Deep Interpretable Early Warning System for the Detection of Clinical Deterioration

Farah E. Shamout, Tingting Zhu, Pulkit Sharma, Peter J. Watkinson, and David A. Clifton

***Abstract*—Assessment of physiological instability preceding adverse events on hospital wards has been previously investigated through clinical early warning score systems. Early warning scores are simple to use yet they consider data as independent and identically distributed random variables. Deep learning applications are able to learn from sequential data, however they lack interpretability and are thus difficult to deploy in clinical settings. We propose the 'Deep Early Warning System' (DEWS), an interpretable end-to-end deep learning model that interpolates temporal data and predicts the probability of an adverse event, defined as the composite outcome of cardiac arrest, mortality or unplanned ICU admission. The model was developed and validated using routinely collected vital signs of patients admitted to the the Oxford University Hospitals between 21st March 2014 and 31st March 2018. We extracted 45 314 vital-sign measurements as a balanced training set and 359 481 vital-sign measurements as an imbalanced testing set to mimic a real-life setting of emergency admissions. DEWS achieved superior accuracy than the state-of-the-art that is currently implemented in clinical settings, the National Early Warning Score, in terms of the overall area under the receiver operating characteristic curve (AUROC) (0.880 vs. 0.866) and when evaluated independently for each of the three outcomes. Our attention-based architecture was able to recognize 'historical' trends in the data that are most correlated with the predicted probability. With high sensitivity, improved clinical utility and increased interpretability, our model can be easily deployed in clinical settings to supplement existing EWS systems.**

***Index Terms*—Early warning system, time-series data, data interpolation, supervised learning, deep learning.**

Manuscript received February 20, 2019; revised June 7, 2019 and July 28, 2019; accepted August 16, 2019. Date of publication September 19, 2019; date of current version February 6, 2020. The work of F. E. Shamout was supported by the Rhodes Trust and the P. J. Watkinson was supported by the NIHR Biomedical Research Centre, Oxford. Disclaimer: This publication presents independent research commissioned by the Health Innovation Challenge Fund (HICF-R9-524; WT-103703/Z/14/Z), a parallel funding partnership between the Department of Health and Wellcome Trust. The views expressed in this publication are those of the author(s) and not necessarily those of the Department of Health or Wellcome Trust. *(Corresponding author: Farah E. Shamout.)*

F. E. Shamout, T. Zhu, P. Sharma, and D. A. Clifton are with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, OX1 2JD Oxford, U.K. (e-mail: farah.shamout@eng; tingting.zhu@eng.ox.ac.uk; pulkit.sharma@eng.ox.ac.uk; david.clifton@eng.ox.ac.uk).

P. J. Watkinson is with the Nuffield Department of Clinical Neurosciences, Oxford University Hospitals NHS Foundation Trust, OX3 9DU Oxford, U.K. (e-mail: peter.watkinson@ndcn.ox.ac.uk).

Digital Object Identifier 10.1109/JBHI.2019.2937803

I. Introduction

IN RECENT years, increased access to Electronic Health Records (EHR) has motivated the development of data-driven systems that detect physiological derangement to secure timely response. Early Warning Score (EWS) systems assess a patient's degree of illness by assigning scores to routinely collected vital-sign measurements based on pre-determined normality ranges. The National Early Warning Score (NEWS), which is currently used in hospitals and recommended by the Royal College of Physicians in the United Kingdom [1], has shown superior performance in comparison to other EWS systems in detecting the composite outcome of unplanned ICU admission, cardiac arrest, and mortality [2]. EWS systems assign an independent score to each vital-sign variable and assume that vital-sign measurements are independent and identically distributed (I.I.D.) random variables. Given their simplistic nature, traditional EWS systems do not learn any spatio-temporal information from the vital signs. We hypothesized that the use of deep learning may improve the accuracy of predicting clinical outcomes by recognizing complex patterns in the data.

Significant improvements over clinical scores and static machine learning models were achieved using deep learning, such as to predict ICU mortality for pediatrics [3] or to detect sepsis [4]. Long Short Term Memory (LSTM) networks in particular have illustrated superior performance when considering various benchmarks [5]–[7]. Most of the relevant work, however, was primarily based in intensive care settings. We designed an EWS framework that could generalize across a heterogeneous patient population in non-critical care wards, from pre-processing sparse vital signs variables to predicting the probability of an outcome.

Additionally, the decision-making process of the previously-proposed deep learning models lacked interpretability, and as such they are viewed as 'black box' models by clinical staff since they do not provide any insight on the patterns learned from the data. We defined interpretability as the ability of the clinician to inspect 'trends' of vital signs that most contribute to the model's predicted probability. Inspired by natural language processing, our approach incorporated an 'attention' mechanism with recurrent architectures that highlights those parts of the input time-series that are most relevant to the output. This is useful since interpretability is considered to be a core component of clinical utility [8].

The physiological data recorded in an EHR is often sparse, noisy, and incomplete, especially when collected in non-critical

care wards, which is challenging for recurrent deep learning architectures that require regularly-spaced data points. Regularly sampled data can be interpolated using naive methods, such as carrying the most recent value forward (CF) and linear interpolation (LI). Although such approaches are computationally inexpensive, they may impose bias and error [9] and do not account for the uncertainty of the imputed data. In a probabilistic approach, Gaussian Process Regression (GPR) was used to model irregularly-sampled physiological time-series data [10]–[12], to interpolate the posterior mean and variance at unseen time points. In our work, we evaluated the benefit of GPR modeling in comparison to CF and LI.

Unlike currently implemented EWS systems that were originally designed in a heuristic fashion, we developed and validated an interpretable end-to-end Deep Early Warning System (DEWS) that alerts for clinical deterioration, defined as the composite outcome of unplanned ICU admission, mortality, and cardiac arrest.

A. Related Works and Contributions

Attention-based deep learning models improve interpretability; they can model extended long-term-dependencies; and they have been used numerously in computer vision and natural language processing [13], [14]. Within clinical settings, attention models have gained limited recognition and have been used for classifying atrial fibrillation in ECG data [15], [16], predicting a future diagnosis [17], [18], or predicting high risk vascular disease, using both diagnosis codes and medication data [19]. The limitation of using diagnosis codes is that they may not be readily available in a real-time setting, as in retrospective databases. We aimed to use information in routinely collected vital-sign data, as in currently implemented EWS systems, constituting multiple sequential inputs.

In traditional sequence-to-sequence modelling problems, attention enables the model to learn deferentially from more and less important parts of the input sequence; i.e., words in the case of sentence translation, or sentences for document classification [20]. Our goal was to learn different content from different time-series signals, and then fuse the information to predict the probability of an outcome. To the best of our knowledge, no existing attention-based deep learning model focuses on learning from a combination of vital-sign time-series data to indicate a patient's health status in real-time.

The primary contribution of this work is a novel deep learning architecture with high clinical utility to predict clinical outcomes. The model learned from regularly-sampled mean and variance features interpolated by modelling sparse vital-sign data via GPR. We evaluated the framework's ability in detecting deterioration prior to the composite outcome, as in previous studies [21], [22], achieving state-of-the-art results in comparison to the clinical benchmark.

