A Knowledge Distillation Ensemble Framework for Predicting Short- and Long-Term Hospitalization Outcomes From Electronic Health Records Data

Zina M. Ibrahim , Daniel Bean , Thomas Searle , Linglong Qian , Honghan Wu, Anthony Shek, Zeljko Kraljevic, James Galloway, Sam Norton, James T. H. Teo , and Richard JB Dobson

Abstract—The ability to perform accurate prognosis is crucial for proactive clinical decision making, informed resource management and personalised care. Existing outcome prediction models suffer from a low recall of infrequent positive outcomes. We present a highly-scalable and robust machine learning framework to automatically predict adversity represented by mortality and ICU admission and readmission from time-series of vital signs and laboratory results obtained within the first 24 hours of hospital admission. The stacked ensemble platform comprises two components: a) an unsupervised LSTM Autoencoder that learns an optimal representation of the time-series, using it to differentiate the less frequent patterns which conclude with an adverse event from the majority patterns that do not, and b) a gradient boosting model, which relies on the constructed representation to refine prediction by incorporating static features. The model is used to assess a patient's risk of adversity and provides visual justifications of its prediction. Results of three case studies show that the model outperforms existing platforms in ICU and general ward settings, achieving average Precision-Recall Areas Under the Curve (PR-AUCs) of 0.891 (95% CI: 0.878-0.939) for mortality and 0.908 (95% CI: 0.870-0.935) in predicting ICU admission and readmission.

Index Terms—Ensemble Learning, Gradient Boost, Imbalanced time-series, Long Short Term Memory networks (LSTM), Clinical Outcome Prediction, Outlier Detection, Machine Learning, Mortality Prediction, Stacked Ensemble.

Zina M. Ibrahim, Daniel Bean, Thomas Searle, Linglong Qian, and Zeljko Kraljevic are with the Department of Biostatistics and Health Informatics, King's College London, SE5 8AF London, U.K. (e-mail: zina.ibrahim@kcl.ac.uk; daniel.bean@kcl.ac.uk; thomas.searle@kcl.ac.uk; linglong.qian@kcl.ac.uk; zeljko.kraljevic@kcl.ac.uk).

Honghan Wu is with the Institute of Health Inforamtics, University College London, WC1E 6BT London, U.K. (e-mail: honghan.wu@ucl.ac.uk).

Anthony Shek is with the Department of Clinical Neuroscience, King's College London, SE5 8AF London, U.K. (e-mail: anthony.shek@kcl.ac.uk).

James Galloway is with Centre for Rheumatic Diseases, King's College London, SE5 8AF London, U.K. (e-mail: james.galloway@kcl.ac.uk).

Sam Norton is with the Department of Psychology and the Department of Inflammation Biology, King's College London, SE5 8AF London, U.K. (e-mail: sam.norton@kcl.ac.uk).

James T. H. Teo is with the NHS Foundation Trust, King's College Hospital, SE5 9RS London, U.K. (e-mail: jamesteo@nhs.net).

Richard JB Dobson is with the Department of Biostatistics and Health Informatics, King's College London, SE5 8AF London, U.K., and also with the Institute of Health Inforamtics, University College London, WC1E 6BT London, U.K. (e-mail: richard.j.dobson@kcl.ac.uk).

Digital Object Identifier 10.1109/JBHI.2021.3089287

I. INTRODUCTION

THE secondary re-use of routinely collected patient data has been a facilitator of innovations aiming to improve patient care. A prominent example is the development of early warning systems that predict adversity from patient physiological measurements. The majority of early warning models take the form of ad-hoc scoring tools [40] such as the National Early Warning Score (NEWS2) widely used in the United Kingdom [54]. Such tools estimate a patient's risk of adversity using aggregates of physiological measurements [27]. Generally, scoring tools suffer from low sensitivity due to overlooking the dependencies among the temporal signatures underlying a patient's physiology. Machine Learning models have been developed to overcome the limitations of scoring tools via sophisticated architectures that capture non-linearities within the multivariate temporal patient data [12].

Despite the promising results of Machine Learning early warning systems, we find that existing approaches bear several shortcomings that adversely affect model performance and adoption potential. First, generic clinical outcome prediction models are scarce. Most existing models are condition-specific,
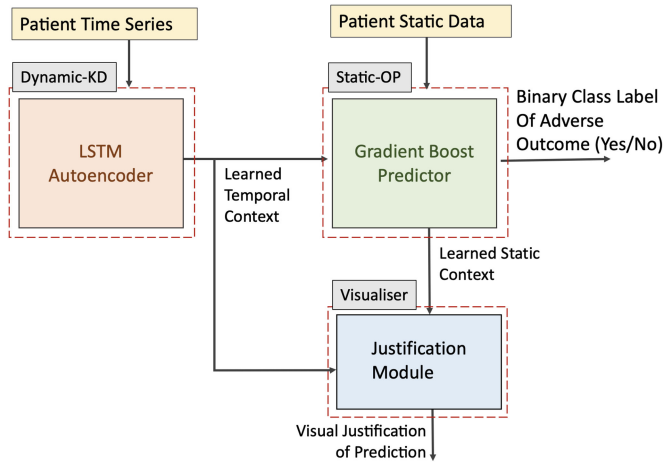
Fig. 1. The Knowledge-Distillation Outcome Predictor *(KD-OP)*. The Dynamic Knowledge Distiller *(Dynamic-KD)* module learns a compressed temporal representation of the no-adversity class. Static Outcome Predictor *(Static-OP)* performs the final prediction by combining *Dynamic-KD*'s output with patient static features. The framework's visualisation module uses both contexts to justify the predictions made.

being developed and evaluated with a single condition in mind, e.g. sepsis [36], cardiac patients [31], COVID-19 [70], brain injury [51]. The few condition-agnostic models available are mostly limited to intensive care (ICU) settings where the magnitude of measurements is high, and the population is more uniform in acuity levels [4], [64]. Second, predicting adverse clinical outcomes is an imbalanced learning problem because any adverse outcome is only present in a minority of the patient sample used to train and evaluate a model. To illustrate, consider the United Kingdom's in-hospital mortality rates, which are around 23% in ICU settings [3] and 4% in secondary wards [21]. Similarly, cardiac arrest incidence is estimated to be 2.3% of ICU admissions [2]. Nevertheless, most current adversity-prediction models are benchmarked using the Area Under the Curve (ROC-AUC) [1], [10], [29], [39], [60], which is known to overestimate model performance on minority outcomes under imbalanced distributions [17], [53]. The result is a general over-optimism in existing models' performance. Finally, in contrast to medical practice, where both the dynamics of a patient's physiology and personal characteristics (e.g. demographics) are used for prognosis, clinical outcome predictions models seldom consider the interplay between the dynamic and static data available about a patient. Existing models either consider the two views descriptively and not a predictive context [7], or distinctly without consideration of the interplay between the two [59].

This paper presents *KD-OP* (Knowledge Distillation Outcome Predictor), an ensemble Machine Learning framework designed to overcome the current difficulties in predicting adverse clinical outcomes from electronic health records data. The framework (Fig. 1) comprises two learner modules. The first, *Dynamic-KD* (Dynamic Knowledge Distiller), learns from the multivariate time-series of a patient's physiology, while the second, *Static-OP* (Static Outcome Predictor), estimates the risk of adversity using static features (e.g. demographics and aggregate measurements). *KD-OP* uses a *stacked* architecture to capture the interplay between the two patient views, using *Dynamic-KD*'s learned context to guide the predictions made

by *Static-OP*. In contrast to existing clinical outcome prediction models, *KD-OP* is designed to befit the relative infrequency of adverse outcomes in real hospital data. This is achieved by implementing *Dynamic-KD* as a Long Short-term Memory (LSTM) Autoencoder, thereby reformulating the prediction task into one of *outlier detection*, whereby adverse outcomes (e.g. mortality =1) are modelled as outliers. We evaluate the framework's predictive power under ICU and general ward settings, using metrics specifically designed for imbalanced classification models [13], [48]. The stacked architecture generates visual justifications of its predictions based on the learned temporal and static context.

II. RELATED WORK

Ensemble models have shown superior performance compared to single-classifier architectures in predicting hospitalisation outcomes [10], [29], [35], [41], [50], [60]. The advantages of ensemble models have been realised either by capturing different data modalities (i.e. static and dynamic views) [59], or by consolidating predictions by several 'weak' classifiers [66]. However, in all existing clinical outcome prediction frameworks, the ensemble's learners are linked by additively combining their respective predictions, e.g. via weighted averaging. This approach is problematic for ensembles operating on different data modalities because the final prediction is not depictive of the possible dependencies between the data's static and the dynamic views [38]. There is, therefore, a missed opportunity in developing models that capture the interplay between a patient's dynamically changing physiology and personal characteristics. Moreover, additive ensembles have been shown to fail to alleviate the individual learners' bias; they are generally outperformed by alternative models that *stack* strong classifiers into an ensemble [14], [38].

The proposed stacked architecture dedicates its time-series prediction task to an LSTM Autoencoder. The model choice is motivated by the natural marginal representation of adverse outcomes in hospital data and the known ability of LSTM AutoEncoders to identify minority outcomes from imbalanced time-series [33], [34]. The idea is that an LSTM Autoencoder encodes the time-series in a low dimensional representation capturing its most informative features. The compact representation then only enables the reconstruction of the representative features from new input without noise. By strictly training the LSTM Autoencoder on the majority (no-adversity) time-series, outliers (time-series corresponding to adverse outcomes) will generate high reconstruction errors [33]. LSTM Autoencoders have been effectively used in fall detection [46], sensor failure prediction [42], fraud detection [19] and video surveillance [68]. LSTM Autoencoders have also shown great potential in healthcare [5], with applications in retinal eye research [57], patient subtyping [6] and healthcare fraud detection [62].

III. METHODOLOGY

In this work, the prediction of adverse clinical outcomes is expressed as a binary classification whose input comprises two types of multivariate data collected for a patient: 1) dynamic data, comprising physiological measurements and laboratory results routinely collected at the bedside either automatically

or by healthcare practitioners (e.g. heart rate), and 2) static data, which is either recorded during admission (e.g. demography), summarised during examination (e.g. comorbidities) or is a dynamic-data aggregate habitually used by healthcare practitioners to evaluate a patient's state (e.g. maximum heart rate during in a given day). For a given adverse outcome, the framework (Fig. 1) assigns each patient a binary class label using the predicted probability of the patient's risk and an adversity threshold $\gamma \in [0, 1]$, which the model also learns.

The dynamic data, which takes the form of a multivariate time-series, is processed by the Dynamic Knowledge Distiller *(Dynamic-KD)* module to learn a vector representation assigning each patient a risk score based on the patient's temporal signatures alone. *Dynamic-KD*'s learned risk is used to guide the prediction process of the Static Outcome Predictor *(Static-OP)* module, a Gradient Boost classifier that uses the static view to predict the final class label. The ensemble model is designed to improve predictive power and adoption potential in the following ways:

1) By adopting a *stacked* architecture to capture the interplay between the temporal signatures of a patient's physiology and static features, thereby enabling the temporal risk learned by *Dynamic-KD* to drive classification based on static using *Static-OP*.

