

B. PERFORMANCE OF EARLY SEPSIS PREDICTION

Along with accurately predicting sepsis, it is crucial to determine whether the patient has it or not as soon as possible. As mentioned earlier, the data set used is [10]. Here, each patient's data is kept in its text file with pipe delimiters. Each file has a consistent header, and each row contains the patient's data for one hour. This study has built numerous data sets where it starts by taking 2 rows at a time, i.e., 2 hours' worth of patient data, to see how long it takes to attain an acceptable accuracy. The different observations on predictions are noted at various time intervals for decision tree, gradient boosting, and random forest algorithms and are depicted in Tables 10 to 13.

TABLE 12. Results after considering 16 hours of data.

Model	Decision Tree	Gradient Boosting	Random Forest
Accuracy	0.856787	0.811844	0.979982
Precision	0.634715	0.857143	0.931559
Recall	0.680178	0.034444	0.546667
ROC AUC Score	0.949816	0.867128	0.988743
Fit time	6.505713	87.973139	60.926660

TABLE 13. Results after considering 24 hours of data.

Metric\Model	Decision Tree	Gradient Boosting	Random Forest
Accuracy	0.987818	0.831759	0.997769
Precision	0.984628	0.858192	0.997585
Recall	0.990398	0.794861	0.997989
ROC AUC Score	0.987818	0.909038	0.999956
Fit time	53.683170	250.774743	815.384522

TABLE 14. Comparison with best performing models of existing work.

Research work	Accuracy	F1-Score	ROC AUC Score
Early Prediction of Sepsis Using Machine Learning [21]	0.9888	0.9834	0.9987
Early Prediction of Sepsis from clinical data via Heterogeneous Event Aggregation [16]	NA	NA	0.8224
Early Prediction of Sepsis: Using State-of-the-art ML Techniques on Vital Sign Inputs [13]	0.866	0.08	0.736
Dynamic Sepsis Prediction for Intensive Care Unit Patients Using XGBoost-Based Model With Novel Time-Dependent Features [26]	NA	0.8772	0.8048
A Comprehensive Machine Learning based Pipeline for an Accurate Early Prediction of Sepsis in ICU (Proposed work)	0.9977	0.9978	0.9999

NA: Not Available

The findings improve when more and more hours of data are collected. We can see that after 24 hours of data, the AUC_ROC value has surpassed 99.9 percent, which is a respectable result. While a clinician would need 2-3 days to conduct numerous tests and diagnose sepsis, our algorithm can do so in as little as 24 hours or less. So, it can be noted that this approach helps with early sepsis prediction.

C. COMPARISON WITH EXISTING WORK

As is evident in Table 14, the accuracy and ROC AUC scores of the model developed in this research study have significantly improved when compared to previous research. It should be emphasized that the study's top-performing model was taken into account for comparison in each situation. This technique certainly helps with a better analysis for sepsis early identification.

V. CONCLUSION

The real-time data set used for this investigation has a large number of null or missing values. Many features have missing values that are greater than 90%. A variety of imputation techniques were examined for patients' early sepsis diagnosis. Each of these methods is tested against different machine learning algorithms. From the results, it can be concluded that applying min-max normalization on data after filling the null values with the proposed outlier based mean-median imputation method improves the prediction performance.Among all the machine learning algorithms that have been experimented with, the Random Forest method yields the best results. The proposed strategies can be used regularly to supplement the usual ICU scoring methods and to plan early treatment for sepsis.

Despite the study's demonstrated improved performance, further external and independent evaluations are necessary to confirm the findings. The results reflect only one dataset population, validation of external, non-trained datasets is necessary before fully generalizing the conclusions. The current study utilises solely structured data. There is scope for experimentation with combining structured and unstructured textual data.

REFERENCES

[1] (2020). World Health Organization. *Sepsis*. Accessed: Feb. 14, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/sepsis

[2] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J. Chiche, C. M. Coopersmith, R. S. Hotchkiss, M. M. Levy, J. C. Marshall, G. S. Martin, S. M. Opal, G. D. Rubenfeld, T. van der Poll, J. Vincent, and D. C. Angus, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *J. Amer. Med. Assoc.*, vol. 315, no. 8, pp. 801–810, Feb. 2016.

[3] J. Gallagher. *'Alarming' One in Five Deaths due to Sepsis*. Accessed: Jun. 30, 2022. [Online]. Available: https://www.bbc.com/news/health-51138859

[4] K. H. Goh, L. Wang, A. Y. K. Yeow, H. Poh, K. Li, J. J. L. Yeow, and G. Y. H. Tan, "Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare," *Nature Commun.*, vol. 12, no. 1, p. 711, Jan. 2021.

