Improving Sepsis Prediction Performance Using Conditional Recurrent Adversarial Networks

MERVE APALAK, (Student Member, IEEE), AND KAMRAN KIASALEH, (Senior Member, IEEE)
Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX 75080, USA
Corresponding author: Merve Apalak (Merve.Apalak@utdallas.edu)

ABSTRACT In this paper, we devise a novel method involving deep neural networks (DNNs) that improves the early prediction of sepsis for patients admitted to the intensive care units (ICUs). It is assumed that the patient data sets are dramatically corrupted by missing information, which negatively impacts the detection of the onset of sepsis. We propose a generative learning framework to estimate the missing information in data. Our model involves Conditional Generative Adversarial Networks (GANs) utilizing Long Short-Term Memory (LSTM) networks as the generator and discriminator when conditioned on class labels. A deep LSTM network is also employed for prediction purposes. The prediction network is trained with an output of the conditional GAN and evaluated on an unseen test set to investigate the performance of the proposed model. Here, we show that the proposed framework not only identifies long-term temporal dependencies but also exploits the missing patterns. We present the performance results and compare them to other well-known techniques. For the 4-hour, 8-hour, and 12-hour prediction of sepsis, the proposed method attains area under the receiver operating characteristic (AUROC) of 94.49%, 93.74%, and 94.01%, respectively. It is shown here that the improvement in imputation and prediction promises a highly effective method that can offer early detection of sepsis in high-risk patients.

INDEX TERMS Sepsis, machine learning, early diagnosis, time-series prediction, generative adversarial networks (GANs), clinical decision support.

I. INTRODUCTION
Sepsis is a life-threatening medical emergency that can rapidly lead to tissue damage, organ failure, and death [1]. Sepsis is considered responsible for more than one-third of the hospital deaths in the United States and the increased incidence have been a growing concern [2]. It is one of the most expensive conditions to treat, representing 13% of the total U.S. healthcare cost. Additionally, statistics show that the average length of stay in hospitals for sepsis patients is nearly 75% longer than that of other medical conditions [3]. It has been reported that the early intervention and recognition of sepsis can significantly reduce the overall mortality and cost burden of sepsis. The importance of early prediction and treatment of sepsis is emphasized in the current clinical and observational studies that show a lower risk of mortality for sepsis patients who received antibiotics and intravenous fluids on time [4], [5]. In another study, it is reported that hourly delays in the initiation of antibiotic therapy can cause an average increase in the mortality rate by 7.6% [6].

In the context of sepsis diagnosis, the Systemic Inflammatory Response Syndrome (SIRS) criteria were considered to be central [7]. Recently, the third international consensus definition for sepsis and septic shock (Sepsis-3) was published. For diagnostic criterion, the Sequential Organ Failure Assessment (SOFA) scoring system was proposed. Furthermore, SIRS criteria have been criticized for inadequate specificity and sensitivity since SIRS may occur in several non-infectious scenarios [8]. The SOFA score is based on the degree of dysfunction of six organ systems, such as respiratory, coagulation, hepatic, cardiovascular, renal, and neurological systems [9]. According to the Sepsis-3 guidelines, patients with a SOFA score of 2 or more are associated with an organ failure consequent to the infection, meaning that a higher SOFA score indicates the increased mortality risk. The Modified Early Warning Score (MEWS) is another...
scoring system used for the determination or prediction of sepsis [10]. These updated definitions and gold standards have been adapted to facilitate the earlier identification and timely management of septic patients. However, sepsis is a dynamic condition and, hence, such criteria may not yield accurate outcomes. Consequently, early prediction of the onset of sepsis remains a challenging problem.

There has been a significant surge in using deep neural networks (DNNs) for solving multivariate, complex, and nonlinear problems. Training such networks requires a significant volume of data. Meanwhile, the intensive care unit (ICU) patients are monitored consistently. This has generated an abundance of data, which allows for training DNN for event prediction or decision support in critical care cases [11]. Recent studies have incorporated such DNN-based approaches using electronic health records (EHRs) for identifying the early stages of complex diseases [12], [13], [14]. In sepsis cases, attention has been placed on creating an accurate and swift prediction model as an extension of the clinical decision since the performance of the deep learning models has been found significantly higher than traditional scoring systems. In [15] and [16], authors have been systematically reviewed and evaluated studies employing machine learning for the prediction of sepsis in the ICU. Desautels et al. [17] compare the performance of the predictive model called InSight versus alternative scoring systems. Authors only employed 8 common measurements for training the model. InSight attains a performance of 4 hours prior to the onset of sepsis with the area under the receiver operating characteristics (AUROC) curve of 0.74. Shashikumar et al. [18] investigated high-resolution blood pressure (BP) and heart rate (HR) dynamics using a multivariate modeling approach for the early sepsis prediction. Laboratory results were not considered for the training model. However, the recent study by Henry et al. [19] showed that laboratory results are powerful attributes. For instance, the ratio of blood urea nitrogen to Creatinine (BUN/Creatinine) was found highly important in sepsis detection [20].