The rest of the paper is organized as follows: Section II describes the methodology pipeline in terms of feature extraction and outcome prediction, Section III describes the datasets used for training and testing, Section IV describes the experimental observations using the proposed models, and Section V discusses findings and presents concluding remarks.

Algorithm 1: Proposed Framework of DEWS that Classifies Whether a Window $\mathcal{D}_W$ of Vital-Sign Measurements is Within N Hours of an Outcome.

Input: Unlabelled window $\mathcal{D}_W = [x_i, \mathbf{y}_i]_{i=1}^{n}$ and regularly sampled time instances $P = [x_t^*]_{t=1}^{T}$ per window.
Outputs: Label $l \in (0, 1)$.
Data Interpolation
For $j = 1$ to m:
  $GPR \leftarrow$ Fit a GPR using $[x_i, y_i]_{i=1}^{n}$
  $\mathbf{y}_{\mu,j}$, $\mathbf{y}_{\sigma,j} \leftarrow GPR([x_t^*]_{t=1}^{T})$
$\mathbf{Y}_\mu = [\mathbf{y}_{\mu,1}, \ldots, \mathbf{y}_{\mu,m}]$
$\mathbf{Y}_\sigma = [\mathbf{y}_{\sigma,1}, \ldots, \mathbf{y}_{\sigma,m}]$
Classification
$\mathbf{c}_\mu \leftarrow Encoder(\mathbf{Y}_\mu)$
$\mathbf{c}_\sigma \leftarrow Encoder(\mathbf{Y}_\sigma)$
$l \leftarrow Decoder(\mathbf{c}_\mu + \mathbf{c}_\sigma)$

II. Proposed Methods

We framed the problem of detecting clinical deterioration as a binary classification task, such that the model assigns a binary label, based on a computed probability and a pre-defined alerting threshold, to each vital-sign measurement. For each measurement, we would like to predict the probability of the composite outcome within the next N hours. An event window was defined as a vital-sign measurement that was within N hours of a composite outcome and its preceding w hours of observations. A non-event window was defined as a vital-sign measurement that was not within N hours of a composite outcome and its preceding w hours window. We set $N = 24$ hours in our study, which is a common evaluation window in the development of EWS systems [21], [22], and we evaluated w at 24, 12, and 6 hours.

Assume $\mathcal{D}_W = [x_i, \mathbf{y}_i]_{i=1}^{n}$ is an unlabelled sample window, where x_i is the ith time instance and $\mathbf{y}_i \in \mathbb{R}^m$ denotes the feature space consisting of m vital signs sequences. The recurrent classification model required the vital signs data to be measured at a fixed set of regularly sampled time instances to compute the output escalation label $l \in (0, 1)$. However, each vital-sign sequence j was temporally irregular due to the nature of EHR data. Hence, we deployed a patient-specific GPR for each vital-sign sequence to interpolate the mean and variance at fixed regularly-spaced time instances $P = [x_t^*]_{t=1}^{T}$.

These posterior mean and variance estimates were concatenated for all the vital signs to obtain: $\mathbf{Y}_\mu = [\mathbf{y}_{\mu,j}]_{j=1}^{m}$ and $\mathbf{Y}_\sigma = [\mathbf{y}_{\sigma,j}]_{j=1}^{m}$, where $\mathbf{Y}_\mu, \mathbf{Y}_\sigma \in \mathbb{R}^{m \times T}$ and $\mathbf{y}_{\mu,j}$ and $\mathbf{y}_{\sigma,j}$ are the GPR mean and variance for the jth vital sign, such that $j = 1, \ldots, m$. Attention-based $encoders$ learned from each interpolated sequence in $\mathbf{Y}_\mu$ and $\mathbf{Y}_\sigma$ to obtain the sequence-level context vectors $[\mathbf{c}_{\mu,j}]_{j=1}^{m}$ and $[\mathbf{c}_{\sigma,j}]_{j=1}^{m}$, respectively. Finally, the summary context vectors $\mathbf{c}_\mu$ and $\mathbf{c}_\sigma$ summed up the sequence-level context vectors and were fed to decoding layers to compute the probability of an outcome. If the probability exceeded the pre-defined alerting threshold, then $\mathcal{D}_W = 1$. The proposed framework is described in Algorithm 1. We now describe each step in more detail.

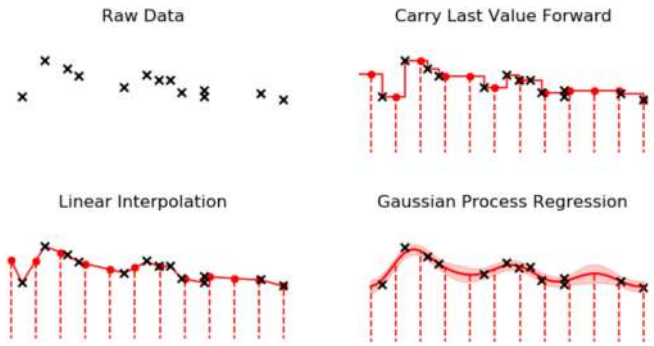

Fig. 1. Visualization of the feature extraction interpolation techniques from the raw data within a w hour window: Carrying Last Value Forward, Linear Interpolation, and Gaussian Process Regression. The y-axis represents the range of any vital sign and the x-axis represents time since admission, with the rightmost data point representing the time of prediction.

A. Patient-Specific Feature Transformation

Each window of length w hours was modelled using GPR. It was also modelled using CF and LI for benchmarking purposes, and an overview of the modelling techniques is shown in Fig. 1. GPR generalizes multivariate Gaussian distributions to infinite dimensionality and offers a probabilistic and non-parametric approach to model a sparse vital-sign time-series as a function of time from admission. We adopted a radial basis function (RBF) with added white noise as our covariance function to map the similarity between pairs of data points x and x', such that

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right) + \sigma_n^2 \delta(x, x') \quad (1)$$

$\delta(x, x')$ is the Kronecker delta function and $\Theta = \{l, \sigma_f, \sigma_n\}$ is the set of hyperparameters, where l is the lengthscale, σ_f is the variance of the RBF, and σ_n is the variance of the added white noise. Since the work involved patient-specific modelling for each sequence per window across a large-scale population, we adopted a Bayesian approach to increase the modelling efficiency. First, we defined the expected value, or the mean function, of each vital-sign GPR as a constant function equivalent to the population mean of patients with the same age and sex. Second, we used lognormal distributions as priors to constrain each hyperparameter to be clinically meaningful. The models were optimized by minimizing the negative log likelihood with respect to the hyperparameters.

After fitting the GPR to the training data $[x_i, y_i]_{i=1}^n$ of vital sign j, the GPR kernel is applied to interpolate missing values at equally-spaced time steps $[x_t^*]_{t=1}^T$ across the input window, such that the posterior mean is

$$\mathbf{y}_\mu = K_* K^{-1} \mathbf{y}, \quad (2)$$

with variance,

$$\mathbf{y}_\sigma = K_{**} - K_* K^{-1} K_*^T, \quad (3)$$

where $\mathbf{y}$ is the training data $[y_i]_{i=1}^n$, K represents the similarity measure between all training values, K_* represents the similarity measure between all training and missing values, and K_{**} represents the similarity measure between all missing values. Finally, we obtained a set of equally-spaced measurements as an input to the neural network.