2) By incorporating a robust mechanism to address the marginal representation of adverse outcomes (e.g. ICU admission = 1) compared to typical outcomes (e.g. no ICU admission) in the overall population. This is achieved by designing the dynamic-learner module, *Dynamic-KD*, using the successful LSTM-Autoencoder outlier detection architecture.

A. Glossary of Terminology and Model Formulation

For a patient k having v dynamic features each measured over T consecutive time windows, and u single-measurement static features, we represent the patient's dynamic view by a matrix $\boldsymbol{X_k} \in /BbbR^{T \times v}$, containing the totality of the patient's dynamic observations. Moreover, the patient's static view is represented by a vector $\boldsymbol{x_k} \in /BbbR^u$ consisting of a sequence of u static and aggregate features. For a population of n patients, the ensemble model therefore accepts two types of inputs: the combined multivariate time-series of the n patients, $\boldsymbol{D^d} = \{\boldsymbol{X}_k\}_{k=1}^n$, and a static patient-feature matrix $\boldsymbol{D^s} = \{\boldsymbol{x}_k\}_{k=1}^n$. Additionally, because *Static-OP* is a classifier, it is trained using $\boldsymbol{D^s}$ and a vector representing the true incidence of the adverse outcome in the cohort $\boldsymbol{y} = \{y_k\}_{k=1}^n$.

For a given adverse outcome, the goal of the framework is to predict the vectors $\hat{\boldsymbol{p}} = \{\hat{p}_k\}_{k=1}^n$ and $\hat{\boldsymbol{y}} = \{\hat{y}_k\}_{k=1}^n$. Each $\hat{y}_k \in \{0, 1\}$ is a binary variable representing the predicted onset of the adverse outcome for a single patient; $\hat{p}_k$ is the predicted probability of the outcome for one patient, which we retain for use during the interpretation stage. Naturally, the class distribution of $\boldsymbol{y}$ is highly imbalanced in favour of non-adverse outcome, as will be demonstrated in the evaluation sections. The framework learns the probability of an adverse outcome for the population of n patients, $\hat{\boldsymbol{p}}_k$ from the two views of the clinical data $\boldsymbol{D}^d$ and $\boldsymbol{D}^s$, using it to estimate $\hat{\boldsymbol{y}_k}$.

Aggregation Function	Variables
Minimum	Albumin, Central Venous Pressure, Haemoglobin
Maximum	Alanine Aminotransferase (ALT), Assured shorthold tenancies(AST), Bilirubin, Creatinine , Fibrinogen C-Reactive Protein , Lactate Dehydrogenase International Normalized Ratio (INR) Partial Pressure of Oxygen Dioxide (PaCO₂), Partial Pressure of Oxygen PaO₂ Partial Venous Pressure of Oxygen Dioxide PvCO2 Peripheral O₂ Saturation, oxygen saturation (SaO₂) Spontaneous Respiratory Rate, Urea
Average	Lymphocytes , Neutrophils, NEWS2, Platelets
Minimum, Maximum	Arterial Oxygen Content (CaO₂) Centralvenous O₂ Saturation, Diastolic Blood Pressure Heart rate, Inspired Fraction of Oxygen (FiO₂) Mean Blood Pressure, PO₂/FiO₂, Systolic Blood Pressure Temperature , White Blood Cell Count

B. Data Processing

Data extracted from Electronic Health Records (EHRs) is generated as a by-product of routine clinical care. As a result, the variables making up a patient's dynamic data view comprise irregularly-sampled time sequences of physiological measurements. Using the extracted data to train *KD-OP* entailed transforming the irregular time series into $\boldsymbol{D^d} = \{\boldsymbol{X_1}, \ldots \boldsymbol{X_n}\}$ for the n patients, where each $\boldsymbol{X_k}$ is a $T \times v$ matrix, of fixed number T of equally-sized observation windows and v vital signs measured at each time window $t \in T$. Furthermore, in order for $\boldsymbol{D^d}$ to be digestible by the two Machine Learning models, interpolation of missing data and scaling are required.

To alleviate missingness while overcoming non-uniform sampling, the number of observation windows T used to construct $\boldsymbol{D^d}$ was optimised iteratively by maximising completeness while minimising the length of the observation windows $t \in T$. If multiple observations are present for a variable during a window, they are binned using a knowledge-based approach that mimics the summaries used by clinicians in practical settings (Table I).

Similarly to [60], we imputed the resulting time-series Gaussian Process Regression (GPR). Gaussian Processes (GPs) extend the assumptions of a Gaussian distribution over functions, whereby a function f is a Gaussian Process if it is entirely characterized by its mean m and covariance o functions: $f \sim \mathcal{GP}(m, o)$ [52]. GPR thus provides a non-parametric method for accommodating statistical uncertainty measures in a regression problem. It has been used in a wide range of applications and has been found to produce superior results when used to interpolate missing data in multivariate time-series [11], [47].

We defined a GPR over the dynamic features in $\boldsymbol{D^d}$ for the T time windows. To ensure the interpolated values are meaningful, we grouped patients into subpopulations based on demographical similarities (age, sex, co-morbidity indices), subsequently defining the mean function of each feature as the subpopulation mean μ of patients within the same demographical group. We used the exponential covariance function, resulting in the following definition of our GPR:

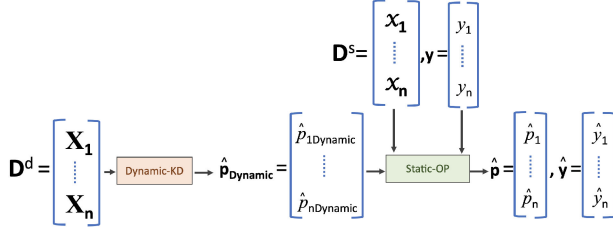$$f \sim \mathcal{GP}(m, o) : m(x) = \mu \ and \ o(x, x') = \exp\left(-\frac{1}{2}(x - x')^2\right)$$

Fig. 2. *KD-OP* Data Flow: $\boldsymbol{D^d}$ contains the totality of the dynamic data for n patients, where each $\boldsymbol{X_k}$ is a $T \times v$ matrix corresponding to the time series of a single patient with v dynamic features over T time windows. $\hat{\boldsymbol{p}}_{Dynamic}$ is a vector of length n containing, for each patient k, the likelihood of the patient k of belonging to the adversity class based on the learned temporal context embedded within $\boldsymbol{X_k}$. *Static-OP* uses $\hat{\boldsymbol{p}}_{Dynamic}$ to guide classification based on data containing static features, $\boldsymbol{D^s}$, to output the final predicted probabilities $\hat{\boldsymbol{p}}$ and adversity classes $\hat{\boldsymbol{y}}$.

Furthermore, the data was normalised using a number of scaling techniques including absolute-value scaling and min-max scaling. Using a min-max scaler with a range of [0,1] achieved the best classification performance and was therefore adopted for all the models. The interpolation and normalisation models were both fitted to the training data, using each model's learned parameters to process the test data.

C. The Architecture

Fig. 2 illustrates the data flow of the proposed architecture. As the figure shows, the combined multivariate time-series of the n patients, $\boldsymbol{D^d}$ is used as input to *Dynamic-KD*, which is an unsupervised LSTM Autoencoder that learns a vector $\hat{\boldsymbol{p}}_{Dynamic} \in [0, 1]$ from the time-series such that for two individuals a and b whose time-series observations $\boldsymbol{X_a}$ and $\boldsymbol{X_b} \in \boldsymbol{D^d}$, if $\hat{y}_a > \hat{y}_b$ then:

$$\log(\hat{p}_{aDynamic}) > \log(\hat{p}_{bDynamic})$$

That is, the vector representation $\hat{\boldsymbol{p}}_{Dynamic}$ creates a separation between time-series corresponding to adverse outcomes and those corresponding to normal outcomes. The use of log transformation captures the difference in the order of magnitude of the resulting representations, rather than possibly non-significant fluctuations within the actual values [15].

The second module, *Static-OP* is a classification ensemble based on gradient boost trees. *Static-OP* is trained using $\boldsymbol{D^s}$ and the true incidence of the outcome in the cohort, $\boldsymbol{y}$ to estimate the final prediction probability $\hat{\boldsymbol{p}}$. During the training process of *Static-OP*, $\hat{\boldsymbol{p}}_{Dynamic}$ is used as sample weights to guide training. The two modules form the bi-level stacked classification system *KD-OP* (Knowledge Distillation Outcome Predictor). The output for the ensemble is:

$$\hat{\boldsymbol{y}} = \begin{cases} 0 & \text{if } \hat{\boldsymbol{p}}_{Static} > \gamma \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

Where the adversity threshold $\gamma \in [0, 1]$ is a learned parameter corresponding to the optimal threshold for classification selected by optimising the mean Area Under the Precision-Recall Curve (PR-AUC). The remainder of this section details the design of *KD-OP*'s individual modules.
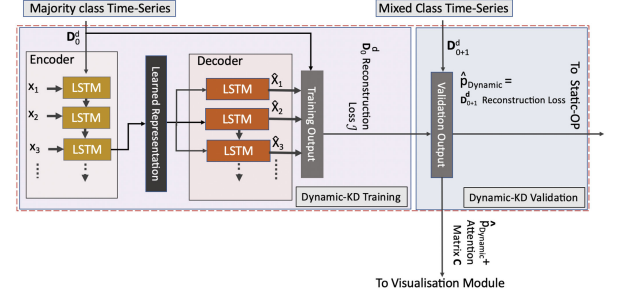

Fig. 3. The Dynamic-KD LSTM Autoencoder Module. $\boldsymbol{D_0^d}$ refers to the portion of the time-series data which corresponds to patients with typical (non-adverse) outcomes used to self-train the LSTM Autoencoder; $\boldsymbol{D_{0+1}^d}$ refers to mixed-class time-series. When provided with $\boldsymbol{D_{0+1}^d}$ as input, the trained LSTM Autoencoder, yields $\hat{\boldsymbol{p}}_{Dynamic}$, which is representative of the probability of each patient k in $\boldsymbol{D_{0+1}^d}$ belonging to the minority class.

1) Dynamic-Kd: The first module is designed to use the data's dynamic view, i.e. the multivariate time-series input $\boldsymbol{D^d}$, to learn a vector-form separation between the majority (non-adverse outcome) time-series and time-series corresponding to adverse events. The module capitalises on the significantly lower frequency of adverse outcomes (e.g. ICU admission = 1) compared to typical outcomes (e.g. no ICU admission) in the overall population, enabling viewing adverse outcomes as outliers.