There is often massive uncertainty in the medical data set because measurements are recorded at different and often irregular times. This can pose unexpected challenges and deterioration in prediction performance for DNN approaches. Therefore, most of the studies have been performed with limited access to data (only vital signs are considered) [17], [21]. Recent literature addressed these drawbacks by directly modeling missing information in clinical time series [22]. Authors have not only imputed the missing data but also have augmented the recurrent neural network (RNN) inputs with binary indicators which address the location of the missing variable in the time series. The problem remains important since the high rates of missing data can result in a potential bias leading to an inaccurate diagnosis and treatment as well as poor modeling and statistical analyses. Common approaches such as omitting the missing values along consecutive terms may result in information loss. It can also dramatically shorten sample size which is not feasible in order to produce reasonable results for DNNs. For instance, mean substitution is a common and simple practice of replacing missing values but it disturbs the variance of completed data and the correlation of other variables. There is literature supporting more sophisticated methods, such as the use of DNNs to estimate missing values from the information available in the data set. Two probabilistic interpretations of bidirectional RNNs were employed as generative models to fill the missing gaps in time series data [23]. For instance, Yoon et al. [24] proposed a hierarchical learning framework based on RNNs for estimating missing information. This method utilizes the correlations within and across streams.

The generative adversarial networks (GAN) introduce an intelligent framework for data generation which can be characterized by training two neural networks, called generator and discriminator [25]. GANs have such a flexible structure that they can be modified to train any neural network, such as RNNs. Unfortunately, RNNs suffer from well-known gradient vanishing and exploding problems [26]. This shortcoming is addressed by Long Short Term Memory (LSTM) network [27], [28]. LSTM networks are capable of capturing long-term temporal dependencies. Since temporal features play a critical role in understanding the changes in a patient’s condition, many studies have investigated the performance of LSTM networks with medical data sets [29], [30], [31], [32]. Furthermore, LSTM has been adopted to GANs. The earliest work that utilized the LSTM network with adversarial training was implemented on continuous sequential data to generate music [33]. For data augmentation purposes, GAN was applied to time series data by instantiating RNN for generator and discriminator [34]. More recently, GANs were adapted to impute the missing information in an incomplete data set [35], [36], [37], [38]. Adversarial imputation with modified gated recurrent unit (GRU) cells was elaborated to process the incomplete multivariate time series data [38]. However, generated data can be constrained by labels to get better performance considering the dependencies and connections between observed and unobserved parts. With the motivation of using conditional GANs [39], Esteban et al. [40], introduced Conditional Recurrent GAN to produce real-valued, realistic time series results to generate the most frequently recorded vital signs in the ICU.

In this article, we are interested in early detection of the onset of sepsis. There are two main drawbacks to the current sepsis prediction models; 1) inadequate performance for longer prediction windows and 2) limited usage of data sets. We explore an adversarial neural network that can alleviate the constraints of missing information to improve the accessibility of time series data. Our strategy is to build a GAN-based preprocessing block that has the ability to learn the underlying dependencies and correlations of the observed time series to estimate the missing values. As opposed to other works, our model requires several modifications to learn mapping to complete the time series. Any differentiable function can be employed as the generator and the discriminator. Thus, we implement simple LSTM network for generator
and discriminator. Next, we customize the traditional GAN inputs where we provide observed parts of time series and labels as additional inputs. The conditioning limits the search space for a more realistic estimations and results in a faster convergence. Deep LSTM network is utilized for the early prediction block. The algorithm is trained and tested on unseen data. We examine the prediction performance of the onset of sepsis in an ICU patient 4 to 12 hours prior to the clinical recognition of sepsis. In conclusion, the primary contributions of our paper can be summarized as follows.

- We propose a novel end-to-end learning framework for early sepsis prediction which can directly deal with the time series with missing values.

- We build a preprocessing block with recurrent GAN framework conditioned on the observed part of the time series and its labels. Hence, we can estimate the missing values by capturing temporal dependencies and correlations in the complete part of the time series.

- We apply deep LSTM network for the prediction block. It is shown here that such an approach achieves highly promising results on prediction with varying prediction windows. We also show that the mutual contributions of preprocessing and prediction block reduce the error propagation from imputation to prediction, thereby improving the prediction performance.

The rest of the paper is organized as follows. An overview of the modeling approach, the data set, preprocessing block utilized in the imputation process, and early prediction block are comprehensively described and formalized in Section II. We highlight the key advantages of each network by reporting the literature. In Section III, we provide results based on experimental observations to evaluate model performance. Subsequently, we compare our results with the recent works that use deep learning methods in this section. We draw conclusions in Section IV.

II. MATERIALS AND PROPOSED EARLY PREDICTION METHOD

A. OVERVIEW OF THE MODELING APPROACH

The summary of the proposed framework is illustrated in Fig.1. It consists of two main blocks, which we call preprocessing and the prediction block. Preprocessing block accepts the data and normalize it by defining a new boundary definition such as (0,1). As a next step, we filter the patients who have an ICU length of stay of fewer than 12 hours. Following this step, we perform binning for age and the ICU length of stay columns. We categorize on the basis of the range of values of the column in which they fall. As it can be seen in Table 1, six groups for age and the ICU length of stay are defined. Following this step, we use the recurrent conditional adversarial network as a data imputation strategy. Imputation network is trained under a generative adversarial learning framework. Finally, preprocessed data is fed into a deep LSTM model to predict sepsis. More detailed information will be given in the following sections regarding the model structures.

| Binning | Age        | ICU Length of Stay (hours) |
|---------|------------|---------------------------|
| 0       | <30        | <25                       |
| 1       | 30-40      | 25-40                     |
| 2       | 40-50      | 40-65                     |
| 3       | 50-60      | 65-100                    |
| 4       | 60-70      | 100-1000                  |
| 5       | >70        | No class                  |

