IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 September 2022, accepted 19 September 2022, date of publication 3 October 2022, date of current version 11 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3211334

RESEARCH ARTICLE

A Multi-Headed Transformer Approach for Predicting the Patient's Clinical Time-Series Variables From Charted Vital Signs

GASPARD HARERIMANA[1], (Member, IEEE), JONG WOOK KIM[2], (Member, IEEE), AND BEAKCHEOL JANG[3], (Member, IEEE)

[1]Department of Information Technology, Carnegie Mellon University, Kigali 6150, Rwanda
[2]Department of Computer Science, Sangmyung University, Seoul 03016, South Korea
[3]Graduate School of Information, Yonsei University, Seoul 03722, South Korea

Corresponding authors: Beakcheol Jang (bjang@yonsei.ac.kr) and Jong Wook Kim (jkim@smu.ac.kr)

This work was supported by the National Research Foundation of Korea Fund under Grant NRF-2022R1F1A1063961.

ABSTRACT Deep learning has progressively been the spotlight of innovations that aim to leverage the clinical time-series data that are longitudinally recorded in the Electronic Health Records(EHR) to forecast the patient's survival and vital signs deterioration. However, their recording velocity, as well as their noisiness, hinder the proper adoption of the recently proposed benchmarks. The Recurrent Neural Networks (RNN) especially the Long-short Term Memory (LSTMs) have achieved better results in recent studies but they are hard to train and interpret and fail to properly capture the long-term dependencies. Moreover, the RNNs suffer greatly with clinical time series due to their sequential processing which cripples the prospect of parallel processing. Recently the Transformer approach was proposed for Natural Language Processing (NLP) tasks and achieved state-of-the-art results. Hence to tackle the drawbacks that are suffered by the RNNs we propose a clinical time series Multi-head Transformer (MHT), which is a transformer-based model that forecasts the patient's future time series variables using the vitals signs. To prove the generalization of the model we use the same model for other critical tasks that describe the Intensive Care Unit (ICU) patient's progression and the associated risks like the remaining Length Of Stay(LoS), the In-hospital Mortality as well as the 24 hours mortality. Our model achieves an Area Under The Curve-Receiver Operating Characteristics(AUC-ROC) of 0.98 and an Area Under the Curve, Precision-Recall (AUC-PR) of 0.424 for vital time series prediction, and an AUC-ROC of 0.875 in the mortality prediction. The model performs well for the frequently recorded variables like the Heart Rate (HR) and performs barely like the LSTM counterparts for the intermittently captured records such as the White Blood Count (WBC).

INDEX TERMS Multi head transformer, clinical time series, natural language processing, self-attention, encoder-decoder attention, interpolation.

I. INTRODUCTION

Clinicians are challenged and overwhelmed with the problem of inferring clinical outcomes from the longitudinal Electronic Health Records (EHR) data. The critical inputs include the most structured ontologies in the form of International Classification of Diseases(ICD) codes which can be leveraged to predict the future of patients including adverse events like mortality, decompensation, and Length of Stay(LoS). With the increasing hospital machinery and the introduction of pervasive medical IoT devices, the future is more dependent on multivariate time series clinical data whose most critical factor is velocity. The ever-dynamic changes of single variables or multivariate variables in the course of admission due to pathophysiological disturbances and therapeutic interventions provide an extensive opportunity for clinical time series analysis. Various challenges like the curse of irregularly recorded data as well as different units and ranges

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

of measurements for time series variables hinder the proper application of learning algorithms to produce actionable clinical insights. The EHR contains various longitudinal records that are recorded at different frequencies and intervals like vital signs which are recorded continuously as well as several records that are measured intermittently. For instance, in various EHR datasets, the Heart Rate(HR) is measured using the Electrocardiogram(ECG) as a continuous wave while the Glucose level is measured occasionally hence the sampling at uniform intervals will result in missing values in the EHR. Various prediction approaches have been proposed including classic systems such as the Simplified Acute Physiology Score (SAPS-II) [1] which uses physiological records to predict mortality and the Acute Physiology and Chronic Health Evaluation (APACHE) [2] which uses statistical equations to predict the patient mortality within 24 hours of admission using the time series physiological measurements. Various studies improved the results by incorporating wearable Internet of Things(IoT) devices [3], [4] and recently key studies used deep learning methods to predict the future risks from clinical time series [5], [6].

The Transformer network [7] is a deep learning method that was proposed to counter the drawbacks suffered by the recurrence-based deep learning models(RNNs) and is built solely on a series of attention mechanisms arranged in an encoder-decoder pattern. The transformer was created for NLP applications and was recently applied in various tasks including Machine Translation [8], Speech recognition [9], Question Answering [10], and Named Entity Recognition [11]. Few recent studies applied only the transformer's components especially the self-attention process for time series analysis [12], [13]. One particular study [14] used the self-attention component of the transformer for clinical time series-based prediction, and to our knowledge, None of the related prior works have ever tried to incorporate the decoder component of the transformer for clinical time series-based predictions.

This study applies the transformer's entire capabilities for predicting the patient's future values of the key clinical time series variables(Vital signs variables). We develop a Multi-Headed Transformer(MHT) model that leverages the Transformer's main building blocks including the self-attention process and the encoder-decoder attention mechanism. By using the same model with little modification to predict other clinical risk factors like the 24hrs mortality, the In-hospital mortality, and the remaining Length-Of-stay(LoS) we demonstrate that the model can generalize and be used for other HER-based benchmark tasks. Our contributions are summarized as follows:

- As depicted in Fig.1 we build a Multi-headed Transformer(MHT) model that uses the multi-variate vital signs records to predict future values of the key time series variables that characterize the Intensive Care Unit (ICU) patient's deterioration.
- We fully investigate the inclusion of the decoder component and the associated encoder-decoder attention. Furthermore, we compare the performance with the other systems that use only the encoder's self-attention mechanism like the one proposed in [14].
- Also in contrast with [15] and [14], rather than modeling the clinical time-series monitoring as a multi-class decompensation problem, we predict future values of the time series variables in a regression mode.
  The major advantage of the Multi-Headed Transformer to the HER-based predictions is the capability for the model to focus on the most important admission in the patient's history as well as the most critical diagnosis in that admission.

To assess the generalization of our approach, we apply the same model with little modification in the output layer to predict the 24hrs mortality and the In-hospital mortality as binary classification problems and the LoS as a regression task. By doing so we draw the inference that our approach can generalize and be used for other vital signs-based prediction tasks like phenotyping and patient-based cohort selection.

In this study, we use the MIMICIII EHR [16] a publicly available critical care database that integrates de-identified, comprehensive clinical data of patients with a total of 53,423 admissions at the ICU of the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts during the period from 2001 to 2012. Our model achieves an Area Under The Curve-Receiver Operating Characteristics (AUC-ROC) of 0.98 and an Area Under the Curve, Precision-Recall(AUC-PR) of 0.424 for vital time series prediction, and an AUC-ROC of 0.875 in the mortality prediction outperforming all the other approaches. Especially the MHT model outperforms Simply Attend and Diagnose(SaND) [14] which is uses self-attention and omits the decoder component. TimeNet [17] which is the most powerful baseline achieves satisfactory results due to the use of pre-trained features. Nevertheless, the generalization of such transfer learning-based models is questionable.

