Deep Learning Based Sepsis Intervention: The Modelling and Prediction of Severe Sepsis Onset

Gavin Tsang
Department of Computer Science
Swansea University, UK
Email: 658679@swansea.ac.uk
http://csvision.swan.ac.uk

Xianghua Xie
Department of Computer Science
Swansea University, UK
Email: X.Xie@swansea.ac.uk
http://csvision.swan.ac.uk

***Abstract*—Sepsis presents a significant challenge to healthcare providers during critical care scenarios such as within an intensive care unit. The prognosis of the onset of severe septic shock results in significant increases in mortality rate, length of stay and readmission rates. Continual advancements in health informatics data allows for applications within the machine learning field to predict sepsis onset in a timely manner, allowing for effective preventative intervention of severe septic shock. A novel deep learning application is proposed to provide effective prediction of sepsis onset by up to six hours prior, involving the use of novel concepts such as a boosted cascading training methodology and adjustable margin hinge loss function. The proposed methodology provides statistically significant improvements to that of current machine learning based modelling applications based off the Physionet Computing in Cardiology 2019 challenge. Results show test F1 scores of 0.420, a significant improvement of 0.281 as compared to the next best challenger results.**

***Index Terms*—Deep Learning, Machine Learning, Electronic Medical Records, Health Informatics, Sepsis.**

I. Introduction

Severe sepsis and septic shock is a life-threatening condition with a high prevalence, resulting in significant mortality and expense in healthcare. Despite modern medical advances in antibiotics and acute care, sepsis remains the primary cause of death from infection [1]. Sepsis presents highly challenging concerns for practitioners within an intensive care unit (ICU) setting. UK prognosis statistics of a patient indicating sepsis show a 35% mortality rate during ICU stay [2], 47% mortality rate during hospital spell [2] and a 63% rate of hospital readmission within the 1st year [3], highlighting the significant dangers of sepsis. With a high prevalence rate of 27.1% of adults meeting severe sepsis criteria within the 24 hours of ICU admission [2], such dangers remain at the forefront of intensive care medicine.

With a still uncertain understanding in the pathophysiology of sepsis, diagnostic procedure for sepsis has undergone significant change over recent years, from the original sepsis definition of the 1980s [4] to the distinction of severe sepsis and septic shock in 1991 using the systematic inflammatory response syndrome (SIRS) definition [5], later renamed to the sequential organ failure assessment (SOFA) criteria, to the modern-day quick SOFA (qSOFA) system of diagnosis, developed in 2016 [1]. The simplicity of the qSOFA system, 2 of 3 indications of low blood pressure, high respiratory rate or altered state of consciousness enabled quick clinical assessment whilst still maintaining effective discriminative quality (AUROC score of 0.72) [1].

The importance in early detection and treatment of sepsis symptoms in both medical definition, diagnosis procedure and treatment strategy is apparent [6]. The UK treatment strategy, the *Sepsis Six*, emphasises intervention within the first hour of initial suspicion [6], [7]. Perhaps the greatest apparent significance of the importance in early detection is a 5-8% increase in mortality per hour for sepsis and septic shock respectively when left untreated [6], [8]. To this end, the early detection of sepsis in advance of clinical indication through the qSOFA guidelines presents potentially improved life-saving intervention and treatment to that of current medical procedures.

A review of recent relevant papers indicating the use of machine learning (ML) based applications on sepsis prediction show multiple studies performed over a period of 2016-2019 highlighting 11 relevant papers. Of all papers reviewed, studies remain solely focused on ICU based settings presumably due to the ubiquity of high-frequency and detailed patient recordings within such a environment. Exceptions include Masino *et al.* [9] using patient records from a neonatal ICU unit (<1 year age) and Le *et al.* [10] using paediatric inpatient and emergency records (2-17 years age).

All studies use similar dataset characteristics of hourly recorded features falling into categories of patient vital signs (Heart rate, oxygen saturation, blood pressure, etc.), laboratory indications (Lactic acid, Platelet count, urine output, etc.) and demographics (age, gender, race, etc.) with a goal of prediction of sepsis onset 3-4 hours early to that of recorded medical suspicion. Dataset populations range drastically from 140 to 32,000 patients with data collection periods spanning back to 2001.

Several studies directly include SOFA or qSOFA [11], [12], the current established system of diagnosis, as a predictive feature highlighting the use of ML applications as a supplementary early warning system as opposed to a replacement to the qSOFA system. In contrast, several studies perform direct comparisons of novel applications against SOFA [12]–[14] or SIRS [10], [12]–[16], all of which indicate performance improvements over SOFA or SIRS.

Faisal *et al.* [11] uses a classic logistic regression model on

transformed features whilst Horng *et al.* [17] uses a classic support vector machine (SVM) on encoded representations of text based records. Both Le *et al.* [10] and Delahanty *et al.* [13] use gradient boosted decision trees with non-transformed features for the former and with the inclusion of 'engineered features' by the latter. Van Wyk *et al.* [18] provides a comparison of predictive performance on random forest (RF), SVM, logistic regression, neural network (NN) and recurrent neural network (RNN) models with RF outperforming the remainder by a statistically significant margin. The most popular methodology was the InSight methodology proposed by Desautels *et al.* [14] with 5 of the 11 studies comparing or involving it's use [9], [12], [14]–[16]. One paper, by Kam *et al.* [16], involves the use of deep learning (DL) based methodologies.