Dynamic-KD (Fig. 3) consists of an unsupervised LSTM-Autoencoder architecture which is 'self-trained' using a subset of the time-series containing only majority outcomes, $\boldsymbol{D_0^d}$. For each patient matrix $\boldsymbol{X_k}$ in $\boldsymbol{D_0^d}$, the training procedure aims to reconstruct $\boldsymbol{X_k}$'s input sequence by minimising a distance-based objective function $\mathcal{J}$. $\mathcal{J}$ is the reconstruction loss measuring the difference between the input vectors in the original series $\boldsymbol{X_k}$ and the vectors of the reconstructed series $\hat{\boldsymbol{X}}_k$. $\mathcal{J}$ is defined as below:

$$\mathcal{J} = \sqrt{\sum_{t=1}^{T} ||\boldsymbol{x_t} - \hat{\boldsymbol{x_t}}||_2^2} \quad (2)$$

where T is the number of multivariate observations for each patient k, $\boldsymbol{x_t}$ and $\hat{\boldsymbol{x_t}}$ are the multi-feature vectors at time t, and $||.||_2$ is the L2-norm. Once trained, the LSTM Autoencoder is run on a subset of the mixed-class time series ($\boldsymbol{D_{0+1}^d}$ in the figure). Because training was performed strictly on majority-class series, the values of the reconstruction loss $\mathcal{J}$ associated with outliers (samples corresponding to the adversity class) in $\boldsymbol{D_{0+1}^d}$ will be significantly higher than the reconstruction loss $\mathcal{J}$ associated with the majority (no adversity) samples of $\boldsymbol{D_{0+1}^d}$. As a result, the vector output $\hat{\boldsymbol{p}}_{Dynamic}$, which is the reconstruction loss associated with $\boldsymbol{D_{0+1}^d}$, will be representative of the probability of each patient k in $\boldsymbol{D_{0+1}^d}$ belonging to the minority (outlier) class.

Dynamic-KD adopts an attention mechanism over the time steps to capture the most important features in each sequence as proposed by [23] and successfully implemented in [58], [60], [61]. Fig. 4 shows a feature-level representation of the attention mechanism in the encoder-decoder architecture of *Dynamic-KD*,
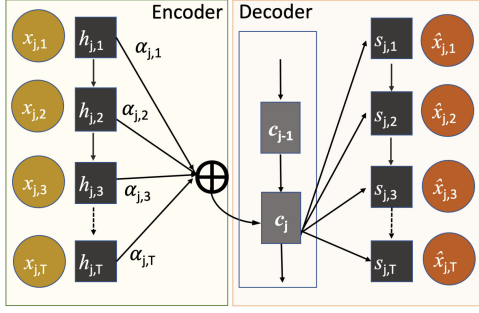
Fig. 4. The attention mechanism of the encoder-Decoder Architecture demonstrated at feature level. X: input sequence; $\hat{X}$: output sequence; h: encoder hidden state; s: decoder hidden state; c attention (context) vector.

reconstructing a multi-variable sequences over T time step batches (i.e. T ordered sequences per patient). For each feature j, a soft attention mechanism is implemented over the encoder's hidden states to obtain a distinct context vector c_j. c_j attenuates the most informative hidden states in $s_{j,1}, \ldots . s_{j,T}$ of the decoder based on the encoder's latent representation and is computed as follows:

For each feature j, the attention probabilities based on the encoded sequence $\alpha = (\alpha_1, \ldots, \alpha_T)$ are calculated using the encoded sequence and the encoder's internal hidden states. First, the importance of information at each time step for feature j is calculated:

$$e_{j,t} = a(\mathcal{U}_j \circledast s_{t-1} + \mathcal{W}_j \circledast h_j + b_j)$$

where $\mathcal{U}_j$ and $\mathcal{W}_|$ are trainable parameters capturing the input-to-hidden and hidden-to-hidden transitions for a given layer j respectively. Terms $\mathcal{W}_j \circledast h_{t-1}$ and $\mathcal{U}_j \circledast x_t$ respectively capture the update from the hidden states at the previous step and the new input. a is the activation function. In the decoder layers, we can measure the importance of the information at each time step for each feature j denoted by $e_{j,t}$ using proximity to $\mathcal{U}_j$. Then $\alpha_{j,t}$ is obtained by normalising $e_{j,t}$ using the softmax operation:

$$\alpha_{j,t} = \frac{exp(e_j)}{\sum_{t=1}^{T} exp(e_t)}$$

Finally, the context vector for each feature c_j is calculated using the weighted sum of the encoded sequence with the attention probabilities. Intuitively, this vector summarizes the importance of the encoded features in predicting t^{th} sequence:

$$c_j = \sum_{t=1}^{T} \alpha_{j,t} h_{j,t} \qquad (3)$$

As algorithm 1 shows, the *Dynamic-KD* LSTM-Autoencoder is trained in batches of D_0^d to minimise the reconstruction loss $\mathcal{J}$ (line 1). The resulting loss is therefore representative of the training errors associated with the negative (majority) class.

Once trained, the LSTM-Autoencoder is run on a non-overlapping subset of the time-series D_{0+1}^d (line 2), which contains mixed data (positive and negative outcomes), using the optimal loss obtained during training. Running the autoencoder yields a reconstruction $\hat{D}_{0+1}^d$ of D_{0+1}^d and an attention matrix C_{0+1}. At the end of the procedure, the reconstruction loss $\hat{p}_{Dynamic}$ (line 3) of $\hat{D}_{0+1}^d$ from the original series D_{0+1}^d

---

Algorithm 1: Static-OP.

Receives: Training and validation subsets of the multivariate, regularly-sampled and batched time-series, D^d: D_0^d containing majority-class time-series for training and D_{0+1}^d containing mixed-class time-series for validation

Returns: Validation reconstruction loss $\hat{p}_{Dynamic}$, and attention matrix C_{0+1}

Train Auto Encoder
1) $\hat{\theta} = \arg\min_\theta \mathcal{J}(D_0^d)$

Run Auto Encoder
2) $\hat{D}_{0+1}^d, C_{0+1} = Decoder(Encoder(D_{0+1}^d), \hat{\theta})$
3) $\hat{p}_{Dynamic} = \hat{D}_{0+1}^d - D_{0+1}^d$

---

augments the original highly-dimensional feature space into a linear representation which is descriptive of the deviation from normality (no adversity) with respect to the temporal interactions embedded within the data. $\hat{p}_{Dynamic}$ is, therefore, discriminatory between the two classes and corresponds to the likelihoods of each batch (patient) k belonging to the positive class. $\hat{p}_{Dynamic}$ is used to complement the learning from static features performed by *Static-OP*, while C_{0+1} is fed into the interpretation component of the framework.

D. Static-Op

The goal of this module is to complement the predictions made from dynamic data by *Dynamic-KD*. *Static-OP* takes as input static information routinely employed by healthcare practitioners to assess a patient's risk factors (e.g. demographics, symptoms, summary statistics of physiology). In other words, instead of using $\hat{p}_{Dynamic}$ as a predictor of the outcome, it is instead used to drive further classification. This way, the overall pipeline has the advantage of capturing the interplay between dynamic physiological measurements and static features in making the final predictions.

Static-OP is a supervised gradient boost model [32] whose overall structure is given in Algorithm 2. The data used to train *Static-OP* comprises static features and outcomes of the same patients used to construct D_{0+1}^d. We denote those by D_{0+1}^s and y_{0+1} respectively. The reconstruction loss $\hat{p}_{Dynamic}$ generated by *Dynamic-KD* serve as sample weights ω (line 1) for each sample in D_{0+1}^s. Because *Dynamic-KD* ensures that $\hat{p}_{Dynamic}$ creates a separation between positive and negative classes, the minority samples of D_{0+1}^s will be the determinant of the decision threshold the model is trained to discover.

Using the model to predict the outcome probabilities from y^{s-test} produces $\hat{p}$, and will also produce the variable importance vector (line 2). The class labels $\hat{y}$, are obtained as in 1 (line 4), using a prediction threshold learned by maximising PR-AUC from the predicted probabilities (line 3).

IV. EXPERIMENTAL EVALUATION ON REAL USE CASES

We critically assess the model's performance using two sources of data. The first is general-ward (non-ICU) hospital data obtained from King's College Hospital and Princess

Algorithm 2: Static-OP.

Receives:

1) D_{0+1}^s and y^s: Static-feature dataset of the same patients used to validate *Dynamic-KD*, with corresponding true-outcome labels.

2) D_{0+1}^{s-test} and y^{s-test}: An independent mixed-class static testing dataset, with corresponding true-outcome labels.

3) Validation reconstruction loss, $\hat{p}_{Dynamic}$ of D_{0+1}^s obtained from *Dynamic-KD*.

Returns: Classification label $\hat{y}, \hat{y} \in \{0, 1\}, \forall \hat{y} \in \hat{\boldsymbol{y}}$, variable importance $\mathcal{I}$

Train Gradient Boost

1    $\mu = \text{TrainGB}(X = D_{0+1}^s, y = y^s, \omega = \hat{p}_{Dynamic})$

Test Gradient Boost

2    $\hat{p}, \mathcal{I} = \mu(D_{0+1}^{s-test}, y^{s-test})$

3    $\gamma = \underset{PR-AUC}{\arg\max}(y^{s-test}, \hat{p})$

4    $\hat{y} \overset{\gamma}{\geq} \hat{p}$

TABLE II
SUMMARY STATISTICS OF THE DATASETS. DATA ARE DISPLAYED AS MEAN
(STANDARD DEVIATION) OR COUNT (PERCENT)

Attribute	COVID-19 (General Ward)	Pneumonia ICU	CKD (ICU)
Patients, n	1,276	2,798	2,822
Females	735(57.6 %)	1,217(45.58%)	1,139 (37.27%)
Age	69.3 (16.79)	50.05 (33.93)	72.61 (22.55)
No. Comorbidities	0.32 (0.14)	0.22 (0.18)	0.29 (0.19)
Mortality			
5-Day	139 (10.88%)	183 (6.48%)	169 (5.98%)
7-Day	187(14.64%)	257 (9.19%)	222 (7.87%)
14-Day	264 (20.68)	412 (14.73%)	310 (10.98%)
30-Day	335 (26.23%)	530 (18.94%)	378 (13.39%)
Days to mortality	12.07 (15.59)	15.38 (12.46)	11.87 (14.44)
ICU	12.07 (15.59)	15.38 (12.46)	11.87 (14.44)
5-Day	105 (8.22 %)	63 (2.36%)	78 (2.81%)
7-Day	112 (8.77%)	102 (3.82%)	80 (2.84%)
14-Day	123 (9.63%)	199 (7.45%)	157 (5.56%)
30-Day	124 (9.71%)	272 (10.18%)	196 (6.95%)
Days to ICU	4.35(8.30)	12.23 (10.46)	11.18 (10.49)

The number of pre-existing conditions was normalised to [0,1] as the two data sources used different scales to measure co-morbidity. ICU refers to ICU admission in the COVID-19 study and ICU re-admission in MIMIC-III.