The remainder of this paper is arranged as follows: in section II, we cover the key related works and background knowledge, and in section III we discuss the structure of the data used in the paper including the Preprocessing and interpolation processes and in IV we cover the method used in our approach. In V we perform the experiments,in VI we present our results as well as performance evaluation, and in VI we conclude.

II. RELATED WORKS AND BACKGROUND KNOWLEDGE

In this section, we introduce the related works and we introduce the key preliminary concepts including the Transformer and the 1D Convolutional Network (1D-CNN) used for embedding and feature extraction.

A. RELATED WORKS

Forecasting the ICU-based adverse events using deep learning methods with temporal clinical data stored in large EHR has recently received considerable research attention.

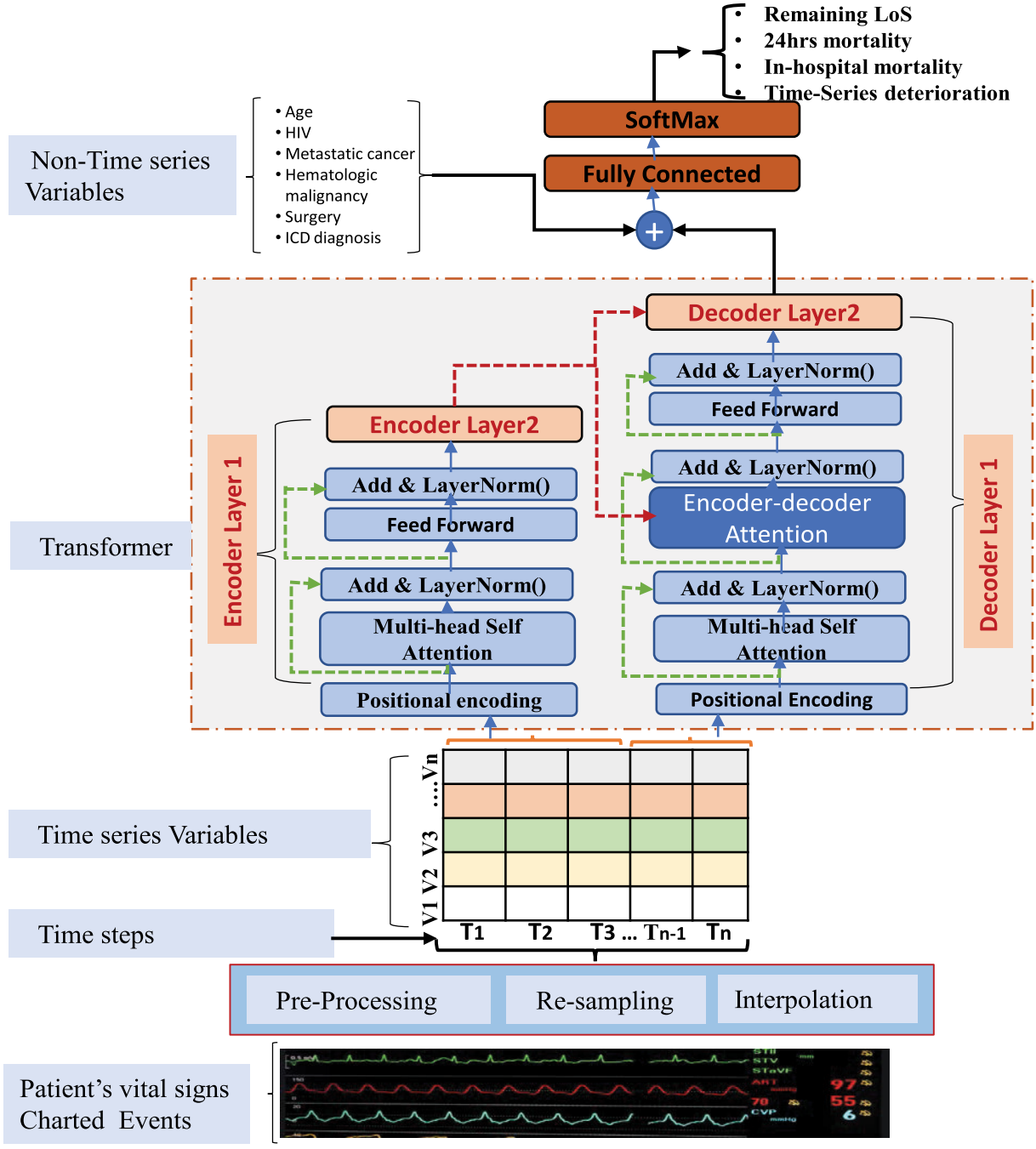


FIGURE 1. The architecture of the proposed MHT model for clinical time series prediction. The model leverages the key components of the transformer architecture. The Encoder's self-attention encodes the time steps across the variables by capturing the latent relationship between variables and between the time steps. The encoder-decoder attention uses these representations to predict future values within a specified lag time. The fully connected layer is structured as per the prediction task.

Lipton et.al [18] is the first study to empirically apply the LSTMs to extract clinical insights and hidden patterns from multivariate temporal clinical measurements. The study covered the prediction of ICD codes [19] grouped under 128 Clinical Classifications Software(CCS) [20] categories from 10, 401 PICU (Pediatric Intensive Care Unit) episodes. Each PICU episode comprised of a multivariate time series of 13 variables including diastolic and systolic blood pressure, end-tidal CO2, etc. The LSTM-based predictive model used sequential target replication and performed well beating the compared approaches.

Razavian et al. [21] proposed a method that applied a Temporal Convolutional Networks (TCN) [22] model to ICU lab tests for the detection of multiple diseases from irregularly measured sparse lab values. The TCN is popularly known to mitigate the drawbacks caused by noisy sparse data, hence the model achieved better results. Song et al. [14] proposed SAnD a multi-task model for multiple clinical predictive

tasks using the MIMIC-III EHR. The study improved the temporality of the measurements and included temporal order into the data representation using positional encoding [23] and a dense interpolation approach. The novelty of the model was to evade the use of the RNNs or CNNs for sequence modelling as it is the case for all popular approaches. The study achieved overall better performances but did not elaborate on how the issue of classification imbalance was dealt with especially for mortality prediction. Moreover, the study did not reveal many important hyper-parameters like the convolution kernel size used in a 1D convolution step to get the multi-dimensional representation of each time step. Also, the important concepts like the masking process used to keep only temporary data of interest are not discussed enough.