B. DATA

The 2019, PhysioNet/ Computing in Cardiology Challenge data set is used to evaluate the performance of the proposed framework [41]. The data set is an ICU patient database sourced from three U.S. hospital systems with three different electronic medical records (EMR). We only use the data set obtained from two hospital systems which are obtained from Beth Israel Deaconess Medical Center and Emory University Hospital. It was stated that the data had been collected over the years with the approval of the institutional review boards. We refer readers to challenge resources for more information [41]. The remaining database of the third hospital system was not shared publicly. It was kept as a hidden data set to evaluate the performance of participants. Data set is composed of anonymous clinical documentation.
from 40,336 patients containing 8 vital signs, 26 laboratory, 6 demographic values, and sepsis labels for each patient. Clinical variables are recorded over time. For each patient, these variables are multivariate time series. For testing and model development, patient features were simplified by condensing the results into hourly measurements. For instance, heart rate measurements in an hourly window were reported as the median heart rate measurements. Sepsis labels describe the onset time of sepsis as reported by the Sepsis-3 criteria where each patient is identified as “sepsis” or “non-sepsis”. Sepsis-3 criteria guidelines were considered for labeling process. Guideline suggests that there must be a suspicion of infection and two-point increase in the SOFA score within a 24 hours period. The onset of sepsis time was decided when there was a two or more points change in the score within the suspicion of infection window [42]. It is important to note that, although the Sepsis-3 criteria is accepted for current standards of sepsis labeling, Sepsis-1 and 2 criteria and other metrics remain in wide use. This also increases feasibility of our approach. It is also important to mention that nearly 70% of the data is missing. If we just use simple imputation methods, it may lead to a poor model performance. Therefore, it becomes important to utilize sophisticated methods to deal with the missing values.

C. RECURRENT CONDITIONAL ADVERSARIAL IMPUTATION ARCHITECTURE

As a general approach, the sampling times are used to capture informative missingness to address the irregular sampling and sampling information that will be fed to DNNs. Previous studies follow different strategies to replace the missing values, such as replacing it with 0, mean values, or latest values [22], [43], [44]. Our proposed GAN-based preprocessing block is the most efficient in circumventing the interruption of the incomplete data for prediction purposes. The main idea of the GAN is based on the Nash equilibrium in game theory. Here, adversaries are generator and discriminator where both models are simultaneously trained to optimize themselves. The generator’s task is to capture real data distribution while the discriminator tries to recognize real data from the generated data. Generation ability and discrimination ability improve gradually by the adversarial training process until the generator produces data that the discriminator cannot recognize from the real data. Proposed network architecture is motivated by a couple of well-known approaches [35], [39], [40]. In order to increase the prediction performance, our motivation is to construct a deep model that is capable of generating a completed time series to make an accurate estimation for the missing part of the data. We assume that it is likely to generate the missing values under an incorrect distribution. Hence, we simulate the distribution of the missing data under the conditional information of observed data and labels, i.e., \( P(\hat{X} \mid X, y) \) denoting the conditional distribution of completed data \( \hat{X} \) given observed variables \( X \), and labels \( y \). So the input to the generator is augmented with conditional information, such as the observed part of the data and labels.

Conditioning allows us to control the data generation process and forces the imputed data to be more realistic. From another aspect, intuitively, a method that feeds class information as additional information has the potential to overcome skewed class distribution. Due to the nature of the time series, it is prone to observe the loss of potentially important information after imputation with the standard GAN; therefore, we substitute both the generator and the discriminator of [35] with LSTM networks. In summary, the proposed preprocessing block can deal with the temporal relation of the data, class label balance, and the case of incomplete data by using the conditional recurrent imputation network with adversarial training. A detailed block diagram for the recurrent conditional adversarial network is illustrated in Fig. 2.

1) GENERATOR

We consider a collection of multivariate time series with \( N \) variables of length \( T \). We denote \( X = (x_1, x_2, \ldots, x_T) \in \mathbb{R}^{N \times T} \) to be multivariate time series. For each \( t \in \{1, 2, 3, \ldots, T\} \), the column vector \( x_t \in \mathbb{R}^{N \times 1} \) represents the measurements taken at the time step \( t \). It is convenient to introduce some additional notation. That is, let \( x_t^0 \) denote the measurement of the \( n \)-th variable taken at the time step \( t \). Suppose a time series \( X \) has unobserved variables which can be denoted by a mask vector, \( m_t \in \{0, 1\}^{N \times 1} \), to record the missing pattern of \( x_t \). We introduce \( M = (m_1, m_2, \ldots, m_T) \in \{0, 1\}^{N \times T} \) to be the mask matrix. More specifically, if a variable of \( x_t \) is unobserved, corresponding \( m_t \) is set to 0, otherwise, it is 1.

Generator \( G \) receives the incomplete (original) time series, \( X \), the corresponding mask matrix, \( M \), the random noise, \( Z \), and the label information, \( y \). We denote random noise as \( Z = (z_1, z_2, \ldots, z_T) \in \mathbb{R}^{N \times T} \). The output of \( G \) is imputed time series, \( \hat{X} \). In traditional GAN, generator tries to map the random noise vector \( Z \) to a time series vector \( X \), which generates an entire time series, \( G(Z) = Z \rightarrow X \). Here, we pass an observed measurements and labels as an additional input by replacing the missing values of original data with random noise. It is important to mention that \( G \) in fact estimates a value for the entire time series including values for the observed variables. We formulate the completed time series, \( \hat{X} \), as follows:

\[
\hat{X} = M \odot X + (1 - M) \odot \overline{X}.
\]

where \( \odot \) indicates element-wise multiplication. The corresponding values from \( \overline{X} \) are used to fill the missing part of the original time series \( X \). As it is illustrated in Fig 2, LSTM network is employed for \( G \) to estimate more realistic outputs. With the help of LSTM, missing values are estimated and filled in a step-by-step manner.