For any point of prediction with historical data spanning less than w hours, we pre-padded the sequence with the population mean and maximum variance of the respective vital sign. Discrete variables were modeled by CF and LI, which only interpolated mean values. The extracted features, through GPR, CF, and LI, were then scaled for the training and testing of the classifiers.

B. Model Architecture

We here describe the architecture of the proposed DEWS method. The interpolated mean and variance features of each vital-sign input were first processed through a Bi-directional LSTM (BiLSTM) network [23], in order to maximize information retrieval in the forward and backward directions. An attention-based BiLSTM model previously performed well for classifying sequential healthcare data [15], and we extended upon it by customizing the mechanism in the attention block and accounting for the uncertainty of the input.

The BiLSTM consisted of two layers which processed each mean and variance input in forward and reverse directions and yielded two hidden layer states $h_{t,f}$ and $h_{t,r}$. The average of $h_{t,f}$ and $h_{t,r}$, denoted as $\overline{h_t}$ served as the input of our attention mechanism. While the definition of attention varies across the literature, we adopted the definition in [15], [20], to learn the most important parts of each sequence. For each vital sign j, $e_{t,j}$ measured the importance of the information at each time step t:

$$e_{t,j} = U_j \, a(W_j \, \overline{h_{t,j}} + b_j), \quad (4)$$

by computing its similarity with U_j, a trainable sequence-level context vector, where a is the rectified linear unit. Next, $e_{t,j}$ was used to derive $\alpha_{t,j}$, or the normalized weights assigned to the hidden states, as:

$$\alpha_{t,j} = \frac{\exp(e_{t,j})}{\sum_{t=1}^T \exp(e_{t,j})}, \quad (5)$$

and $\alpha_{t,j}$ was further employed to derive the final context vector $\mathbf{c}_j$:

$$\mathbf{c}_j = \sum_{t=1}^T \alpha_{t,j} \overline{h_{t,j}} \quad (6)$$

Equations (4)-(6) were applied to the BiLSTM outputs of each mean and variance input of each vital sign j. The sequence-level context vectors obtained from all vital signs were then aggregated by the trainable weights V_μ for the mean features:

$$\mathbf{c}_\mu = V_\mu \sum_j \mathbf{c}_{\mu,j}, \quad (7)$$

and V_σ for the variance features:

$$\mathbf{c}_\sigma = V_\sigma \sum_j \mathbf{c}_{\sigma,j} \quad (8)$$

The aggregated context vectors were then summed and processed by two dense layers consisting of a rectified linear unit

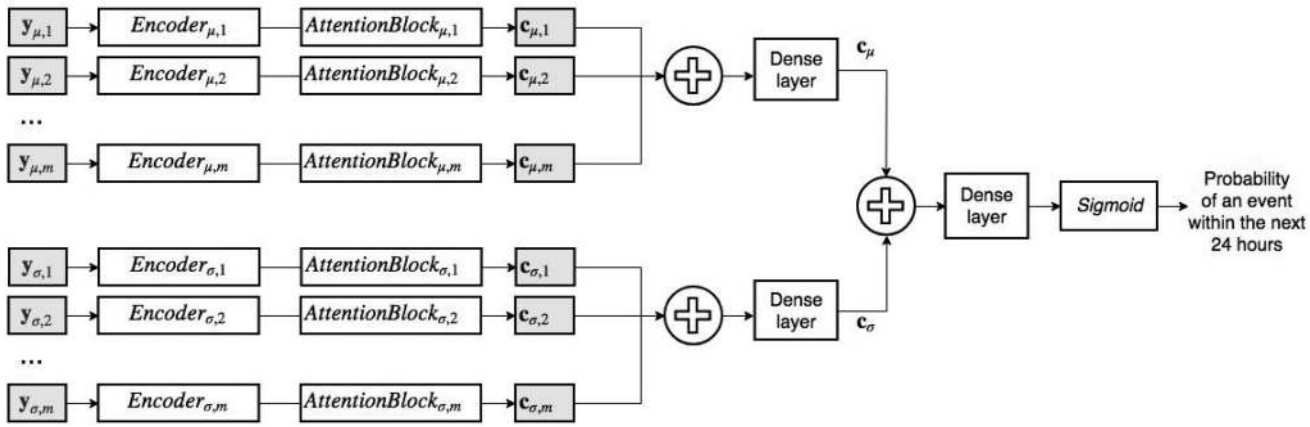

Fig. 2. Schematic diagram of the classification architecture DEWS which learns the attention weights from the mean and variance of each vital-sign variable and produces a binary output to indicate whether an observation set of a patient is within 24 hours of a composite outcome (cardiac arrest, mortality or unplanned ICU admission). Each regularly sampled input $\mathbf{y}$ $(T \times 1)$ is processed by an $encoder$ unit consisting of a BiLSTM with e output units at each time step to produce the hidden states $(T \times e)$. The attention blocks then compute the context vectors $(e \times 1)$ that summarize their respective inputs. The context vectors of the mean and variance of the vital signs are finally aggregated using dense layers.

and a sigmoid function, respectively. Finally, the computed probability of an outcome within the next 24 hours was compared to a pre-defined alerting threshold and a binary label was assigned to the window. The model schematic is shown in Fig. 2.

During training, the weights U_k, W_k, b_k, V_μ and V_σ were optimized, where V_μ and V_σ accounted for the correlations across the vital signs since they combined their respective context vectors. We chose the other hyperparameters, such as number of output nodes e per encoder timestep shown in Fig. 2, and the alerting threshold through experimentation.

C. Model Evaluation

We evaluated our model using training and testing sets. First, we assessed the modelling quality of GPR, CF, and LI. For each sequence in the training windows, we randomly held out 20% of the data as test points and modelled the rest using each interpolation technique. We then calculated the root mean squared error (RMSE) comparing the true values and the interpolated values at the held out test points.

We evaluated the performance of our classifiers using the area under receiver-operating characteristics (AUROC), sensitivity, and specificity on the testing set. All metrics were performed using a bootstrapping technique without replacement [24] with a fixed number of bootstraps (nb). We compared the performance of the models across 16-45 years old patients and >45 years old patients, and across the three outcomes independently.

To assess the clinical utility of DEWS in comparison to the clinical benchmark, we plotted the percentage of generated 'triggers', or windows at or above a given EWS score, on the y-axis against sensitivity on the x-axis [22]. We also assessed the proposed model's decision-making process by visualizing the attention weights computed for a case study. Finally, we compared the average normalized NEWS score and the average DEWS probability for the first 120 hours from admission and the last 24 hours prior to an outcome.

III. Dataset

This section describes the data retrieved from a retrospective large database of routinely collected observations from concluded hospital admissions between 21st March 2014 and 31st March 2018 within the Hospital Alerting Via Electronic Noticeboard (HAVEN) project (REC reference: 16/SC/0264 and Confidential Advisory Group reference 08/02/1394). The database included the vital-sign measurements of adult patients admitted to four Oxford University Hospitals: the John Radcliffe Hospital, Horton General Hospital, Churchill Hospital, and the Nuffield Orthopaedic Hospital, collected by the System for Electronic Notification and Documentation (SEND, Sensyne Health) [25]. We extracted the vital-sign measurements and the occurrences of outcomes to develop and validate a model that is analogous to EWS systems.