Royal University Hospital in London, U.K.. The compiled dataset comprised inpatients diagnosed with COVID-19 for which *KD-OP* was used to predict (a) mortality and (b) ICU admission. The second source is an ICU database, namely the freely-available Medical Information Mart for Intensive Care III (MIMIC-III) [30]. Here, we extracted two datasets corresponding to Pneumonia and Chronic Kidney Disease (CKD) patients. Since MIMIC-III strictly contains ICU data, *KD-OP* was used to predict (a) mortality and (b) ICU re-admission from the MIMIC-III cohorts. *KD-OP* was trained to predict adverse outcomes using time-series corresponding to the first 24 hours of hospital admission (in the COVID-19 use case) or ICU admission (in the Pneumonia and CKD use cases). Moreover, the risk of adverse outcomes (mortality, ICU admission or ICU re-admission) was predicted at intervals of 5, 7, 14 and 30 days within admission.

A. Datasets

1) COVID-19 Case Study: Data was collected from 1276 adult ($\geq$ 18 years old) inpatients testing positive for severe acute respiratory syndrome coronavirus 2 (COVID-19) by reverse transcription polymerase chain reaction (RT-PCR) between the 1st of March and 31st April 2020 at two acute hospitals (King's College Hospital and Princess Royal University Hospital) in South East London (U.K.). Only symptomatic patients who required inpatient admission were included. Presenting symptoms included but were not limited to fever, cough, dyspnoea, myalgia, chest pain, or delirium. Static data collected include age, sex, ethnic background, the length of the period from symptoms onset to hospital admission, and pre-existing conditions (specifically, chronic obstructive pulmonary disease (COPD), Asthma, heart failure, diabetes, ischemic heart disease (IHD), hypertension and chronic kidney disease). For training and risk prediction, pre-existing conditions were aggregated into one ordinal feature describing the number of comorbidities at the time of admission. The dynamic features included 14 of the routinely

collected vital signs and laboratory tests of Table I. Details of the variables used in the COVID-19 study and their missingness ratios in the data collected are available on our online repository.[1]

2) Pneumonia and CKD Case Studies: We used the data of ICU stays between 2001 and 2012 obtained from the MIMIC-III database, which is a freely-available anonymised ICU database and is the largest resource of time-series hospital data available worldwide [30]. We extracted admission details, demographics, time-stamped vital signs and laboratory test results obtained over the first 24 hours of admission of adults having ICD-9 code = 482.9 (pneumonia, cause not otherwise specified) and 585.9 (CKD, cause not otherwise specified) as the primary diagnoses in the ICU admission notes. Our choice of relying on ICD-9 codes to construct the two cohorts is a pragmatic one. ICD-9 codes are highly-specific but exhibit low sensitivity in extracting confirmed diagnoses from EHRs, including those of pneumonia [25], and CKD [65]. Since our aim is to evaluate the developed model, ICD-9 codes are sufficient to extract correct yet possibly incomplete cohorts for the two conditions.

Since the MIMIC-III database is structured such as each hospital admission may correspond to multiple ICU stays, we extract the time-series pertaining to the first ICU stay of each admission, and used subsequent ICU admission to ascertain readmission outcomes. The resulting datasets comprise 509 323 records corresponding to 2798 pneumonia ICU stays and 702 813 records corresponding to 2822 CKD ICU stays (SQL and python scripts for recreating the dataset using the MIMICIII database are available on our online repository[2]).

3) Data Description and Characteristics: Table II provides statistical summaries of the three datasets, including demographics and outcome distributions in the training and test sets of the three cohorts. The datasets vary in size, where pneumonia and CKD are much larger than COVID-19. The difference in size is

---

[1]https://github.com/zibrahim/MIMICTimeSeriesProcessing/blob/main/VitalAggregation.pdf

[2]https://github.com/zibrahim/MIMICTimeSeriesProcessing

a direct consequence of the mode of collection. The pneumonia and CKD datasets were extracted from the largest publicly-available ICU time-series database [30], while the COVID-19 data was locally collected over a short time span. Females were only the majority of cases in the COVID-19 dataset (females = 57.6 %), but sex distribution only significantly differed from the CKD dataset (females = 37.27%) and not from the pneumonia dataset (females = 45.58%). The pneumonia cohort was significantly younger and less co-morbid than the other two. The pneumonia cohort also showed a wider distribution of age compared to COVID-19 and CKD. In addition, the table shows that the number of pre-existing conditions varied greatly in the pneumonia and CKD cohorts, while the distribution of pre-existing conditions was more uniform in the COVID-19 dataset. The different distributions in age and pre-existing conditions is quite reasonable and align with the nature of the use cases: CKD is an age-related chronic illness [43], with previous studies showing that the rate of comorbidities is around 41% [37]. COVID-19 hospital admissions are more likely in the elderly with pre-existing conditions such as hypertension and diabetes, where symptoms are likely to be more severe as opposed to the young healthy individuals [49]. In contrast, although both older age and pre-existing conditions increase the risk of acquiring pneumonia, they have not been found to be associated with the severity of the condition and subsequent intensive care needs [45], [63].

Across all prediction intervals, the COVID-19 dataset had higher rates of mortality, while mortality rates of CKD were significantly lower than in the other two cohorts. In addition, the CKD cohort had significantly lower rates of ICU re-admissions across all prediction intervals compared pneumonia, and compared to ICU admission in COVID-19. However, the time to ICU admission in COVID-19 was much lower than the time to ICU re-admission in CKD and pneumonia, where the average duration was 4.35 days, compared to 12.23 days in pneumonia and 11.18 days in CKD. In all three cohorts, the training and test sets are comparable in age, percentage of females and the distribution of outcomes (5,7, 14 and 30-day mortality, ICU admission or re-admission).

B. Implementation Details

Applying the regularisation procedure described in Section III-B produced $T = 48$ for pneumonia and CKD (aggregation into half-hourly windows) and $T = 12$ for the COVID-19 use-case (aggregation into 2-hourly windows) for the 24-hour data extracted. The resulting datasets comprised 30 624 samples with 12 variables for the COVID-19 dataset, 134 304 and 135 456 samples with 30 variables for the pneumonia and CKD cases respectively.

The datasets have two notable properties: 1) the outcomes are skewed, with positive outcomes being highly under-represented in the time series, and 2) the temporal ordering is defined over T batches for each patient. In order to retain the natural distribution of outcomes and temporal ordering during training and to prevent information leakage, we used stratified grouped k-fold cross-validation,[3] with k = 3, to split the data as shown

[3]https://www.kaggle.com/jakubwasikowski/stratified-group-k-fold-cross-validation
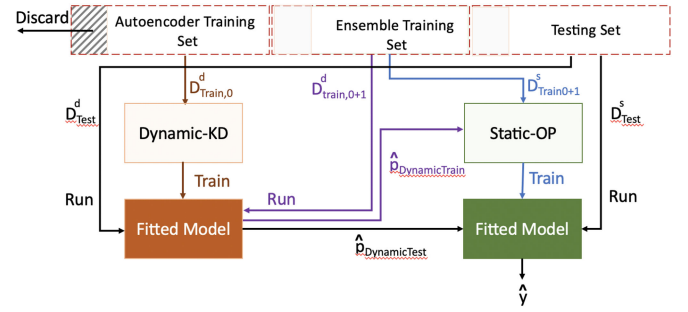

Fig. 5. The procedure followed for the three-way split of the data in *KD-OP*. Each group shows an approximation of the proportion of positive-outcome (shaded) to negative-outcome samples (non-shaded). The positive outcomes have been marked as discarded in the autoencoder training set because the unsupervised *Dynamic-KD* is trained solely on positive-class time series. The positive class samples of the ensemble training and testing sets have been highlighted for visualisation purposes only.
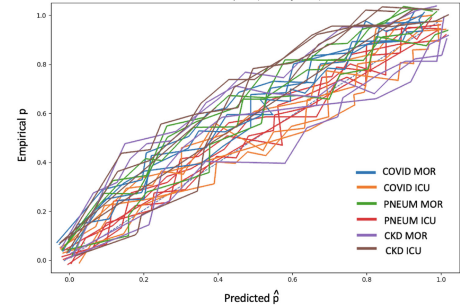

Fig. 6. Reliability plot of *KD-OP*'s predicted $\hat{p}$ in the test set, where empirical p is the fraction of patients with adverse events within the interval (e.g. 30-day mortality).

in Fig. 5. At each iteration, the data used to train *Dynamic-KD* is obtained using one fold (Autoencoder Training Set in the figure), discarding the samples corresponding to patients with positive outcomes to yield $D^d_{Train,0}$. The second fold (Ensemble Training Set in the figure) is used to run *Dynamic-KD* to predict $\hat{p}_{DynamicTrain}$ and further to train *Static-OP*, using $\hat{p}_{DynamicTrain}$ as sample weights. The third fold (Testing Set in the figure) is used to predict the testing $\hat{p}_{DynamicTest}$ using *Dynamic-KD*, using it as sample weights in *Static-OP*, where the final model prediction is made to produce $\hat{y}$.

We used the Python language and the Keras library with Tensorflow backend.[4] For *Dynamic-KD*, model hyperparameters were optimised through empirical evaluation, by carefully observing the prediction performance using a set of candidate values of the hyperparameters; those included the number of neurons per layer, the number of hidden layers, dropout rates and activation function. The final design included bi-layered encoder and decoder, with the outmost layers having neurons in the order of $2 \times n\_features$, where $n\_features$ is the number of dynamic variables used (14 for COVID-19 and 30 in the pneumonia and CKD studies). A dropout rate of 0.5 was used between each two layers to prevent the autonecoder from overfitting the training data and an adaptive learning rate was used using the Adam optimizer and starting at 0.001. The

[4]The source is available at: https://github.com/zibrahim/StackedPredictor/

TABLE III
AVERAGE PERFORMANCE PER OUTCOME FOR COVID-19, PNEUMONIA AND CKD

	COVID-19		Pneumonia		CKD	
	Mortality	ITU Admission	Mortality	ITU Readmission	Mortality	ITU Readmission
Avg Macro Prec	0.918 (0.815-0.931)	0.939 (0.902-0.946)	0.881 (0.814-0.902)	0.873 (0.737-0.931)	0.916 (0.894-0.921)	0.868 (0.760-0.905)
Avg Macro Rec	0.938 (0.907-0.969)	0.970 (0.963-0.984)	0.884 (0.865-0.892)	0.869 (0.825-0.926)	0.953 (0.853-0.906)	0.889 (0.802-0.985)
Avg Macro F1	0.936 (0.913-0.963)	0.977 (0.976-0.978)	0.901 (0.90-0.984)	0.881 (0.774-0.951)	0.912 (0.883-0.935)	0.919 (0.781-0.985)
Avg PR-AUC	0.922 (0.881-0.959)	0.948 (0.851-0.965)	0.871 (0.830-0.923)	0.881 (0.727-0.967)	0.896 (0.869-0.938)	0.895 (0.756-0.998)
Avg ROC-AUC	0.931 (0.896-0.942)	0.971 (0.969-0.987)	0.884 (0.865-0.892)	0.870 (0.825-0.926)	0.878 (0.853-0.908)	0.895 (0.826-0.945)

number of epochs was 1000, which was selected via cycles of experiments and careful monitoring of the validation loss. An early stopping criteria was used to retain the best model by minimising the validation loss with a patience parameter of 50 epochs. All layers of the autoencoder used ReLU as their activation function, which performed best during our evaluation. The *Static-OP* module was implemented using the XGBoost algorithm. The parameters were chosen through a grid-search over the hyperparameter space. *Static-OP*'s sample weights were set to *Dynamic-KD*'s prediction loss. We used calibrated classification using the nonparametric isotonic method included in Python's scikit-learn package. The calibration plot is shown in Fig. 6.