[5] N. Kijpaisalratana, D. Sanglertsinlapachai, S. Techaratsami, K. Musikatavorn, and J. Saoraya, "Machine learning algorithms for early sepsis detection in the emergency department: A retrospective study," *Int. J. Med. Informat.*, vol. 160, Apr. 2022, Art. no. 104689.

[6] K. Ackermann, J. Baker, M. Green, M. Fullick, H. Varinli, J. Westbrook, and L. Li, "Computerized clinical decision support systems for the early detection of sepsis among adult inpatients: Scoping review," *J. Med. Internet Res.*, vol. 24, no. 2, Feb. 2022, Art. no. e31083.

[7] P. E. Marik and A. M. Taeb, "SIRS, qSOFA and new sepsis definition," *J. Thoracic Disease*, vol. 9, no. 4, pp. 943–945, Apr. 2017.

[8] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. M. H. Schein, and W. J. Sibbald, "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis," *Chest*, vol. 101, no. 6, pp. 1644–1655, Jun. 1992.

[9] R. Lyu, "Improving treatment decisions for sepsis patients by reinforcement learning," M.S. thesis, Univ. Pittsburgh, Pittsburgh, PA, USA, Mar. 2020.

[10] A. M. Reyna, C. Josef, S. Seyedi, R. Jeter, S. P. Shashikumar, M. B. Westover, A. Sharma, S. Nemati, and D. G. Clifford. (2019). *Early Prediction of Sepsis From Clinical Data: The Physionet/Computing in Cardiology Challenge 2019*. [Online]. Available: https://physionet.org/content/challenge-2019/1.0.0/#challenge-data