Harutyunyan *et al.* [15] proposed a multitask learning model that uses a Channel-wise LSTM followed by deep supervision [24] using MIMICIII EHR for four benchmark tasks; Phenotype classification, LoS prediction, decompensation prediction, and In-hospital mortality prediction using clinical time-series data. The model tries to train a single network for multiple tasks simultaneously by capturing latent features that are generic across these different tasks. However, as in the previous task, some of the classification benchmarks like mortality classification contain unbalanced classes but the model did not tackle this important issue. In this study, we perform a performance comparison to assess the superiority of our proposed approach over this model. Shamout et.al [25] proposed a Deep Early Warning System (DEWS), a system that uses a Bidirectional LSTM(BiLSTM) with an attention mechanism to classify whether an observation is within 24 hours of an adverse outcome (cardiac arrest, mortality or unplanned ICU admission).

Yu et al. [5] built a multi-task learning RNN model with attention to predicting hospital mortality, using the reconstruction of physiological time series as an auxiliary task. The model was compared with the standard SAPS-II [1] and achieved the best sensitivity. Gupta et al. [17] proposed a novel approach that leverages TimeNet [26] a transformer [7] based pre-trained deep learning model for phenotypes classification and in-hospital mortality tasks from multivariate clinical time series. The model uses generic features for clinical time series using an RNN pre-trained on diverse time series across different domains, hence making the model more robust and more efficient than other previous approaches. Lin et al. [27] proposed a Hierarchical Attention-based Temporal Convolutional Networks for the prediction of Myotonic Dystrophy Diagnosis. The hierarchical model outperformed other machine learning models and the more advanced Temporal Convolutional networks. However, the study cannot generalize to other EHR predictive tasks.

B. BACKGROUND KNOWLEDGE

1) 1D CONVOLUTION FOR TIME SERIES

the 2D CNNs which are usually used in computer vision use filters (kernel) of variable sizes to a 2D image for feature learning. The features are extracted from the image's pixels and colour channels. Various applications of the 2D CNN for NLP and sentiment classification have been proposed [28]. The 1D convolution is a CNN variant that uses a 1D convolution and has been used in various applications that involve time series data as well as NLP tasks [29]. In this work, we apply a 1D convolution across the time steps for all the variables to obtain a distributed representation of the time steps. The main advantage of 1D CNN is that it can allow the use of a bigger kernel size to extract rich features.

2) THE TRANSFORMER

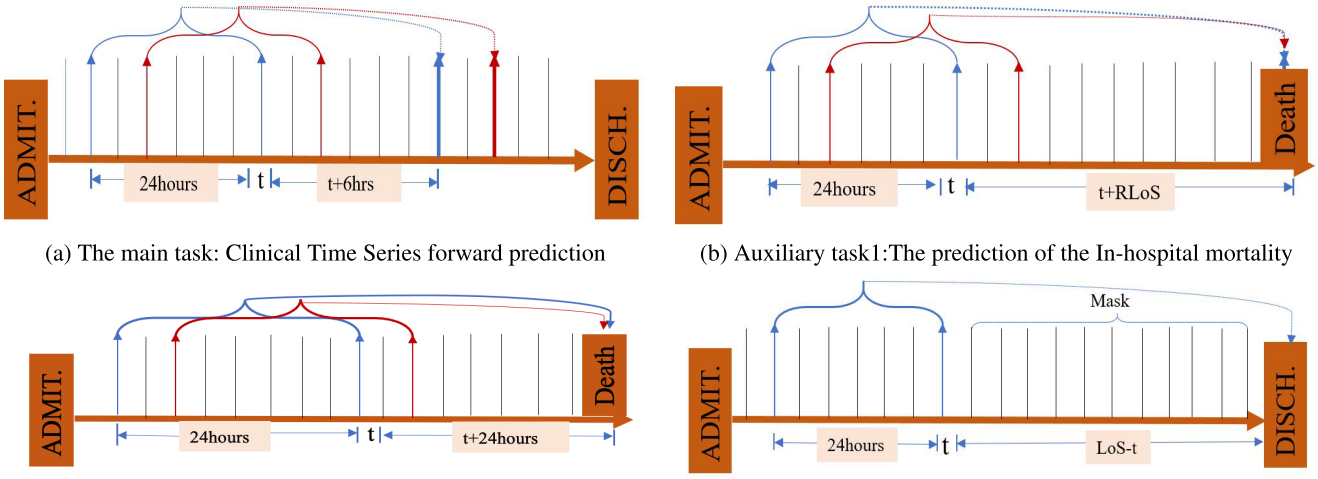
The transformer [7] is made of a series of attention-based encoder-decoder networks. It has performed excellently for NLP applications, especially for Neural Machine Translation and Named Entity Recognition (NER). For such applications, It has particularly scored the best results against the LSTMs and other sequence-based recurrence models. Only a handful of time series applications have been proposed with some of them opting to use its components than using the whole transformer network. As in language translation, the transformer is used to predict the next item given the previous items(words in a sentence and time steps in a time series). The encoder uses an attention layer that helps to encode a certain time step in a form where its relationship with other time steps is reflected. The decoder uses the same components but adds an encoder-decoder attention layer that focuses on relevant past time steps to predict the values in the future time-steps [30].

III. EXPLORATORY DATA ANALYSIS

In this section, we describe the dataset used, the prediction scenarios, the data pre-processing as well as the interpolation processes.

A. DATA DESCRIPTION

The patient's admission is characterized by various constants including demographic records (Age, Marital status,...), clinical structured ontologies (Ex: ICD codes for diagnosis) and the time series charted events that are fetched from various measuring equipment such as the Electrocardiogram, Pulsemetric devices, other medical IoT wearable devices as well as lab results (also recorded in a coded ontology). Table 1 describes the variables used in this study. We included the variables that are used in SAPS II as well as additional variables and constants that are critical to the patient's outcome. The time series variables are recorded in a continuous pattern (Ex: HR) or intermittent patterns such as White Blood Count (WBC). From the table, we observe that many time series variables are routine vital signs while others are indicators that may be recorded less frequently. Non-time series variables comprise the usual demographic constants as well as other physiological records that may determine how the time series variables affect the prediction. For instance, the HIV status during admission drives the effect of the WBC on the mortality prediction. The MIMICIII ID is a local ID used to identify various variables. We use a total of 30 variables including 24 longitudinal variables, 4 static physiological



(a) The main task: Clinical Time Series forward prediction

(b) Auxiliary task1:The prediction of the In-hospital mortality

(c) Auxiliary task2:The prediction of the forward 24 hours mortality

(d) Auxiliary task3, Prediction of the Remaining LoS(RLoS)

FIGURE 2. The benchmark prediction tasks for the MHT model. As per (a) the model is built for the prediction of the forward values for the key clinical time-series variables. In (b)(c)(d) the final layer of the model is tuned to perform the auxiliary tasks.

variables (Ex: HIV status), and 2 demographic variables (age and type of admission). The recording time of several variables and the length of records are not synchronized, hence creating a dataset that has many missing values.