Modern information technology provides a unique opportunity to incorporate machine learning applications with continually monitored patient physiological data to produce an effective risk analysis application to identify septic patients in an earlier time-frame than that of the established diagnosis systems. As such, we propose a DL based application using said patient data. The goal of which is to effectively predict patient onset of Sepsis at least six hours earlier than official medical diagnosis. Whilst such applications already exist, as indicated previously, we contribute a highly novel methodology towards prediction with statistically significant improved performance as compared to state-of-the-art technologies. Said novel methodologies include a new boosted cascading sub-network training procedure, specialised in effective prediction on highly imbalanced datasets. The novel shifting-margin hinge loss function provides adaptive over-fitting protection on the continually increasing parameter-space complexity of the aforementioned boosted cascade training procedure. Further novelty comes from the tailored problem-based approach of the Critical Diagnosis-Point Penalty loss function and the Negative Reversal Penalty loss function. The combination of said methodologies provide a highly effective sepsis prediction application, outperforming current state-of-the-art methodologies with statistical significance.

II. Dataset

Evaluation will be based off the PhysioNet Computing in Cardiology Challenge 2019 dataset and associated challenge participants [19]. Consequently, evaluation methodology and metrics will follow as specified within the study by Reyna *et al.*, allowing for direct comparison of our proposed methodology to the published challenge participant leaderboard.

Said dataset, sourced from ICU patient records from two separate hospital units, henceforth labelled as dataset A and B respectively, contain individual patient level timelines providing 40 unique features categorized into vital signs, lab test results and demographic information as shown in table I. Patient timelines are organised as hourly snapshots representing all available patient data recorded within said hour time window. Consequentially, proportion of missing data within each hourly recording is extreme. As highlighted in table I, lab test result features and vital sign features average 32.4% and 94.9% missing data respectively, whilst static per patient demographic features contain no missing data.

TABLE I
Data attributes and missing data percentage of the PhysioNet CinC 2019 challenge dataset

Attribute	Missing Data (%)	Attribute Details
	Vital Signs	
HR	9.9	Heart rate (beats per minute)
O2Sat	13.1	Pulse oximetry (%)
Temp	66.2	Temperature (Deg C)
SBP	14.6	Systolic BP (mm Hg)
MAP	12.5	Mean arterial pressure (mm Hg)
DBP	31.3	Diastolic BP (mm Hg)
Resp	15.4	Respiration rate (breaths per minute)
EtCO2	96.3	End tidal carbon dioxide (mm Hg)
	Laboratory Values	
BaseExcess	95.8	Measure of excess bicarbonate (mmol/L)
HCO3	95.8	Bicarbonate (mmol/L)
FiO2	91.7	Fraction of inspired oxygen (%)
pH	93.1	N/A
PaCO2	94.4	Partial pressure of carbon dioxide from arterial blood (mm Hg)
SaO2	96.5	Oxygen sat from arterial blood (%)
AST	98.4	Aspartate transaminase (IU/L)
BUN	93.1	Blood urea nitrogen (mg/dL)
Alkalinephos	98.4	Alkaline phosphatase (IU/L)
Calcium	94.1	(mg/dL)
Chloride	95.5	(mmol/L)
Creatinine	93.9	(mg/dL)
Bilirubin_direct	99.8	Bilirubin direct (mg/dL)
Glucose	82.9	Serum glucose (mg/dL)
Lactate	97.3	Lactic acid (mg/dL)
Magnesium	93.7	(mmol/dL)
Phosphate	96.0	(mg/dL)
Potassium	90.7	(mmol/L)
Bilirubin_total	98.5	Total bilirubin (mg/dL)
TroponinI	99.0	Troponin I (ng/mL)
Hct	91.1	Hematocrit (%)
Hgb	92.6	Hemoglobin (g/dL)
PTT	97.1	partial thromboplastin time (seconds)
WBC	93.6	Leukocyte count (count*10^3/μL)
Fibrinogen	99.3	(mg/dL)
Platelets	94.1	(count*10^3/μL)
	Demographics	
Age	0.0	Years (100 for patients 90 or above)
Gender	0.0	Female (0) or Male (1)
Unit1	0.0	Admin identifier for ICU unit (MICU)
Unit2	0.0	Admin identifier for ICU unit (SICU)
HospAdmTime	0.0	Hours between hospital and ICU admit
ICULOS	0.0	ICU length-of-stay (hours)
	Class Labels	
SepsisLabel	0.0	Positive sepsis (1) otherwise (0)

Class labels are defined for each hourly timestep as binary label, positive or negative sepsis. A positive sepsis event is defined within the challenge by Reyna *et al.* [19] as a clinical suspicion of infection, portrayed by medical events dictating antibiotic administration with blood lab culture testing within a certain time period, or an official designation of sepsis via a two point deterioration in SOFA score. As the objective of this study is a six hour early prediction of sepsis, positive class labels are given six hours prior to said positive sepsis event, and continue onwards to end of patient timeline.

Overall population demographics indicate an average age of 61.6 years with a 16.5 year standard deviation and minimum and maximum ages of 14 and 90+. Gender proportions within the considered population are 22,566 males to 17,770 females, providing a total considered population of 40,336 ICU patients. In regards to sepsis prevalence, 2932 patients have a sepsis positive event compared to the remaining 37,404 ICU patients indicating no signs of sepsis.

