Deep Learning with Diverse Objectives Improves ARDS Prediction

Carson Lam (✉ clam@dascena.com)

Dascena (United States)

Rahul Thapa

Dascena (United States)

Jenish Maharjan

Dascena (United States)

Keyvan Rahmani

Dascena (United States)

Chak Foon Tso

Dascena (United States)

Jacob Calvert

Dascena (United States)

Qingqing Mao

Dascena (United States)

Ritankar Das

Dascena (United States)

Research Article

Keywords: neural networks (NNs), Acute Respiratory Distress Syndrome (ARDS), gradient boosted trees

Posted Date: May 21st, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-534947/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Deep Learning with Diverse Objectives Improves ARDS Prediction

Carson Lam^{*,†}, Rahul Thapa[†], Jenish Maharjan[†], Keyvan Rahmani[†], Chak Foon Tso, Jacob Calvert, Qingqing Mao, Ritankar Das

[†]equal contribution

Dascena, Inc. ^{All Authors}

12333 Sowden Rd Ste B PMB 65148
Houston, Texas 77080-2059

***Materials & Correspondence:**

Email: clam@dascena.com
12333 Sowden Rd Ste B PMB 65148
Houston, Texas 77080-2059

Abstract

When training neural networks (NNs) on time-series inpatient data, as the number of outcomes predicted diversifies, the NN both generalizes better on external validation and reaches higher performance in similar numbers of training epochs. We demonstrated this in the context of predicting decompensation in Acute Respiratory Distress Syndrome (ARDS). The NN outperformed gradient boosted trees, achieving an area under the receiver operating characteristic of 0.86 on an external hold out test set of hospitals not included in the training set. We estimated real world benefit by comparing mortality rates between similarly at risk patients when diagnosed before or after the time of algorithm evaluation. We showed that, among similarly at risk patients, earlier diagnosis of ARDS nearly doubles the rate of in-hospital survival. Using cluster analysis of the algorithm's internal representations, we identified distinct ARDS sub-groups, some of which had similar mortality rates but different clinical presentations.

Background

Artificial Intelligence (AI) has shown great promise in medicine with potential applications across multiple domains (1). However, despite the advancements in AI, there has been slow progress in incorporating AI into the healthcare delivery system. This is true despite the promise of lower cost of care, expanded accessibility, earlier diagnosis, more accurate diagnosis and precision treatment. A wide variety of such algorithms have been studied and implemented by groups in both academia and industry (2–8). In real world clinical settings, the task is to anticipate multiple diseases or clinical states. Our work is related to past studies on multi-label

classification tasks, which predict multiple outcomes simultaneously. For example, Maxwell *et al.* use a Deep Neural Network to predict multiple chronic diseases such as hypertension and diabetes and Lipton *et al.* use Recurrent Neural Networks (RNN) to classify 128 different diagnoses (9–11). We perform a deeper exploration of how multi-label classification in the clinical setting can take advantage of the underlying dependencies between different diseases to allow for improved performance (12).

To this end, we focus on detection and early prediction of sepsis, COVID-19, hypoxemia, in-hospital mortality and varying severities of an acute condition often considered to have broad and subjective diagnostic criteria but is associated with significant mortality and morbidity, Acute Respiratory Distress Syndrome (ARDS) (13–20). Previous studies have suggested that physiological states that exist early in the presentation of ARDS can be used to predict ARDS before the confirmatory tests required by the gold standards such as the Berlin’s Criteria (21,22), which require radiology reports, which are often delayed due to logistical issues or delays in recognition of ARDS (23). Early prediction is desirable because it would lead to earlier intervention, more time for the careful administration of treatment, or modification of the ongoing treatment, which in turn should lead to improved outcomes such as reduced morbidity and mortality. For example, early warning of impending ARDS should allow implementation of lower tidal volumes in breathing support and more careful fluid management, two strategies to prevent or reduce severity of ARDS (24,25). Furthermore, whereas far more research has been done showing that early disease prediction is *possible*, we present estimates supporting that early prediction is *beneficial*. Lastly, we explore an interesting additional benefit of using NNs in hospitals, the identification of distinct disease phenotypes.

Methods

Data Description

The dataset includes 40,703 patient encounters whose care settings included the emergency department, inpatient, or intensive care units. The encounters were EHR data from 7 different hospitals between 4/20/2018 and 3/17/2021 **Figure 1**. Radiology data was not available. This prevented direct measurement of the Berlin Criteria for ARDS (22). The information collected from each hospital includes discharge disposition, demographic data such as age and sex, as well as time varying data including vital signs, lab values, oxygen delivery method, medications, and diagnosis times of any conditions present in the health record. This data was extracted as an unordered record of data type, data value, data units, and data collection time, also known as datetime. Preprocessing of this data first involved reordering the data in chronological order under each data type, with 3 equal length arrays representing the values, units and datetimes of each measurement.

Figure 1. Flow Chart of Patients. Among 7 hospitals, 40,703 patients met 3 criteria: 1. admission within the date range 4/20/2018 - 3/17/2021, 2. length of stay within the range of 2 hours - 3 months and 3. availability of basic vitals (blood pressure, heart rate, temperature, respiratory rate and peripheral oxygen saturation) and labs (complete blood count, CBC, and basic metabolic panel, BMP) in the EHR. These patients were separated into training, validation and test sets based on their hospital sites for external validation. The test set was limited to those patients with the required features listed in Table 1 consisting of age, sex and basic labs plus CBC with differential.