C. Results

We evaluate *KD-OP*'s performance across four dimensions. First, we evaluate the performance under different settings presented by the three datasets; these include cohort heterogeneity with respect to individual characteristics and outcome distribution with respect to the minority (positive) cases. Here, we initially report metrics averaged across the different prediction intervals for each setting to obtain an overall view, and subsequently evaluate the model's robustness across different prediction intervals. We then evaluate the contributions of the two modules *Dynamic-KD* and *Static-OP* to the overall performance, validating those empirically and against clinical knowledge. Finally, we compare the predictive power of *KD-OP* with existing outcome prediction models as reported in the literature. After evaluating *KD-OP*'s performance, we demonstrate its visualisation capability in section IV-C5.

Throughout the experiments, we report the Precision-Recall Area Under the Curve (PR-AUC) to capture the model's performance with respect to the minority cases, as well as the widely-used Receiver-Operator Area Under the Curve (ROC-AUC). Despite our knowledge of ROC-AUC's impartial assessment of the model's performance for positive and negative outcomes [24], we choose to show it here due to its wide usage in the literature. Specifically, we use ROC-AUC to compare our model's performance with state-of-the-art models in section IV-C4. We also report the macro-averaged precision, recall and F1-score. We used macro averages to understand the modules' true performance with respect to the under-represented outcomes [16].

1) Overall Performance and Sample Diversity: We first evaluate the overall performance across the three case studies. For each dataset, Table III shows the model's performance averaged across the prediction intervals of 5, 7, 14 and 30 days for each outcome. The performance is high overall. However, better performance across prediction intervals was achieved using the
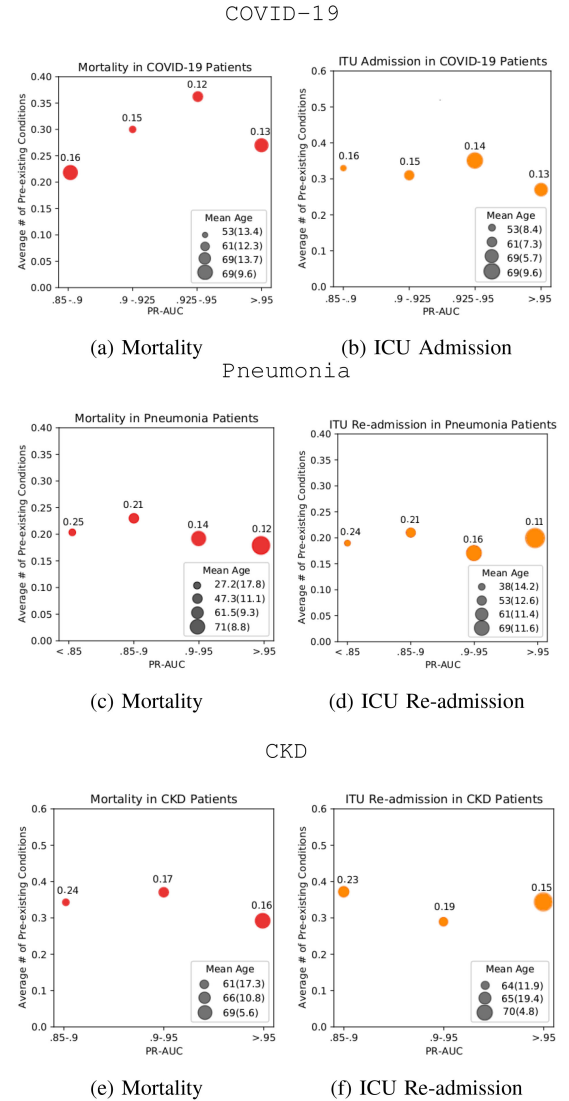
Fig. 7. Grouped performance plot. The x-axis corresponds to performance grouped by PR-AUC. The mean comorbidity per performance group is given by the y-coordinates (circles), while the comorbidity standard deviation is shown as text annotations over each circle. The circle size indicates the mean age within the performance range, while the numerical age means and standard deviations are shown in the legend.

COVID-19 dataset compared to pneumonia and CKD, despite the latter two being larger datasets with a higher resolution of observations (half-hourly windows as opposed to two-hourly windows used to construct the COVID-19 time-series). A close examination is shown in Fig. 7, where higher performance ranges appear to be closely correlated with sample homogeneity

Fig. 8. Performance measured in PR-AUC (top) and ROC-AUC (bottom) for COVID-19, Pneumonia and CKD. In each plot, the respective performance is shown over 5, 7, 14 and 30 days for mortality and ICU admission (in the COVID-19 case) or ICU re-admission (in the pneumonia and CKD cases).
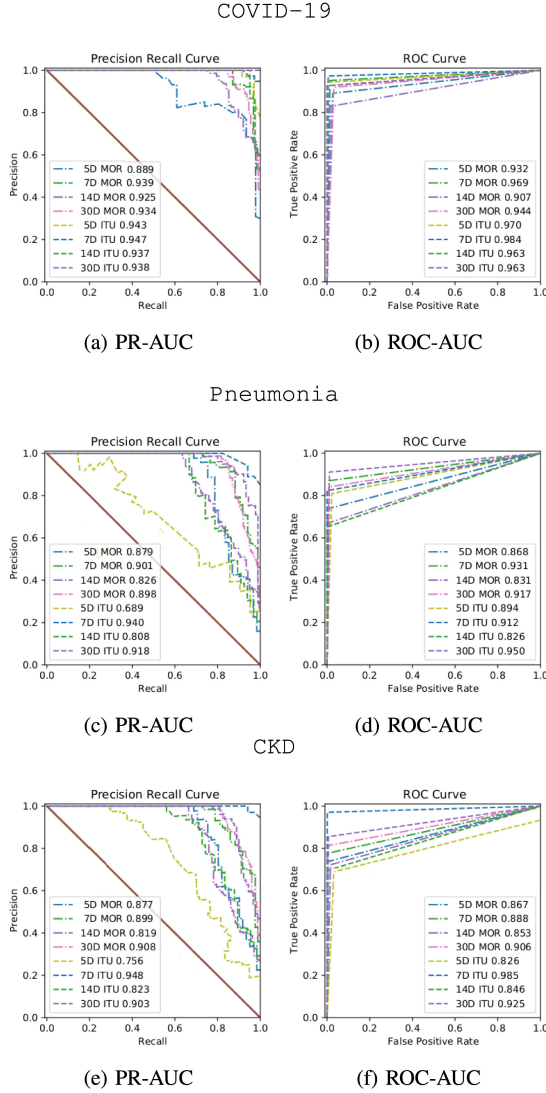
TABLE IV
COMPARISON OF DYNAMIC-KD'S PREDICTIONS WITH THE FINAL PREDICTION MADE BY KD-OP ACROSS THE CASE STUDIES, GROUPED BY PREDICTION INTERVAL (5, 7, 14 AND 30 DAYS) AND BY OUTCOME (MORTALITY, ICU ADMISSION/RE-ADMISSION)

Dataset	Outcome	Model	Averge Macro Pr	Average Macro Rec	Average Macro F1	PR-AUC	ROC-AUC
General (COVID)	5MOR	Dyn-KD	0.917	0.717	0.786	0.553	0.683
		KD-OP	0.865	0.932	0.913	0.889	0.932
	7MOR	Dyn-KD	0.931	0.882	0.927	0.855	0.887
		KD-OP	0.922	0.984	0.978	0.939	0.969
	14MOR	Dyn-KD	0.898	0.824	0.871	0.782	0.823
		KD-OP	0.896	0.907	0.925	0.925	0.907
	30MOR	Dyn-KD	0.915	0.893	0.922	0.885	0.895
		KD-OP	0.899	0.944	0.943	0.934	0.944
	Avg Δ		**-0.019**	**+0.128**	**+0.065**	**+0.153**	**+0.119**
	5ICU	Dyn-KD	0.918	0.897	0.925	0.853	0.932
		KD-OP	0.933	0.970	0.976	0.943	0.970
	7ICU	Dyn-KD	0.944	0.932	0.961	0.885	0.887
		KD-OP	0.933	0.970	0.976	0.947	0.984
	14ICU	Dyn-KD	0.929	0.826	0.950	0.782	0.823
		KD-OP	0.946	0.963	0.979	0.937	0.963
	30ICU	Dyn-KD	0.933	0.962	0.972	0.994	0.999
		KD-OP	0.946	0.963	0.979	0.938	0.963
	Avg Δ		**+0.008**	**+0.245**	**+0.026**	**+0.063**	**+0.059**
ICU (Pneum)	5MOR	Dyn-KD	0.938	0.736	0.896	0.72	0.8
		KD-OP	0.904	0.891	0.919	0.879	0.868
	7MOR	Dyn-KD	0.934	0.841	0.997	0.765	0.862
		KD-OP	0.911	0.892	0.923	0.901	0.931
	14MOR	Dyn-KD	0.934	0.841	0.897	0.73	0.802
		KD-OP	0.894	0.865	0.90	0.826	0.831
	30MOR	Dyn-KD	0.922	0.878	0.916	0.861	0.885
		KD-OP	0.917	0.888	0.921	0.898	0.894
	Avg Δ		**-0.026**	**+0.035**	**-0.011**	**+0.107**	**+0.044**
	5ICU	Dyn-KD	0.648	0.684	0.679	0.587	0.643
		KD-OP	0.737	.825	0.774	0.689	0.894
	7ICU	Dyn-KD	0.944	0.853	0.999	0.967	0.926
		KD-OP	0.931	0.926	0.951	0.940	0.912
	14ICU	Dyn-KD	0.950	0.742	0.816	0.561	0.713
		KD-OP	0.894	0.835	0.866	0.808	0.826
	30ICU	Dyn-KD	0.816	0.818	0.899	0.764	0.828
		KD-OP	0.931	0.894	0.932	0.918	0.950
	Avg Δ		**+0.034**	**+0.096**	**+0.032**	**+0.119**	**+0.086**
ICU (CKD)	5MOR	Dyn-KD	0.938	0.836	0.896	0.766	0.799
		KD-OP	0.909	0.897	0.915	0.887	0.867
	7MOR	Dyn-KD	0.934	0.841	0.915	0.728	0.855
		KD-OP	0.931	0.901	0.947	0.899	0.888
	14MOR	Dyn-KD	0.917	0.807	0.863	0.728	0.806
		KD-OP	0.909	0.863	0.902	0.819	0.853
	30MOR	Dyn-KD	0.918	0.879	0.916	0.862	0.903
		KD-OP	0.924	0.906	0.935	0.908	0.906
	Avg Δ		**-0.009**	**+0.050**	**+0.027**	**+0.107**	**+0.0375**
	5ICU	Dyn-KD	0.612	0.646	0.665	0.658	0.733
		KD-OP	0.727	0.772	0.739	0.756	0.826
	7ICU	Dyn-KD	0.944	0.853	0.911	0.778	0.890
		KD-OP	0.935	0.985	0.985	0.948	0.985
	14ICU	Dyn-KD	0.898	0.743	0.816	0.705	0.714
		KD-OP	0.898	0.846	0.920	0.823	0.846
	30ICU	Dyn-KD	0.934	0.861	0.941	0.785	0.859
		KD-OP	0.912	0.925	0.942	0.903	0.925
	Avg Δ		**+0.027**	**+0.106**	**+0.064**	**+0.126**	**+0.095**

(lower standard deviation) for both age and the number of pre-existing conditions. COVID-19 admissions are more uniform in age and comorbidity, consequently influencing changes in physiological states. In pneumonia, the standard deviation of comorbidity is higher in younger patients where the model achieves the lowest performance.