2) DISCRIMINATOR

We introduce a discriminator \( D \) which is simply a binary classifier and helps adversely to train \( G \). However, in this framework, \( G \) is not completely fake, actually it compromises some observed (real) and some imputed (fake) values.
FIGURE 2. Block diagram of the proposed deep neural network architecture. (a) Detailed recurrent conditional imputation network with adversarial training. The model receives incomplete times series data $X$, random noise variable $Z$, mask matrix $M$ and labels $y$. It outputs the completed time series $\hat{X}$. As a next step, hint matrix $H$ and $\hat{X}$ are fed to the discriminator. Once the generator outputs the optimal completed data, training is halted and $\hat{X}$ is ready to be used for predictions. (b) Deep LSTM network is accepting $\hat{X}$ to execute early prediction of sepsis.

Therefore, D is used to identify which values are imputed by G. That is, D attempts to maximize the probability of predicting $M$, while G tries to minimize the probability of predicting $M$. Simply, D forces G to learn the appropriate data distribution. We borrow the idea of hint mechanism which is introduced in [35]. The idea is to provide partial information about $M$ to the discriminator by defining hint matrix $H$. It is defined as

$$H = B \odot M + 0.5(1 - B)$$

where $B$ is a random variable taking values in $(0, 1)^{N \times T}$ and follows the Bernoulli distribution. $B$ is defined to control the amount of information included in $H$ about $M$. If we pass too much information about $M$ to $H$, D would overfit to the $H$. In contrast, if we simply omit the hint mechanism, then there are various distributions that G can learn. Hence, it may not replicate the desired distribution.

Formally, D is a function implementing a binary classifier $D : \hat{X} \rightarrow [0, 1]^N$. Here, $H$ is combined with $\hat{X}$, so that D becomes a function $D : \hat{X} \times H \rightarrow [0, 1]^N$.

3) OBJECTIVE

Training of the traditional GAN requires obtaining the parameters of a discriminator that maximize the probability of assigning the correct label to both real data and fake samples and obtaining the parameters of a generator that minimize the probability of D making incorrect decisions. This is also considered as a min-max game between the discriminator and the generator. The cost of a typical GAN training is evaluated by the objective function given as

$$\min_G \max_D V(D, G)$$

where

$$V(D, G) = \mathbb{E}_{X \sim p_{data}(X)}[\log D(X)] + \mathbb{E}_{Z \sim p_{z}(Z)}[\log(1 - D(G(Z))].$$

Here, the real data $X$ and random noise $Z$ are samples drawn from $p_{data}(X)$ and $p_{z}(Z)$, respectively. When G is optimal, $p_{data}(X) = p_{z}(Z)$, implying that the distribution of the generated data is equivalent to that of the real data.

Note that we perform the conditional GAN in our proposed model. The estimation of the missing part of the data is based on the prior information of the observed part of the data and labels. Therefore, inspired by [36], the standard GAN objective function is adapted for our conditional model. It is described as

$$\min_G \max_D V(D, G) = \mathbb{E}_{\hat{X} \sim p_{data}({\hat{X}} | X, y)}[M^T \log D(\hat{X}, H) + (1 - M^T \log(1 - D(\hat{X}, H))]$$

where $\hat{X}$ is completed data generated by G and sampled from $p_{data}(\hat{X} | X, y)$. 
It is important to emphasize that D is designed to distinguish between the observed and the imputed values rather than identifying the entire time series. In other words, D is designed to predict the mask matrix $M$. Discriminator loss is used to evaluate the confidence level of the imputed values. We use binary cross entropy to measure the difference between $M$ and $\hat{M}$. The discriminator loss, $L_D$, is defined as follows:

$$L_D(M, \hat{M}) = \sum_{t=1}^{T} \sum_{n=1}^{N} [m_t^n \log (\hat{m}_t^n) + (1 - m_t^n) \log (1 - \hat{m}_t^n)] \quad (5)$$

where $m_t^n$ and $\hat{m}_t^n$ are the $n$th elements of $m_t$ and $\hat{m}_t$, respectively. Note that $\hat{M}$ is the estimation of the mask matrix $M$ and the output of D. D is trained with a fixed G according to the following criterion:

$$\min_{D} (-L_D(M, \hat{M})). \quad (6)$$

As the next step, the optimization process of G yields an updated D. In fact, G outputs an estimation for the entire time series. This ensures two main points. First, G can produce a genuine imputed values to fool D. Second, it promises that the values estimated for the observed part are close to those actually observed. Therefore, to achieve this, we define two loss functions for the training process of G such as generator loss and imputation loss. Generator loss, $L_G$, is designed to be

$$L_G(M, \hat{M}) = -\sum_{t=1}^{T} \sum_{n=1}^{N} [(1 - m_t^n) \log(\hat{m}_t^n)]. \quad (7)$$

