2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871500

Using Gated Recurrent Unit Networks for the Prediction of Hemodynamic and Pulmonary Decompensation

Christian Mandel[1], Kathrin Stich[2], Serge Autexier[1], Christoph Lüth[1], Ariane Ziehn[3], Karin Hochbaum[2], Rolf Dembinski[2] and Christoph Int-Veen[4]

***Abstract*— This paper presents a new medical severity scoring system, used to assess the risk of hemodynamic and pulmonary decompensation for patients being treated in intensive care units. The score presented here includes drug circulatory support and ventilation mode data for the evaluation of the patient's biosignals and laboratory values. It is shown that Gated Recurrent Unit-based neural networks are able to predict the maximal severity class within a 24 hour prediction time-frame (hemodynamic: 0.85 AUROC / pulmonary: 0.9 AUROC), and can estimate the underlying decompensation score for prediction times of up to 24 hours with mean errors of 6.3% of the maximal possible pulmonary, and 9.6% of the hemodynamic score. These results are based on 60h observation period.**

***Clinical relevance*—Hemodynamic and pulmonary decompensation are life threatening, dynamic events that can lead to death of patients. Early detection of these incidents is essential in order to intervene therapeutically and to improve survival chances. In everyday intensive care physicians are confronted with a vast number of laboratory values and vital parameters. There is a risk that early stages of hemodynamic and pulmonary decompensation are misjudged. The implementation of robust warning systems could support physicians in detecting these critical events and initiate therapeutical intervention in time, which would achieve significant reduction of patient mortality.**

I. INTRODUCTION

Hemodynamic and pulmonary decompensation (DEC) describe the rapid or slow progressive deterioration of the heart's or lung's functionality respectively. While the effects of an underlying disease could be compensated for a certain time, the incipient DEC might lead to life threatening situations for the patient. Since the dynamics of these processes pose a challenging task for clinicians, it is desirable to support the diagnosis by robust prediction tools.

For this purpose we describe classification- and prediction-systems that have been trained on a real-world data set containing more than 10k anonymized patient records, all collected between 2013 and 2021 at an intensive care unit (ICU) of the Klinikum Bremen-Mitte in Germany. A patient's record contains a set of multiple time series over assessment data, i.e. biosignals, medications, lab results, total balance of excretions, as well as treatment and demographic information. In order to predict the maximal hemodynamic and pulmonary severity class for a certain prediction time interval, we need to map the patient data to two numerical scores which allow the severity to be gauged. The scores allow automatic labelling of the data, and thus training of the prediction network. They are defined by two sets of rules that map relevant medical data to DEC scores describing the seriousness of DEC for each point in time, discretized into three classes, i.e. *none*, *beginning-moderate*, and *severe* DEC. The machine learning part trains deep Gated Recurrent Unit (GRU) network models on observation periods over the patients' assessment and treatment data, or alternatively over the DEC score time series. For both approaches, we can calculate the ground truth labels of subsequent prediction time windows (classification task), or the ground truth score itself (score prediction task).

The rest of this paper is structured as follows. In Sec. II, we present established medical scoring systems, and give a brief overview on time series classification with recurrent neural networks applied in medical data analysis. Sec. III describes the proposed classification-/ and prediction-systems in detail, such as data preprocessing, the GRU models, and (hyper-)parameter selection. In Sec. IV, we present and discuss results of our approach w.r.t. a retrospective data set containing $>$10k anonymized patient records. In Sec. V we conclude and lay out future work ideas.

II. RELATED WORK

Scoring systems are established and routinely used tools in the field of intensive care medicine. They depict the complex clinical condition of each patient at a certain time, based on treatment and biometric data, vital parameters, laboratory values and create a one-dimensional scale. This reduction of clinical data provides an objective, comparative assessment of patients condition, disease severity and allows the outcome to be predicted. Additionally, scores are used in research, health economics, quality assurance and medical education.

Specific scores like the Glasgow Coma Scale [14] or the Sequential Organ Failure Assessment [15] quantify special diseases or organ failure. Severity scores like the Acute Physiology and Chronic Health Evaluation (APACHE IV) [18]

This work has been conducted in context of the project *RIDIMP – Risikoindikatoren für cardiopulmonale Dekompensation auf Intensivstationen durch Monitoring von Vitalparametern*. RIDIMP is partially funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWi) as part of the joint project *KI-SIGS – Spaces für Intelligente Gesundheitssysteme* under grant agreement 01MK20012P.