Each vital-sign measurement was recorded manually by hospital staff and consisted of 5 continuous variables: heart rate (HR), systolic blood pressure (SBP), respiratory rate (RR), temperature (TEMP), and oxygen saturation (SPO_2), and 2 discrete variables: Alert, Voice, Pain and Unconscious (AVPU) score and a binary indicating whether supplemental oxygen was provided. We defined the time of a composite outcome as the time of the first occurring event of unplanned ICU admission, mortality and cardiac arrest. In the case of multiple occurrences of adverse events, we removed observations recorded after the first event.

We split the dataset by time as recommended by TRIPOD guidelines [26]: $\mathcal{D}_1$ (21 March 2014–31 October 2017) for training and validation, and $\mathcal{D}_2$ (1 November 2017–31 March 2018) for testing, roughly corresponding to 85% and 15% of the overall dataset, respectively. We labelled each vital-sign measurement as an event or non-event window. To overcome class imbalance, we performed random under-sampling of the non-event windows in $\mathcal{D}_1$ to match the maximum number of event windows. $\mathcal{D}_2$ remained imbalanced to mimic a real-life testing set, yet it excluded patients who were well enough to be discharged on the day of admission, elective admissions with

TABLE I
CHARACTERISTICS AND DEMOGRAPHICS OF $\mathcal{D}_1$, USED FOR TRAINING AND VALIDATION, AND $\mathcal{D}_2$ USED AS A TESTING SET

Characteristic	$\mathcal{D}_1$	$\mathcal{D}_2$
Age, mean (SD)	68 (18)	69 (19)
Admissions, n	21,512	15,772
Females, n (%)	10.703 (49.8)	8,276 (52.5)
16-45 yrs old, n (%)	3,462 (16.1)	3,568 (22.6)
> 45 yrs old, n (%)	18,050 (83.9)	12,204 (77.4)
Vital-Sign measurements, n	45,314	359,481
Non-event, n (%)	22,657 (50)	356,498 (99.2)
Event, n (%)	22,657 (50)	2,983 (0.8)
Unplanned ICU	10,942 (48.3)	811 (27.2)
Cardiac Arrest	1,314 (5.8)	223 (7.5)
Mortality	10,401 (45.9)	1,949 (65.3)

TABLE II
DISTRIBUTIONS OF THE VITAL SIGNS IN TERMS OF MEDIAN AND INTERQUARTILE RANGE ACROSS $\mathcal{D}_1$ AND $\mathcal{D}_2$

Vital Sign, unit	$\mathcal{D}_1$	$\mathcal{D}_2$
HR, beats/min	86 (73-101)	81 (70-92)
SBP, mm Hg	121 (105-139)	126 (112-142)
TEMP, °C	36.4 (36.0-36.9)	36.4 (36.0-36.8)
RR, breaths/min	18 (16-21)	17 (16-18)
SPO_2, %	96 (94-98)	96 (95-98)
AVPU Score		
Alert, *n* (%)	40,642 (89.6)	352,367 (98.0)
Voice, *n* (%)	2884 (6.4)	5,895 (1.6)
Pain, *n* (%)	981 (2.2)	940 (0.3)
Unresponsive, *n* (%)	807 (1.8)	279 (0.08)
Supplemental Oxygen	18,845 (41.6)	60,037 (16.7)

scheduled visits, and admissions with no vital-sign measurements collected in the last 24 hours prior to an outcome because such patients are likely to be on terminal care pathways. This is the same exclusion criteria adopted in related previous works [21], [22], as EWS systems aim to assess acutely-ill patients.

IV. EXPERIMENTAL OBSERVATIONS

In this section, we summarize the main findings of our study pertaining to experimental and design choices and performance evaluation.

A. Experimental Setup

1) Data Modelling: The patient admissions had varying lengths of stay, ranging between 0.01 and 1,165 days, and the number of timestamped observations per admission ranged between 1 and 1,901 observations. Across the extracted vital-sign measurements, the missing values for HR, SBP, TEMP, RR, SPO_2, AVPU and supplemental oxygen were 1.94%, 1.79%, 10.29%, 3.23%, 1.99%, 4.74%, and 3.63%, respectively. The characteristics and demographics of $\mathcal{D}_1$ and $\mathcal{D}_2$ are shown in Table I and the distributions of their vital signs are shown in Table II. Both datasets have a similar mean age and proportion of females. Since $\mathcal{D}_1$ was balanced while $\mathcal{D}_2$ was imbalanced, we observe differences between the distributions of some variables.

TABLE III
MEAN AND STANDARD DEVIATION OF THE ROOT MEAN SQUARED ERROR (RMSE) FOR MISSING DATA INTERPOLATION USING THE DIFFERENT MODELLING TECHNIQUES FOR ALL VITAL SIGNS: CF, LI, AND GPR. THIS WAS DONE BY MODELLING 80% OF EACH TRAINING WINDOW AND HOLDING OUT 20% OF THE TRAINING WINDOW AS TEST POINTS TO CALCULATE THE RMSE

Vital sign	CF	LI	GPR
HR	8.24 ±8.08	7.20 ±7.15	7.49 ±6.96
RR	2.24 ±2.63	2.07 ±2.36	1.97 ±2.15
TEMP	0.41 ±0.37	0.35 ±0.31	0.36 ±0.31
SBP	13.72 ±11.58	12.15 ±10.46	12.79 ±10.21
SPO_2	2.16 ±2.81	2.03 ±2.70	1.97 ±2.46
AVPU	0.10 ±0.33	0.10 ±0.32	-
Supplemental Oxygen	0.12 ±0.29	0.12 ±0.26	-

Lognormal priors over the hyperparameters were selected to ensure that the modelled vital signs fell within the expected ranges in a clinical setting. The lognormal distributions chosen as priors for l were $(\mu = 1.0, \sigma = 0.1)$ for HR, RR, TEMP, and SPO_2, and $(\mu = 1.5, \sigma = 0.1)$ for SBP. The lognormal distributions chosen as priors for σ_f were $(\mu = 0.0, \sigma = 0.1)$ for HR, SBP, and SPO_2, $(\mu = 1.5, \sigma = 0.1)$ for RR, and $(\mu = 3.5, \sigma = 0.1)$ for TEMP. The lognormal distributions chosen as priors for σ_n were $(\mu = 0.0, \sigma = 4.0)$ for HR, SBP, and SPO_2, $(\mu = 0.0, \sigma = 0.1)$ for RR, and $(\mu = 1.5, \sigma = 0.1)$ for TEMP. Applying population-based lognormal distributions to the priors of the three hyperparameters enabled us to efficiently fit patient-specific vital signs GPR models in a large-scale dataset. All GPR models were built using GPy (v 1.9.6) [27].

Using the optimized GPR models, we interpolated the posterior mean and variance at every 2 hours across the w hours long window, in keeping with national guidelines of alerting at least at every other hour. We modelled a truncated window of up to w hours to reduce the number of timesteps per input for the recurrent neural network. This reduces the complexity of the architecture, which is essential given our dataset size. Additionally, modelling a large number of timesteps would require the storage of a subsequently large GPR kernel matrix for each window, which would impose computational complexity. The model performed best with windows of length 24 hours i.e. $w = 24$ hours, after evaluating its performance for lengths of 6, 12, and 24 hours.