As can be seen from its definition, $L_G$ only considers the missing values ($m_t = 0$). If $L_G$ is kept to a minimum, D will have difficulty to recognize the imputed values. Moreover, minimizing $L_G$ forces G to capture the distribution of the original time series. Here, G is trained to keep the difference between the original time series and the completed time series as small as possible, which also ensures a reasonable completed data. Secondly, we define the imputation loss, $L_{imp}$. It is defined as the difference between the original time series and the completed time series and is given as follows:

$$L_{imp}(X, \hat{X}) = \sum_{t=1}^{T} \sum_{n=1}^{N} [m_t^n(x_t^n - \hat{x}_t^n)] \quad (8)$$

Finally, G is updated according to the following criterion:

$$\min_{G} \left( L_G(M, \hat{M}) + \gamma \ L_{imp}(X, \hat{X}) \right) \quad (9)$$

where $\gamma$ is a hyper-parameter.

Finally, discriminator and generator networks consist of an input layer, the LSTM layer including 150 cells, and an output layer. The training phase carried out through Adam optimization function with 50 epochs. To avoid possible overfitting problems, last hidden state of the LSTM layers is utilized with a 50% dropout rate. We examine the performance of the proposed model for training and test sets. We randomly split the data set into 80% training and 10% validation sets. We then report the classification error on the held-out test set. There is no patient overlap between the sets to avoid possible data leakage.

**D. PREDICTING WITH LSTM NETWORK**

We exploit LSTM not only for imputation network but also for the early prediction network. Early prediction networks rely on several stacked hidden layers to capture a more complex patterns of sequential data, which results in a generalization that is superior to that of the shallow networks. Indeed, there are new and complex advances in time-series prediction. Still, since we used a complex imputation method in the early pre-processing network, we carried on with a primary but efficient prediction network. A detailed stacked LSTM example is given in Fig. 3. It consists of an input layer, four stacked LSTM layers, in which the output of a LSTM hidden layer will be an input for the next LSTM hidden layer, three fully connected layers (FC), and an output layer. The input layer of an LSTM network receives the pre-processed data. Hyperbolic tangent function, given by $S(x) = (\exp(x) - \exp(-x))/\exp(x) + \exp(-x)$, is preferred as an activation function in hidden layers to learn nonlinear mapping between the input features and the output labels. Sigmoid activation function, given by $S(x) = 1/(1 + e^{-x})$, is considered for the output layer. Each LSTM cell includes 200 neurons.

![FIGURE 3. Stacked LSTM structure.](image-url)
when no longer improvement is seen in the validation loss. Patience, the number of the epochs with no improvement, is set to 10. We examine the performance of the proposed model for training and test sets. We split the data set wherein 90% of the data was used for the training and 10% for the testing. We assess our results in terms of estimation accuracy, sensitivity, area under the receiver operating characteristics (AUROC) metrics which are routinely used to evaluate model performance with imbalanced data sets. We assess the quality of predictions in terms of estimation accuracy, sensitivity, area under the receiver operating characteristics (AUROC) metrics. AUROC provides a more accurate performance profile of models for a highly imbalanced data set. Due to the nature of EHR-derived time-series data, our data set is highly skewed to negative examples.

The model is implemented in Python using Keras framework 2.4.0 with Tensorflow 2.4.1 as the backend. The hardware configuration used was an NVIDIA GeForce GTX 980 on an AMD Ryzen 3600 processor with 16 GB of RAM. The implementation of our work will be available at https://github.com/MerveApalak/AdversarialSepsisPrediction.

III. RESULTS AND DISCUSSIONS

A. DATA ANALYSIS

We discover that 26 features have more than 90% of their values missing. Additionally, for 6 features more than 98% of the values are missing. Fig. 4 illustrates the missing value rate for each column of the data set. There are only 9 features with missing values less than 20%, and only 3 features (age, gender, ICULOS, i.e., hours since ICU admit) are with complete data. Considering the data set involves hourly measurements, it can be accepted as highly corrupted with the missing measurements, which is certainly causing a bias. The data set is found to be highly imbalanced in which the number of negative labels outweighs the number of positive labels by a significant margin. Only 2932 of the 40336 patients develop sepsis (7.27%). The data set has a length of more than 1,000,000 h of recordings, and sepsis patients include only 1.8% of the recordings.

We also run some analyses based on patient demographics. The data set includes a total of 40,336 patients; 1521 (6.7%) who develop sepsis in ICU are male patients. Female patients who develop sepsis hold 5.8% of the population. In this case, 59% of population is over 60 years old.

B. EVALUATION OF ADVERSARIAL IMPUTATION MODEL

Even though GANs have achieved state of art results, common over-fitting problem is not detailed and evaluated. We addressed this issue by using conditional GANS, and dropout layer in generator and discriminator network. It is expected that if the generator learn an implicit distribution, it does not overfit to the training data. In order to analyze the feasibility of the adversarial imputation network and address the aforementioned problem, we evaluate the quality of the imputed data with 3 simple tests suggested in [29]. Since the ground truth of the missing data points are unknown, using the original data with all features is not feasible for testing. Therefore, we limit our search on patients having no missing value for HR and the O2Sat features for the evaluation of the adversarial imputation network. We picked these features because they are the least missing 2 features in the data set. We intentionally removed data points randomly to control the “missingness” rate. We apply Principle Component Analysis (PCA) on both the original and the imputed data set in order to perform a qualitative assessment. Figure 5 visualizes the original and the imputed data in a 2-dimensional space for varying missing rates including 80%, 50%, and 10%. Results are illustrating a matching distribution between the original and the imputed database. Furthermore, we depict the reconstruction errors in terms of RMSE in Table 2. We report that the distribution of the reconstruction errors is not significantly different from those of training and test sets in any of these cases, and that the model does not appear to be biased towards reconstructing the training set examples. Figure 6 illustrates the feature importance for after imputation. Since we are aiming to increase the availability of the data set, we do not consider feature elimination except Unit 1 and Unit 2 features. These two features are involving missing data points and not time varying features. Instead of updating our loss function for this issue, we simply exclude Unit 1 and Unit 2 features only.
C. EVALUATION OF PREDICTION MODEL