[1]Research Department Cyber-Physical Systems, German Research Center for Artificial Intelligence, 28359 Bremen, Germany. {Christian.Mandel, Serge.Autexier, Christoph.Lueth}@dfki.de

[2]Gesundheit Nord gGmbH, Klinikverbund Bremen, 28211 Bremen, Germany. Kathrin.Stich@klinikum-bremen-mitte.de, {Karin.Hochbaum, Rolf.Dembinski}@gesundheitnord.de

[3]Research Department Intelligente Analytik für Messdaten, German Research Center for Artificial Intelligence, 10559 Berlin, Germany. Ariane.Ziehn@dfki.de

[4]Philips GmbH Market DACH, 22335 Hamburg, Germany. christoph.int-veen@philips.com

This work is licensed under a Creative Commons Attribution 3.0 License.
For more information, see http://creativecommons.org/licenses/by/3.0/

and the Simplified Acute Physiology Score (SAPS III) [11] categorize the common physiological status. Therapeutic and care procedures are evaluated by the Therapeutic Intervention Scoring System (TISS-28) [10]. In contrast to these scoring systems, we focused on hemodynamic and pulmonary DEC. In order to define these two events, we assessed commonly used basic parameters, as well as medication and ventilatory support information. Parameters that describe a specific entity or advanced monitoring, which is used in late, severe phase of DEC, are left out in order to capture early stages of DEC and different entities.

Machine learning-based prediction of cardiac or respiratory failure is a well studied field (cf. [16] for a literature review). For example, Kim et al. [7] describe LSTM networks for the prediction of cardiac arrest (AUROC 0.886) and respiratory failure (AUROC 0.869) $1h$ to $6h$ prior to its occurrence. These results are based on medical time series collected over the patients' whole stay on the ICU with a sampling frequency $f = 1$ datapoint/hour. Kwon et al. [9] report similar LSTM-results for predicting cardiac arrest (AUROC 0.85) up to $24h$ before the event by using $8h$ of input data ($f = 6$ datapoints/hour).

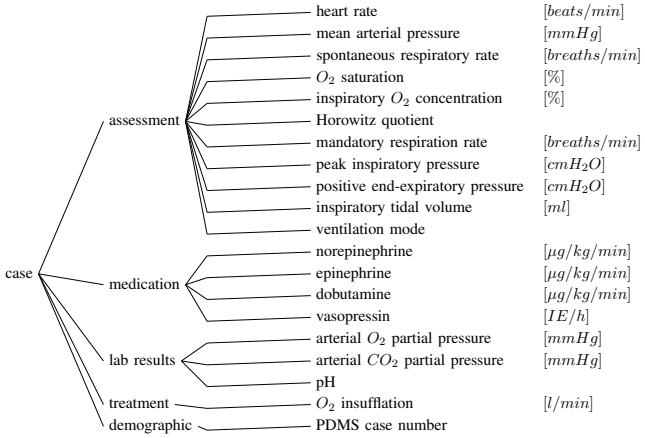


Fig. 1: Subset of information from the patient data management system that is used for the prediction of hemodynamic and pulmonary DEC. Note that the PDMS case number served as a case identifier for the medical partners that had access to non-anonymized data for verification of derived DEC scores, while the data processing team members worked solely on anonymized data.

III. METHODS

A. Patient Data Management System

The basis of machine learning models used for prediction and classification in the medical domain, is given by knowledge stored in patient data management systems (PDMS). In this work, we import data from the *IntelliSpace Critical Care and Anesthesia Data* system, developed by Philips [5].

Each *case*, i.e., a patient data record, contains the following *data groups* amongst others: assessment, medication, lab result, treatment and demographic information. Data groups itself provide *data entries*, such as time series of observations made during the patient's hospital stay. According to a white paper defined by the medical project partners, and approved for this retrospective study by the ethical committee, we received up to 106 anonymized data entries from six different data groups for each case. For the classification and prediction tasks under scope, we focussed on 19 data entries out of 5 data groups (cf. Fig. 1), required to calculate the DEC scores defined in Sec. III-C.