B. PREDICTION SCENARIOS

The prediction scenarios are used to generalize on critical clinical tasks as depicted in Fig.2. Furthermore, for an accurate comparative evaluation of the model we use some of the prediction scenarios described in [15] and in [14]as follows:

1) CLINICAL TIME SERIES FORECASTING

This is the main prediction task for the model and its is displayed in Fig. 2a. It involves a regression prediction problem that predicts the time series variables at different time windows. At each precise time step t in the admission period the model forecasts the values for the same variables in the forward time of h hours using the records of p past hours. In our implementation, we use the records of the last 24 hours to predict the values at the forward 6 hours. In contrast with [15], we consider both the patients whose conditions are deteriorating as well as patients whose conditions are stable because those cases can add more insights to the model. Since some variables occur in a handful of admissions and at sporadic intervals, the prediction dataset will contain missing values.

2) IN-HOSPITAL MORTALITY PREDICTION

As depicted in Fig.2b The in-hospital mortality prediction task aims at predicting whether the patient will die during the current admission. This is a key prediction task which is the main task performed by the SAPS-II scoring mechanism. It is modelled as a binary classification problem. The fraction of patients who die during admission is about 10% in MIMIC-III resulting in a class imbalance. To address this problem we apply the Synthetic Minority Oversampling Technique (SMOTE) balancing technique [31], [32]. Also for this task, we use the records of the last 24 hours to predict future mortality.

3) THE 24 HRS MORTALITY PREDICTION

This is another version of mortality prediction also modelled as a binary classification as shown in Fig. 2c. At any particular time step, we wish to predict if the mortality is within the future 24 hours period. This prediction is important for fast deteriorating patients and can help draw the necessary care by the clinicians. For this task, we consider only the patients whose outcome is mortality. Also for this task, we use the records of the last 24 hours.

4) LoS PREDICTION

The remaining LoS shown in Fig.2d is different from the overall LoS. At each time step, we aim to predict the remaining number of days until discharge. This is a regression task that tries to predict the exact number of days than grouping the LoS under the bins of fixed intervals. Also for this task, we use the records of the last 24 hours.

C. DATA PREPROCESSING

We extracted a total of 330,712,483 records associated with 12,487 charted time series variables (Charted events items) from the Charted events table. We retained only the records that are associated with the 24 time-series variables described in Table 1. For the non-time series variables, we built a separate dataset from the admission and patient tables to extract the ICD9 codes associated with these variables. For each patient, we considered every admission independently and obtained a total of 58976 admissions. For each benchmark task, a separate dataset was built. For LoS we removed the admissions that led to in-hospital mortality and we remained with 43029 admissions. With 47796 admissions for the

mortality dataset, only 10% resulted in mortality outcomes hence the reason to use SMOTE for balancing the data.

1) RE-SAMPLING

The clinical time series are recorded in regular and irregular fashions. As mentioned earlier some variables are routine records while others are intermittently recorded. As we wish to use the records observed in the last 24 hours, certain variables will have over 1000 records (Ex: HR) while others will not have any record observed during that period such as WBC. The re-sampling process involves the transformation of the original data into discrete intervals. The re-sampling process reduces the sheer size of the frequently recorded variables and increases the size of the rarely observed variables hence producing a training dataset with many missing instances. In this implementation, we use a time step of 30 minutes partitioning an admission day into 48 observations for all variables.

2) INTERPOLATION STRATEGY

From Fig.3a and Fig.3b we observe that the re-sampling produces a dataset that has many missing values(56%). The interpolation and gap-filling process helps in filling the missing values while avoiding bias and preserving the data. We used Linear interpolation method [33] and with this method, we filled the gaps by persisting the recently recorded value for less frequently recorded values and use the median of the values recorded in the current time-step (30 mins) for high-frequency variables.

IV. METHOD

In this section, we describe the proposed transformer model for clinical time-series data by basing our intuitions on the original NLP-based transformer components described in the original study [7]. We formulate the task of clinical time series forecasting as a multivariate multi-step prediction problem given that the target variables to be predicted depend on one or maybe more input variables and other target variables. The overall architecture of the proposed system is depicted in Fig.1. Let there be V unique variables described in Fig.1 and Table1 for each patient in the dataset. The pre-processing stage outputs a 3D dataset with one dimension comprised of input variables, another dimension comprising of patients and the last dimension to keep track of the time steps from the start of individual variable's recording up to discharge or the In-hospital mortality. Each patient p is associated with a set of static demographics (i.e Age) and clinical variables(static per a given admission) like the history of metastatic cancer and the reason of current admission, as well as time-dependent inputs $V_{pt} \in R^{m_v}$ with $R^{m_v}$the size of the vector of all valliables m for patient v. The prediction targets variables include the input variables shifted for 6 hours as well as scalar output targets $y_{p,t} \in R$ at each time-step $t \in [0, Ti]$. The output targets are formulated as per the prediction needs as described earlier. The transformer uses the encoder to build a representation of the input vector and the decoder to construct the output values of the time-series. To achieve this the encoder and the decoder use the following steps:

3) TIME SERIES EMBEDDING

We treat each time step in a patient's admission as a word in a sequence of text. At each time step t is characterized by $x_t = v_{1t}, v_{2t}, v_{3t}, v_{4t} \ldots, v_{nt}$ with n representing the number of time series variables. However, some of the clinical time series values lack boundaries in their magnitude, hence representing each time step with time series measurements with varying scales will result in erroneous results. The biggest issue in the interpolated data is that the admission lengths are not uniform for all patients in the dataset. Also, not all patients have data for all the 30 variables, and the imputation strategy does not fully address these issues. Hence we apply a 1D CNN to extract homogeneous features. The advantage of 1D CNNs against other time series techniques is that it can learn the proper internal representations from the raw time series data directly without manually engineering the input features. Given a time series recording $y(t)$ (Ex: HR) for the course of the patient's admission we convolve several 1D filters to the input. The convolution is followed by a pooling layer and a *tanh* activation function. Finally, we extract the output of the activation function as a dense representation of the time series.

$$y_{conv} = Convolve(y(t), Weights) \tag{1}$$

$$y_{pool} = pool(y_{conv}) \tag{2}$$

$$\vec{y}_t = tanh(y_{pool}) \tag{3}$$

$\vec{y}_t$ is the resultant dense representation of a time series variable.The final representation of a patient admission is obtained by concatenating the individual $\vec{y}_t$ for the 24 time series variables represented in Table 1. The positive attribute for the obtained dense vector representing a patient is that the vectors have the same size $S \in R^{E \times n}$ for all patients irrespective of the admission size.