[11] F. Mas-Celis, J. Olea-López, and J. A. Parroquin-Maldonado, "Sepsis in trauma: A deadly complication," *Arch. Med. Res.*, vol. 52, no. 8, pp. 808–816, Nov. 2021.
[12] N. Heming, E. Azabou, X. Cazaumayou, P. Moine, and D. Annane, "Sepsis in the critically ill patient: Current and emerging management strategies," *Expert Rev. Anti-Infective Therapy*, vol. 19, no. 5, pp. 635–647, 2021.
[13] M. Nakhashi, A. Toffy, P. V. Achuth, L. Palanichamy, and C. M. Vikas, "Early prediction of sepsis: Using state-of-the-art machine learning techniques on vital sign inputs," in *Proc. IEEE Comput. Soc.*, Sep. 2019, p. 1.
[14] V. Abromavičius, D. Plonis, D. Tarasevičius, and A. Serackis, "Two-stage monitoring of patients in intensive care unit for sepsis prediction using non-overfitted machine learning models," *Electronics*, vol. 9, no. 7, p. 1133, Jul. 2020.
[15] X. Li, G. André Ng, and F. Schlindwein, "Convolutional and recurrent neural networks for early detection of sepsis using hourly physiological data from patients in intensive care unit," in *Proc. Comput. Cardiology Conf. (CinC)*, Dec. 2019, pp. 1–4.
[16] L. Liu, H. Wu, Z. Wang, Z. Liu, and M. Zhang, "Early prediction of sepsis from clinical data via heterogeneous event aggregation," in *Proc. Comput. Cardiol. Conf. (CinC)*, Dec. 2019, pp. 1–4.
[17] N. Nesaragi and S. Patidar, "Early prediction of sepsis from clinical data using ratio and power-based features," *Crit. Care Med.*, vol. 48, no. 12, pp. E1343–E1349, 2020.
[18] N. Nesaragi, S. Patidar, and V. Thangaraj, "A correlation matrix-based tensor decomposition method for early prediction of sepsis from clinical data," *Biocybernetics Biomed. Eng.*, vol. 41, no. 3, pp. 1013–1024, Jul. 2021.
[19] N. Nesaragi, S. Patidar, and V. Aggarwal, "Tensor learning of pointwise mutual information from EHR data for early prediction of sepsis," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104430.
[20] S. M. Lauritsen, M. E. Kalør, E. L. Kongsgaard, K. M. Lauritsen, M. J. Jørgensen, J. Lange, and B. Thiesson, "Early detection of sepsis utilizing deep learning on electronic health record event sequences," *Artif. Intell. Med.*, vol. 104, Apr. 2020, Art. no. 101820.
[21] A. Shankar, M. Diwan, S. Singh, H. Nahrpurawala, and T. Bhowmick, "Early prediction of sepsis using machine learning," in *Proc. 11th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence)*, Jan. 2021, pp. 837–842.
[22] A. D. Bedoya, J. Futoma, M. E. Clement, K. Corey, N. Brajer, A. Lin, M. G. Simons, M. Gao, M. Nichols, S. Balu, K. Heller, M. Sendak, and C. O'Brien, "Machine learning for early detection of sepsis: An internal and temporal validation study," *JAMIA Open*, vol. 3, no. 2, pp. 252–260, Jul. 2020.
[23] M. Selcuk, O. Koc, and A. S. Kestel, "The prediction power of machine learning on estimating the sepsis mortality in the intensive care unit," *Informat. Med. Unlocked*, vol. 28, 2022, Art. no. 100861.
[24] R. R. Watkins, R. A. Bonomo, and J. Rello, "Managing sepsis in the era of precision medicine: Challenges and opportunities," *Expert Rev. Anti-infective Therapy*, vol. 20, no. 6, pp. 871–880, Jun. 2022.
[25] D. Zhang, C. Yin, K. M. Hunold, X. Jiang, J. M. Caterino, and P. Zhang, "An interpretable deep-learning model for early prediction of sepsis in the emergency department," *Patterns*, vol. 2, no. 2, Feb. 2021, Art. no. 100196.
[26] S. Liu, B. Fu, W. Wang, M. Liu, and X. Sun, "Dynamic sepsis prediction for intensive care unit patients using XGBoost-based model with novel time-dependent features," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 8, pp. 4258–4269, Aug. 2022.
[27] T. C. Do, H. J. Yang, S. B. Yoo, and I.-J. Oh, "Combining reinforcement learning with supervised learning for sepsis treatment," in *Proc. 9th Int. Conf. Smart Media Appl.*, New York, NY, USA, Sep. 2020, pp. 219–223.
[28] Z. He, L. Du, P. Zhang, R. Zhao, X. Chen, and Z. Fang, "Early sepsis prediction using ensemble learning with deep features and artificial features extracted from clinical electronic health records," *Crit. Care Med.*, vol. 48, no. 12, pp. E1337–E1342, 2020.
[29] Z. Wang and B. Yao, "Multi-branching temporal convolutional network for sepsis prediction," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 2, pp. 876–887, Feb. 2022.
[30] J. Singh, M. Sato, and T. Ohkuma, "On missingness features in machine learning models for critical care: Observational study," *JMIR Med. Informat.*, vol. 9, no. 12, Dec. 2021, Art. no. e25022.
[31] L. Zhang, Z. Wang, Z. Zhou, S. Li, T. Huang, H. Yin, and J. Lyu, "Developing an ensemble machine learning model for early prediction of sepsis-associated acute kidney injury," *iScience*, vol. 25, no. 9, Sep. 2022, Art. no. 104932.
[32] S. Roy, P. Sharma, K. Nath, D. K. Bhattacharyya, and J. Kalita, "Pre-818 processing: A data preparation step," Tech. Rep., 2018.
[33] W. K. Bowyer, V. N. Chawla, O. L. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," 2011, *arXiv:1106.1813*.
[34] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

B. C. SRIMEDHA received the B.Tech. degree in information technology from the Manipal Institute of Technology, Manipal, Karnataka, India, in 2022.

She has worked on computer intelligence for medical systems (heart rate detection) as a part of an Internship, in 2021. Her current research interests include machine learning, artificial intelligence, and computer vision.

RASHMI NAVEEN RAJ received the B.E. degree in electrical and electronics engineering from Mangalore University, Karnataka, India, in 2001, and the master's degree in technology in digital electronics and advanced communication and the Ph.D. degree in cognitive radio ad-hoc networks from the Manipal Academy of Higher education, Manipal, Karnataka, in 2009 and 2021, respectively.

She is working as an Associate Professor with the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal. She has 16 years of teaching experience to engineering students at National Institute of Technology, Karnataka, and the Manipal Institute of Technology, Manipal. She has published many articles in good journals. Her research interests include cognitive radio networks, machine learning, and autonomous driving vehicles.

VEENA MAYYA (Senior Member, IEEE) received the M.Tech. degree in software engineering from the Manipal Institute of Technology, Manipal University, India, and the Ph.D. degree in healthcare analytics from the National Institute of Technology Karnataka, Surathkal, India.

She has more than 15 years of experience, which includes both academic and industrial experience in IT. She is currently with the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education (MAHE), India. She has published more than 15 research articles in reputed journals and conference proceedings. Her current research interest includes multimodal data-driven clinical decision-support systems.