B. Data Filtering and Preprocessing

Due to missing or corrupt information (e.g. malfunction in sensors, data persistence issues) and the asynchronous data recording process (e.g. high frequent blood pressure readings vs. low frequency blood gas analysis), the patients' clinical data records need to be cleaned before they can be used by classification algorithms. In the following, we describe preprocessing steps yielding clean and synchronized time series that can be fed into the classification/prediction algorithms described in Sec. III-D.

1) Padding: The first step in the preprocessing pipeline is to search for empty data entries. According to data recording procedures, missing values for medication group entries indicate that the corresponding drug has not been applied. Since this situation is indistinguishable from cases where recordings of drug dispensing have been lost, we pad missing medication entries by dedicated values termed NaN.

2) Corrupt Case Removal: We define a case as being corrupt, and thus remove it from further processing, if one or more of the following conditions hold:

- The case is missing a complete data entry that is required to compute the DEC scores (except the medication group that has been padded in step (1)).
- The case is missing a PDMS case number. Predictions of our system cannot be verified by the clinicians without this number.
- The time spanned by the case's data entries' time series, i.e., the documented time of hospital stay, is shorter than the observation period and prediction time frame required by the classification algorithm (cf. Sec.III-D).

3) Aligning Start- and Endtime: With the goal to equalize the lengths of a case's time series, we define the case's start time s and end time e as the earliest and latest point in time at which its data entries (cf. Fig.1) contain a piece of information. For time series that start after s, or end before e, we insert a NaN value at the corresponding point in time.

4) Resampling: Retrospective data imported from the PDMS is available with a maximal rate of approx. one data point per hour. Even data recorded at high frequency, such as arterial pressure or heart rate, is down-sampled for storage reasons several days after the patient is discharged from the hospital. The more important reason for resampling is the asynchronous structure of the time series. By starting at a case's aligned start time s, we resample each time series with a frequency of one data point per hour by averaging. Missing input values are set to NaN.

5) Missing Value Imputation: Considering the prediction tasks at hand, i.e., establishing a mapping between time series

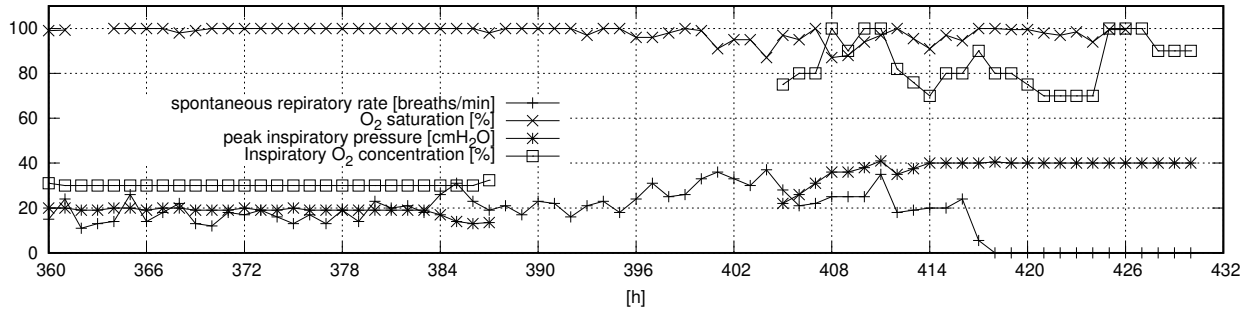


Fig. 2: Exemplary assessment time series of case 944, which are required for the calculation of the pulmonary DEC score (cf. Sec.III-C). The graphs depicted result from preprocessing and filtering steps described in Sec.III-B (except standardization), and zoom into the time frame containing the second pulmonary DEC peak in Fig. 3.

and DEC scores/classes, it becomes clear that gaps in the input data disturb training and inference of the networks. Our missing data imputation scheme includes two steps. First, we apply linear interpolation for 2 hour gaps in assessment entries (except ventilation mode), as well as for treatment and lab result time series. After standardization (cf. next paragraph), and directly before feeding the data into the GRU network, we replace remaining NaN values (gaps of > 2 hour length) by 0 (the mean of standardized features).

6) Standardization: Recurrent neural networks work best if the different types of input data are rescaled to a common domain. This behavior is due to the fact that input data adjust the network's layer weights during the training process, and certain dimensions with higher value ranges might exert a higher influence, even if they are less important to the network's task. Therefore, we standardize the neural network's input features, i.e., the single time series, to 0-centered mean (μ) and ± 1 for the standard deviation (σ). Note that standardization is executed after the calculation of DEC scores (cf. Sec III-C). We further strictly split training and test data sets, i.e., the features' μ and σ are calculated separately for training and test data.

