Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240529

A Comparative Analysis of Federated and Centralized Learning for SpO_2 Prediction in Five Critical Care Databases

Johanna SCHWINN[a,1], Seyedmostafa SHEIKHALISHAHI[a], Matthaeus MORHART[a], Mathias KASPAR[a] and Ludwig Christian HINSKE[a]
[a] *Digital Medicine, University Hospital of Augsburg, Augsburg, Germany*
ORCiD ID: Johanna Schwinn https://orcid.org/0009-0000-3107-1619

Abstract. This study explores the potential of federated learning (FL) to develop a predictive model of hypoxemia in intensive care unit (ICU) patients. Centralized learning (CL) and local learning (LL) approaches have been limited by the localized nature of data, which restricts CL approaches to the available data due to data privacy regulations. A CL approach that combines data from different institutions, could offer superior performance compared to a single-institution approach. However, the use of this method raises ethical and regulatory concerns. In this context, FL presents a promising middle ground, enabling collaborative model training on geographically dispersed ICU data without compromising patient confidentiality. This study is the first to use all five public ICU databases combined. The findings demonstrate that FL achieved comparable or even slightly improved performance compared to local or centralized learning approaches.

Keywords. Federated Learning, Regression, LSTM, Flower, MIMIC-IV, eICU-CRD, AmsterdamUMCdb, HiRID, SICdb

1. Introduction

In critical care settings, peripheral oxygen saturation (SpO2) is routinely monitored as a critical tool to assess a patient's oxygenation status. Low blood oxygen saturation, known as hypoxemia, is associated with higher mortality rates [1]. The prediction of hypoxemia has received recent attention in the field [2]. Machine learning (ML) approaches have been explored to predict hypoxemia or SpO2 in hospitalized patients. Studies have employed artificial neural networks [3] and deep learning models with time series analysis [4] to predict future SpO2 values or classify hypoxemia events.

Federated learning (FL) offers a stronger privacy guarantee compared to centralized learning (CL) and the opportunity to create more generalizable models compared to local learning (LL). Unlike CL, which gathers data centrally, and LL, which trains models solely within institutions, FL enables collaborative learning while keeping data distributed [5]. While prior work has compared FL and CL for classification tasks [6] and explored FL for different domains, e.g. vital sign classification using MIMIC-IV [7], no research has investigated FL for SpO2 prediction across multiple datasets. This study addresses this gap by evaluating a FL-based approach for SpO2 prediction in the ICU.

[1] Corresponding Author: Johanna Schwinn; E-mail: Johanna.schwinn@uk-augsburg.de.

The objective of this study is to investigate the effectiveness of various ML approaches – CL, FL, and LL – in predicting SpO2 solely from retrospective SpO2 data for critically ill patients in five intensive care databases comprising real-world data of more than 200 hospitals.

2. Methods

In this retrospective study, we utilized five publicly available ICU databases, namely eICU-CRD [8], HiRID [9], MIMIC-IV [10], SICdb [11], and UMCdb [12] to evaluate SpO2 prediction for different ML approaches. All datasets included are IRB-exempt. In the LL setting, four LL models were trained and validated individually using data from four databases, i.e., eICU-CRD, HiRID, MIMIC-IV, and UMCdb. Similarly, a CL model was trained using the combined data from these four datasets. Finally, the initial four datasets were integrated into a federated learning framework for FL model development. The model's performance and generalizability were evaluated through two validation steps. Internal validation was conducted on four separate datasets and the centralized dataset, while external validation utilized the SICdb dataset.

2.1. Data Preprocessing

Accounting for inherent limitations in pulse oximeter design, SpO_2 values were restricted to the clinically relevant range (70%-100%) [13]. Values below 70% were excluded due to unreliable accuracy, while those exceeding 100% were removed as outliers. SpO_2 values were scaled using the method described by [4]. Data was systematically sampled at five-minute intervals, with only the first value within each interval included. A minimum of three consecutive intervals with at least one valid SpO_2 measurement were required for prediction. Only patients with at least 100 minutes of continuous monitoring (20 consecutive observations) were included.

2.2. Machine Learning Model

The model predicts the SpO_2 value for the next five-minute interval. The prediction is based on the two most recent observations from each of the two preceding five-minute intervals in the patient's stay. A recurrent neural network (RNN) with Long Short-Term Memory (LSTM) [14] units was adopted for SpO_2 prediction [4]. The architecture consisted of two LSTM layers (16 and 2 units) followed by a dense layer. Batch normalization and dropout (0.1) addressed training stability and overfitting, respectively. The Adam optimizer (learning rate = 0.001) with mean squared error (MSE) loss function was used. A learning rate scheduler (20% reduction per epoch) and early stopping (patience: 4 epochs) were implemented for optimization. The model architecture remained consistent across all experimental settings.