4) POSITIONAL ENCODING

After the embedding process, we obtain the dense vectors representing each time step and apply the positional encoding to represent the values of input variables at various time steps in a way that their positional information in the patient's admission journey is reflected. The original transformer study [7] proposed a cos/sin based intuition to encode the time steps. The intuition is to use sine and cosine waves at different frequencies to encode the position of a time-step vector. The position is translated from an integer position value and expressed by a vector of the same size as the embedding size. For a given time step at position p, of the patient's admission timeline, the position encoding vector $\vec{x}_t$ is given by:

$$\vec{x_p}^{(i)} = f(p)^{(i)} := \begin{cases} \sin(\omega_k . p), & \text{if } i = 2k \\ \cos(\omega_k . p), & \text{if } i = 2k+1 \end{cases} \tag{4}$$

with $\omega_k = \frac{1}{10000^{2k/p}}$ and k represents individual element in the position vector. with $\vec{x}_t \in \mathbb{R}^d$ and d is the positional encoding

TABLE 1. Description of physiological variables(time series and non-time series) and key demographic variables used for prediction.

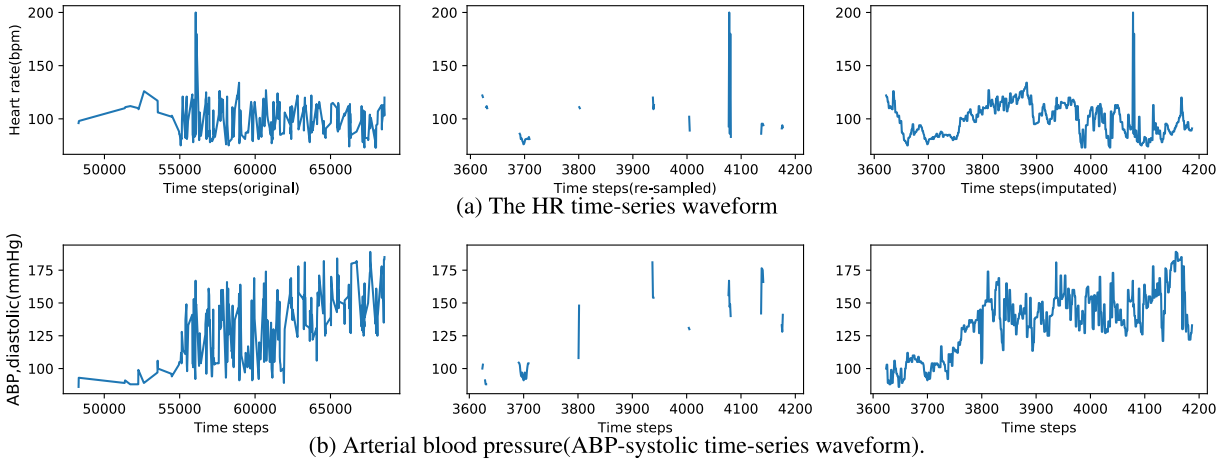
Variable (units)	Description	MIMIC ID	Parameter Value
Time Series physiological variables			
HR(bpm)	Routine vital signs	220045	Numeric
Arterial Blood Pressure systolic (mmHg)	Routine vital signs	220050	Numeric
Arterial Blood Pressure diastolic (mmHg)	Routine vital signs	220051	Numeric
Temperature($^{\circ}C$)	Routine vital signs	223762	Numeric
Glascow coma scale eye opening	Neurological	220739	Text
Glascow coma scale motor response	Neurological	223901	Text
Glascow coma scale verbal response	Neurological	223900	Text
Mechanical ventilation or CPAP	Respiratory	227583	Text
PaO2(mmHg)	Partial Pressure Of Oxygen(Respiratory)	490	Numeric
FiO2(%)	Fraction of inspired oxygen(Respiratory)	223835	Numeric
Urine output ApacheIV	Scores APACHE IV	227519	Numeric
BUN	Blood Urea Nitrogen(Labs)	225624	Numeric
Sodium	Labs(Sodium ,whole blood)	226534	Numeric
Potassium	Labs(Potassium ,whole blood)	44711	Numeric
Bicarbonate(mL)	total amount of carbon dioxide (CO2) in the blood	46362	Numeric
Bilirubin_ApacheIV	ScoresAPACHE IV (2)	226998	Numeric
WBC	Labs	220546	Numeric
RBC	red blood cell count(Hematology)	833	Numeric
Capillary refill R	Cardiovascular (Pulses)	223951	Text
Glascow coma scale total	Neurological	198	Score
PAR-Oxygen saturation	Routine Vital Signs	228232	Text
Respiratory rate(insp/min)	Respiratory	220210	Numeric
PH (Arterial)	Body PH for acidity status	223830	Numeric
Blood Glucose	Blood Glucose	3744	Numeric
Non-Time Series physiological and demographic variables			
Age	Age at the time of record	-	Numeric
HIV	presence of HIV ICD code on Admission	ICD9(042)	binary,1/0
Metastatic cancer	presence of metastatic cancer on or during admission	ICD9(140-239)	binary,1/0
Type of admission	Elective or emergency admission	-	binary,1/0
Intubated	presence of ICD code for Continuous mechanical ventilation of unspecified duration	ICD9(96.70)	binary,1/0
Hematologic Malignant Neoplasm	Presence of ICD9 codes for Hematologic Neoplasm in previous admissions	ICD9(286.7-286.4)	binary,1/0

dimension. In other words $f(p)^{(i)}$ takes an integer position p value and generate a vector position $\vec{x_t}$. Finally the resulting representation of a given time-step will be given by adding the dense embedding vector and the positional vector;

$$\vec{Z_t} = \vec{y}_t + \vec{x_p} \tag{5}$$

5) MULTI-HEAD SELF ATTENTION

During the patient's hospitalization, we want our encoding to discover and associate certain related events. For example, an HR event recorded at a certain time step may be associated with a Transient Ischemic Attack (TIA) event recorded within another time step. Luckily using the transformer language



(a) The HR time-series waveform

(b) Arterial blood pressure(ABP-systolic time-series waveform).

FIGURE 3. **The wave forms that show the original,the re-sampled and the interpolated version for a typical patient's HR and the Systolic Arterial Blood Pressure.**

model we make use of the transformer's self-attention module that can associate the TIA event with an HR event during the encoding of the time step that contains the TIA record. To learn long-term relationships across various time steps, we use self-attention. At each time step t We train 3 matrices and multiply each with the vector $\vec{Z}_t$ of t time-step to generate a Query vector $Q \in \mathbb{R}^{N\times d_{att}}$ vector, a Key vector $K \in \mathbb{R}^{N\times d_{att}}$, and a Value vector $V \in \mathbb{R}^{N\times dv}$. The dimensions of Q, K, and V are fixed (usually at 64). Given the keys and input vector obtained from the positional encoding we get:

$$Z = Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_{att}}}).V \qquad (6)$$

where $Z \in \mathbb{R}^{N\times d_{att}}$ and d_{att} is the dimension of the key vectors. The Softmax is a normalization function that will express the importance of each of the preceding admission time steps while we are encoding a given time step. Unlike in the original transformer paper where the words can have an arbitrary interdependence irrespective of their positions in the sentence, for clinical time series we choose the preceding measurements because we believe that future records cannot affect the records that preceded them. Again unlike the NLP applications where the words appear together more frequently and their hidden relationship can be inferred easily using the described self-attention process, there is a little latent relationship between various time steps in a patient timeline. Hence to improve the representations we use 8 self-attention heads in parallel and concatenate their results afterward. Hence,

$$MultiHeadAtt(Q, K, V) = [Head_1, \ldots, Head_H]WH \qquad (7)$$

$$Head_h = Attention(QW_Q^h, KW_K^h, VW_V^h) \qquad (8)$$

With W_Q^h,W_K^h,W_V^h the weights matrix of head H_h for Q,K,V respectively. Also W_Q^h,$W_K^h \in \mathbb{R}^{d_m\times d_{att}}$ and $W_V^h \in \mathbb{R}^{d_m\times d_v}$ and $W_H^h \in \mathbb{R}^{d_{att}\times d_m}$.