2) Performance Across Prediction Intervals and Outcome Distribution Settings: Fig. 8 shows *KD-OP*'s performance in predicting mortality and ICU admission/re-admission over 5, 7, 14 and 30-day intervals on COVID-19, pneumonia and CKD. Two observations can be made when examining this figure in conjunction with the distribution of the outcomes of Table II. Apart from 5-day ICU re-admission in pneumonia and CKD, (a) *KD-OP* shows high performance across short and long-term intervals, and (b) *KD-OP*'s performance remains high when the minority (positive) samples constitute $< 10\%$ of the overall population, which confirms the merit of relying on outlier detection

to construct the temporal representation used in the pipeline. These findings are in-line with the demographic diversity results of Section IV-C1, as the mean and standard deviations of age and the number of pre-existing conditions of those re-admitted to the ICU within 5 days were 33.76 (35.1) and 0.23 (0.2) for pneumonia and 66.12(21.3) and 0.28(0.17) for CKD respectively, showing a highly-diverse sample, deviating almost as highly as the overall pneumonia and CKD populations as shown in table II. In contrast, the 5-day ICU admission sample in the COVID-19 study had a mean and standard deviation of age and the number of pre-existing conditions of 63.3 (9.79) and 0.32 (0.05) respectively, showing a narrow range of demographical variation compared to pneumonia and CKD.

3) The Contribution of Static-OP vs Dynamic-KD: We now turn to compare the relative contribution of the two modules to *KD-OP*'s overall prediction across the three use cases, outcomes and four intervals. The detailed comparison is provided in table IV. In the table, we list the contribution of each module per

TABLE V
COMPARISON OF KD-OP'S PERFORMANCE WITH EXISTING LITERATURE OF GENERAL OUTCOME PREDICTION TESTED IN GENERAL WARDS AND ICU
SETTINGS. ADDITIONALLY, A COMPARISON WITH CURRENT STATISTICAL MODELS OF TO PREDICT DETERIORATION IN COVID-19 IS GIVEN

Model	Primary Outcome	Outcome Distribution	AUC	PR-AUC	Sensitivity	Specificity
General Ward						
NEWS2 [27]	Deterioration (1 day)	(94.8%, 5.2%)	0.78 (0.73 - 0.83)	NA	0.28 (0.21 - 0.37)	0.80 (0.79 0.82)
eCART [10], [29]	Unplanned ICU Admission,	(95%, 5%)	0.75 (0.74–0.75)	NA	NA	NA
	Mortality	(98.8%, 1.2%)	0.93 (0.93–0.93)	NA	NA	NA
	Cardiac Arrest	(99.5%, 0.05%)	0.89 (0.88–0.91)		0.89 (0.88–0.91)	0.52 (0.52–0.52)
DEWS [60]	Unplanned ICU Admission	(72.8%, 27.2%)	0.811 (0.811 - 0.822)	NA	0.555 (0.554 - 0.557)	0.90 (0.90 - 0.90)
	Mortality	(35.7%, 65.3%)	0.926 (0.926 - 0.927)	NA	0.831 (0.831 - 0.832)	0.888 (0.888 - 0.888)
LightGBM [35]	Mortality	(94.9%, 5.1%)	0.961 (NA)	NA	NA	0.641
KD-OP (General Ward)	**Mortality**	(82.9%, 18.1%)	0.927 (0.9 - 0.932)	0.922 (0.881 - 0.959)	0.905 (0.89 - 0.91)	0.94 (0.935 - 0.945)
	Unplanned ICU Admission	(91.19%, 8.81%)	**0.981 (0.98 - 0.984)**	0.948 (0.903 - 0.962)	0.966 (0.901-0.980)	0.914 (0.902 -0.934)
ICU Settings						
SANMF [41]	Mortality	(89.98%, 10.02%)	0.848 (0.846, 8.84)			
SICULA [50]	Mortality	(88.76%, 12.24%)	0.88 (0.87–0.89)	NA	NA	NA
KD-OP (ICU)	**Mortality**	(89.05%, 10.95%)	**0.881 (0.880 - 0.885)**	0.8841 (0.841 - 0.913)	0.953 (0.903 - 0.969)	0.918 (0.88 - 0.931)
COVID-19						
Carr [8]	Deterioration (14 days)	-	0.78 (0.74 - 0.82)			
Guo [22]	Deterioration (14 days)	-	0.67 (0.61 - 0.73)			
Liu [28]	Mortality & Deterioration	-	0.74 (0.69 - 0.79)			
Galloway [18]	Deterioration	-	0.72 (0.68 - 0.77)			
Gong [20]	Deterioration	-	0.853 (0.790- 0.916)			

prediction interval for each outcome using macro-averaged precision, recall and F1-score, as well as PR-AUC and ROC-AUC. We also show **avg** Δ, the average change in each metric's value between *Dynamic-KD* and the final prediction made by *KD-OP*. It is clear that the two modules complement each other to reach a high performance that is not otherwise achievable by the time-series predictor alone. This effect is especially noticeable in recall, where *Static-OP* often increases **avg** Δ. **avg** Δ of mortality outcomes also show that the stacked model slightly decreases *Dyanmic-KD*'s precision, indicating that the temporal signatures are somewhat better suited for identifying more patients who are at risk of mortality. However, the magnitude of the decrease in the overall precision is insignificant compared to the magnitude of increased recall, leading to a higher decrease in false alarms, which is a common bottleneck in clinical outcome prediction models [27].

Examining the performance from a domain angle, *Static-OP*'s contribution to the overall performance appears to be more pronounced in short-term outcomes. A highly noticeable difference is in the case of COVID-19 5-day mortality, where the average macro F1-score increases by 0.127 (from 0.786 using *Dynamic-KD* alone to 0.913 using the full pipeline). In contrast, the increase in F1 goes down to 0.021 (from 0.922 to 0.943) when examining 30-day mortality. This observation is consistent with current knowledge and recent findings that demographic information (e.g. age, pre-existing conditions) are highly predictive of short-term mortality in COVID-19 patients [69]. Similarly, for ICU readmission, replicated studies have found co-morbidities to be highly predictive of intensive care readmission during the same hospitalization [26].

4) Comparison With Existing Outcome Prediction Models: Here, we compare *KD-OP*'s performance with the reported performance of relevant adverse outcome prediction models, showing the results in Table V. We restrict the comparison to models that a) predict adverse outcomes (e.g. deterioration, mortality, re-admission, cardiac arrest), and b) report performance exceeding the NEWS2 baseline of ROC-AUC = 0.78. Having gone through the literature, the only machine learning frameworks found to have been validated in non-ICU settings were DEWS [60], eCART [10], [29], and LightGBM [35] so we list those first. As the NEWS2 score is widely used to predict deterioration, we include the latest evaluation of its performance

(from [27]) in the table for comparison. For these models, we compare their performance against the average performance of *KD-OP* when applied to the COVID-19 use-case, as it is a general ward population. For each model, we highlight the class distribution of the target outcome as reported by each model. For *KD-OP*, the class distribution was taken as the average distribution of the outcomes over the intervals evaluated (5, 7, 14 and 30 days), as shown in Table II. As none of the models reports PR-AUC, we compiled a list of all reported metrics, using them to compare against *KD-OP*'s performance in conjunction with the widely-used ROC-AUC.

For mortality, LightGBM achieved a high ROC-AUC of 0.961 with a class distribution of 5.1% of the outcome. However, the only other available metric for LightGBM is specificity, which is at a low 0.641 and entails a high rate of false alarms generated by the framework. eCart's ROC-AUC is equally distinctive at 0.93 with extremely low frequencies of mortality cases in the data used (1.2 % of the cases). However, no information on the recall or specificity of the model is available. We, therefore, draw attention the only outcome for which eCart's sensitivity and specificity are investigated, which is cardiac arrest. For this outcome, despite the high ROC-AUC achieved by the model (0.89) at a very low distribution rate of the outcome variable (0.05%), specificity is at 0.52, which once again shows the over-optimism of the ROC-AUC reported by the model. *KD-OP* achieved a ROC-AUC of 0.978, with mortality averaging at 18.1% of the samples (ranging between 10.88-26.23%). DEWS' AUC score was also high at 0.926. However, the class distribution reported in their study is highly skewed in favour of the outcome (65.3% mortality). With respect to unplanned ICU admissions, *KD-OP* achieved the highest AUC of 0.981 with an average class distribution of 8.81% (ranging between 8.22 −9.71%) for the four intervals. DEWS was the closest competitor at 0.811 AUC, albeit with a significantly higher distribution of the positive outcome (27%). Overall, *KD-OP* shows the highest performance stability across the two outcomes, rendering it a better candidate for general hospitalisation outcome prediction; especially given the lack of thorough assessment of competitive models using metrics suitable for the problem under study.