C. Decompensation Scores

In comparison to established scoring systems (cf. Sec. II), we define two separate scores to describe the event of hemodynamic or pulmonary DEC during treatment on the ICU: The hemodynamic DEC score includes the vital parameters mean arterial pressure and heart rate in relation to drug circulatory support (cf. TABLE I). The pulmonary DEC score includes spontaneous respiratory rate, peripheral oxygen saturation and end-tidal carbon dioxide as measured vital parameters, different laboratory values taken from blood gas analysis and oxygenation or support by a ventilator (cf. TABLE II). On the basis of these parameters, we define the pulmonary and hemodynamic DEC scores as given in eqn. (1)-(2), as well as severity thresholds as shown in TABLE III.

In addition to the DEC score time series, we calculated confidence time series. For each point in time t of a single case, they describe the fraction of available parameters at t, and thus support the classifier when being trained solely on DEC score time series (cf. Sec. IV-A).

TABLE I: Hemodynamic DEC scores for different parameter.

$s(hp_i(t))$: / $hp_i(t)$:	0	1	2	3	4
$hp_1(t)$ =heart rate	50-90	45-49|91-100	40-44|101-110	40-44|101-110	<40|>110
$hp_2(t)$ =mean arterial pressure	65-80	60-64	50-59	50-59	<50
$hp_3(t)$ =catecholamine therapy	none	singular	singular	combined	singular or combined (in high dose)
norepinephrine	0	0.01-0.09	0.1-0.4	0.1-0.4	>0.4
epinephrine	0	0.01-0.09	0.1-0.4	0.1-0.4	>0.4
dobutamine	0	1-3	3.1-5	3.1-5	>5
vasopressin	0	0	0	0	>0.01

TABLE II: Pulmonary DEC scores for different parameter.

$s(pp_i(t))$: / $pp_i(t)$	0	1	2	3
$pp_1(t)$ =spontaneous respiratory rate	10-25	26-30	31-35	>35
$pp_2(t)$ =O_2 saturation	96-100	95-90	85-89	<85
$pp_3(t)$ =arterial O_2 partial pressure	70-100	69-65	64-60	< 60
$pp_4(t)$ =arterial CO_2 partial pressure	35-45	30-34|46-49	25-29|50-58	<29|>59
$pp_5(t)$ =pH	7.35-7.45	7.46-7.49| 7.26-7.34	7.5-7.55|7.16-7.25	>7.55|<7.15
$pp_6(t)$ =inspiratory O_2 concentration	30-35	36-49	50-60	61-100
$pp_7(t)$ =O_2 insufflation	0	2-5	6-8	>8
$pp_8(t)$ =Horowitz quotient	400-600	200-399	100-199	<100
$pp_9(t)$ =mandatory respiration rate	10-20	21-23	24-26	>26
$pp_{10}(t)$ =peak inspiratory pressure	10-25	26-28	29-30	≥31
$pp_{11}(t)$ =positive end-expiratory pressure	5-8	9-11	12-15	16-25
$pp_{12}(t)$ =inspiratory tidal volume	401-500	301-400	201-300	<200
$pp_{13}(t)$ =ventilation mode	spontaneous breathing	oxygen insufflation	assisted spontaneous breathing	bivent

$$hds(t) = \sum_{i=1}^{3} s(hp_i(t)) \tag{1}$$

$$pds(t) = \sum_{i=1}^{13} s(pp_i(t)) \tag{2}$$

TABLE III: Score intervals for different DEC classes.

class of decompensation	$hds(t)$	$pds(t)$
none	$0-3$	$0-4$
beginning-moderate	$4-5$	$5-20$
severe	>5	>20

D. Recurrent Neural Networks

Time series classification (TSC) is a long studied field in Artificial Intelligence that maps one or more time ordered lists of data to probabilities over classification labels [1]. Classical approaches include the *Dynamic Time Warping* algorithm [13], shapelet-based methods [17], or nearest neighbor calculations based on Mahalanobis-distances between time series and class representatives [12].