Input Features

Model inputs were a defined set of data types, or features, across all hospitals regardless of the data availability at a particular hospital. **Table 1** includes all the features used to train the Machine Learning Algorithms (MLAs) in this study. The required features are the subset of features including age, sex and basic labs plus CBC with differential used to determine the time the algorithm makes it's prediction. These data values were next organized into a matrix with features along the first dimension (rows) and discrete time in 20 minute intervals along the second dimension (columns). The first column, column index 0, contains the first time-point of any vital or lab measurement and was considered to be the start of care. The first row was normalized age. The second and third rows were binary indicators for male and female sex. The remaining set of rows were the time varying features and their corresponding mask to distinguish missing values from actual zeros **Table 2**.

Table 1. Input Features to the Machine Learning Algorithm. Features with an asterisk (*) are “required” features.

Demographics	
Age*	Sex*
Vital Signs	
Systolic blood pressure*	Respiratory rate*
Diastolic blood pressure*	Peripheral oxygen saturation (SpO ₂)*
Heart rate*	Temperature*
Partial Pressure of Oxygen (PaO ₂)	Fraction of Inspired Oxygen (FiO ₂)
Laboratory Results	
Glucose	Blood urea nitrogen (BUN)
Bilirubin	Creatinine*
White blood cell count (WBC)*	Platelet
Red blood cell count (RBC)	Neutrophils*
Lymphocytes*	Monocytes
Alanine Transaminase (ALT)	Hematocrit (HCT)
D-Dimer	Lactate
pH	Aspartate (AST)
Monocyte Volume Distribution Width (MDW)	Positive end-expiratory pressure (PEEP)
International Normalized Ratio (INR)	
Other Measurements	
Systemic Inflammatory response syndrome (SIRS)	

To normalize the features, we did a coarse approximation of the mean and standard deviation based on the normal range of these features in the lab reports. The center of the normal range was used as the approximate mean value and the half the range was used to approximate the standard deviation. If a feature was missing or not measured, it is set to 0. To let the model distinguish between null values and real values, a new set of features representing the availability mask was vertically appended to the matrix. Each feature row had a corresponding binary mask

vector which contains 0s and 1s, representing null values and non-null values respectively. During batch training, these matrices were zero-padded on the left side into equally sized tensors of [batch size, 58 features, 64 timesteps].

Table 2. RNN Features. The first 3 rows are the raw values, the next three rows are the corresponding normalized features, and the last 3 rows are the corresponding availability mask to distinguish missing values from actual zeros. The mask is a Boolean vector that is 0 if that measurement is missing and 1 if that measurement is present. Note: this is a subset of the total features used, for example purposes only.

	0 minutes	20 minutes	40 minutes	60 minutes
Raw Data				
SpO ₂	90%	100%	99%	80%
Creatinine	None	1.8 mg/dL	=>1.8 mg/dL	1.0 mg/dL
WBC	None	None	None	12 x10 ⁹ /L
Normalized				
SpO ₂	-0.5	+0.02	0	-2.3
Creatinine	0	+0.2	+0.2	-0.11
WBC	0	0	0	+0.15
Mask				
SpO ₂	1	1	1	1
Creatinine	0	1	1	1
WBC	0	0	0	1

Model Output and Targets

Additional target labels were chosen for the model that are distinct from ARDS, yet clinically related to it such that a collective representation in the NN is justified. These labels are shown in **Table 3**. The model was trained to predict the following target labels, also referred to as outcomes, using a binary cross entropy loss function:

Table 3. Clinical Outcomes used as Target Labels for the Machine Learning Algorithm. Thirteen output labels mapped to their respective definition. These output labels will be used throughout the paper hereafter.

<i>Output Label</i>	<i>Definition</i>
ARDS-1	ARDS defined as having an ICD code for ARDS as well as a drop in SpO ₂ below 91%.
ARDS-2	ARDS defined as having an ICD code for ARDS as well as a drop in SpO ₂ below 96%. A direct but broader criteria for ARDS,
ARDS-3	ARDS defined as having an ICD code for ARDS as well as a drop in SpO ₂ below 91% and no mention of a heart failure related ICD code among prior diagnoses.
ARDS-4	ARDS defined as having an ICD code for ARDS as well as a drop in SpO ₂ below 96% and no mention of a heart failure related ICD code among prior diagnoses
ARDS-5	ARDS defined as having an ICD code for ARDS. A direct and simple definition of ARDS.
Sepsis-6	Sepsis defined as having an ICD code for sepsis, or septic shock as well as a systemic inflammatory response syndrome (SIRS) score of greater than 2.
Sepsis-7	Sepsis defined as having an ICD code for sepsis, or septic shock. A direct and simple definition of Sepsis.
Hypoxemia-8	A drop in SpO ₂ below 91% any time during hospitalization.
Hypoxemia-9	A drop in SpO ₂ below 96% any time during hospitalization.
Hypoxemia-10	A drop in SpO ₂ below 91% after algorithm evaluates
Hypoxemia-11	A drop in SpO ₂ below 96% after algorithm evaluates
Death-12	In-hospital-mortality
Covid-13	Covid-19 positivity defined as in-hospital Covid-19 positive PCR or new ICD diagnosis within 7 days before or after admission