III. Methodology

Due to high proportions of data sparsity within certain features, missing values for each patient were artificially generated by linear interpolation of surrounding available data. Instances of patient's lacking any data for specific features were artificially set to a constant average value based off a population subset of 50 individuals of the same gender and closest similar age, with no consideration for sepsis status of said individuals. Data normalization was performed afterwards to soft limit each feature's overall population value range to between 0 and 1.

A. Long Short-Term Memory

The foundational basis of our proposed methodology, the long short-term memory (LSTM) [20] provides the capability to maintain a memory of previous samples called the cell state, C_t. Such memory provides effective application on variable time-series based data problems by enabling consideration of multiple previous timestep data embeddings within the prediction of the current timestep.

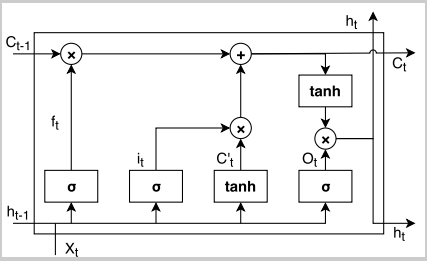

Fig. 1. A diagram of a singular LSTM cell. As seen, the LSTM is comprised of multiple activation, weight pairs to form the four components intrinsic to the LSTM. Input data is passed into the LSTM to the input gate controlling data transformation into an embedded state for the update and forget gate to modify memory cell state. The output gate takes in said modified cell state in addition to the input data to produce the final activation output of the LSTM cell.

Let w and b, be learnable parameters unique to each LSTM component and let x_t be the incoming input from a previous layer or data in addition to the activation output of the considered cell during the previous timestep, h_{t-1}. The concatenated vector of both x_t and h_{t-1} are passed into the first LSTM component, the update gate, which controls the update procedure of the incoming previous timestep cell state, C_{t-1} using the signals generated by the forget and input gate, f_t and $i_t C'_t$ respectively, through the following equation

$$C_t = f_t C_{t-1} + i_t C'_t \tag{1}$$

Each component of (1) is dictated by the remaining components of a LSTM memory cell, learnt through a combined input vector $[x_t, h_{t-1}]$. The removal of irrelevant memory-components within the cell state is dictated by the learned filtering function of the forget gate,

$$f_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f) \tag{2}$$

The input gate combines both input encoding, C'_t and selective filtering, i_t of the input vector to produce the new cell state:

$$C'_t = tanh(W_c \cdot [h_{t-1}, X_t] + b_c) \tag{3}$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \tag{4}$$

Finally, memory cell output is controlled by incorporating all previous components to produce the final activation of the LSTM, h_t:

$$h_t = O_t \cdot tanh(C_t) \tag{5}$$
$$O_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \tag{6}$$

As seen, a singular LSTM cell increases learnable parameters, w and b, significantly as compared to a RNN cell or traditional neuron. Such increase in overall parameter capacity however, provides significant improvements towards time-series based modelling applications.

B. Boosted Cascading Sub-networks

A major consideration is the general imbalance of classes, skewed towards the condition negative class. With a patient time-window between patient admittance into ICU to just after diagnosis of sepsis, the resulting timeline is defined by mostly negative timesteps. The resulting dataset consists of 1,524,294 negative timesteps compared to 27,916 positive timesteps. Traditional methodologies used to remedy class imbalance such as over-sampling and under-sampling result in limited effectiveness [21]. As such, we propose a novel combined model architecture and training strategy, influenced by a combination of cascade and boosting ensemble training algorithms, which seeks to improve upon current class imbalance prediction methodologies. The cascade approach of our methodology aims to iteratively eliminate easy-to-predict samples belonging to the imbalanced negative class. The ranking of said samples is controlled by the boosting aspect of our methodology which seeks to highlight difficult edge-case samples for later cascade stages. The remaining increasingly difficult samples are thus matched with an increasing model parameter capacity with larger architectures as cascades are built.

Let the initial cascade sub-model, $m = 0$ be trained in a traditional manner as indicated in (7), the minimization of overall model loss L^m:

$$L^m = \arg\min_{L} \sum_{i=0}^{N} \left(w_i^m \sum_{t=0}^{T} (L(f(x_{t,i})) \right) + L_p^m \quad (7)$$

Of note, is the included per-patient, i adaptive weighting factor, $w_i^{m=0}$ dictating overall patient importance to the model loss. Training emphasis on said initial cascade sub-model is predominantly placed on patients containing a positive sepsis diagnosis event at any point during their timeline via a manually defined high initial patient weighting factor $w_i^{m=0}$. The goal of which, is to produce an initial cascade sub-model emphasising high to near perfect true positive rate (TPR) and negative predictive value (NPV) at the cost of low positive predictive value (PPV). Such prediction properties from said model, represents the first course screening of certain negative predictions in the cascade.

Subsequent cascade sub-models are identically trained using an updated patient weighting factor based off of previous model performance as indicated:

$$w_i^m = (1 - \lambda_w) w_i^{m-1} + \frac{\lambda_w}{T} \sum_{t=1}^{T} \left| y_{t,i} - \hat{y}_{t,i}^{m-1} \right| \quad (8)$$

Individual patient weightings are slowly pushed towards greater or lesser importance based off the mean absolute error of the previous model prediction, $\hat{y}_{t,i}^{m-1}$ on the entire patient timeline, T and true label, $y_{t,i}$ representing the boosting aspect of said training strategy.