We also list high-performing machine learning models that have only been strictly validated in ICU settings; those include SANMF [41], SICULA (a.k.a. the super learner) [50] and [44].
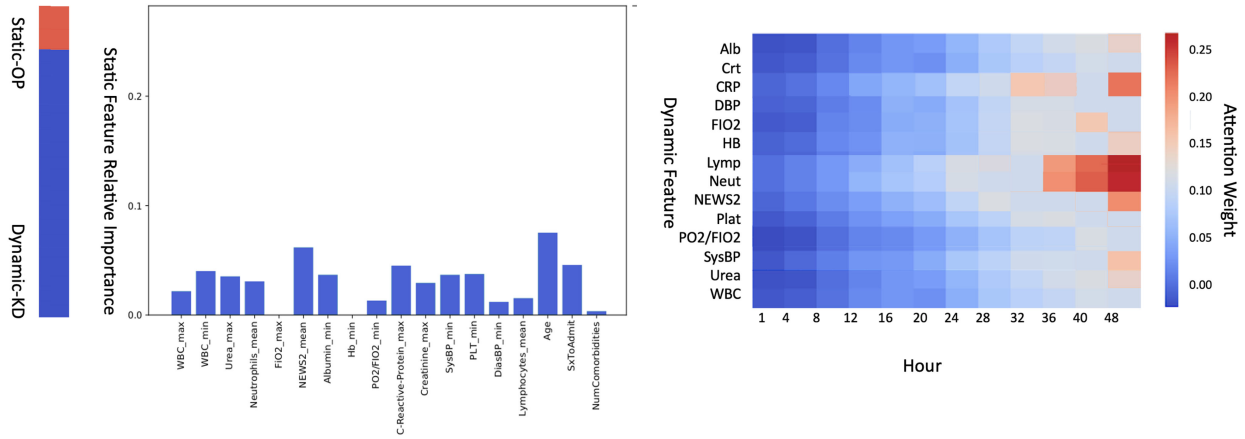
Fig. 9.   Visual Justification of KD-OP's Predictions.

It is worth noting that none of the models predicts ICU re-admission. We, therefore, resort to comparing with *KD-OP*'s average performance in predicting mortality when applied to pneumonia and CKD using the MIMIC-III ICU dataset. Also, apart from [44], which reports sensitivity, the models strictly rely on ROC-AUC to report their performance. We will, therefore, resort to comparing with *KD-OP*'s performance using ROC-AUC. As the table shows, *KD-OP* is the best predictor of mortality in an ICU setting, marginally exceeding SICULA's performance (ROC-AUC of 0.881 vs 0.880). Given that the SICULA's performance is the current benchmark for mortality prediction in the ICU, *KD-OP*'s performance is well-aligned with existing prediction potential.

Finally, as the literature now contains several statistical models aiming to make prognostic predictions of COVID-19 hospital admissions, we compare those with *KD-OP* applied to the COVID-19 case. It is worth noting that all of the models listed under the COVID-19 section of Table V are scoring systems aiming to mimic or exceed the performance of NEWS2 in predicting COVID-19 deterioration. Hence, *KD-OP* presents a novel contribution to the COVID-19 use case in being a scalable end-to-end machine learning architecture for predicting hospitalisation outcomes for COVID-19 admissions.

5) Visual Justification of Predicted Outcomes: The stacked nature of *KD-OP* naturally enables visualising its predictions using the built-in visualisation properties of each module and obtaining the relative contributions of each module's prediction to the outcome. For *Dynamic-KD*, the feature attention weights at each time window embedded in C (see output to Visualisation module in Fig. 3 and line 2 of Algorithm 1) make up the relative importance of the temporal signatures of each feature. On the other hand, the gradient boost implementation of *Static-OP* provides a feature importance capability, which we use to understand the relative contribution of each static feature. Since the relative contribution of each module to the final prediction is dependent on the outcome and prediction interval, including those variables in the visualisation of the output is highly essential for clinical utility as it directs the attention to the most contributing view (static or dynamic) of the patient. We

define the contribution of each module using the ratios of the respective modules' PR-AUC.

An example of the generated visualisation of a positive 30-day mortality outcome of a COVID-19 patient is shown in figure 9. In the figure, the left-most bar shows the relative contribution of the individual modules. In this scenario, *Dynamic-KD* is a significant contributor ($\approx 93\%$ of the overall contribution). Examining the attention weights generated by the framework shows that the highest weights are of lymphocytes and neutrophils (Lymph and Neut in the figure) at hours 36-48 as well as C-reactive protein (CRP in the figure) at time-step 48 (24th hour). We use this information as a justification for the prediction made by *Dynamic-KD*. On the other hand, the patient's age and the mean NEWS2 score show the highest importance among the static features, followed by the length of the period from symptoms to admission (SxToAdmit) and the maximum C-reactive protein level over the 24 hours.

It is essential to view this justification in relation to current findings. C-reactive protein, lymphocytes and lactic dehydrogenase have been recently found to be highly correlated with adverse outcomes in COVID-19 patients [70]. Although lactic dehydrogenase was not part of our COVID-19 dataset, the temporal signatures of both C-reactive protein and lymphocytes have been accurately identified by *Dynamic-KD* as predictors. This, in addition to age being marked as an important static feature, agrees with recent findings [69], showing that the signals jointly picked up by the framework's modules are coherent and well-aligned with clinical findings.

V. Discussion

We developed and validated *KD-OP*, an end-to-end pipeline for predicting adversity during hospitalisation. The pipeline comprises two stacked modules, each making predictions from a view of the patient's data: dynamic time-series and static features. The stacking of the pipeline's modules enables mimicking a clinician's approach to making prognostic decisions, by taking into account the interplay between the temporal signatures of a patient's physiology as well as time-invariant characteristics. By

design, the pipeline is cognizant of the class imbalance natural to hospitalisation outcome data. It is trained and validated using stratified data that retains the original distribution of the real outcomes within the population.

Model interpretability has been extensively linked with clinical utility and trust in clinical risk prediction systems [56], where clinicians are generally reluctant to accept decisions guided by 'black-box' Machine Learning models [55], [67]. *KD-OP* enables the post-hoc interpretation of its predictions by extending the idea of using attention weights to interpret neural network outcomes [9]. *KD-OP*'s visualisation component accounts for the interplay between dynamic and static features in justifying the predictions made by the pipeline, which is a feature that derives directly from the stacked architecture. To our knowledge, this feature is not available in any existing hospitalisation outcome predictor.

We evaluated *KD-OP*'s performance using real hospital data on three use cases representative of the diversity of electronic health records data. Using the pipeline to predict mortality and ICU admission/re-admission over 5-day, 7-day, 14-day, and 30-day intervals resulted in prediction accuracies exceeding 90% in all mortality outcomes and most of the ICU admission/re-admission outcomes.

KD-OP is among the few models available in the literature that have been validated in ICU and non-ICU settings. Adversity prediction in ICU settings is a less challenging endeavour due to the high volume of frequently-recorded variables and the relative uniformity of patients with respect to the level of acuity. In a non-ICU setting, *KD-OP* outperforms all existing models when considering the wide range of metrics needed to make an informed judgement about the models' predictive power. More importantly, *KD-OP* is the only model that has been validated using PR-AUC, which measures the model's ability to predict minority (adverse) outcomes.

The high performance achieved by *KD-OP* in predicting adverse outcomes in three diverse patient populations is a confirmation of the model's capacity in recognising *context*. The model's ability to identify the most relevant features for a given diagnosis reflects what one would like to see in a general outcome prediction model, where the primary diagnosis serves as a guide to making a prognosis about a given patient. A disease-agnostic model such as *KD-OP* could be built into the visual display of an EHR for all clinicians to use. The present challenge is that each hospital department has its outcome prediction scoring system, subsequently making it unrealistic to build over 30 distinct models into an EHR system. The generic nature of *KD-OP*, coupled with high performance and visualisation capability, gives it a broader potential for integration in ICU and non-ICU settings.

KD-OP can be extended in a number of ways. First, it would be interesting to project the progression of the risk of adversity; we are currently developing a temporal risk score model to predict and visualise the risk of a given outcome on an individual level over time, using *KD-OP* as the base model. Second, the current pipeline only supports classification outcomes. Existing targets include the prognosis of continuous outcomes such as worsening oxygenation or cardiac function. Also, the current framework strictly uses routinely collected clinical variables as

predictors. Other types of data can be of high relevance to a given use case. For example, ECG signals are the predictors of choice for cardiology-related outcomes; X-ray images can positively improve predictive power in the case of COVID-19, etc. Although the stacked architecture has proven to be highly robust compared to parallel ensembles, it is intrinsically less flexible towards incorporating additional models, which renders extending the stacked model an interesting research problem. Finally, we are currently extending the framework's justification component to incorporate both the magnitude and direction of variable contribution to *Static-OP* and *Dynamic-KD*'s predictions. However, our ultimate goal is to extend beyond a correlation-based visualisation and into a decision-theoretical framework enabling the contextual selection of predicted risk based on the modular and overall framework performance.

REFERENCES

[1] M. Aczon *et al.*, "Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks," *CoRR*, vol. abs/1701.06675, 2017.

[2] R. Armstrong *et al.*, "The incidence of cardiac arrest in the intensive care unit: A systematic review and meta-analysis," *Pediatr. Crit. Care Med.*, vol. 20, no. 2, pp. 144–154, 2019.

[3] ICNARC, "Audit: Key statistics from the case mix programme, Apr. 2011–Mar. 2012," 2013.

[4] A. Awad *et al.*, "Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach," *Int. J. Med. Informat.*, vol. 108, pp. 185–195, Oct. 2017.

[5] T. Bai, S. Zhang, B. Egleston, and S. Vucetic, "Interpretable representation learning for healthcare via capturing disease progression through time," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 43–51.

[6] I. Baytas, C. Xiao, X. Zhang, F. Wang, A. Jain, and J. Zhou, "Patient subtyping via time-aware LSTM networks," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2017, pp. 65–74.

[7] N. Brajer *et al.*, "Prospective and external evaluation of a machine learning model to predict in-hospital mortality of adults at time of admission," *JAMA Netw. Open*, vol. 3, no. 2, 2020, Art. no. e1920733.

[8] E. Carr *et al.*, "Evaluation and improvement of the national early warning score (news2) for COVID-19: A multi-hospital study," *BMC Biomed.*, vol. 19, Art. no. 23, 2021.

[9] 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. 30th Conf. Neural Inf. Process. Syst.*, 2016.

[10] M. M. Churpek *et al.*, "Multicenter development and validation of a risk stratification tool for ward patients," *Amer. J. Respir. Crit. Care Med.*, vol. 190, no. 6, pp. 649–55, 2014.

[11] L. Clifton, D. A. Clifton, M. A. F. Pimentel, P. J. Watkinson, and L. Tarassenko, "Gaussian processes for personalized e-health monitoring with wearable sensors, *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 193–197, Jan. 2013.

[12] R. Das and D. Wales, "Machine learning landscapes and predictions for patient outcomes," *Roy. Soc. Open Sci.*, vol. 4, no. 7, 2017, Art. no. 170175.

[13] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. Mach. Learn.*, 2006, pp. 233–240.

[14] S. Dzeroski and B. Zenko, "Is combining classifiers with stacking better than selecting the best one?," *Mach. Learn.*, vol. 54, pp. 255–273, 2004.

[15] C. Feng *et al.*, "Log-transformation and its implications for data analysis," *Shanghai Arch. Psychiatry*, vol. 26, no. 2, pp. 105–109, 2014.