Timing of Algorithm Evaluation

For simplicity we evaluated the algorithm at a single point in time. This time is two timesteps after the first time at which all required features have been measured at least once. At this time, which we call the **Algotime**, the model predicts all of the target outcomes of **Table 3**. In the training and validation sets, we used the required features to determine **Algotime**, but in the case

of missing features, it defaults to 8 hours after admission. In the test set, we included only the patients that have all of the required features. As can be seen in **Figure 2**, on average, the Algotime occurred 31 hours after admission and ARDS was clinically diagnosed 139 hours after admission, hence the average number of hours between the Algotime and the clinical diagnosis of ARDS was 108 hours.. It should be noted that the performance statistics reported in this paper corresponds to the Algotime, not the time of clinical ARDS diagnosis or the end of the stay.

Figure 2. Timing of Algorithm and Diagnosis. A. Histogram of number of hours between start of care and the first time-point when all required features have been measured at least once. B. Histogram of number of hours from start of care until new diagnosis of ARDS is placed into the EHR.

Benefit Estimation

Thus far we have defined our machine learning objective as the early prediction of conditions. However, more important than the early prediction, is the impact of early prediction on patient health outcomes. To approximate this improvement in the outcome of mortality, we compared mortality rates between patients who received early and late diagnosis of ARDS. We defined early diagnosis and late diagnosis as early or late relative to when the algorithm was likely to provide its prediction. In other words, “early” and “late” were defined relative to Algotime, the minimum set of features became available in the EHR. If a clinical diagnosis of ARDS, using the ARDS-1 definition of ARDS ICD and SpO₂ below 91% , was made before Algotime, this was considered an early diagnosis of ARDS. If a diagnosis was made after Algotime, this was considered a late diagnosis of ARDS.

Machine Learning Models

We used RNN as the main ML model for our research. RNN is a class of artificial neural networks in which connections between nodes form a directed graph along a temporal sequence. RNNs can use their internal memory to process variable length sequences of inputs. The network is capable of learning a mapping function from the inputs over time to an output. It can even learn temporal dependence from the data. All these properties make RNN a well suited model for time series data like the one that is used in this study. The model schema of the RNN used in this research is presented in **Figure 3**. The RNN was implemented with the PyTorch package version 1.40 in Python 3.6 (26). For the RNN, the sequence module that was used was a 4-layer gated recurrent unit (GRU) (27) with 128 hidden units. Before the sequence of vectors was fed to the GRU, it passed through a normalization layer:

$$n(v) = a\left(\frac{v - \mu}{\sigma + \epsilon}\right) + b \quad (1)$$

Equation (1) is a normalization function that learns the parameters mean μ , standard deviation σ , scaling a and translation factor b used to normalize the sequence of vectors containing inputs age, sex boolean, vitals, labs, and systemic inflammatory response syndrome (SIRS) score. before entering the RNN. A soft attention module was used to assign scores to each timestep in the sequence. Scores are intended to be positively correlated with the importance of its respective timestep. A weighted sum of the sequence's hidden activations was called the context vector. We concatenated the context vector to the final GRU embedding and passed this to a 2-layer feed forward neural network for classification. The intermediate layer before the output logits was a 128 dimensional representation of each patient, referred to as the penultimate embedding. Similar to Bahdanau et al. 2014 (28), the score equation (2) of the attention neural network was parameterized by a feed forward neural network of the form:

$$\text{score}(h_i, h_i) = K^T \tanh(W_a \text{Prelu}(W_b[h_i, h_i])) \quad (2)$$

where $\tanh$ and Prelu denote the hyperbolic tangent function and parameterized rectified linear unit non-linearity functions, respectively; h_i denotes the last hidden activation in the GRU; h_i denotes each hidden activation in the sequence; i denotes the timestep; $[\cdot]$ denotes concatenation of separate vectors into one, and K , W_a , W_b denote learned parameters of the neural network. The whole GRU-RNN, attention module and classification module were end-to-end differentiable, which enabled optimization from input to output. The attention neural network was a mechanism of the RNN that allowed for higher quality learning. Instead of summarizing a time series of vectors, the attention neural network assigned each vector a score according to how important the vector was in allowing the model to make a prediction. In this way, the attention network mechanism allowed the RNN to focus on specific parts of the input, therefore enabling improved model performance.

Figure 3. RNN Model Schema. The Inputs have been simplified for diagrammatic purposes. Different time steps of normalized inputs are fed into GRU layers. The context vector from attention layers and encoder output from last GRU is concatenated before feeding into first fully connected (FC) layers.