The importance factor of subsequent cascades is regulated by the weight hyperparameter coefficient, $\lambda_w \in \{0 < \mathbb{N} < 1\}$. With consideration of the initial patient weighting factor contributing entirely to a cascade style course screen, subsequent weight updates push said weighting factor towards a boosting style weighting application. Consequently, the weight coefficient hyperparameter dictates the proportion of influence by subsequent weight updates where larger values push the weighting factor more rapidly towards a pure boosting strategy with focus on misclassification whilst small values maintain the cascading approach.

Repetitions of the cyclical cascade generation and iterative patient weight factor updates are dictated by a maximum cascade hyperparameter or the reduction of validation loss improvement between subsequent cascades to a minimum improvement level controlled by hyperparameter.

Model prediction follows a cascading screening approach with samples passed through each cascade, m in order with positive predictions, $\hat{y}_{t,i}^m > 0.5$, passed to subsequent cascades whilst negative predictions filter out from each cascade as negative overall predictions. Remaining samples evaluated by the bottom-most cascade layer as negative, are placed with the previous set of negative predictions whilst positive samples finally represent the final positive predictions of the overall cascade model.

The generation of multiple cascading sub-networks within our training procedure affords opportunity for the adjustment of sub-model complexity. Under the assumption that deeper cascades gradually focus on greater discrimination of the between class boundaries across complex edge case patient samples, through the aforementioned boosting and cascade sample weighting technique, greater model parameter capacity can potentially be afforded to improve upon said discrimination. Sub-model architecture is consequently constructed as two hidden layers of initially 16 and 8 LSTM nodes, each additionally containing a batch normalisation and 25% dropout layer, for the initial top cascade. Each subsequent cascade increases LSTM node count of both hidden layers by an additional 25% of the previous cascade respectively each time. As such, our overall architecture of six generated cascade sub-models, results in the final cascade sub-model containing hidden layers of 31 and 61 LSTM nodes respectively.

C. Shifting Margin Hinge Loss

The continually increasing model capacity of each cascade with the continually reducing sample diversity and size provides great over-fitting risk within further cascades. As such, this study introduces the novel shifting margin hinge loss to further manage discrimination complexity across the class boundary whilst simultaneously dampening potential over-fitting issues from the increased model capacity.

The shifting margin hinge loss is an adaption to the traditional concept of constructing linear decision boundaries based off the maximisation of distance between differing classes. Let the considered linear decision boundary be defined as the hyperplane, $x_i^T\beta + \beta_0$ defined by parameters β and β_0 where sample feature vectors are indicated by $x_i, i = 1, \ldots, N$. The separation of samples can thus be indicated by $y_i^+(x_i^T\beta + \beta_0) > 0$ and $y_i^-(x_i^T\beta + \beta_0) < 0$ where y_i^+ and y_i^- indicate positive and negative class samples respectively.

Since, within a linearly separable dataset, there are infinitely many combinations of β which allows for correct separation, constraints must be defined to limit solutions to a unique and optimal boundary. This is traditionally done by defining the optimal separating hyperplane to maximise the distance between the margin and the closest samples of each class. However, such simplistic problem spaces are rare, with the vast majority being of the non linearly separable type with generally overlapping class distributions. As a result, slack is generally built into the formulation of the optimisation problem, in order to allow for slight overlap of samples between the decision boundary. As such, the optimisation problem with margin constraints and slack becomes:

$$\max_{\beta, \beta_0, \|\beta\|} M \quad (9)$$
$$\text{subject to } y_i(x_i^T\beta + \beta_0) \geq M - \xi_i, i = 1, \ldots, N$$

where $\xi = \{\xi_1, \ldots, \xi_N\}, \xi_i \geq 0, \sum_{i=1}^{N} \xi \leq C$, the slack, measures the distance overlap of incorrect samples from the margin. The constant C bounds the total proportional distance allowed by predictions to lie on the wrong side of the margin.

With consideration of $\|\beta\| = 1$ and that any positively scaled multiple of β and β_0 will satisfy (9), $\|\beta\|$ can be arbitrarily set to $1/M$. Thus arriving at the minimisation problem of:

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i \quad (10)$$
$$\text{subject to } \xi_i \geq 0, y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i$$

resulting in the final form of the maximum margin classifier, generally associated with the SVM methodology. The reformulation of (10) into the Lagrangian dual form, hence becoming a convex optimisation problem, can thus be easily solved using standard quadratic programming solutions.

In order to facilitate the increase or decrease of the overall margin size to formulate the proposed shifting margin hinge loss, a further hyperparameter coefficient, λ_m, is defined dictating margin size proportion, $M' = \lambda_m M$. As such, (9) and the resulting optimisation problem, (10) becomes:

$$y_i(x_i^T\beta + \beta_0) \geq \lambda_m M - \xi_i \quad (11)$$

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i \quad (12)$$
$$\text{subject to } \xi_i \geq 0, y_i(x_i^T\beta + \beta_0) \geq \lambda_m - \xi_i$$

respectively. The use of the aforementioned Lagrangian dual form to solve said optimisation problem, would simply result in the maximisation of the decision boundary margin eliminating any influence of the shifted margin. However, the optimisation of the decision boundary via stochastic gradient descent does maintain the influence of the shifted margin. As such, the final shifting margin hinge loss function can be derived in a similar manner to that of standard hinge loss. With consideration of $\xi_i \geq 0$ indicating ξ_i must be a positive real value, the constraints defined by (12), $y_i(x_i^T\beta+\beta_0) \geq \lambda_m - \xi_i$ can be redefined as a loss function to be part of the overall minimisation problem:

$$\xi_i = \max(0, \lambda_m - y_i(x_i^T\beta + \beta_0)) \quad (13)$$

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\max(0, \lambda_m - y_i(x_i^T\beta + \beta_0)) \quad (14)$$