In this study, the conditional recurrent adversarial network and deep LSTM network are used to automatically detect temporal trends for predicting sepsis patients admitted to the ICU up to 12 hours in advance. Table 2 presents the results of prediction performance based on AUROC, sensitivity, specificity, and accuracy. It is apparent that the highest AUROC, accuracy, and sensitivity of 94.5%, 90.87%, and 98.10% are observed, respectively, at the 4-hour prediction window. For unbiased comparison purposes, sensitivity is set to 85%, which has been fixed to values between 80% and 90% in prior studies. When sensitivity was fixed at 85%, specificity in the test set attains the highest value of 90.79%.

Over longer time windows, we were expecting to see a significant decrease in the model performance; however, AUROC change from 94.5% best performance of 4-hour prediction window to 94.10% at 12-hour prediction window is hardly noticeable.

Table 3 summarizes the state-of-the-art research that has been done on sepsis prediction. For a fair comparison, we include the most recent methods providing the information of the database, prediction time, and the most common performance metrics used in each study. Our method exhibits a distinguishable high-performance over varying prediction windows and it repeatedly outperforms the recent approaches in different aspects. The AISE algorithm [47] was built based on Weibull-Cox proportional hazards model. Using the same sepsis definition and the data set, they reported performance results for 12-hour, 8-hour, and 4-hour in advance detection with AUROC of 83.0%, 84.0%, 85.0%, respectively. In this study, we also studied varying the prediction windows to emphasize the performance of our method at various time frames to discuss potential utility. It is worth mentioning that our method has achieved AUROC of 94.01%, 93.74%, 94.49% for 12-hour, 8-hour, 4-hour ahead prediction. Recently, Shashikumar et al. [52] have enhanced AISE algorithm to Deep AISE algorithm. In addition to the hazards model, the authors implemented the Gated Recurrent Unit (GRUs) to capture the clinical trajectory of patients over time.
Although they have improved prediction performance with AUROC score between 87% and 90%, the results presented here remain superior. Al-Mualemi and Lu [50] have applied the same database for the early prediction of sepsis. While we have also investigated the LSTM-based model similar to [50], but the two papers discuss different strategies. They proposed an adaptive CNN (ACNN) classifier. They also validate their results with LSTM-RNN and SVM classifier. Although ACNN archived an accuracy of 93.84%, accuracy is not an appropriate assessment measure given the class imbalance. AUROC is reported for SVM classifier to be 78%. However, for the purpose of reproducibility and comparison, implementation details of LSTM-RNN are not discussed. The authors have also omitted the demographic features, stating that data including these features may cause generalization problems on model performance. However, we have found that the ICU length of stay and hospital administration time plays an important role in identifying the dynamic nature of sepsis. Compared to our results, in [49] and [51], RNN based deep learning models attained highly poor prediction performances for 3 hours prediction window. In [53], authors exploited the same data set and they try to deal with the missing value and imbalance issues by capturing the temporal pattern and heterogeneous variable interactions. Even though this is the closet approach to our work, our model outperforms it. As the secondary type involving LSTM networks [54], achieved a higher AUROC value of 92.9%, but a direct comparison of their result is not applicable. To elaborate, authors aimed to estimate the first hour of 5 continuous hours of SIRS. Therefore, they must exclude a huge amount of data from the MIMIC-II database. Hence, larger data sets are required to train deep learning models. Fig. 5 shows the ROC plot derived from our study. It is important to state that this study can be followed by a prospective clinical validation, such as in [55], to build confidence in the proposed algorithms.

**IV. CONCLUSION**

In this paper, we presented and studied a novel prediction model with the conditional recurrent adversarial network and deep LSTM for septic patients. Our initial findings reveal that the available data contained a considerable amount of missing values. Thus, to mitigate the negative effect of the missing information on the performance of the prediction model, we proposed a novel preprocessing and early prediction network. Our model not only discovers the missing patterns to improve the prediction results, but also can cooperate with broader detection windows. We concluded that capturing the uncertainty in the time series is specifically important in the medical settings to mitigate the propagation of error for the purpose of prediction. Our proposed method showed superior results, and it was shown to be applicable for any applications involving infrequently recorded health records. To our knowledge, this is the latest study to demonstrate a sepsis prediction algorithm over incrementally longer time windows with a significant performance involving adversarial training. Our data set includes two different hospitals with
different treatment protocols, which results in a more general-ized model. Yet the data set includes only the ICU recordings. For future work, data recorded in a non-ICU environment can be included to explore the performance of the proposed model for different data sets.