Each point in the RNN model schema is representative of a neuron. At each layer, the RNN combined the information from the current and previous timesteps to update the activations of the deepest GRU hidden layer. The last of and deepest RNN activation is concatenated from the context vector provided by the attention network. The context vector is an importance weighted average of the deepest layer activations generated by the attention neural network. This concatenated vector is passed through 2 fully connected (FC) layers to generate an output (e.g.. prediction of ARDS onset). With this RNN schema, the model was trained to predict several target labels simultaneously and to evaluate a loss function based on all targets. We implemented a deep learning method where a single network was trained to output one logit per label utilizing a binary cross entropy loss function (29). Each output logit was independently passed through a sigmoid activation function to give the final multi-label output (30). Early stopping was used and was based on the ARDS-1 validation performance measured in area under the Receiver Operating Characteristic Curve (AUROC). To explore the relationship between the objective function's diversity of targets and final model performance, the lowest 2 AUROC targets were removed successively from each version of the RNN such that the RNN was trained using 13, 11, 9, 7 and 5 targets.

Tree based models frequently outperform deep learning in many clinical applications (31). To ensure this is not the case here, for comparison, XGBoost (XGB) models for each of the target labels were trained using XGBoost v0.81 in Python 3.6 package (27,28) and the same feature matrix as the RNN model. The XGB models were trained in a 1 versus all fashion for each target.

Clustering

In order to explore the representations used by the model and to reveal distinct phenotypes among ARDS patients, we collected the 64 dimensional activations produced by the first fully connected layer (FC1) as a compressed representation, or embedding, for each patient. To visually display these embeddings, we used Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) to reduce the embedding to two dimensions (34). We then used k-means clustering to group each ARDS patient into their unique cluster. Clusters of embeddings are not inherently interpretable; therefore, to confer clinical interpretability to these clusters, a new XGB model was trained using original features as inputs and cluster labels as output. The XGB model's SHAP (SHapley Additive exPlanations) plots (35) were used to interpret the clinical properties of each cluster.

Statistics

To compare different algorithms and training objectives, we computed the 95% confidence interval around the AUROC using the bootstrapping method (36,37). These intervals are with respect to the test set (n=951).

Ethics

All patient data was de-identified in compliance with the Health Insurance Portability and Accountability Act (HIPAA). This study was considered to be of minimal risk for human subjects as data collection was passive and did not pose a threat to the subjects involved. The project was approved with a waiver of informed consent (protocol study number 20-DASC-122) by an independent institutional review board, Pearl IRB (Indianapolis, Indiana; www.pearlirb.com).

Results

Comparison of RNN model with XGB

The XGB and RNN models were compared across all 13 outputs **Table 3**. The performance metric used for comparison is the AUROC. The single RNN model trained to classify 13 13 outputs outperformed the XGB models trained separately to classify each of the outputs in ARDS and oxygen related outcomes. For ARDS-5, RNN struggled to generalize well on Sepsis and COVID-19 related outcomes. For our use case, where ARDS is the primary target (ARDS-1 to 5), the single RNN model's 5 output's performance is similar or better compared to 5 separately trained XGB models.

Table 3. Comparison of performances of the XGB models and RNN13 model on each of the outcomes. The table shows the AUROC of each model for each of the labels and the sensitivity (Sens) and specificity (Spec). We note that the RNN13 model performs better than the XGB model on 7/13 outputs.

Labels	XGB			RNN13		
	AUROC (95% CI)	Sens (95% CI)	Spec (95% CI)	AUROC (95% CI)	Sens (95% CI)	Spec (95% CI)
ARDS-1	0.794 (0.737 - 0.849)	0.659 (0.519 - 0.799)	0.749 (0.721 - 0.777)	0.857 (0.814 - 0.899)	0.659 (0.519 - 0.799)	0.835 (0.81 - 0.859)
ARDS-2*	0.700 (0.631 - 0.768)	0.654 (0.525 - 0.783)	0.644 (0.613 - 0.675)	0.816 (0.770 - 0.858)	0.654 (0.525 - 0.783)	0.795 (0.769 - 0.822)
ARDS-3	0.797 (0.728 - 0.866)	0.667 (0.524 - 0.809)	0.785 (0.759 - 0.812)	0.855 (0.813 - 0.897)	0.667 (0.524 - 0.809)	0.822 (0.797 - 0.847)
ARDS-4	0.730 (0.661 - 0.801)	0.653 (0.52 - 0.786)	0.723 (0.694 - 0.752)	0.827 (0.786 - 0.87)	0.653 (0.52 - 0.786)	0.822 (0.797 - 0.846)
ARDS-5*	0.684 (0.618 - 0.749)	0.66 (0.533 - 0.788)	0.661 (0.63 - 0.692)	0.816 (0.770 - 0.856)	0.66 (0.533 - 0.788)	0.818 (0.793 - 0.844)
Sepsis-6	0.718 (0.622 - 0.805)	0.682 (0.487 - 0.876)	0.682 (0.652 - 0.712)	0.657 (0.562 - 0.742)	0.682 (0.487 - 0.876)	0.521 (0.489 - 0.553)
Sepsis-7	0.706 (0.584 - 0.808)	0.682 (0.487 - 0.876)	0.767 (0.74 - 0.794)	0.606 (0.504 - 0.704)	0.682 (0.487 - 0.876)	0.469 (0.437 - 0.501)
Hypoxemia-8	0.726 (0.690 - 0.764)	0.654 (0.596 - 0.712)	0.664 (0.629 - 0.699)	0.737 (0.706 - 0.767)	0.651 (0.592 - 0.709)	0.655 (0.62 - 0.69)
Hypoxemia-9	0.824 (0.796 - 0.849)	0.65 (0.616 - 0.684)	0.839 (0.788 - 0.891)	0.830 (0.807 - 0.851)	0.659 (0.625 - 0.692)	0.828 (0.775 - 0.881)
Hypoxemia-10	0.651 (0.581 - 0.722)	0.651 (0.581 - 0.722)	0.562 (0.528 - 0.597)	0.660 (0.627 - 0.695)	0.655 (0.585 - 0.726)	0.6 (0.565 - 0.634)
Hypoxemia-11	0.880 (0.860 - 0.9)	0.651 (0.602 - 0.7)	0.862 (0.834 - 0.89)	0.873 (0.855 - 0.891)	0.67 (0.622 - 0.719)	0.863 (0.835 - 0.89)
Death-12	0.747 (0.664 - 0.832)	0.68 (0.497 - 0.863)	0.666 (0.636 - 0.696)	0.696 (0.612 - 0.772)	0.68 (0.497 - 0.863)	0.618 (0.586 - 0.649)
Covid-13	0.831 (0.799 - 0.862)	0.654 (0.592 - 0.715)	0.851 (0.825 - 0.877)	0.708 (0.675 - 0.745)	0.658 (0.596 - 0.719)	0.676 (0.642 - 0.71)