6) LAYER NORMALIZATION

In deep learning applications, the layer normalization technique [34] is adopted to overcome the limitations of the batch normalization [35] which is a technique used to keep the mean and variance remain the same irrespective of the hidden layers parameters update. This keeps the network stable in subsequent layers in a network because the technique ensures that no activation is gone really high or too low. Unlike batch normalization, layer normalization does not depend on a batch. Also while batch normalization computes the mean and variance across the batch and these values remain unchanged for each example in the batch, layer normalization computes these values across each feature and is independent of other examples. In our implementation, the layer normalization operation is applied after the multi-head attention,encoder-decoder attention, and feed-forward layers.

7) THE DECODER AND THE ENCODER-DECODER ATTENTION

The decoder uses all the components used by the encoder and mainly, the decoder's self-attention process attends only to previous positions by masking future positions. In the clinical time series, we want to predict the values of the variables in the future according to a time lag. Hence the decoder takes the encoded time steps and uses them to generate values of the future time step. This process is done through an encoder-decoder attention process. The encoder-decoder attention process allows every position to focus on all of the most influential past time steps to predict the future time step.

8) LINEAR AND SOFTMAX LAYERS

before predicting the time steps of interest using the linear layer, we introduce the static information and the patient's history, and depending on each benchmark task, the linear softmax layer is used to perform the final prediction. Both mortality prediction tasks are binary classification tasks,

hence the loss function is obtained using the Sigmoid function while Time-series future values prediction and the remaining LoS prediction tasks are modelled in a multi-class regression mode hence the final layer of this model uses a SoftMax function [36].

9) MASKING OF FUTURE TIME-STEPS AND ANCIENT TIME-STEPS

To predict the future time steps without using them in the input, we perform a masking operation that blocks the model from attending to future time steps during training. Unlike the other approaches, we use a rolling mask to mask both the ancient past values as well as future measurements to keep only the values for the last 24 hours. This masking improves the computational speed of the training process and does not affect the results. The masking operation uses an attention mask that allows the model to only look at these previous time steps.

V. EXPERIMENTS

In this section, we describe the experimental process. We cover the related approaches, the training process, and various implementation details.

A. THE RELATED APPROACHES

We compare the performance of our model against the most recent methods. For each method due to limitations associated with the dataset's access, we only use the main building blocks while the pre-processing, embedding, and the dataset remain the same for all of them and similar to our model. hence the obtained results might not corroborate the ones reported in respective studies.

- **Temporal Convolutional Network(TCN)** [37]: This is a method that predicts the ICU adverse events using TCN using the variables that are similar to these used in SAPS II. For a fair evaluation, we implemented this model and used the same dataset and similar evaluation metrics as our model. Hence only the learning algorithm is implemented, while the materials and the pre-processing steps are similar to the current study.
- **SAnD** [14]: This study uses the self attention component of the transformer by omitting the encoder-decoder attention process. Though this reasoning has achieved good results for other benchmark tasks, it can fail to perform well for predicting the time series variables as a regression task.
- **LSTM [15]**: This approach used the LSTM for various benchmarks. We use the multitask LSTM versions in our comparisons.
- **TimeNet for clinical time series** [17]: This work leverages an RNN based model named TimeNet [26] using transfer learning. TimeNet was pre-trained on a big number of various public time series from UCR Repository [38]

B. TRAINING AND IMPLEMENTATION DETAILS

The splitting method used to generate the train/test sets in the current study is based on the patient split. For each patient, the first 70% of the records since admission was used for training and the following 30% used for testing the model. We adopted a recursive prediction approach by generating time steps in an auto-regressive manner because the previously predicted time steps are input in predicting the next. We used a mask that takes into consideration only limited past values by taking only the last 24 hours. This masking process makes it possible that the predictions at time-step i will only depend on the values at positions less than i with only 24 hrs backward window. All layers in the encoder and the decoder, including the embedding layer's output were fixed at d_m=512. For simplicity and faster training, the encoder and the decoder sides were made of 2 similar cascaded components. The models were trained for 50 epochs and optimized using the Adam optimizer [39] with a learning rate α=0.01

VI. RESULTS AND DISCUSSIONS

A. PERFORMANCE EVALUATION METRICS

To evaluate the performance of the model we use various metrics as depicted in Table 2. For Time series prediction, we use AUC-ROC as well as the AUC-PR [40]. These metrics are appropriate for timeseries prediction. These metrics use True Positives and True negatives. We used them because we If we care about true negatives as much as we care about true positives in our predictions.The remaining LoS is a regression task of predicting the remaining time in terms of hours until the patient discharge. We used the MSE, Kappa, and MAPE [41] as performance metrics. The MSE [42] represents the average of the squared errors that resulted from the difference between the LoS values predicted by the model the and true labels of each time-step. The MSE gives a perfect measure of the spread of our results around the true values.

B. TIME-SERIES PREDICTION

Fig.4 depicts the AUC-ROC values obtained for the task of time series forward prediction per each baseline and the transformer-based proposed model. To allow a clear visualization, the figure displays only a sample of 8 time-series variables from the 30 variables considered in the current study. A general observation is that the model achieves better AUC-ROC scores for the frequently sampled variables while achieving low performance for the intermittently sampled variables. For instance, the HR variable contains an average of 80 records per hour of the ICU admission while the WBC has an average of 0.2 records in an ICU admission hour(Per a certain patient). Hence from the figure, MHT achieves an AUC-ROC of 0.95 while for the WBC variable prediction the model achieves an AUC-ROC of 0.64. Hence we can deduct that though the interpolation process can fill the missing records, the process results in the wrong prediction, especially for intermittently recorded variables. We note that for this main prediction task, our transformer