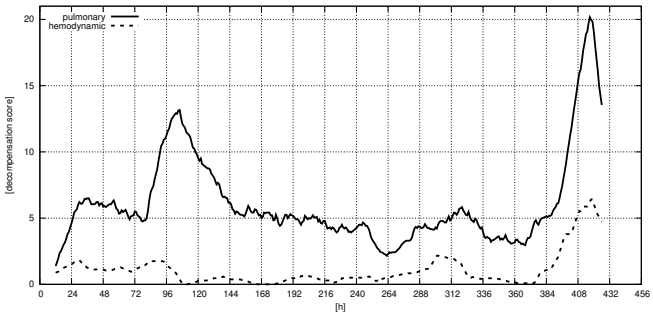
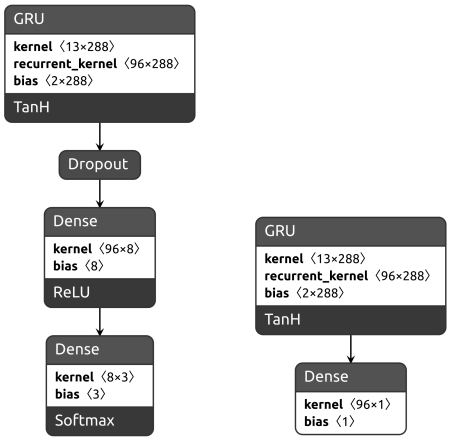


Fig. 3: One day moving averaged DEC scores over time of stay for case 944. Note that both pulmonary peaks build up within approx. 24h.



(a) classification network (b) prediction network

Fig. 4: Illustration of GRU-based network architectures used for severity of DEC classification and DEC score prediction, both for the pulmonary case. Networks for the hemodynamic case differ in the size of the GRU-layers' kernel, i.e. $\langle 6 \times 288 \rangle$, reflecting the different number of input parameter. For networks that solely work on DEC score and confidence time series, the kernel size is given by $\langle 2 \times 288 \rangle$.

The successful use of deep artificial neural networks (DNN) in the computer vision domain has triggered substantial research that investigated the usefulness of DNNs for TSC [3]. The basic principle is to apply recurrent neural networks (RNNs) that not only propagate the input data in a unidirectional way through the net of weights, but feed (parts of) a layer's output to itself or a preceding layer. Although RNNs are able to model time-dependencies in the input data, they are prone to exploding error gradients calculated during the network's training phase. This problem can be tackled by utilizing *Long Short-Term Memory* (LSTM) network cells [6]. Instead of common network cells that maintain a single state, LSTM cells maintain two internal states, i.e. the *long-term* and the *short-term memory*. These states are updated while the time-ordered input data passes through the unfolded LSTM-network via so-called gates. The *forget gate* controls which parts of the long-term memory (representing already processed time steps) can be forgotten. The *input gate* controls which parts of the input data's current time step and the long term memory of the previous time step enter into the current long-term memory. The *output gate* finally controls the output of the long term memory, either to be used as an input for the subsequent LSTM fold, or as the output of the overall LSTM network layer.

The work presented in this paper uses *Gated Recurrent Unit* (GRU) network cells [2]. In contrast to LSTM cells, GRU cells maintain a single internal state, and require 4 instead of 6 update equations. Thus, they contain fewer weights to optimize (weight terms W and bias terms b in equations (3)-(6)), making them less prone to overfitting, and showing faster convergence compared to LSTMs. Equations (3)-(6) describe a GRU cell's update rules [4].

$$z_{(t)} = \sigma\left(W_{xz}^T x_{(t)} + W_{hz}^T h_{(t-1)} + b_z\right) \quad (3)$$

$$r_{(t)} = \sigma\left(W_{xr}^T x_{(t)} + W_{hr}^T h_{(t-1)} + b_r\right) \quad (4)$$

$$g_{(t)} = \tanh\left(W_{xg}^T x_{(t)} + W_{hg}^T \left(r_{(t)} \odot h_{(t-1)}\right) + b_g\right) \quad (5)$$

$$h_{(t)} = z_t \odot h_{(t-1)} + \left(1 - z_{(t)}\right) \odot g_{(t)} \quad (6)$$

Here, $z_{(t)}$ describes the GRU cell's gate controller that controls the input gate and the forget gate. The gate controller $r(t)$ controls which parts of the GRU's previous fold output is forwarded to the cell's main layer $g_{(t)}$. The cell's output is given by $h_{(t)}$, and serves either as an input for the subsequent GRU fold, or as the output for the overall GRU layer.