After sampling equally-spaced measurements, we experimented with standard scaling, min-max scaling and scaling by the maximum absolute value. The best classification performance was achieved through min-max scaling of mean features into the range $[-1,1]$ and maximum absolute scaling of variance features into the range [0,1]. During training, we used 20% of $\mathcal{D}_1$ as a validation set, and so the scaling and shifting operations were obtained through the other 80% and then applied to the validation and test set $\mathcal{D}_2$.

The mean and standard deviation of the RMSE of the training windows are summarized in Table III to compare the data interpolation quality of GPR, CF, and LI. In DEWS, AVPU and supplemental oxygen were interpolated using CF with no time-limit. If no previous value was available, then we assumed 'Alert' for AVPU and that supplemental oxygen was not provided. We

used CF because it is less computationally expensive than LI, considering that they both resulted with similar mean RMSE.

2) Model Variants: We designed several deep learning architectures to compare to DEWS. The first set of models consisted of simple architectures, namely Logistic Regression (LR), a single-layer LSTM network, and a single-layer BiLSTM network. $\mathbf{Y}_\mu$ was the input to the models' first single layer. Despite its simplicity, the BiLSTM architecture lacked interpretability.

Therefore, the second set of models included attention mechanisms applied to $\mathbf{Y}_\mu$. BiLSTM-ATT-1 consisted of a simple BiLSTM followed by one attention (ATT) module. This approach was similar to language modelling, since it consisted of a single input feature space, however we were unable to identify the individual contributions of each vital sign. BiLSTM-ATT-2 thus processed each vital sign independently using a dedicated BiLSTM and attention mechanism. The context vectors of the vital signs were then summed and decoded.

The third set of models consisted of 'Uncertainty-Aware' (UA) models since attention was not only applied to the mean features, but also to the variance features $\mathbf{Y}_\sigma$. UA-BiLSTM-ATT-1 consisted of two BiLSTM layers, where one BiLSTM processed $\mathbf{Y}_\mu$ and the other processed $\mathbf{Y}_\sigma$. Each BiLSTM was then followed by one ATT module. Finally, UA-BiLSTM-ATT-2, our proposed model DEWS, had one BiLSTM-ATT per mean and variance features of the vital signs as shown in Fig. 2. We compared all models to NEWS and a simple logistic regression (LR) which used I.I.D. features as inputs; i.e., the last recorded set of vital-sign measurements.

3) Deep Learning Experiments: We tried training the model using data from emergency admissions only, as in $\mathcal{D}_2$, yet the model performed best when the exclusion criteria was not applied. Despite differences in characteristics of vital signs in $\mathcal{D}_1$ and $\mathcal{D}_2$, the model was able to learn complex patterns since we utilized patient-specific modelling.

The hyperparameters of the models, including the number of hidden layers, units per layers, and activation functions were optimized empirically using the training and validation set $\mathcal{D}_1$. Within the encoder units, the optimal number of output nodes (e) at each timestep of the BiLSTM that resulted with the best performance was 12. The dense layers that aggregated all context vectors consisted of 5 units, while the final dense layer consisted of 1 unit. For the similarity function of the attention block, we compared the hyperbolic tangent function and ReLU, and the latter performed better for our application.

All deep learning architectures were trained with early stopping by monitoring the loss on the validation set to avoid overfitting, and a batch size of 128, after experimenting with a batch size of 32, 64, and 128. Each batch consisted of sequences of the same length, whereby length of sequence refers to the number of sampled equidistant data points prior to padding. The models were optimized using the Adam optimizer. All deep learning models were implemented using Keras (v 2.2.2) [28] with a TensorFlow backend (v 1.5.0) [29].

B. Performance Evaluation

Table IV shows the performance results of all models on $\mathcal{D}_2$. DEWS performed best compared to all models with an 0.880 AUROC [95% CI 0.880-0.880] and 0.729 sensitivity [95% CI 0.728-0.729], exceeding NEWS, LR, and all model variants. The decision threshold of DEWS that achieved a similar specificity to NEWS was 0.66. All LR-based models performed worse than DEWS and NEWS, while all other model variants, with the exception of LI-LSTM, performed similar to NEWS. When only the mean features of the GPR were considered in the BiLSTM-variant models (i.e. BiLSTM, BiLSTM-ATT-1, and BiLSTM-ATT-2), no improvement was observed when compared to other interpolation methods (CF and LI). However, when the variance of the GPR was considered, it further improved the results of GPR-based methods in comparison to other interpolation methods, such as in UA-BiLSTM-ATT-1 and DEWS.

DEWS performed better than NEWS for both age groups, as shown in Table V, especially for 16-45 years old patients, with 0.820 AUROC [95% CI 0.818-0.822] compared to 0.760 AUROC [95% CI 0.757-0.762], respectively. DEWS also performed better than NEWS across all outcomes, both in terms of AUROC and sensitivity.

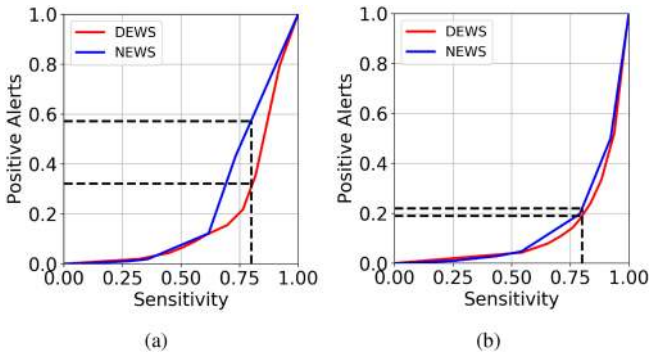

Fig. 3. Efficiency curves plotting sensitivity (x-axis) against the trigger rate, or the percentage of observations (y-axis) with a DEWS probability or normalized NEWS score greater than or equal to a decision threshold, with the decision threshold ranging between 0 and 1, for (a) 16–45 years old patients and (b) >45 years old patients. The vertical black dashed line represents a fixed sensitivity of 80%.

C. Clinical Utility

Fig. 3 shows the percentage of triggers, or positive alerts, produced by our best performing model, DEWS, and NEWS at different sensitivity values (x-axis). Across the 16-45 years old patients, Fig. 3(a), NEWS approximately had a 59% trigger rate while DEWS had a 37% trigger rate, at a fixed sensitivity of 80%. This shows that DEWS reduced the trigger rate by approximately 22%, which could directly ease staff burden. Across the >45 years old, DEWS reduced the trigger rate by approximately 3%.