2.3. Federated Learning

In the context of FL, we utilized the Flower framework [15] within a Docker containerized environment. Each participating site (n=4) contributed data from a critical care database (eICU-CRD, MIMIC-IV, HiRID, UMCdb) encapsulated in a separate container. Furthermore, the central server was deployed in a separate container. This setup facilitated strict data privacy throughout the FL process. Experiments are done with Flower 1.5, Python 3.10, and TensorFlow 2.14. The FL training involved five global rounds with FedProx aggregation [16]. Each round included five-fold cross-validation with up to 20 local epochs and early stopping (patience=4). The process yielded four client-specific models and one globally aggregated model. We utilized a five-fold cross-validation for model training with 10% of the training data used for validation. Internal performance is reported as the mean and standard deviation (SD) across folds. For external validation, a bootstrapped (n=50) sample from the SICdb is used. Results for external validation mirror those for internal validation, using micro-averages across all predictions.

3. Results

A total of 98,238 ICU encounters were selected for this study from four databases originating in the United States, the Netherlands, and Switzerland. Specifically, 66,941 encounters were selected from eICU-CRD, 64 from MIMIC-IV, 25,711 from HiRID, and 5,522 from UMCdb. For external testing, 7,481 patient ICU encounters from SICdb were included.

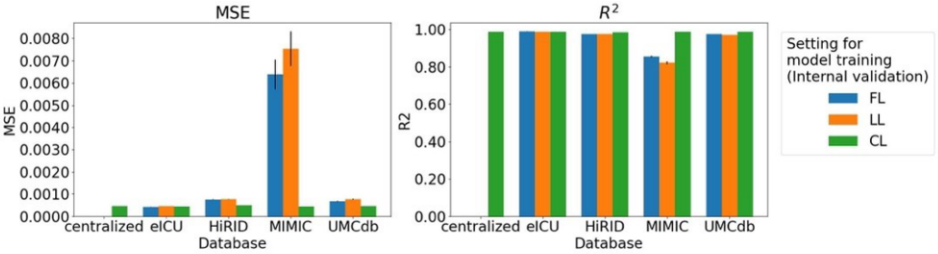

Figure 1. This figure shows the performance metrics for the models trained in three different settings: CL, FL, and LL. The color coding of each bar corresponds to the respective training setting. The x-axis shows the training datasets. Validation was conducted on the same dataset used for training. The CL model was validated on each dataset independently.

Figure 1 shows the results for three training settings: CL, LL, and FL. Across all four databases, the MSE was consistently the lowest in CL (≤0.0005) and the highest in LL, with values ranging from 0.00046 (eICU-CRD) to 0.0075 (MIMIC-IV). In FL, the MSE values are slightly better than the LL results and range from 0.00042 (eICU-CRD) to 0.0064 (MIMIC-IV). The results in MIMIC-IV are worse than those in the other databases for LL and FL, but similar for CL. Similarly, the R^2 results were best in the CL setting. Overall, values were greater than 0.97 for all experiments, except for FL and LL in MIMIC-IV. The latter only had an R^2 of 0.822±0.0089 and 0.856±0.0047 for LL and FL, respectively.

Figure 2 illustrates the results of the external validation on SICdb. The MSE was 0.0007 for all models of the LL and the FL setting except for MIMIC-IV (0.0014 (LL) and 0.0011 (FL)). The R^2 values were greater than 0.963±0.0011 (MIMIC-IV) across all settings, with FL being marginally better.

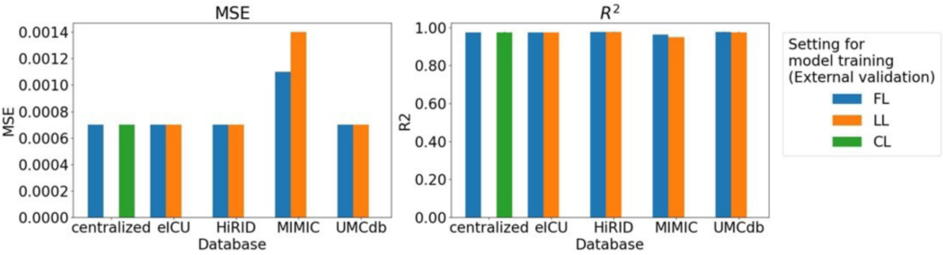

Figure 2. This figure shows the performance metrics for the models trained in three different settings: CL, FL, and LL. The color coding of each bar corresponds to the respective training setting. The x-axis shows the training datasets. Testing was performed on SICdb for all models.