E. Selected (Hyper-)Parameter and Implementation Details

The ability of neural networks to generalize from training data heavily depends on the selected optimizer and further (hyper-)parameter. In this work, we used the Adam optimizer [8] with a static learning rate $lr = 0.04$, exponential decay rates $\beta_1 = 0.9$, $\beta_2 = 0.999$, and an early stopping of 100 training epochs if the loss does not decrease over 3 epochs. Further major configuration is given by batchsize $bs = 512$, the number of GRU layer units $glu = 96$, and for the classification networks the number of dense layer units $dlu = 8$, and a dropout rate of $dr = 0.33$.

The methods presented in Sec. III have been implemented on an Intel(R) Xeon(R) Gold 5218 CPU with 64 GB of available system RAM. For training and inference of the network models, we used a Nvidia Tesla V100 GPU with 32GB of available memory. This research-server has been setup within the protected IT-environment of the *Gesundheit Nord* (GeNo) hospital association, and was reachable for the data scientists via dedicated VPN access. We implemented the described system inside a *Docker* container, running on top of the Ubuntu 20.04.2 LTS OS. For data import, preprocessing, network training/inference, and analyses we used *pandas*, *NumPy*, *SciPy*, *scikit-learn*, and *TensorFlow*.

IV. EVALUATION

A. Research Question and Approach

The evaluation investigated the question of how well the GRU networks described in Sec. III-D are suitable to predict hemodynamic and pulmonary DEC, given a certain timeframe of observation data for training and inference cases.

To answer this question we 1) trained classification-networks (cf. Fig. 4a) that estimate the maximal severity-class (cf. Tab III) within a given prediction timeframe, and 2) trained prediction-networks that estimate the numerical DEC scores for some fixed points in time within the prediction interval. While the former might be used as an alarming system that indicates whether a severe DEC event can be expected, the latter might be used for the concrete prediction of a certain point in future time.

In order to evaluate both tasks, we applied a repeated (stratified) k-fold cross-validation training and inference scheme with $r = 10$ repeats, $k = 10$ folds, and a training-set validation split of 0.25. Stratification, which equalizes class distributions over all folds, has only been applied for classification, since it is not applicable in score prediction.

We chose *categorical accuracy* and the *AUROC*-score (area under the true positive rate vs. false positive rate curve for varying operational points) as performance metrics for classification. Since AUROC is usually defined for binary classification tasks, the AUROC-score presented here is based on an averaged one-versus-all calculation for all three classes (cf. TABLE III). For the DEC score prediction task, we used the *mean squared error* between ground truth and estimated DEC score as the performance criterion.

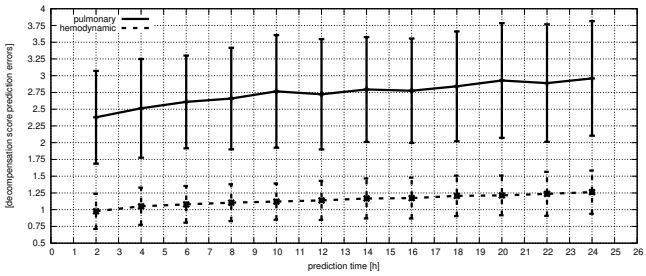
While it would have been possible to train and test the networks on a data set that contains a single random sample from each of the preprocessed cases (cf. Sec. III-B), we decided to oversample each case by completely non-overlapping samples of length of the observation period and prediction time interval. Even if two samples from a single case fall into the training and test dataset, this approach does not introduce data leakage, since we do not classify case identities, but the severity of future DEC status instead.

Finally, we investigated whether classification and prediction results depend on the kind of input data used, i.e., preprocessed vital sign time series vs. derived DEC score and confidence time series (cf. Sec. III-C). This question seemed reasonable, since the neural network has to learn the functional dependencies given in TABLE I and II when training on the former, but not when working on the latter.