D. Case Studies

The attention weights of two windows are visualized in Fig. 4, where the box with blue borders shows the vitals signs after feature transformation through GPR modelling and scaling. The two windows belonged to the same patient, where the first row was a non-event window, since the time of prediction was not within 24 hours of an outcome, while the second row was an

TABLE IV

PERFORMANCE EVALUATION OF DEWS IN COMPARISON TO NEWS AND ALL MODEL VARIANTS, USING I.I.D., CF, LI, OR GPR INTERPOLATED FEATURES. THE DECISION THRESHOLD OF ALL CLASSIFIERS WAS ADJUSTED TO ACHIEVE A SPECIFICITY SIMILAR TO THAT OF NEWS (≈ 0.89). MEAN AND CONFIDENCE INTERVALS WERE EVALUATED USING A BOOTSTRAPPING TECHNIQUE ($nb = 1,000$) ON $\mathcal{D}_2$

Model	AUROC	Sensitivity	Specificity
I.I.D.			
NEWS	0.866 (0.865-0.866)	0.702 (0.702-0.703)	0.888 (0.888-0.888)
LR	0.847 (0.846-0.847)	0.651 (0.650-0.652)	0.888 (0.888-0.888)
CF			
LR	0.834 (0.834-0.835)	0.641 (0.640-0.642)	0.888 (0.888-0.888)
LSTM	0.867 (0.866-0.867)	0.691 (0.690-0.691)	0.887 (0.887-0.887)
BiLSTM	0.867 (0.867-0.867)	0.699 (0.699-0.700)	0.888 (0.887-0.888)
BiLSTM-ATT-1	0.868 (0.868-0.868)	0.715 (0.714-0.716)	0.887 (0.887-0.887)
BiLSTM-ATT-2	0.869 (0.868-0.869)	0.702 (0.701-0.703)	0.888 (0.888-0.888)
LI			
LR	0.845 (0.845-0.846)	0.670 (0.669-0.670)	0.887 (0.887-0.887)
LSTM	0.857 (0.857-0.858)	0.667 (0.667-0.668)	0.887 (0.887-0.887)
BiLSTM	0.860 (0.860-0.860)	0.690 (0.689-0.691)	0.888 (0.888-0.888)
BiLSTM-ATT-1	0.868 (0.868-0.868)	0.712 (0.711-0.713)	0.887 (0.887-0.887)
BiLSTM-ATT-2	0.866 (0.865-0.866)	0.707 (0.706-0.707)	0.888 (0.887-0.888)
GPR			
LR	0.845 (0.844-0.845)	0.667 (0.666-0.668)	0.886 (0.886-0.886)
LSTM	0.866 (0.866-0.866)	0.696 (0.696-0.697)	0.887 (0.887-0.887)
BiLSTM	0.872 (0.872-0.872)	0.701 (0.700-0.702)	0.887 (0.887-0.887)
BiLSTM-ATT-1	0.866 (0.866-0.867)	0.715 (0.715-0.716)	0.889 (0.889-0.889)
BiLSTM-ATT-2	0.865 (0.865-0.866)	0.693 (0.692-0.694)	0.888 (0.888-0.888)
UA-BiLSTM-ATT-1	0.872 (0.872-0.873)	0.717 (0.716-0.718)	0.887 (0.887-0.887)
DEWS	0.880 (0.880-0.880)	0.729 (0.728-0.729)	0.887 (0.887-0.887)

TABLE V

PERFORMANCE EVALUATION OF DEWS IN COMPARISON TO NEWS ACROSS SUB-POPULATIONS OF INTEREST I.E. 16-45 YEARS OLD, >45 YEARS OLD, AND EACH OF THE THREE EVENTS IN THE COMPOSITE OUTCOME. THE RECOMMENDED DECISION THRESHOLD FOR NEWS IS 5 AND THE ADJUSTED DECISION THRESHOLD FOR DEWS WAS 0.66, TO ACHIEVE A SIMILAR OVERALL SPECIFICITY (≈0.89) ON $\mathcal{D}_2$. MEAN AND CONFIDENCE INTERVALS WERE EVALUATED USING A BOOTSTRAPPING TECHNIQUE ($nb = 1,000$) FOR THE RESPECTIVE SUB-POPULATION

Model	AUROC	Sensitivity	Specificity
16-45 years old			
DEWS	0.820 (0.818-0.822)	0.536 (0.532-0.539)	0.925 (0.925-0.925)
NEWS	0.760 (0.757-0.762)	0.522 (0.519-0.526)	0.935 (0.934-0.935)
> 45 years old			
DEWS	0.881 (0.881-0.882)	0.740 (0.739-0.741)	0.881 (0.881-0.881)
NEWS	0.869 (0.869-0.870)	0.713 (0.712-0.714)	0.880 (0.880-0.880)
Unplanned ICU			
DEWS	0.811 (0.811-0.812)	0.555 (0.554-0.557)	0.900 (0.900-0.900)
NEWS	0.772 (0.771-0.773)	0.528 (0.527-0.530)	0.900 (0.900-0.900)
Cardiac Arrest			
DEWS	0.767 (0.765-0.769)	0.457 (0.454-0.461)	0.900 (0.900-0.900)
NEWS	0.752 (0.751-0.754)	0.426 (0.423-0.430)	0.900 (0.900-0.900)
Mortality			
DEWS	0.926 (0.926-0.927)	0.831 (0.831-0.832)	0.888 (0.888-0.888)
NEWS	0.922 (0.922-0.922)	0.805 (0.805-0.806)	0.888 (0.888-0.888)