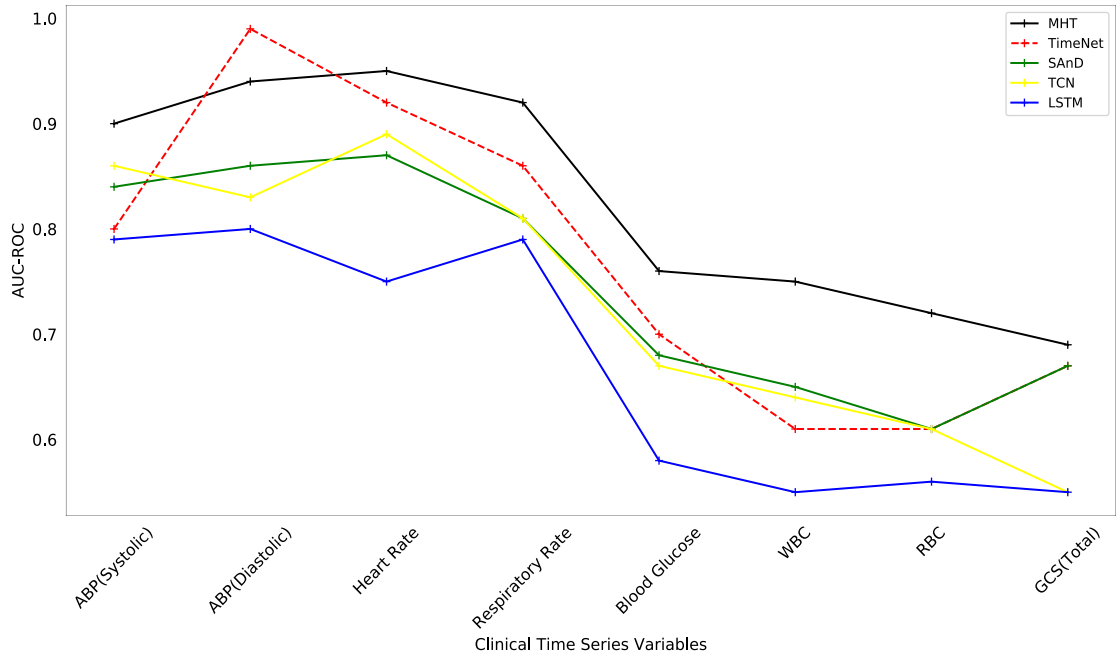


FIGURE 4. AUC-ROC values obtained by the model per each variable for the time series prediction task.

TABLE 2. Performance of the MHT model against the related approaches.

Metrics	Approach				
	TCN [37]	LSTM-Multi [15]	SAnD-Multi [14]	TimeNet-48 [14]	Our Approach
Task 1:Time Series Prediction					
AUC-ROC	0.895	0.880	0.901	0.903	0.908
AUC-PR	0.324	0.319	0.327	0.419	0.427
Task 2:Remaining LoS					
MSE	45432	42131	39918	35611	34232
Kappa	0.456	0.426	0.469	0.480	0.470
MAPE	167.5	188.5	157.8	156.0	185.9
Task 3:In-Hospital Mortality					
AUC-ROC	0.795	0.863	0.859	0.865	0.867
AUC-PR	0.472	0.517	0.519	0.519	0.623
min(Se,+P)	0.484	0.499	0.504	0.510	0.514
Task 4: 24Hrs Mortality					
AUC-ROC	0.675	0.845	0.698	0.859	0.875
AUC-PR	0.434	0.510	0.431	0.519	0.453
min(Se,+P)	0.425	0.435	0.448	0.559	0.415

model outperforms all the recent approaches followed by TimeNet and SAnD. MHT improvement is owing to the use of encoder-decoder attention which can use the most influential past time steps for prediction. On the other hand, the improved performance achieved by TimeNet is due to its features extraction step that leverages transfer learning using pre-trained features. While other works and MHT handcraft their features from the custom embedding layer, TimeNet overcomes this step and leverages the off-the-shelf feature extractor. In [37] TCN achieves better results for proposed

prediction scenarios which are usually binary classification problems like the prediction of intubation risk and the risk of a fluid challenge. Using TCN for the forward time series variables prediction achieves worse results because CNN is not able to capture the long-term dependencies as well as attend properly to the ancient past. Table 2 reports the AUC-ROC performance of each model for all the tasks and all variables. For the time series prediction task our approach achieved an AUC-ROC value of 0.908 and an AUC-PR of 0.903 followed again by TimeNet with LSTM-Multi achieving the lowest values.

C. REMAINING LoS

Table 2 depicts the overall results for all models.The overall observation from the table is that our model outperforms the other related approaches. The model with a smaller MSE is said to be more efficient. MHT(MSE:34232) scored the best values followed by TimeNet and TCN performed poorly (MSE:45432). With the LoS, the regression approach is more prone to errors than when the LoS is binned in the range of days. However, it is clinically important to try to predict the exact days than intervals.

D. IN-HOSPITAL MORTALITY PREDICTION

We use binary classification for prediction. For the In-hospital mortality prediction task, our model outperforms other approaches in AUC-ROC and AUC-PR values as well as the minimum precision and sensitivity (Min(Se, P+)). The precision measures the ratio of the True Positive observations(TP) over the total observations while the sensitivity(also called Recall) measures the model's capability to predict the true negatives of each of the 2 categories. We calculated the sensitivity and precision and retained the minimum of these two metrics. TimeNet achieves considerable results due to its capability of using the previous episode's time series data.

E. THE 24-HOUR MORTALITY PREDICTION

For the 24 hours mortality auxiliary task, the same evaluation metrics are used. though MHT scores the best AUC-ROC values, TimeNet outperforms all models with a bigger margin(+0.11 for Min(Se, P+)). The reason behind this improved performance exhibited by TimeNet is again due to pre-trained features that do not require handcrafted features.

VII. CONCLUSION

In this paper, we presented a method that leverages the recent discovery in NLP for clinical time-series prediction. We applied the transformer architecture to the longitudinal clinical time series and demographic static data to predict the patient's future. For each admission time-step, we predict the future events exactly at the subsequent 6 hours. Unlike the other works that use the whole of records and mask only the future records, we only use the past 24hr records by using a rolling mask. The model is further used to predict other key benchmark tasks that describe the patient's survival including the mortality and the remaining LoS. Our MHT model outperforms other recent related approaches including the LSTM, TCNN, and the Transformer based self-attention. The time series variables that are recorded frequently like the HR and the Respiratory Rate achieve better prediction results than the variables that are recorded intermittently like the WBC. The use of the transformer and its attention mechanisms boosts the performance because the self-attention process encodes the past time steps where the latent relationships between various variables are reflected. The performance is further boosted by the encoder-decoder attention process which helps in identifying the most influential past time steps for the prediction of the future time steps in the course of the patient's admission.

ACKNOWLEDGMENT

The authors would like to thank Prof. Roger G. Mark, the Laboratory for Computational Physiology, Philips Healthcare, Massachusetts Institute of Technology, and the Beth Israel Deaconess Medical Center for giving them full access to the MIMIC-III database.

REFERENCES

[1] A. Haq, S. Patil, A. L. Parcells, and R. S. Chamberlain, "The simplified acute physiology score III is superior to the simplified acute physiology score II and acute physiology and chronic health evaluation II in predicting surgical and ICU mortality in the 'oldest old,'" *Current Gerontol. Geriatrics Res.*, vol. 2014, pp. 1–9, Oct. 2014.