[16] H. Fernandez *et al.*, *Learning From Imbalanced Data Sets*. Berlin, Germany: Springer, 1st ed., 2018, pp. 25.

[17] C. Ferri, J. Hernandez-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognit. Lett.*, vol. 30, no. 1, pp. 27–38, 2009.

[18] J. Galloway *et al.*, "A clinical risk score to identify patients with covid-19 at high risk of critical care admission or death: An observational cohort study," *J. Infect.*, vol. 81, no. 2, pp. 282–288, 2020.

[19] Z. Ghrib, R. Jaziri, and R. Romdhane, "Hybrid approach for anomaly detection in time series data," in *Proc. Int. Joint Conf. Neural Netw.*, 2020, pp. 1–7.

[20] J. Gong *et al.*, "Correlation analysis between disease severity and inflammation-related parameters in patients with COVID-19 pneumonia," *BMC Infect. Dis.*, vol. 20, Art. no. 963, 2020.

[21] P. Griffiths *et al.*, "Nurse staffing, nursing assistants and hospital mortality: Retrospective longitudinal cohort study," *BMJ Qual. Saf.*, vol. 28, no. 8, pp. 609–617, 2019.

[22] T. Guo *et al.*, "Cardiovascular implications of fatal outcomes of patients with coronavirus disease 2019 (COVID-19)," *JAMA Cardiol.*, vol. 5, no. 7, pp. 811–818, Jul. 2020.

[23] T. Guo, T. Lin, and N. Antulov-Fantulin, "Exploring interpretable LSTM neural networks over multi-variable data," in *Proc. Mach. Learn. Res.*, 2019, pp. 2494–2504.

[24] H. Haibo and M. Yunqian, *Imbalanced Learning: Foundations, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley-IEEE Press, 1st ed., 2013, p. 27.

[25] T. L. Higgins *et al.*, "Assessment of the accuracy of using ICD-9 diagnosis codes to identify pneumonia etiology in patients hospitalized with pneumonia," *JAMA Netw. Open*, vol. 3, no. 7, 2020, Art. no. e207750.

[26] K. Ho *et al.*, "The effect of comorbidities on risk of intensive care readmission during the same hospitalization: A linked data cohort study," *J. Crit. Care*, vol. 24, no. 1, pp. 101–107, 2009.

[27] L. Hodgson *et al.*, "NEWS 2 - too little evidence to implement?," *Clin. Med.*, vol. 18, no. 5, pp. 371–373, 2018.

[28] J. Liu *et al.*, " Clinical outcomes of COVID-19 in Wuhan, China: A large cohort study," *Ann. Intensive Care*, vol. 10, pp. 282–288, Dec. 2020.

[29] J. Rojasm, K. Carey, D. Edelson, L. Venable, M. Howell, and M. Churpek "Predicting intensive care unit readmission with machine learning using electronic health record data," *Ann. Amer. Thoracic Soc.*, vol. 15, no. 7, pp. 846–853, 2018.

[30] A. Johnson *et al.*, "MIMIC-III, A freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, 2016, Art. no. 160035.

[31] J. Johnson *et al.*, "Artificial neural networks improve early outcome prediction and risk classification in out-of-hospital cardiac arrest patients admitted to intensive care," *Crit. Care*, vol. 24, no. 1, Art. no. 474, 2020.

[32] G. Ke *et al.*, " LightGBM: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 3146–3154, 2017.

[33] T. Kieu, B. Yang, C. Guo, and C. Jensen, "Outlier detection for time series with recurrent autoencoder ensembles," in *Proc. Int. Joint Conf. Artif. Intell.*, vol. 7, 2019, pp. 2725–2732.

[34] T. Kieu, B. Yang, and C. S. Jensen, "Outlier detection for multidimensional time series using deep neural networks," in *Proc. IEEE Conf. Mobile Data Manage.*, 2018, pp. 125–134.

[35] J. Kobylarz *et al.*, "A machine learning early warning system: Multicenter validation in brazilian hospitals," in *Proc. IEEE Int. Symp. Comput.-Based Med. Syst.*, vol. 06 2020, pp. 321–326.

[36] G. Kong, K. Lin, and Y. Hu, "Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU," *BMC Med. Informat. Decis. Mak.*, vol. 20, no. 1, Art. no. 251, 2020.

[37] W. C. Lee *et al.*, "The number of comorbidities predicts renal outcomes in patients with stage 3-5 chronic kidney disease," *J. Clin. Med.*, vol. 7, no. 12, Art. no. 493, 2018.

[38] A. Leontjeva and I. Kuzovkin, "Combining static and dynamic features for multivariate sequence classification," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal.*, vol. 10, 2016, pp. 21–30.

[39] X. Li, D. Zhu, and P. Levy, "Predicting clinical outcomes with patient stratification via deep mixture neural networks," in *Proc. Amer. Med. Informtics Annu. Symp.*, 2020, pp. 367–376.

[40] V. Liu *et al.*, "Comparison of early warning scoring systems for hospitalized patients with and without infection at risk for in-hospital mortality and transfer to the intensive care unit," *JAMA Netw. Open*, vol. 3, no. 5, 2020, Art. no. e205191.

[41] Y. Luo *et al.*, "Predicting ICU mortality risk by grouping temporal trends from a multivariate panel of physiologic measurements," in *Proc. AAAI Conf. Artif. Intell.*, 2016,pp. 42–50.

[42] P. Malhotra *et al.*, "LSTM-Based encoder-decoder for multi-sensor anomaly detection," *CoRR*, vol. abs/1607.00148, 2016.

[43] M. Mallappallil *et al.*, " Chronic kidney disease in the elderly: Evaluation and management," *Clin. Pract.*, vol. 11, no. 5, pp. 525–535, 2014.

[44] M. Moreno-García *et al.*, "Machine learning methods for mortality prediction of polytraumatized patients in intensive care units - dealing with imbalanced and high-dimensional data," in *Proc. Intell. Data Eng. Automated Learn.*, 2014,pp. 309–317.

[45] E. Muthumbi *et al.*, "Risk factors for community-acquired pneumonia among adults in Kenya: A case-control study," *Pneumonia*, vol. 9, Art. no. 17, 2017.

[46] J. Nogas, S. Khan, and A. Mihailidis, "Fall detection from thermal camera using convolutional LSTM autoencoder," in *Proc. Workshop AI Aging, Rehabil. Independent Assist. Living, Easy Chair*, 2019.

[47] E. Oh and H. Lee, "An imbalanced data handling framework for industrial big data using a Gaussian process regression-based generative adversarial network," *Symmetry*, vol. 12, no. 4, 2020.

[48] B. Ozenne, F. Subtil, and D. Maucort-Boulch, "The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases," *J. Clin. Epidemiol.*, vol. 68, no. 8, 855–859, 2015.

[49] F. Perrotta *et al.*, "COVID-19 and the elderly: Insights into pathogenesis and clinical decision-making," *Aging Clin. Exp. Res.*, vol. 32, no. 8, pp. 1599–1608, 2020.

[50] R. Pirracchio, *Mortality Prediction in the ICU Based on MIMIC-II Results From the Super ICU Learner Algorithm (SICULA) Project*. Berlin, Germany: Springer, Sep. 2016, pp. 295–313.

[51] R. Raj *et al.*, "Machine learning-based dynamic mortality prediction after traumatic brain injury," *Nature Sci. Rep.*, vol. 9, 2019, Art. no. 17672.

[52] C. E. Rasmussen, *Gaussian Processes in Machine Learning*. Berlin Heidelberg: Springer, 2004, pp. 63–71.

[53] S. Romero-Brufau, J. Huddleston, G. Escobar, and M. Liebow, "Why the c-statistic is not informative to evaluate early warning scores and what metrics to use," *Crit. Care*, vol. 19, no. 1, Art. no. 285, 2015.

[54] Royal College of Physicians, "National early warning score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS," Royal College of Physicians, London, U.K., Working Party Rep., 2017.

[55] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, pp. 206–215, 2019.

[56] S. Wongvibulsin, K. Wu, and S. Zeger, "Improving clinical translation of machine learning approaches through clinician-tailored visual displays of black box algorithms: Development and validation," *J. Med. Informat. Res.*, vol. 8, no. 6, 2020, Art. no. e15791.

[57] U. Schmidt-Erfurth *et al.*, "Artificial intelligence in retina," *Prog. Retinal Eye Res.*, vol. 67, pp. 1–29, 2018.

[58] P. Schwab *et al.*, "Beat by beat: Classifying cardiac arrhythmias with recurrent neural networks," *Comput. Cardiol.*, 2017, pp. 1–4.

[59] F. Shamout *et al.*, "An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department," *Nature Digit. Med.*, vol. 4, Art. no. 80, 2021.

[60] F. E. Shamout, T. Zhu, P. Sharma, P. J. Watkinson, and D. A. Clifton, "Deep interpretable early warning system for the detection of clinical deterioration," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 437–446, Feb. 2020.

[61] S. P. Shashikumar *et al.*, "Detection of paroxysmal atrial fibrillation using attention-based bidirectional recurrent neural networks," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 715–723.

[62] V. Snorovikhina and A. Zaytsev, "Unsupervised anomaly detection for Â discrete sequence healthcare data," in *Proc. Anal. Images, Social Netw. Texts,* 2021, pp. 391–403.

[63] A. D. Storms *et al.*, "Rates and risk factors associated with hospitalization for pneumonia with ICU admission among adults," *BMC Pulmonary Med.*, vol. 17, Art. no. 208, 2017.

[64] H. Thorsen-Meyer *et al.*, "Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: A retrospective study of high-frequency data in electronic patient records," *Lancet Digit. Health*, vol. 2, no. 4, pp. E179–E191, 2020.

[65] S. LekhaTummalapalli and C. A. Peralta, "An electronic CKD phenotype: A step forward in improving kidney care," *Clin. J. Amer. Soc. Nephrol.*, vol. 14, no. 9, pp. 1277–1279, 2019.

[66] L. Turgeman and J. May, "A mixed-ensemble model for hospital readmission," *Artif. Intell. Med.*, vol. 72, pp. 72–82, Aug. 2016.

[67] J. van der Waa, T. Schoonderwoerd, J. van Diggelen, and M. Neerincx, "Interpretable confidence measures for decision support systems," *Int. J. Hum.-Comput. Stud.*, vol. 144, 2020, Art. no. 102493.

[68] X. Xie *et al.*, "Real-time illegal parking detection system based on deep learning," in *Proc. ACM Deep Learn. Technol.*, 2017, pp. 23–27.

[69] A. Yadaw *et al.*, "Clinical features of COVID-19 mortality: Development and validation of a clinical prediction model," *Lancet Digit. Health*, vol. 2, no. 1, pp. E516–E525, 2020.

[70] L. Yan, H. Zhang, and Y. Yuan, "An interpretable mortality prediction model for COVID-19 patients," *Nature Mach. Intell.*, vol. 2, pp. 283–288, 2020.