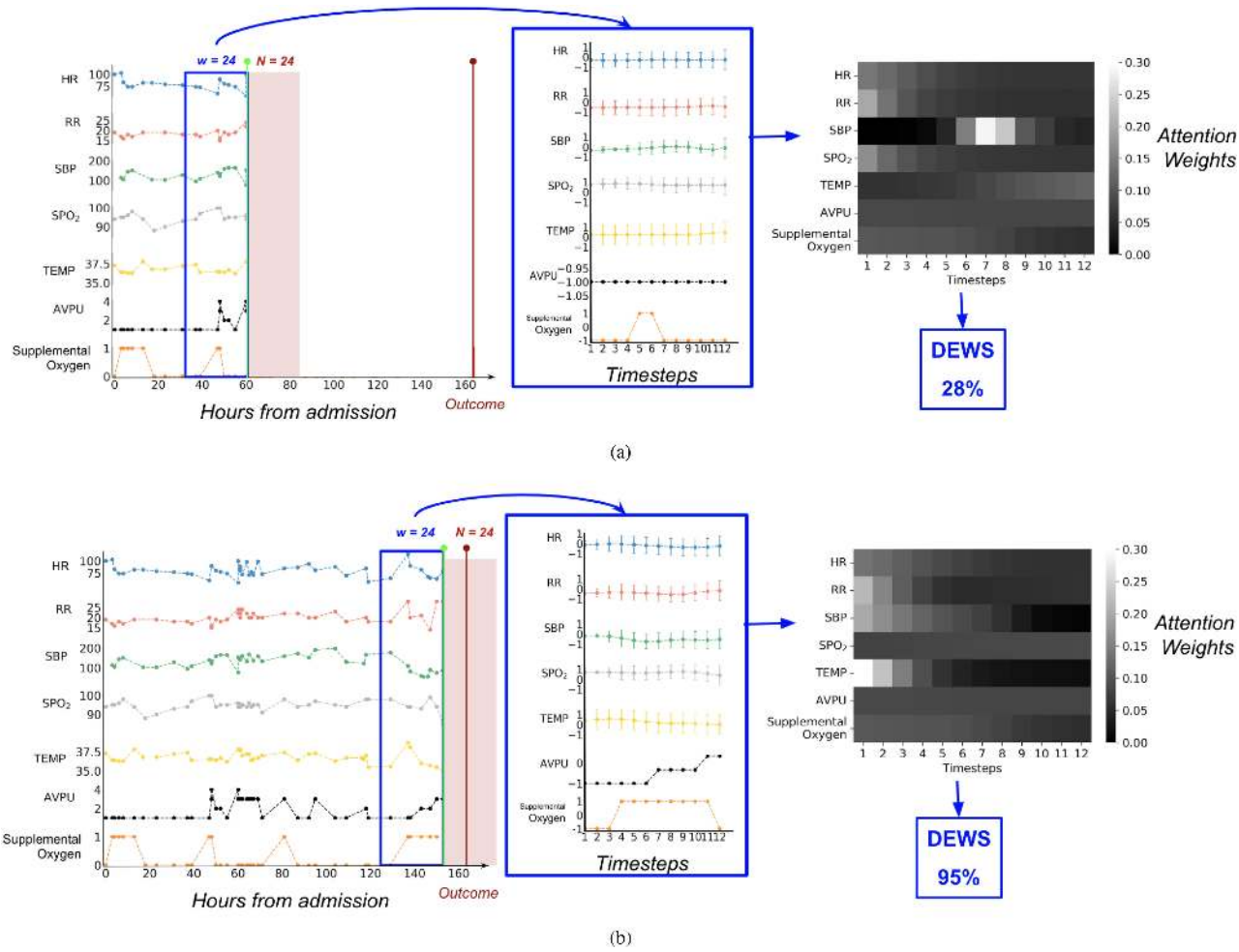

Fig. 4. Visualization of the attention weights learned from the mean and variance of all vital signs sequences using DEWS, where (a) is a non-event window and (b) is an event window, sampled from the same patient. The leftmost figure shows the raw data, where the green vertical line marks the time at which a prediction is made and the red vertical line marks the time of an outcome. The middle blue bordered figure shows the interpolated features within the window after GPR modelling and scaling.

event window since the time of prediction was within 24 hours of an outcome. In the attention weights of the non-event window, we observe that time steps 6–10 gained more importance than other time steps for SBP. When compared to the raw data, we notice that this trend corresponds to an increase followed by a decrease in SBP across the respective time steps. All other uniform distributions indicate that the model equally values each time step. The probability of an event produced by DEWS for this window was 28%, compared to a score of 6 by NEWS. This window was thus classified as a true negative by DEWS and a false positive by NEWS.

As for the second row, the event window, RR, SBP, and TEMP varied similarly, with a decreasing attention from left to right. In the original data, RR and TEMP sharply increased in the earlier time steps. SBP, on the other hand, decreased from a high value. In this scenario, DEWS produced a probability of 98%, while NEWS produces a score of 15, and as such both models produced a true positive.

In Fig. 5, the mean probability produced by DEWS model and the normalized NEWS score appear to be aligned in terms

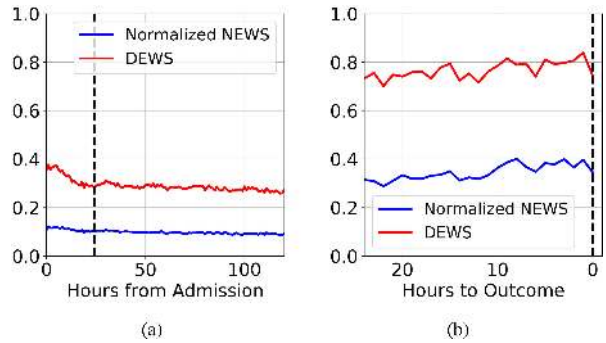

Fig. 5. Investigations of the mean probability of DEWS and the mean normalized NEWS score for (a) the first 120 hours from admission where the dashed line indicates 24 hours from admission, and (b) the last 24 hours prior to an outcome where the dashed line indicates the time of the composite outcome.

of overall trends. We notice in (a) that in the first 24 hours window from admission, where the thick dashed black line represents 24 hours from admission, the probability produced

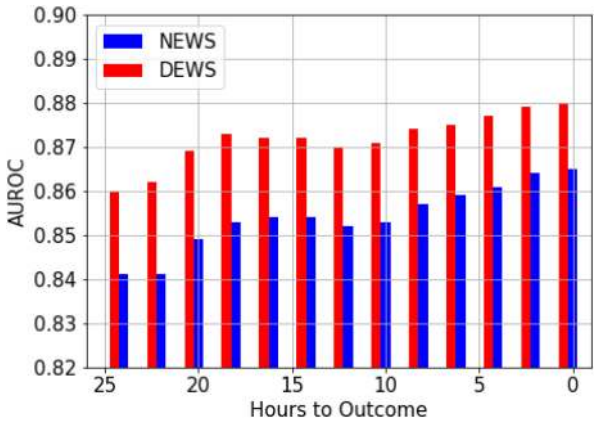

Fig. 6. Performance in terms of AUROC (y-axis) at every two hours to an outcome (x-axis) for DEWS and NEWS, in the last 24 hours window prior to the composite outcome.

by DEWS decreases, which suggests that DEWS gains confidence as the patient's length of stay increases. (b) In the last 24 hours window prior to an outcome, DEWS maintained a mean probability greater than its alerting threshold of 0.66, while NEWS maintained a mean normalized score at around 0.3. When comparing the models' AUROC against time to outcome, we observe that DEWS performed better than NEWS across the 24 hours window prior to outcome, as shown in Fig. 6.

V. Conclusion and Discussion

We propose an attention-based neural network that learns from historical trends of vital signs through interpolated mean and variance features to alert for clinical deterioration. Our proposed architecture DEWS achieved state-of-the-art performance, even while considering a limited set of features. DEWS decreased the number of triggers in comparison to NEWS, especially amongst younger patients. This would ease the burden on clinical staff in such a demanding environment.

Furthermore, our model performed best across the composite outcome and the three individual outcomes (unplanned ICU admission, cardiac arrest, and mortality). Improving the performance to alert for each outcome independently will be further investigated in future work by training outcome-specific models. However, in this paper, we chose a composite outcome to avoid further class imbalance and as what is being done in the literature [21].

Existing EWS systems only assess the most recently collected vital signs, as I.I.D. data. We demonstrate how historical trends of vital signs can provide potentially beneficial supplementary information. By examining the attention weights assigned for each vital sign, in Fig. 4, we were able to demystify the decision-making process of our deep learning model. For example, the clinician could examine why DEWS was alerting by inspecting the time frames where the attention weights were highest in Fig. 4, such as increasing RR or decreasing SBP. Such trend analysis can support designing interventions on hospital wards. The alignment of scores between NEWS and DEWS, which is further illustrated in Fig. 5, emphasizes their supplementary purposes. We envisage the system to provide the NEWS score, DEWS probability, and an overview of the importance of historical trends to the clinicians.

We also accounted for the correlations across vital signs using the trainable weights V_μ and V_σ, which learned the relationships across the aggregated context vectors of the vital signs. This incurred further computational complexity during training of the deep learning model, but only represented a forward pass during testing. Future work includes experimenting with multi-task GPR (MGP) to account for correlations during feature transformation. However, the computational cost of MGP is $O(m^3n^3)$, which is prohibitive for low-resource settings in comparison to $m \times O(n^3)$ for the univariate GPR [30].