*statistically significant p-value < 0.05

Benefit of Diverse Training Objectives

An intermediate number of output targets between 1 and 13 were also used to re-train the RNN. **Figure 4** shows the maximum AUROC with different subsets of the 13 outcomes used as training targets. For most targets there is a general trend towards overall improvement of

AUROC. For ARDS-2 and ARDS-5 the improvement is monotonic with the increasing number of targets. This demonstrates that there is some underlying dependency between some of the labels.

Figure 4. Model Performance Varies with the Number of Outcomes Predicted During Training. Validation AUC plotted against the number of targets in the RNN output (e.g. RNN9 refers to an RNN with 9 outputs). From right to left, the worst performing 2 targets in terms of AUROC are removed to train the next RNN with a smaller number of targets.

Training on Diverse Objectives Converges in a Comparable Number of Epochs

Learning quality and efficiency of single versus multiple outcome models were evaluated in terms of the rate of improvement of AUROC on the validation set per each stochastic gradient descent training epoch. We compared the rate of learning between RNNs trained with single targets and RNNs trained with multiple targets, to demonstrate that training on diverse learning objectives does not empirically require longer durations in training compared to single learning objectives. The rate of learning was measured as the AUROC of the validation set on for each epoch. In **Figure 5** the plots of the AUROC of three separate randomly initialized training

episodes for 15 epochs are shown for ARDS-1 and ARDS-2. For these two outcomes the time to reach the maximum validation AUROC in terms of the number of epochs is comparable between single and multiple target models.

Figure 5. Training on Multiple Targets Converge in Similar Time to Single Targets. Learning progress measured using the AUROC on the validation set. Each line is a different training run with new randomized initial weights and training batches. **(A)** AUROC predicting ARDS-1 shows vs number of training epochs. AUROC predicting ARDS-2 vs number of training epochs.

ARDS Clustering

Figure 6 shows the results of applying the K-means clustering algorithm to find clusters of patients as mentioned in the *machine learning models* section. The K-means algorithm was set to identify 3 clusters as this was the number of clusters that was observed from the visualization of the UMAP results on the output of the FC1 layer in the neural network. In **Figure 7**, we see that two clusters, A and B, have similarly high mortality rates but different clinical presentations. Cluster B was characterized by features that clearly signal suspicion for ARDS, such as low oxygen saturation, especially to levels less than 91%, neutrophilia, fever, and tachypnea. Cluster A however had a surreptitious presentation, oxygen saturation that is moderately low but higher than 91%, and leukocytosis present without fever. The model predicted for these patients that although desaturation below 91% has not happened yet, desaturation below 91% is likely to occur in the future.

Figure 6. RNN Representations Separate into Unique Clusters. Clustering ARDS population into 3 different groups as given by k-means clustering. The dimension of the embedding vector is reduced using the UMAP clustering algorithm. The red dots represent the patients with ARDS who died in hospital. The percentage included with the legend is the mortality percentage of each group.

A

B

C

Figure 7. Clusters Represent Distinct ARDS Phenotypes. SHAP plots for the top 15 features for 3 different XGB models trained using the input features to predict each of the 3 clusters of Fig. 7. Unlike the other features, names ending in *future* are not inputs to the models, but rather components of the target outcomes for these patients. For example, *SpO2 < 91 future* indicates that these patients will desaturate in the future after Algotime. **(A)** considered the cluster A as positive class and other two clusters as negative class; **(B)** considered cluster B as positive class and other two clusters as negative class; **(C)** considered cluster C as positive class and other two clusters as negative class.

Benefit Estimation

From our benefit estimation case study, we found that the mortality rate for patients who were diagnosed early with ARDS was 14/266 (5.26%), whereas for those diagnosed late with ARDS was 116/995 (11.66%). The Fisher exact test statistic value was 0.002 for the 6.3% mortality benefit of early diagnosis. For reference, the baseline mortality of patients without ARDS was 656/39442 (1.66%) and of patients with ARDS was 130/1261 (10.309%).