REFERENCES

[1] A. Rhodes, L. E. Evans, W. Alhazzani, M. M. Levy, M. Antonelli, R. Ferrer, A. Kumar, J. E. Sevransky, C. L. Sprung, and M. E. Nunnally, “Surviving sepsis campaign: International guidelines for management of sepsis and septic shock: 2016,” *Critical Care Med.*, vol. 44, no. 3, pp. 304–377, 2017.

[2] D. C. Angus, W. T. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, and M. R. Pittsky, “Epidemiology of severe sepsis in the United States: Analysis of incidence, outcome, and associated costs of care,” *Crit. Care Med.*, vol. 29, no. 7, pp. 1303–1310, Jul. 2001.

[3] C. J. Pooli, M. A. Reynolds, M. Sinha, M. Gitlin, and E. Crouser, “Epidemiology and costs of sepsis in the United States—An analysis based on timing of diagnosis and severity level,” *Crit. Care Med.*, vol. 46, no. 12, p. 1889, 2018.

[4] V. X. Liu, V. Fielding-Singh, J. D. Greene, J. M. Baker, T. I. Iwashyna, J. Hoffmann, C. R. Jutzeler, and K. Borgwardt, “Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics,” *J. Electrocardiol.*, vol. 50, no. 6, pp. 739–743, Nov. 2017.

[5] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, “A targeted real-time early warning score (TREWScore) for septic shock,” *Sci. Transl. Med.*, vol. 7, no. 299, Aug. 2015, Art. no. 299ra122.

[6] J.-S. Rachouin, R. Daher, C. Moussallem, B. Milcarek, K. Hunter, C. Schorr, M.-S. Abdou, P. Henry, and L. S. Weisberg, “The fallacy of the bun: Creatinine ratio in critically ill patients,” *Nephrol. Dialysis Transplantation*, vol. 27, no. 6, pp. 2248–2254, 2012.

[7] Z. Q. Mao, M. Jay, J. L. Hoffman, J. Calvert, C. Barton, D. Shimabukuro, L. Shieh, U. Chettipally, G. Fletcher, Y. Kerem, Y. Zhou, and R. Das, “Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU,” *BMJ Open*, vol. 8, no. 1, Jan. 2018, Art. no. e017833.

[8] Z. C. Lipton, D. C. Kale, and R. Wolfel, “Modelling missing data in clinical time series with RNNs,” *Proc. Mach. Learn. Healthcare*, vol. 56, pp. 253–270, Jun. 2016.

[9] M. Berglund, T. Raiko, M. Honkala, L. Karkkainen, A. Vetek, and J. T. Karhunen, “Bidirectional recurrent neural networks as generative models,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–15.

[10] J. Yoon, W. R. Zame, and M. van der Schaar, “Estimating missing data in temporal data streams using multi-directional recurrent neural networks,” *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1477–1490, May 2019.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–13.

[12] S. Hochreiter, “The vanishing gradient problem during learning recurrent neural nets and its problem solution,” Int. J. Uncertainty, Fuzziness Knowl.-Based Syst., vol. 6, no. 2, pp. 107–116, 2003.

[13] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] R. Jozefowicz, W. Zaremba, and I. Sutskever, “An empirical exploration of recurrent network architectures,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2342–2350.

[15] J. Yoon, D. Jarrett, and M. V. D. Schaar, “Time-series generative adversarial networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–9.

[16] H. Harutyunyan, H. Khachatryan, D. C. Kale, G. Ver Steeg, and A. Galstyan, “Multitask learning and benchmarking with clinical time series data,” *Sci. Data*, vol. 6, no. 1, pp. 1–18, Jun. 2019.

[17] Z. He, X. Chen, Z. Fang, W. Yi, C. Wang, L. Jiang, Z. Tong, Z. Bai, Y. Li, and Y. Pan, “Early sepsis prediction using ensemble learning with features extracted from LSTM recurrent neural network,” in *Proc. Comput. Cardiol. Conf. (CINC)*, Dec. 2019, p. 1.

[18] A. Rafiei, A. Rezaee, F. Haajati, S. Gheisari, and M. Golzan, “SSP: Early sepsis detection with generative adversarial networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, Art. no. e017833.

[19] O. Mogren, “C-RNN-GAN: Continuous recurrent neural networks with generative adversarial networks,” in *Proc. Int. Conf. Mach. Learn.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] M. Berglund, T. Raiko, M. Honkala, L. Karkkainen, A. Vetek, and J. T. Karhunen, “Bidirectional recurrent neural networks as generative models,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–15.

[21] Z. C. Lipton, D. C. Kale, and R. Wolfel, “Modelling missing data in clinical time series with RNNs,” *Proc. Mach. Learn. Healthcare*, vol. 56, pp. 253–270, Jun. 2016.

[22] S. Hochreiter, “The vanishing gradient problem during learning recurrent neural nets and its problem solution,” Int. J. Uncertainty, Fuzziness Knowl.-Based Syst., vol. 6, no. 2, pp. 107–116, 2003.

[23] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] R. Jozefowicz, W. Zaremba, and I. Sutskever, “An empirical exploration of recurrent network architectures,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2342–2350.

[25] J. Yoon, D. Jarrett, and M. V. D. Schaar, “Time-series generative adversarial networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–9.

[26] H. Harutyunyan, H. Khachatryan, D. C. Kale, G. Ver Steeg, and A. Galstyan, “Multitask learning and benchmarking with clinical time series data,” *Sci. Data*, vol. 6, no. 1, pp. 1–18, Jun. 2019.