[2] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, A. Damiano, and F. E. Harrell, Jr., "The APACHE III prognostic system: Risk prediction of hospital mortality for critically III hospitalized adults," *Chest*, vol. 100, no. 6, pp. 1619–1636, 1991.

[3] T. Poongodi, R. Krishnamurthi, R. Indrakumari, P. Suresh, and B. Balusamy, "Wearable devices and IoT," in *A Handbook of Internet of Things in Biomedical and Cyber Physical System*. Cham, Switzerland: Springer, 2020, pp. 245–273.

[4] Y. Zheng, C. C. Y. Poon, B. P. Yan, and J. Y. W. Lau, "Pulse arrival time based cuff-less and 24-H wearable blood pressure monitoring and its diagnostic value in hypertension," *J. Med. Syst.*, vol. 40, no. 9, p. 195, Sep. 2016.

[5] R. Yu, Y. Zheng, R. Zhang, Y. Jiang, and C. C. Y. Poon, "Using a multitask recurrent neural network with attention mechanisms to predict hospital mortality of patients," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 486–492, Feb. 2020.

[6] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *J. Amer. Med. Informat. Assoc.*, vol. 24, no. 2, pp. 361–370, Mar. 2017.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[8] K. Ahmed, N. S. Keskar, and R. Socher, "Weighted transformer network for machine translation," 2017, *arXiv:1711.02132*.

[9] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5884–5888.

[10] T. Shao, Y. Guo, H. Chen, and Z. Hao, "Transformer-based neural network for answer selection in question answering," *IEEE Access*, vol. 7, pp. 26146–26156, 2019.

[11] M. Arkhipov, M. Trofimova, Y. Kuratov, and A. Sorokin, "Tuning multilingual transformers for language-specific named entity recognition," in *Proc. 7th Workshop Balto-Slavic Natural Lang. Process.*, 2019, pp. 89–93.

[12] J. Ma, Z. Shou, A. Zareian, H. Mansour, A. Vetro, and S.-F. Chang, "CDSA: Cross-dimensional self-attention for multivariate, geo-tagged time series imputation," 2019, *arXiv:1905.09904*.

[13] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, and X. Yan, "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5244–5254.
[14] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.
[15] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. V. Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," 2017, *arXiv:1703.07771*.
[16] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W.-H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, Dec. 2016, Art. no. 160035.
[17] P. Gupta, P. Malhotra, L. Vig, and G. Shroff, "Using features from pre-trained TimeNet for clinical predictions," in *Proc. KHD@IJCAI*, 2018, pp. 38–44.
[18] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–18.
[19] World Health Organization. (2006). *International Classification of Diseases (ICD)*. [Online]. Available: http://www.who.int/classifications/icd/en/
[20] *Clinical Classifications Software (CCS) for ICD-9-CM Fact Sheet*, Healthcare Cost Utilization Project, Rockville, MD, USA, 2018.
[21] N. Razavian and D. Sontag, "Temporal convolutional neural networks for diagnosis from lab tests," 2015, *arXiv:1511.07938*.
[22] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 47–54.
[23] S. Takase and N. Okazaki, "Positional encoding to control output sequence length," 2019, *arXiv:1904.07418*.
[24] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. Artif. Intell. Statist.*, 2015, pp. 562–570.
[25] F. E. Shamout, T. Zhu, P. Sharma, P. J. Watkinson, and D. A. Clifton, "Deep interpretable early warning system for the detection of clinical deterioration," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 437–446, Feb. 2020.
[26] P. Malhotra, T. V. Vishnu, L. Vig, P. Agarwal, and G. Shroff, "TimeNet: Pre-trained deep recurrent neural network for time series classification," 2017, *arXiv:1706.08838*.
[27] L. Lin, B. Xu, W. Wu, T. Richardson, and E. A. Bernal, "Medical time series classification with hierarchical attention-based temporal convolutional networks: A case study of myotonic dystrophy diagnosis," 2019, *arXiv:1903.11748*.
[28] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*.
[29] D. Kuang, "A 1D convolutional network for leaf and time series classification," 2019, *arXiv:1907.00069*.
[30] A. Zeyer, P. Bahar, K. Irie, R. Schlüter, and H. Ney, "A comparison of transformer and LSTM encoder decoder models for ASR," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 8–15.
[31] B. S. Raghuwanshi and S. Shukla, "SMOTE based class-specific extreme learning machine for imbalanced learning," *Knowl.-Based Syst.*, vol. 187, Jan. 2020, Art. no. 104814.
[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.
[33] T. Blu, P. Thévenaz, and M. Unser, "Linear interpolation revitalized," *IEEE Trans. Image Process.*, vol. 13, no. 5, pp. 710–719, May 2004.
[34] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.
[36] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," 2017, *arXiv:1702.05659*.
[37] F. J. R. Catling and A. H. Wolff, "Temporal convolutional networks allow early prediction of events in critical care," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 3, pp. 355–365, Mar. 2020.
[38] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, and E. Keogh, "The UCR time series archive," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 6, pp. 1293–1305, 2019.
[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[40] S. Narkhede, "Understanding AUC—ROC curve," *Towards Data Sci.*, vol. 26, no. 1, pp. 220–227, 2018.
[41] U. Khair, H. Fahmi, S. A. Hakim, and R. Rahim, "Forecasting error calculation with mean absolute deviation and mean absolute percentage error," *J. Phys., Conf. Ser.*, vol. 930, Dec. 2017, Art. no. 012002.
[42] T. Chai and R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)," *Geosci. Model Develop. Discuss.*, vol. 7, no. 1, pp. 1525–1534, 2014.

GASPARD HARERIMANA (Member, IEEE) received the B.S. degree in computer engineering from Ethiopian Defense University, in 2008, the M.S. degree in information technology from Carnegie Mellon University, in 2015, and the Ph.D. degree in computer science from Sangmyung University, Seoul, South Korea, in 2020. He is currently a Visiting Lecturer at the African Center of Excellence in Data Science and a Visiting Scholar at Carnegie Mellon University Africa Campus. His research interests include machine learning, big data, and health analytics.

JONG WOOK KIM (Member, IEEE) received the Ph.D. degree from the Computer Science Department, Arizona State University, in 2009. He was a Software Engineer with the Query Optimization Group, Teradata, from 2010 to 2013. He is currently an Assistant Professor in computer science with Sangmyung University. His primary research interests include the areas of data privacy, distributed databases, and query optimization. He is a member of the ACM.

BEAKCHEOL JANG (Member, IEEE) received the B.S. degree from Yonsei University, in 2001, the M.S. degree from the Korea Advanced Institute of Science and Technology, in 2002, and the Ph.D. degree from North Carolina State University, in 2009, all in computer science. He is currently an Associate Professor with the Graduate School of Information, Yonsei University. His primary research interests include big data analytics, artificial intelligence, natural language processing, and the Internet of Things. He is a member of the ACM.