Conclusion

In this study, we described the development of a deep learning model for predicting multiple outcomes simultaneously using the same set of input features. We showed that the RNN13 model, trained to predict 13 outcomes simultaneously generalizes better on most outcomes than XGB models trained to predict individual outputs. We showed that this improvement is proportional to the number of targets predicted by the RNN. Additionally, the RNN13 generalized better in ARDS prediction and in some cases, reached its highest validation AUROC performance in fewer training epochs. This reinforces our conclusion that training the RNN model on a larger set of outcomes improves generalization. We hypothesize that diverse training objectives generalize better in part due to parameter sharing, which has a regularizing effect, and information sharing across outcomes, which learns richer representations (38).

We used RNN in this research because of its ability to use its internal memory to process variable length sequences of inputs, learn temporal dependence from the data, and share representations for an arbitrary number of outputs. We used a generic RNN with 4 GRU layers, an attention module, and 2 fully-connected layers for all numbers of outputs. We experimented with various RNN architectures, varying parameters such as number of layers and hidden units. From our light grid search, we found that the RNN model architecture used in this paper performs best for our use case. Additionally, the attention module seems to be an important part of the architecture in making the prediction because without it, the performance of the RNN dropped significantly.

To compare RNN with other algorithms, we used XGB because of its ability to handle missing or null values and its current dominance in industrial applications. Because EHR data often has a high level of missing values due to variability in data acquisition and recording habits in the live clinical environment, this quality is appealing. We trained multiple XGB models separately on the same input to classify different outcomes independently. We performed a grid search for hyperparameter optimization, tuning parameters such as tree depth and learning rate.

We also demonstrated an application of cluster analysis to probe deep learning models for clinical insights. Our analysis of the total ARDS population uncovered 3 distinct populations, two of which have similarly high mortality rates but different clinical presentations. One population of ARDS patients had apparent signs and symptoms of ARDS, while the other had much less obvious signs, yet similar future trajectories into respiratory decompensation and death. Recent studies have corroborated similar results in the covid-19 population in which two distinct phenotypes of ARDS were found with similar respiratory dynamics but 2-fold difference in odds of 28 day mortality (39). With the methods outlined in this study, phenotype discovery would be an additional benefit automatically applied to an arbitrarily large number of outcomes predicted.

To connect our machine learning findings with real world clinical effects, we compared the mortality rates between patients diagnosed earlier with ARDS and patients diagnosed later with ARDS relative to the algorithm's prediction time. Our estimation showed that the mortality rate in the population diagnosed early with ARDS was almost half the rate in the population who were diagnosed late with ARDS.

Limitations and future research

This study has several limitations. In many hospital systems, radiology images and radiology reports are kept in a separate software system from the EHR. Ideally we would prefer to confirm ARDS ICD codes by verifying the presence of bilateral lung infiltrates on chest imaging. Our inputs only included demographics, vital signs, and lab information. Future work should therefore incorporate EHR as well as imaging data. Our dataset spans the ED, inpatient and ICU settings and prescribes a single early time point for prediction. This could be one factor in the low AUROC for Sepsis predictions, which prior studies have shown to be reliably predictable in the ICU setting (3,8). This discrepancy warrants future investigation. Additionally, we did not have reliable data on race and ethnicity of the patient population. Future studies should explore the correlation between “early” or “late” diagnoses and the onset of standard ARDS criteria, this may provide evidence to suggest why ARDS diagnosed early might benefit from improved mortality, such as due to diagnosis earlier in the disease progress. Furthermore, this is a retrospective study, therefore, we are not able to determine the performance of our algorithm in a prospective clinical setting. Prospective validation is essential to determine how clinicians will respond to predictions of various outcomes. It is also important to determine whether our predictions can impact patient outcomes or resource allocation. Our work here is meant to serve as a reference for future research directions in establishing the most beneficial role for MLAs in the healthcare ecosystem and expanding the capabilities of machine learning in healthcare.

Competing Interests:

The authors are employees of Dascena, Inc (Houston, TX).

Acknowledgements:

Anna Siefkas, for discussion, comments and edits on this manuscript.

Zohora Iqbal, for comments and edits on this manuscript.

Gina Barnes, for comments and edits on this manuscript.

Abigail Green Saxena, for comments and edits on this manuscript.

Author contributions:

CL contributed data processing, designed the machine learning models and experiments

RT contributed to running machine learning experiments, designing clustering algorithm, and generating plots/diagrams

JM contributed to running machine learning experiments, architecture search, analysis of results and generating plots

KR contributed to exploring clustering algorithms, interpretation of clusters and visualizations

CFT contributed to data collection and analysis

JC contributed statistical analysis, writing and literature review

QM contributed to experimental design and data collection

RD contributed to experimental design and data collection

References:

1. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. *Nat Med*. 2019 Jan;25(1):24–9.
2. Radhachandran A, Garikipati A, Zelin NS, Pellegrini E, Ghandian S, Calvert J, et al. Prediction of short-term mortality in acute heart failure patients using minimal electronic health record data. *BioData Min*. 2021 Mar 31;14(1):23.
3. Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open*. 2018 Jan 1;8(1):e017833.
4. Mohamadlou H, Lynn-Palevsky A, Barton C, Chettipally U, Shieh L, Calvert J, et al. Prediction of Acute Kidney Injury With a Machine Learning Algorithm Using Electronic Health Record Data. *Can J Kidney Health Dis*. 2018;5:2054358118776326.
5. Ryan L, Mataraso S, Siefkas A, Pellegrini E, Barnes G, Green-Saxena A, et al. A Machine Learning Approach to Predict Deep Venous Thrombosis Among Hospitalized Patients. *Clin Appl Thromb Off J Int Acad Clin Appl Thromb*. 2021 Dec;27:1076029621991185.
6. Lam C, Siefkas A, Zelin NS, Barnes G, Dellinger RP, Vincent J-L, et al. Machine Learning as a Precision-Medicine Approach to Prescribing COVID-19 Pharmacotherapy with Remdesivir or Corticosteroids. *Clin Ther [Internet]*. 2021 Mar 29 [cited 2021 May 7]; Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8006198/>
7. Burdick H, Lam C, Mataraso S, Siefkas A, Braden G, Dellinger RP, et al. Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial. *Comput Biol Med*. 2020 Sep 1;124:103949.
8. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Med Inform*. 2016 Sep 30;4(3):e28.
9. Maxwell A, Li R, Yang B, Weng H, Ou A, Hong H, et al. Deep learning architectures for multi-label classification of intelligent health risk prediction. *BMC Bioinformatics*. 2017 Dec 28;18(Suppl 14):523.
10. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to Diagnose with LSTM Recurrent Neural Networks. *ArXiv151103677 Cs [Internet]*. 2017 Mar 21 [cited 2021 Apr 16]; Available from: <http://arxiv.org/abs/1511.03677>
11. Zhang X, Zhao H, Zhang S, Li R. A Novel Deep Neural Network Model for Multi-Label Chronic Disease Prediction. *Front Genet [Internet]*. 2019 [cited 2021 Apr 14];10. Available from: <https://www.frontiersin.org/articles/10.3389/fgene.2019.00351/full>
12. Read J, Perez-Cruz F. Deep Learning for Multi-label Classification. *ArXiv150205988 Cs [Internet]*. 2014 Dec 17 [cited 2021 Apr 9]; Available from: <http://arxiv.org/abs/1502.05988>
13. Bellani G, Laffey JG, Pham T, Fan E, Brochard L, Esteban A, et al. Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries. *JAMA*. 2016 Feb 23;315(8):788–800.

14. Matthay MA, Ware LB, Zimmerman GA. The acute respiratory distress syndrome. *J Clin Invest.* 2012 Aug;122(8):2731–40.
15. Fröhlich S, Murphy N, Doolan A, Ryan O, Boylan J. Acute respiratory distress syndrome: underrecognition by clinicians. *J Crit Care.* 2013 Oct;28(5):663–8.
16. Maley JH, Thompson BT. Embracing the Heterogeneity of ARDS. *Chest.* 2019 Mar;155(3):453–5.
17. Nanchal RS, Truitt JD. Recent advances in understanding and treating acute respiratory distress syndrome. *F1000Research.* 2018;7.
18. Rubenfeld GD, Caldwell E, Peabody E, Weaver J, Martin DP, Neff M, et al. Incidence and outcomes of acute lung injury. *N Engl J Med.* 2005 Oct 20;353(16):1685–93.
19. Villar J, Blanco J, Kacmarek RM. Current incidence and outcome of the acute respiratory distress syndrome. *Curr Opin Crit Care.* 2016 Feb;22(1):1–6.
20. Chiumello D, Marino A, Cammaroto A. The Acute Respiratory Distress Syndrome: Diagnosis and Management. *Pract Trends Anesth Intensive Care* 2018. 2018 Oct 1;189–204.
21. Coudroy R, Frat J-P, Boissier F, Contou D, Robert R, Thille AW. Early Identification of Acute Respiratory Distress Syndrome in the Absence of Positive Pressure Ventilation: Implications for Revision of the Berlin Criteria for Acute Respiratory Distress Syndrome. *Crit Care Med.* 2018 Apr;46(4):540–6.
22. The ARDS Definition Task Force*. Acute Respiratory Distress Syndrome: The Berlin Definition. *JAMA.* 2012 Jun 20;307(23):2526–33.
23. Rubenfeld GD, Caldwell E, Granton J, Hudson LD, Matthay MA. Interobserver variability in applying a radiographic definition for ARDS. *Chest.* 1999 Nov;116(5):1347–53.
24. Seitz KP, Caldwell ES, Hough CL. Fluid management in ARDS: an evaluation of current practice and the association between early diuretic use and hospital mortality. *J Intensive Care.* 2020 Oct 12;8(1):78.
25. Malhotra A. Low-Tidal-Volume Ventilation in the Acute Respiratory Distress Syndrome. *N Engl J Med.* 2007 Sep 13;357(11):1113–20.
26. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, et al. Automatic differentiation in PyTorch. 2017 Oct 28 [cited 2021 Jan 27]; Available from: <https://openreview.net/forum?id=BJJsrnfCZ>
27. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *ArXiv14123555 Cs [Internet].* 2014 Dec 11 [cited 2021 Jan 27]; Available from: <http://arxiv.org/abs/1412.3555>
28. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. *ArXiv14090473 Cs Stat [Internet].* 2016 May 19 [cited 2021 Jan 27]; Available from: <http://arxiv.org/abs/1409.0473>
29. Good IJ. Some terminology and notation in information theory. *Proc IEE - Part C Monogr.* 1956 Mar 1;103(3):200–4.
30. Ben-Baruch E, Ridnik T, Zamir N, Noy A, Friedman I, Protter M, et al. Asymmetric Loss For Multi-Label Classification. *ArXiv200914119 Cs [Internet].* 2021 Mar 18 [cited 2021 Apr 16]; Available from: <http://arxiv.org/abs/2009.14119>
31. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell.* 2020 Jan;2(1):56–67.
32. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet].* San Francisco California USA: ACM; 2016 [cited 2021 Mar 11]. p. 785–94. Available from: <https://dl.acm.org/doi/10.1145/2939672.2939785>
33. xgboost: XGBoost Python Package [Internet]. [cited 2021 Jan 19]. Available from: <https://github.com/dmlc/xgboost>

34. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv180203426 Cs Stat [Internet]. 2020 Sep 17 [cited 2021 Apr 13]; Available from: <http://arxiv.org/abs/1802.03426>
35. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. ArXiv170507874 Cs Stat [Internet]. 2017 Nov 24 [cited 2021 May 3]; Available from: <http://arxiv.org/abs/1705.07874>
36. Bertail P, Clmencon SJ, Vayatis N. On Bootstrapping the ROC Curve. :8.
37. Liu H, Li G, Cumberland WG, Wu T. Testing Statistical Significance of the Area under a Receiving Operating Characteristics Curve for Repeated Measures Design with Bootstrapping. :22.
38. Nam J, Loza Menca E, Kim HJ, Frnkranz J. Maximizing Subset Accuracy with Recurrent Neural Networks in Multi-label Classification. Adv Neural Inf Process Syst [Internet]. 2017 [cited 2021 Apr 22];30. Available from: <https://proceedings.neurips.cc/paper/2017/hash/2eb5657d37f474e4c4cf01e4882b8962-Abs tract.html>
39. Ranjeva S, Pinciroli R, Hodell E, Mueller A, Hardin CC, Thompson BT, et al. Identifying clinical and biochemical phenotypes in acute respiratory distress syndrome secondary to coronavirus disease-2019. EClinicalMedicine [Internet]. 2021 Apr 15 [cited 2021 Apr 16];0(0). Available from: [https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370\(21\)00109-7/abstract](https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(21)00109-7/abstract)

Figures

Figure 1

Flow Chart of Patients. Among 7 hospitals, 40,703 patients met 3 criteria: 1. admission within the date range 4/20/2018 - 3/17/2021, 2. length of stay within the range of 2 hours - 3 months and 3. availability of basic vitals (blood pressure, heart rate, temperature, respiratory rate and peripheral oxygen saturation) and labs (complete blood count, CBC, and basic metabolic panel, BMP) in the EHR. These patients were separated into training, validation and test sets based on their hospital sites for external validation. The test set was limited to those patients with the required features listed in Table 1 consisting of age, sex and basic labs plus CBC with differential.

Figure 2

Timing of Algorithm and Diagnosis. A. Histogram of number of hours between start of care and the first time-point when all required features have been measured at least once. B. Histogram of number of hours from start of care until new diagnosis of ARDS is placed into the EHR.

Figure 3

RNN Model Schema. The Inputs have been simplified for diagrammatic purposes. Different time steps of normalized inputs are fed into GRU layers. The context vector from attention layers and encoder output from last GRU is concatenated before feeding into first fully connected (FC) layers.

Figure 4

Model Performance Varies with the Number of Outcomes Predicted During Training. Validation AUC plotted against the number of targets in the RNN output (e.g. RNN9 refers to an RNN with 9 outputs). From right to left, the worst performing 2 targets in terms of AUROC are removed to train the next RNN with a smaller number of targets.

Figure 5

Training on Multiple Targets Converge in Similar Time to Single Targets. Learning progress measured using the AUROC on the validation set. Each line is a different training run with new randomized initial weights and training batches. (A) AUROC predicting ARDS-1 shows vs number of training epochs. AUROC predicting ARDS-2 vs number of training epochs.

Figure 6

RNN Representations Separate into Unique Clusters. Clustering ARDS population into 3 different groups as given by k-means clustering. The dimension of the embedding vector is reduced using the UMAP clustering algorithm. The red dots represent the patients with ARDS who died in hospital. The percentage included with the legend is the mortality percentage of each group.

Figure 7

Clusters Represent Distinct ARDS Phenotypes. SHAP plots for the top 15 features for 3 different XGB models trained using the input features to predict each of the 3 clusters of Fig. 7. Unlike the other features, names ending in future are not inputs to the models, but rather components of the target outcomes for these patients. For example, SpO2 < 91 future indicates that these patients will desaturate in the future after Algotime. (A) considered the cluster A as positive class and other two clusters as

negative class; (B) considered cluster B as positive class and other two clusters as negative class; (C) considered cluster C as positive class and other two clusters as negative class.