[27] Z. He, X. Chen, Z. Fang, W. Yi, C. Wang, L. Jiang, Z. Tong, Z. Bai, Y. Li, and Y. Pan, “Early sepsis prediction using ensemble learning with features extracted from LSTM recurrent neural network,” in *Proc. Comput. Cardiol. Conf. (CINC)*, Dec. 2019, p. 1.

[28] A. Rafiei, A. Rezaee, F. Haajati, S. Gheisari, and M. Golzan, “SSP: Early prediction of sepsis using fully connected LSTM-CNN model,” *Comput. Biol. Med.*, vol. 128, Jan. 2021, Art. no. 104110.

[29] O. Mogren, “C-RNN-GAN: Continuous recurrent neural networks with adversarial training,” 2016, arXiv:1611.09904.

[30] G. Ramponi, P. Protopapas, M. Brambilla, and R. Janssen, “T-CGAN: Conditional generative adversarial network for data augmentation in noisy time series with irregular sampling,” 2018, arXiv:1811.08295.

[31] J. Yoon, J. Jordan, and M. Schara, “Gain: Missing data imputation using generative adversarial nets,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5689–5698.

[32] S. Yang, M. Dong, Y. Wang, and C. Xu, “Adversarial recurrent time series imputation,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 4, 2020, doi: 10.1109/TNNLS.2020.3010524.

[33] S. X.-C. Li, B. Jiang, and B. Marlin, “MisGAN: Learning from incomplete data with generative adversarial networks,” 2019, arXiv:1902.09599.

[34] Y. Luo, X. Cai, Y. Zhang, and J. Xu, “Multivariate time series imputation with generative adversarial networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–12.

[35] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” 2014, arXiv:1411.1784.
C. Esteban, S. L. Hyland, and G. Rätsch, “Real-valued (medical) time series generation with recurrent conditional GANs,” 2017, arXiv:1706.02633.

M. Reyna, S. P. Shashikumar, B. Moody, P. Gu, A. Sharma, S. Nemati, and G. Clifford, “Early prediction of sepsis from clinical data: The PhysioNet computing in cardiology challenge 2019,” in Proc. Comput. Cardiol. Conf. (CinC), Dec. 2019, p. 1.

W. C. Seymour, X. V. Liu, J. T. Iwashyna, M. F. Brunkhorst, D. T. Rea, A. Scherag, G. Rubenfeld, M. J. Kahn, M. Shankar-Hari, M. Singer, S. C. Deutschman, J. G. Escobar, and C. D. Angus, “Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (sepsis-3),” JAMA, vol. 315, no. 8, pp. 762–774, 2016.

Z. Wang and B. Yao, “Multi-branching temporal convolutional network for multivariate time series with missing values,” Sci. Rep., vol. 8, no. 1, pp. 1–12, Apr. 2018.

E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, “Doctor AI: Predicting clinical events via recurrent neural networks,” in Proc. Mach. Learn. Healthcare Conf., 2016, pp. 301–318.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

R. Caruana, S. Lawrence, and C. Giles, “Overfitting in neural nets: Back-propagation, conjugate gradient, and early stopping,” in Proc. Adv. Neural Inf. Process. Syst., vol. 13, 2000, pp. 1–15.

S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. DClifford, and T. G. Buchman, “An interpretable machine learning model for accurate prediction of sepsis in the ICU,” Crit. Care Med., vol. 46, no. 4, p. 547, 2018.

M. Moor, M. Horn, B. Rieck, D. Roqueiro, and K. Borgwardt, “Early recognition of sepsis with Gaussian process temporal convolutional networks and dynamic time warping,” in Proc. Mach. Learn. Healthcare Conf., 2019, pp. 2–26.

M. Scherf, F. Grüller, H. Malberg, and S. Zaunseder, “Predicting sepsis with a recurrent neural network using the MIMIC III database,” Comput. Biol. Med., vol. 113, Oct. 2019, Art. no. 103395.

B. Y. Al-Muallemi and L. Lu, “A deep learning-based sepsis estimation scheme,” IEEE Access, vol. 9, pp. 5442–5452, 2021.

S. M. Lauritsen, M. E. Kalor, E. L. Kongsgaard, K. M. Lauritsen, M. J. Jørgensen, J. Lange, and B. Thießen, “Early detection of sepsis utilizing deep learning on electronic health record event sequences,” Artif. Intell. Med., vol. 104, Apr. 2020, Art. no. 101820.

S. P. Shashikumar, C. S. Josef, A. Sharma, and S. Nemati, “DeepAISE—An interpretable and recurrent neural survival model for early prediction of sepsis,” Artif. Intell. Med., vol. 113, Mar. 2021, Art. no. 102036.

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.

H. J. Kam and H. Y. Kim, “Learning representations for the early detection of sepsis with deep neural networks,” Comput. Biol. Med., vol. 89, pp. 248–255, Oct. 2017.

L. F. Roggeveen, L. M. Fleuren, T. Guo, P. Thoral, H. J. de Grooth, E. L. Swart, T. L. T. Klausch, P. H. J. van der Voort, A. R. J. Girbes, R. J. Bosman, and P. W. G. Elbers, “Right dose right now: Bedside data-driven personalized antibiotic dosing in severe sepsis and septic shock—Rationale and design of a multicenter randomized controlled superiority trial,” Trials, vol. 20, no. 1, pp. 1–13, Dec. 2019.