From a deployment perspective, the score can be easily incorporated into existing hospital devices, such as bedside monitors or hand held devices since it uses the same data streams as EWS systems, i.e. NEWS, but it calculates the score differently. Displaying the attention weights on the screen would require ergonomics specific analysis, which is beyond the scope of this paper.

We focused on utilizing physiological time-series data to establish analogous grounds with EWS systems currently deployed in clinical settings. We considered incorporating diagnosis codes as in previous studies [17], [31], however we decided that the inclusion of diagnosis codes may be impractical in real-life clinical settings because they are usually assigned at discharge for billing purposes. One study introduced a relevant attention-based model to predict in-hospital mortality in ICU [32]. However, their proposed attention-based multivariate LSTM model did not achieve a better AUROC than its variant without attention for predicting mortality.

Other scores have incorporated laboratory tests [33], [34], yet the main objectives of our work were to inspect vital signs trends in real-time as in EWS and to use routinely-collected variables. We hypothesize that incorporating laboratory tests may marginally improve performance, and this is an area of future study.

Our proposed model was developed and validated on a private dataset as there are currently no publicly available datasets for non-ICU settings. Further assessment of the proposed methodology's generalisability is required with larger datasets and other toy classification problems using a multivariate input. We would also like to test our methods on other clinical prediction tasks through transfer learning.

References

[1] Royal College of Physicians, "National Early Warning Score (NEWS) - Standardising the assessment of acute-illness severity in the NHS. Report of a working party," Royal College of Physicians, London, U.K., Tech. Rep. July, 2012. [Online]. Available: https://www.ombudsman.org.uk/sites/default/files/National%20Early%20Warning%20Score%20%28NEWS%29%20-%20Standardising%20the%20assessment%20of%20acute illness%20severity%20in%20the%20NHS_0.pdf

[2] G. B. Smith, D. R. Prytherch, P. Meredith, P. E. Schmidt, and P. I. Featherstone, "The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death," *Resuscitation*, vol. 84, no. 4, pp. 465–470, 2013.

[3] M. Aczon *et al.*, "Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks," 2017. [Online]. Available: http://arxiv.org/abs/1701.06675

[4] J. Futoma, S. Hariharan, and K. Heller, "Learning to detect sepsis with a multitask Gaussian process RNN classifier," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1174–1182.
[5] N. Razavian, J. Marcus, and D. Sontag, "Multi-task prediction of disease onsets from longitudinal lab tests," in *Proc. 1st Mach. Learn. Healthcare Conf.*, 2016, pp. 1–27.
[6] H. Suresh, N. Hunt, A. Johnson, L. A. Celi, P. Szolovits, and M. Ghassemi, "Clinical intervention prediction and understanding using deep networks," 2017. [Online]. Available: http://arxiv.org/abs/1705.08498
[7] J. van der Westhuizen and J. Lasenby, "Bayesian LSTMs in medicine," 2017. [Online]. Available: http://arxiv.org/abs/1706.01242
[8] M. A. Ahmad, A. Teredesai, and C. Eckert, "Interpretable machine learning in healthcare," in *Proc. IEEE Int. Conf. Healthcare Inform.*, 2018, pp. 559–560.
[9] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[10] T. A. Lasko, J. C. Denny, and M. A. Levy, "Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data," *PLoS ONE*, vol. 8, no. 6, 2013, Art. no. e66341.
[11] D. A. Clifton and M. Pimentel, "Gaussian processes for personalized e-health monitoring with wearable sensors gaussian processes for personalized e-health monitoring with wearable sensors," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 193–197, Jan. 2013.
[12] M. Ghassemi, T. Naumann, T. Brennan, D. A. Clifton, and P. Szolovits, "A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 446–453.
[13] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 6000–6010.
[14] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Learn. Representations*, 2015.
[15] P. Schwab, G. Scebba, J. Zhang, M. Delai, and W. Karlen, "Beat by Beat: Classifying cardiac arrhythmias with recurrent neural networks," 2017. [Online]. Available: http://arxiv.org/abs/1710.06319
[16] S. P. Shashikumar, A. J. Shah, G. D. Clifford, and S. Nemati, "Detection of paroxysmal atrial fibrillation using attention-based bidirectional recurrent neural networks," 2018. [Online]. Available: http://arxiv.org/abs/1805.09133
[17] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 1903–1911.
[18] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "GRAM: Graph-based attention model for healthcare representation learning," *23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 787–795. [Online]. Available: http://arxiv.org/abs/1611.07012
[19] Y. J. Kim, Y.-G. Lee, J. W. Kim, J. J. Park, B. Ryu, and J.-W. Ha, "High risk prediction from electronic medical records via deep attention networks," 2017.
[20] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 1480–1489.
[21] P. J. Watkinson, M. A. Pimentel, D. A. Clifton, and L. Tarassenko, "Manual centile-based early warning scores derived from statistical distributions of observational vital-sign data," *Resuscitation*, vol. 129, pp. 55–60, 2018.
[22] D. R. Prytherch, G. B. Smith, P. E. Schmidt, and P. I. Featherstone, "ViEWS-Towards a national early warning score for detecting adult inpatient deterioration," *Resuscitation*, vol. 81, no. 8, pp. 932–937, 2010.
[23] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[24] D. C. Montgomery and R. C. George, *Applied Statistics and Probability for Engineers*, 6th ed. Hoboken, NJ, USA, Wiley, 2014.
[25] D. Wong, T. Bonnici, J. Knight, L. Morgan, P. Coombes, and P. Watkinson, "SEND: A system for electronic notification and documentation of vital sign observations," *BMC Med. Inform. Decis. Making*, vol. 15, 2015, Art. no. 68.
[26] K. G. Moons *et al.*, "Transparent reporting of a multivariable prediction model for individual prognosis or disagnosis (TRIPOD): Explanantion and elaboration," *Ann. Internal Med.*, vol. 162, no. 1, pp. W1–W74, 2015.
[27] GPyOpt, "A Bayesian Optimization framework in python," 2016. [Online]. Available: http://github.com/SheffieldML/GPyOpt
[28] F. Chollet and others, "Keras," 2015. [Online]. Available: https://keras.io
[29] A. Martin and Others, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: http://tensorflow.org/
[30] R. Dürichen, M. A. F. Pimentel, L. Clifton, A. Schweikard, and D. A. Clifton, "Multitask Gaussian processes for multivariate physiological time-series analysis," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1, pp. 314–322, Jan. 2015.
[31] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Proc. NIPS*, 2016, pp. 3504–3512.
[32] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: clinical time series analysis using attention models," 2017. [Online]. Available: http://arxiv.org/abs/1711.03905
[33] A. Rajkomar and Others, "Scalable and accurate deep learning for electronic health records," *Nature Digit. Med.*, no. 1, pp. 1–10, 2018.
[34] M. M. Churpek *et al.*, "Multicenter development and validation of a risk stratification tool for ward patients," *Am. J. Respiratory Crit. Care Med.*, vol. 190, pp. 649–655, 2014.