In regards to the proposed model being a deep NN architecture, the original SVM model function, $\lambda_m - y_i(x_i^T\beta + \beta_0)$ can be replaced and obfuscated to $\hat{y}_i$ indicating the overall probability prediction of said NN. Through conversion to a standard NN loss function minimisation, the initial term within (14), $\frac{1}{2}\|\beta\|^2$ becomes traditional L2 weight regularisation, where β is the equivalent of model learnable parameters w. C becomes a constant factor dictating proportional weight of model loss compared to said weight regularisation and can be rearranged in favour of the traditional L2 weight regularisation factor and thus removed from overall consideration. Finally, the proposed shifting margin hinge loss can be defined as follows:

$$L = \sum_{i=1}^{N}\max(0, \lambda_m - y_i\hat{y}_i) \quad (15)$$

where y are timestep class labels with positive and negative class values being: $y^+ = 1, y^- = -1$. Model predictions are defined as $\hat{y}$ and λ_m is the hyperparameter coefficient dictating margin size where $\lambda_m \in \{\mathbb{R} \geq 0\}$.

D. Critical Diagnosis Point Penalty

The previous methodology proposals within this study have thus focused on general applications on an unbalanced dataset, however there remains opportunity for tailored problem-specific approaches based upon our problem space of sepsis prediction on time-series data. Various unique intricacies exist within our problem space. By acknowledging and adapting such intricacies into our application, further performance improvements can be made.

For an individual sepsis positive patient's timeline, a patient will present a continual degradation of vitals until sepsis is apparent and medically diagnosed. Sepsis can thus be considered a gradual progression up until a distinct medical diagnosis event within the patient's timeline. A patient will be sepsis negative up until said event, at which point, sepsis is continually present. The application objective can alternatively be expressed as a pseudo regression based task, predicting the critical sepsis diagnosis event. A traditional binary classification approach would be unable to take into account such a concept, instead applying binary predictions across a timeline without consideration of this singular critical diagnosis point.

A regression based methodology however is also inappropriate for such an application. With patient vitals being continually streamed per timestep with a requirement to indicate sepsis as a binary decision at current timestep, this prohibits a regression style prediction of time of sepsis occurrence. The traditional timestep per timestep binary classification is consequently the only practical application available. However, said regression based rationalization can still be leveraged into a binary classification application.

The critical diagnosis point penalty function is proposed to emphasise said critical diagnosis point. By introducing a penalty to true positive predictions based on distance away from the critical diagnosis point. Additionally, a secondary penalty is introduced greatly penalising late starts to initial positive predictions past the critical point based off false negative predictions. Through the balancing of both penalties, , we seek to drive the initial start of positive predictions closer towards said critical diagnosis point.

Let $\hat{y}_t$ be the model class prediction (probability output rounded towards the closest class), the true positive penalty function (17) is dictated by four continuous piecewise linear functions with joining knots dictated by the hyperparameter time periods, t_{early}, t_{opt} and t_{late}. Where t_{sepsis} indicates the clinical diagnosis point within a condition positive patient timeline allowing for the function to be shifted to the correct time period. Time period t_{opt}, indicates the point of no penalty and indicates the critical diagnosis point, six hours early to t_{sepsis}. Consequently, timesteps t extending away from t_{opt}

incur a linearly increasing penalty.

$$C_C = \lambda_C \sum_{t=0}^{N(\hat{y}_t)} \begin{cases} C_{TP}(t - t_{sepsis}), & \text{if } \hat{y}_t \text{ is TP} \\ C_{FN}(t - t_{sepsis}), & \text{if } \hat{y}_t \text{ is FN} \end{cases} \quad (16)$$

$$C_{TP}(t) = \begin{cases} \lambda_{TP} + \lambda_e, & \text{if } t < m_1(\lambda_{TP} + \lambda_e) + b_1 \\ m_1(t) + b_1, & \text{else if } t < t_{opt} \\ m_2(t) + b_2, & \text{else if } t < t_{late} \\ \lambda_{TP}, & \text{otherwise} \end{cases} \quad (17)$$

$$C_{FN}(t) = \begin{cases} \lambda_{TP}, & \text{if } t < t_{opt} \\ m_3(t) + b_3, & \text{else if } t < t_{late} \\ 1, & \text{otherwise} \end{cases} \quad (18)$$

where

$$m_1 = \frac{-\lambda_{TP}}{(t_{opt} - t_{early})}, \qquad b_1 = -m_1 t_{opt},$$

$$m_2 = \frac{\lambda_{TP}}{(t_{late} - t_{opt})}, \qquad b_2 = -m_2 t_{opt},$$

$$m_3 = \frac{1 - \lambda_{TP}}{(t_{late} - t_{opt})}, \qquad b_3 = -m_3 t_{late} + 1$$

The false negative penalty function (18) is similarly dictated by two continuous piecewise functions with a zero constant piecewise penalty function attached to a joining knot at t_{opt}, at which point a linearly increasing penalty is induced penalising a continually later and later initial positive prediction.