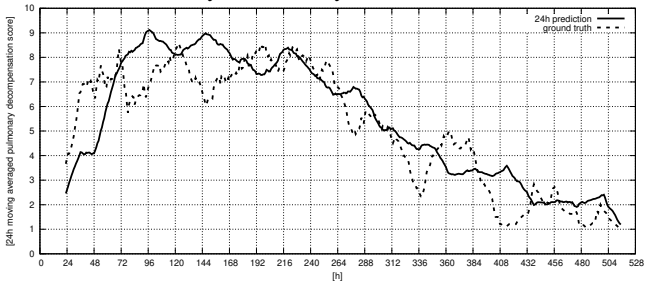
B. Results

From 10283 cases included in our dataset, filtering and preprocessing yielded 8140 usable cases. After removing cases that do not contain any information for at least one of the required data entries, we found 7552 cases to be usable for hemodynamic classification and prediction, and 3495 cases for pulmonary processing respectively. Due to the minimal required sample length of observation period and prediction time frame, we could finally draw $84h$ long samples from 3911 cases for hemodynamic assessment, and from 2455 cases for pulmonary assessment.

TABLE IV shows the results of the cross-validated classification of DEC classes. It can be seen that both in terms of categorical accuracy and AUROC-score the pulmonary classifier outperforms the hemodynamic one by a small margin. Although accuracy values of 72.65% and 83.68% appear moderate w.r.t. majority classes representing 58.0% and 75.4% of all samples, AUROC-values of 0.85 and 0.9 indicate a good classifier performance.



(a) Mean absolute errors and standard deviations for pulmonary and hemodynamic DEC score prediction. The underlying networks have been trained on preprocessed vital sign time series. We omit results for networks trained solely on DEC score and confidence time series, since they are virtually identical.



(b) Exemplary comparison of predicted pulmonary DEC scores (solid) and ground truth values (dashed) for case 2082. Each point on the solid curve is predicted 24h before its abscissa. Although, the predicted curve reasonably follows the ground truth, certain gaps can be observed, e.g. at 336h and 408h. However, we do not consider this as critical, since these errors are relatively small compared to the max. possible pulmonary DEC score of 39.

Fig. 5: The DEC score prediction results depicted above are based on a 60h observation period. For classifier training and inference 9199 samples from 2455 cases have been included for pulmonary prediction, and 13348 samples from 3911 cases for hemodynamic prediction.

Fig. 5a illustrates the networks' cross-validated DEC score prediction capabilities. With a forecast period ranging between $2h$ and $24h$ after the query-time, mean absolute errors are given by approx. 6.3% of the maximal pulmonary, and 9.6% of the maximal hemodynamic DEC score. For both cases we observe that prediction errors increase with the forecast period. In addition, Fig. 5b shows good comparability between $24h$ predictions and ground truth values for pulmonary DEC scores for an exemplary case.

Finally, classification and prediction networks only marginally benefit from training on DEC score and confidence time series, instead of parameter time series, if at all. An example for slightly better network performance is given by categorical accuracy and AUROC values for the severity classification of pulmonary DEC (cf. TABLE IV).

C. Discussion

In contrast to the related work presented, which requires time-consuming labelling by the medical staff, i.e., whether

TABLE IV: Decompensation Class Prediction: mean and standard deviation of categorical accuracy and AUROC.

DEC classifier	input time series	#cases	#samples	observation time	prediction time	no / moderate / severe DEC	categorical accuracy	AUROC
hemodynamic	vital signs	3911	13348	60h	24h	58.03% / 32.27% / 9.70%	$72.65\% \pm 0.94\%$	0.85 ± 0.01
	score and confidence						$71.37\% \pm 0.98\%$	0.84 ± 0.01
pulmonary	vital signs	2455	9002			22.65% / 75.44% / 1.91%	$83.05\% \pm 1.19\%$	0.89 ± 0.02
	score and confidence						$83.68\% \pm 1.22\%$	0.90 ± 0.02

or not cardiac arrest and respiratory failure occur at a given point in time, results given in Sec. IV-B are achieved by neural networks with training processes targeted to patients' fine-grained hemodynamic and pulmonary conditions. These DEC states are expressed as functions over data collected in the ICU. In doing so, we can not only classify finer subdivided DEC conditions, instead of a patient's ultimate heart or lung system failure, but can also predict the further development of the patient's condition by dedicated scores. The achieved classification performance is in terms of AUROC-scores on a similar level when compared to results found in literature. This is noteworthy in that our approach implements ternary instead of binary classification.

Looking at the rather moderate categorical accuracy results, we tried to improve accuracy by oversampling of the minority classes, as well as by class-dependent sample-weighting in the networks' loss functions. Both approaches are usually applied to mitigate problems arising with imbalanced datasets (cf. ground truth class distribution given in column 7 of TABLE IV). However, here they did not lead to noteworthy improvements, so we surmise that the issue is due to samples with DEC scores oscillating around the limit between two classes within the prediction window.