The hyperparameter coefficient, $\lambda_{TP} \in \{0 \leq \mathbb{R} \leq 1\}$ dictates the proportion of the overall penalty afforded to the true positive penalty as opposed to the false negative penalty and is used to balance the driving force in either direction of the initial positive prediction. The hyperparameter coefficient, λ_C dictates overall weight of the critical diagnosis point penalty within the overall training minimisation function.

E. Negative Reversal Penalty

Within an application standpoint, the self recovery of severe sepsis within a patient without medical intervention is not possible. Since, the considered dataset focuses on a time period of before and just after sepsis onset, no such sepsis recovery event may reasonably occur within the time window. As such, the proposed negative reversal penalty imposes a linearly increasing penalty on a reversion to a negative prediction state on any timestep post initial sepsis prediction, indicating a sepsis recovery event. Said linearly increasing penalty to the model loss is based on the time interval between first positive prediction and subsequent reversal to negative, if any. Accordingly, the penalty of each patient timeline is the summation of values resulting from the following condition.

$$C_N = \lambda_N \sum_{t=0}^{N(\hat{y}_t)} \begin{cases} t' - t, & \text{if } \hat{y}_t = 0 \wedge \exists\ \hat{y}_{t'} = \\ & \quad 1 \in \{\forall \hat{y}_{t'} : t' < t\} \\ 0, & \text{otherwise} \end{cases} \quad (19)$$

For each timestep model prediction rounded to the closest positive 1, or negative 0, prediction $\hat{y}_t$, if there is a negative prediction and any previous timestep prediction $\hat{y}_{t'}$ was positive, a linearly increasing penalty is imposed, $t' - t$, based on the time interval between initial positive prediction timestep t', and current negatively predicted timestep t. The penalty weight imposed on the overall model loss is dictated by hyperparameter coefficient λ_N, allowing for the tuning of model behaviour to enforce a continuous positive prediction post critical point whilst dampening the effect of erroneous early false positives. Marginal movement of the critical point per training iteration is possible by the relatively small penalty to early negative prediction reversals as opposed to the heavy penalisation of late negative predictions.

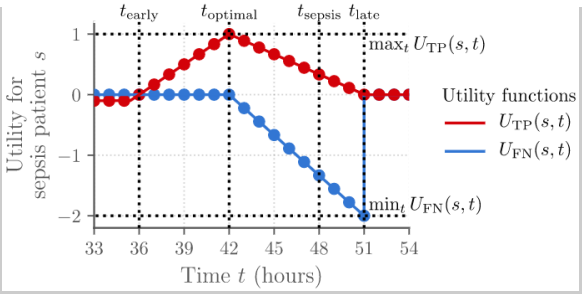

Fig. 2. Graph indicating the scoring system of the 'utility score' evaluation metric proposed by Reyna *et al.* [19] emphasising closeness of initial indication of sepsis by an application to that of the optimal early clinical diagnosis point. Model predictions are mapped to said utility score function to indicate score gain or loss at each timestep prediction. Of note, a true negative prediction accrues no score whilst a false positive prediction accrues a constant -0.05 utility score.

IV. Results

Experimental procedure and evaluation follows those proposed by the original Physionet 2019 challenge. The aforementioned two datasets detailed in section II, dataset A and B, form the training and testing sets respectively and are alternated to produce two training-testing pairs. Consequently, direct comparisons in model performance can be made between our proposed methodology and submission entry results detailed within the study by Reyna *et al.* [19].

Reyna *et al.*proposes a custom 'utility score' metric, emphasising the importance of condition positive sepsis patients with a linearly weighted score impact on early and late initial indications of sepsis as compared to the optimal six hour early clinical diagnosis point, t_{optimal}. As shown in Fig. 2, true positive initial indications of sepsis onset close to t_{optimal} are encouraged with a linearly increasing positive score contribution within a certain time window of said point. Simultaneously, too early or too late of an indication of sepsis by an application are penalised in a linearly increasing fashion. Subsequently, true negative predictions do not count towards utility score whilst false positives induce a small constant score penalty.

The total utility score U_{total}, of an evaluated model is thus the sum utility score of every patient timestep before normalisation. Utility score is normalised to the upper theoretical maximum total utility score of an optimal classifier with

perfect accuracy and a lower theoretical minimum total utility score of an inactive classifier (all predictions are negative):

$$U_{\text{normalised}} = \frac{U_{\text{total}} - U_{\text{inactive}}}{U_{\text{optimal}} - U_{\text{inactive}}} \tag{20}$$

In regards to training procedure hyperparameters, the NN architecture is an initial cascade architecture of 40 input layer followed by two LSTM node hidden layers of 16 and 8 nodes respectively, including batch normalisation and dropout layers at 25% dropout proportion after each hidden layer. Output consists of a single neuron with a linear activation function. Hidden layer node count increases by 25% of the previous cascade, each cascade. Training was performed using the Adam stochastic gradient descent methodology with parameters, lr $= 0.001$, $\beta_0 = 0.9$, $\beta_1 = 0.999$, $\epsilon = 1 \times 10^{-7}$ and batch size $= 1000$. Training was repeated till convergence, indicated by minimum delta model loss improvement of 0.0001, followed by selection of epoch model weights based on best validation loss results. Cascade generation stopping criteria was dictated by an overall minimum delta model validation loss improvement of 0.0001 arriving at a consistent six cascade sub-models for every experimental run.

A. Experimental Results

Following on are the overall results of the proposed methodology trained on dataset A and tested on B, henceforth labelled as 'Set A' and *vice versa* for 'Set B' respectively. A 5 k-fold cross validation procedure was performed on the combined A and B dataset, labelled 'Set A&B'. Table III provides challenge results of the top 5 ranking teams with associated metrics in addition to the relevant metrics of this study's proposed methodology. Of note, challenge team results show top scoring metrics of the submitted trial runs, whereas the proposed methodology metrics are mean results of 10 runs. AUROC and AUPRC results for the top ranked team were not provided.

As shown, the proposed methodology surpasses all top team results except in set A utility score. Significant improvements can be seen in F1 score and AUPRC across both test sets with a near three times improvement. In reference to table II, statistically significant performance improvements can be seen in AUPRC, accuracy and F1 score metrics for both test sets with other team results lying outside of the two standard deviation range. AUROC shows potential statistical improvement for both test sets with team results lying within one to two standard deviations. In regards to utility score, set B shows improved results as compared to the other teams whilst set A score remains comparable, placing fourth in said category.

V. Evaluation

Within this study, we demonstrate a high-performing prediction model of sepsis onset six hours prior to official clinical diagnosis by a human healthcare provider within an ICU setting. In comparison to challenge participants attempting the same prediction application and dataset, our proposed methodology shows statistically significant improvements to performance in traditional model evaluative metrics (AUPRC, accuracy & F1 scores greater than two standard deviations of participant methodologies).

TABLE II
Overall evaluative metrics of the proposed methodology across the datasets

Metric	Set A	Set B	Set A&B
True Pos. Rate	0.480±0.093	0.533±0.006	0.470±0.105
True Neg. Rate	0.982±0.010	0.985±0.002	0.977±0.019
False Pos. Rate	0.018±0.010	0.015±0.002	0.023±0.019
False Neg. Rate	0.520±0.093	0.467±0.006	0.530±0.105
Pos. Predictive Value	0.374±0.092	0.336±0.038	0.341±0.130
Neg. Predictive Value	0.988±0.002	0.993±0.000	0.990±0.003
False Omission Rate	0.012±0.002	0.007±0.000	0.010±0.003
False Discovery Rate	0.626±0.092	0.664±0.038	0.659±0.130
Accuracy	0.971±0.008	0.979±0.002	0.968±0.017
F1 Score	0.420±0.008	0.412±0.021	0.363±0.058
AUROC	0.855±0.032	0.893±0.026	0.737±0.142
AUPRC	0.391±0.010	0.351±0.042	0.258±0.051

The most significant improvements can be seen in both AUPRC and F1 score metrics with results reaching close to three times the improvement. Meanwhile, accuracy and AUROC metrics show comparatively less improvement. Such behaviour stems from significant differences in model precision between the methodologies. With such a significant imbalance in class distributions, a near 13 to 1 negative to positive class balance, model precision suffers greatly due to the proportionally large amount of potential false positive errors in comparison to even total condition positive samples let alone correct, true positive predictions. Even with significant precision performance improvements, such an issue still remains present within our methodology with an individual dataset best precision of 0.374. Class balance issues such as the case, will remain an ever present issue and major consideration within the field of medical informatics, with populations generally skewed towards condition negative individuals regarding a considered condition or disease. The use of cascading sub-models provides a significant step towards improving predictive performance on such class balance issues.

Major differences be seen in comparisons against the combined dataset performance versus the individual datasets. Standard deviation values indicate increased variation in precision and recall across the test folds as compared to the individual datasets, with a resulting overall reduction in model performance. Said results can stem from the k-fold cross validation and class imbalance resulting in folds not representative of overall population characteristics in both features and classes, especially with the distinct lack of positive class samples. Across the two individual datasets, overall performance indicated by accuracy, F1 score, AUROC and AUPRC show statistically similar results. However, in regards to true positive rate and positive predictive value, set B shows greater true positive rates and lower positive predictive values than that of set A. Such results highlights the balancing act between

TABLE III
PHYSIONET CINC 2019 CHALLENGE TOP 5 FINAL LEADERBOARD WITH THE ADDITION OF OUR PROPOSED METHODOLOGY RESULTS

Team Name	Utility Score		AUROC		AUPRC		Accuracy		F1 Score	
	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B
Proposed Methodology	0.415	**0.450**	**0.855**	**0.893**	**0.391**	**0.351**	**0.971**	**0.979**	**0.420**	**0.412**
Can I get your signature?	**0.433**	0.434	0.000	0.000	0.000	0.000	0.828	0.888	0.139	0.140
Sepsyd	0.409	0.396	0.811	0.853	0.105	0.119	0.819	0.901	0.131	0.142
Separatrix	0.422	0.395	0.814	0.844	0.102	0.110	0.803	0.882	0.128	0.130
FlyingBubble	0.420	0.401	0.813	0.855	0.108	0.117	0.798	0.878	0.126	0.129
CTL-Team	0.401	0.407	0.806	0.846	0.101	0.116	0.797	0.891	0.122	0.137

precision and sensitivity across the two models. Influenced perhaps by the differing class distribution ratios between the two datasets (1:45 class ratio in set A, 1:70 for set B).