The problem described can also be alleviated as a result of the quantitative DEC score prediction. With mean $24h$-prediction errors of 6.3% (pulmonary), and 9.6% (hemodynamic) of the maximal achievable score, physicians could use this system to estimate the course of the DEC in the following 24 hours with sufficient accuracy.

V. CONCLUSIONS

Our work demonstrates that it is possible to predict hemodynamic and pulmonary DEC of patients being monitored in ICUs by using GRU-networks. Both in terms of severity classes and DEC scores, the system presented here shows encouraging results which lead us to believe that it can serve as a basis for an on-site warning system in the future.

Before such a real-world application, more immediate future work has to address further validation and explainability questions. In order to consider different etiologies of hemodynamic and pulmonary DEC, we will systematically explore combinations of so far unused data entries of patient records. Furthermore, we will collect new patient data during an upcoming second evaluation phase. Since new data will be available with higher sampling rate, we aim for validation of the current system, improved classification and prediction results as well as shorter observation periods. Beside further investigation of suitable network architectures, we also will focus on DEC events that show rapid deterioration, since these cases require special medical attention.

ACKNOWLEDGMENT

Special thanks go to R. App and J.-U. Czirr for the safe integration of the research server into the clinical IT-infrastructure, and to K. Jarosz for her administrative support.

REFERENCES

[1] A. Bagnall, J. Lines, A. Bostrom, et al. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31:606–660, 2017.

[2] K. Cho, B. v. Merrienboer, Ç. Gülçehre, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv (CoRR)*, abs/1406.1078, 2014.

[3] H. I. Fawaz, G. Forestier, J. Weber, et al. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33:917–963, 2019.

[4] A. Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc., Sebastopol, CA, USA, 2019.

[5] Philips Healthcare. IntelliSpace Critical Care and Anesthesia. https://www.philips.de/healthcare/product/HCNOCTN332/intellispace-critical-care-and-anesthesia. Accessed: 2022-04-06.

[6] S. Hochreiter and J. Schmidhuber. Long Short-term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[7] J. Kim, M. Chae, H.-J. Chang, et al. Predicting Cardiac Arrest and Respiratory Failure Using Feasible Artificial Intelligence with Simple Trajectories of Patient Data. *J. Clin. Med.*, 8(9):1–14, 2019.

[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv*, 2014.

[9] J.-M. Kwon, Yo. Lee, Ye. Lee, et al. An Algorithm Based on Deep Learning for Predicting In-Hospital Cardiac Arrest. *Journal of the American Heart Association*, 7(13):1–11, 2018.

[10] D.R. Miranda, A. de Rijk, and W. Schaufeli. Simplified Therapeutic Intervention Scoring System: the TISS-28 items–results from a multicenter study. *Critical Care Medicine*, 24(1):64–73, 1996.

[11] R. P. Moreno, P. G. H. Metnitz, E. Almeida, et al. SAPS 3–From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Medicine*, 31(10):1345–1355, 2005.

[12] Z. Prekopcsàk and D. Lemire. Time series classification by class-specific Mahalanobis distance measures. *Advances in Data Analysis and Classification*, 6:185–200, 2012.

[13] S. Seto, W. Zhang, and Y. Zhou. Multivariate time series classification using dynamic time warping template selection for human activity recognition. *CoRR*, 2015.

[14] G. Teasdale and B. Jennet. Assessment of coma and impaired consciousness. A practical scale. *The Lancet*, 304(7872):81–84, 1974.

[15] J.L. Vincent, R. Moreno, J. Takala, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Medicine*, 22(7):707–710, 1996.

[16] A.-K. Ian Wong, P. C. Cheung, R. Kamaleswaran, et al. Machine Learning Methods to Predict Acute Respiratory Failure and Acute Respiratory Distress Syndrome. *frontiers in Big Data*, 3:579774, 2020.

[17] L. Ye and E. Keogh. Time series shapelets: a new primitive for data mining. In *Proc. of the 15th ACM SIGKKD intl. conf. on Knowledge discovery and data mining*, pages 947–956. ACM DL, 2009.

[18] J.E. Zimmermann, A.A. Kramer, D.S. McNair, et al. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Critical Care Medicine*, 34(5):1297–1310, 2006.