VI. CONCLUSION

Within this study, we propose and demonstrate a novel methodology to predict the onset of sepsis for a patient six hours earlier than clinical diagnosis within the ICU setting. The application of a novel boosted cascading training procedure, augmented with a shifting margin hinge loss function and tailored penalty functions, our proposed methodology was able to significantly outperform current methodologies on the PhysioNet CinC 2019 challenge dataset. Whilst our proposed methodology shows great promise, there remains many avenues of potential future improvements. There remains opportunity in the adaption of the λ_m hyper-parameter of the shifting margin loss function into a dynamic learned parameter to automate and adaptively limit over-fitting issues on increasingly complex model architectures. The limited test dataset of our study warrants further validation on varied datasets of alternative population characteristics to ensure performance validity. Of course, there exists many potential avenues of adaption of our proposed methodology into alternative time-series based critical-moment prediction problems within and outwith medical settings. Within the context sepsis prediction, such an application affords great opportunity in shifting sepsis care from reactionary based, to that of intervention.

37 pt
0.514 in
13.1 mm

ACKNOWLEDGEMENTS

This work was supported by The Engineering and Physical Sciences Research Council (EP/N028139/1).

REFERENCES

[1] M. Singer, C. S. Deutschman, C. W. Seymour *et al.*, "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, p. 801, feb 2016.
[2] A. Padkin, C. Goldfrad, A. R. Brady *et al.*, "Epidemiology of severe sepsis occurring in the first 24 hrs in intensive care units in England, Wales, and Northern Ireland," *Critical Care Medicine*, vol. 31, no. 9, 2003.
[3] J. Hajj, N. Blaine, J. Salavaci *et al.*, "The "Centrality of Sepsis": A Review on Incidence, Mortality, and Cost of Care," *Healthcare*, vol. 6, no. 3, p. 90, jul 2018.
[4] R. A. Balk, "Systemic inflammatory response syndrome (SIRS): Where did it come from and is it still relevant today?" *Virulence*, vol. 5, no. 1, pp. 20–26, 2014.
[5] R. C. Bone, R. A. Balk, F. B. Cerra *et al.*, "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis," *Chest*, vol. 101, no. 6, pp. 1644–1655, 1992.
[6] E. C. Bishop, *Early Identification and Treatment of Sepsis: Clinical Guideline*. NHS Trust, 2017.
[7] R. Daniels, "Surviving the first hours in sepsis: getting the basics right (an intensivist's perspective)," *Journal of Antimicrobial Chemotherapy*, vol. 66, no. Supplement 2, pp. 11–23, apr 2011.
[8] C. W. Seymour, F. Gesten, H. C. Prescott *et al.*, "Time to treatment and mortality during mandated emergency care for sepsis," *New England Journal of Medicine*, vol. 376, no. 23, pp. 2235–2244, 2017.
[9] A. J. Masino, M. C. Harris, D. Forsyth *et al.*, "Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data," *PLOS ONE*, vol. 14, no. 2, feb 2019.
[10] S. Le, J. Hoffman, C. Barton *et al.*, "Pediatric Severe Sepsis Prediction Using Machine Learning," *Frontiers in Pediatrics*, vol. 7, pp. 1–8, oct 2019.
[11] M. Faisal, A. Scally, D. Richardson *et al.*, "Development and External Validation of an Automated Computer-Aided Risk Score for Predicting Sepsis in Emergency Medical Admissions Using the Patient's First Electronically Recorded Vital Signs and Blood Test Results*," *Critical Care Medicine*, vol. 46, no. 4, pp. 612–618, apr 2018.
[12] Q. Mao, M. Jay, J. L. Hoffman *et al.*, "Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU," *BMJ Open*, vol. 8, no. 1, jan 2018.
[13] R. J. Delahanty, J. Alvarez, L. M. Flynn *et al.*, "Development and Evaluation of a Machine Learning Model for the Early Identification of Patients at Risk for Sepsis," *Annals of Emergency Medicine*, vol. 73, no. 4, pp. 334–344, apr 2019.
[14] T. Desautels, J. Calvert, J. Hoffman *et al.*, "Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach," *JMIR Medical Informatics*, vol. 4, no. 3, p. 28, sep 2016.
[15] J. S. Calvert, D. A. Price, U. K. Chettipally *et al.*, "A computational approach to early sepsis detection," *Computers in Biology and Medicine*, vol. 74, pp. 69–73, jul 2016.
[16] H. J. Kam and H. Y. Kim, "Learning representations for the early detection of sepsis with deep neural networks," *Computers in Biology and Medicine*, vol. 89, no. April, pp. 248–255, oct 2017.
[17] S. Horng, D. A. Sontag, Y. Halpern *et al.*, "Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning," *PLOS ONE*, vol. 12, no. 4, apr 2017.
[18] F. van Wyk, A. Khojandi, A. Mohammed *et al.*, "A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier," *International Journal of Medical Informatics*, vol. 122, no. August 2018, pp. 55–62, feb 2019.
[19] M. A. Reyna, C. S. Josef, R. Jeter *et al.*, "Early Prediction of Sepsis From Clinical Data," *Critical Care Medicine*, vol. 48, no. 2, pp. 210–217, feb 2020.
[20] K. Greff, R. K. Srivastava, J. Koutnik *et al.*, "LSTM: A Search Space Odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.
[21] R. Barandela, R. M. Valdovinos, J. S. Sánchez *et al.*, "The Imbalanced Training Sample Problem: Under or over Sampling?" in *Lecture Notes in Computer Science*, no. February, 2004, pp. 806–814.