4. Discussion

Employing three distinct ML approaches – CL, FL, and LL –, this study explores the application of ML models based on the SWIFT approach [4] that predicts the next SpO2 value using only prior SpO2 values. FL allows geographically dispersed ICUs to collaboratively train a model without sharing raw patient data and thereby protecting patient data privacy. An analysis of such an algorithm with different approaches has not been done before for this domain. The results of this study show that FL slightly outperformed the LL. Across most datasets, FL achieved performance comparable to CL, the ideal scenario with unconstrained data access. However, in the LL setting, the performance did not exhibit substantial deviations. Our results are consistent with [17], who observed only minimal variation across settings for mortality prediction in a single database. The smallest dataset, MIMIC-IV, did benefit the most from the CL and FL settings. While MIMIC-IV achieved substantially worse results in the LL, the model performance improved when trained in the FL setting. Additionally, the results indicate an improved MSE in the CL (≤0.0005) compared to the MSE in [4] (≤0.0007).

5. Conclusions

The results suggest that especially hospitals managing relatively limited datasets may benefit substantially from FL while preserving patient data privacy. While this study analyzed a very simple ML task with only a single variable, future work could explore the application of more complex prediction tasks and models, potentially unlocking the full potential of FL.

References

[1] Grimaldi D, Hraiech S, Boutin E, Lacherade JC, Boissier F, Pham T, et al. Hypoxemia in the ICU: prevalence, treatment, and outcome. Ann Intensive Care 2018;8:82. doi:10.1186/s13613-018-0424-4.

[2] Pigat L, Geisler BP, Sheikhalishahi S, Sander J, Kaspar M, Schmutz M, et al. Predicting Hypoxia Using Machine Learning: Systematic Review. JMIR Med Inform 2024;12:e50642. doi:10.2196/50642.

[3] Ghazal S, Sauthier M, Brossier D, Bouachir W, Jouvet PA, Noumeir R. Using machine learning models to predict oxygen saturation following ventilator support adjustment in critically ill children: A single center pilot study. PLOS ONE 2019;14:e0198921. doi:10.1371/journal.pone.0198921.

[4] Annapragada AV, Greenstein JL, Bose SN, Winters BD, Sarma SV, Winslow RL. SWIFT: A deep learning approach to prediction of hypoxemic events in critically-ill patients using SpO2 waveform prediction. PLoS Comput Biol 2021;17:e1009712. doi:10.1371/journal.pcbi.1009712.

[5] McMahan B, Moore E, Ramage D, Hampson S, Arcas BA y. Communication-Efficient Learning of Deep Networks from Decentralized Data. Proc. 20th Int. Conf. Artif. Intell. Stat., PMLR; 2017, p. 1273–82.

[6] Sadilek A, Liu L, Nguyen D, Kamruzzaman M, Serghiou S, Rader B, et al. Privacy-first health research with federated learning. Npj Digit Med 2021;4:132. doi:10.1038/s41746-021-00489-2.

[7] Rakhmiddin R, Lee K. Federated Learning for Clinical Event Classification Using Vital Signs Data. Multimodal Technol Interact 2023;7:67. doi:10.3390/mti7070067.

[8] Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018;5:180178. doi:10.1038/sdata.2018.178.

[9] Faltys M, Zimmermann M, Lyu X, Hüser M, Hyland S, Rätsch G, et al. HiRID, a high time-resolution ICU dataset 2021. doi:10.13026/323R-NK04.

[10] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023;10:1. doi:10.1038/s41597-022-01899-x.

[11] Rodemund N, Wernly B, Jung C, Cozowicz C, Koköfer A. The Salzburg Intensive Care database (SICdb): an openly available critical care dataset. Intensive Care Med 2023;49:700–2. doi:10.1007/s00134-023-07046-3.

[12] Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, et al. Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example. Crit Care Med 2021;49:e563. doi:10.1097/CCM.0000000000004916.

[13] Chan ED, Chan MM, Chan MM. Pulse oximetry: Understanding its basic principles facilitates appreciation of its limitations. Respir Med 2013;107:789–99. doi:10.1016/j.rmed.2013.02.004.

[14] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput 1997;9:1735–80. doi:10.1162/neco.1997.9.8.1735.

[15] Beutel DJ, Topal T, Mathur A, Qiu X, Fernandez-Marques J, Gao Y, et al. Flower: A Friendly Federated Learning Research Framework 2022.

[16] Li Q, Diao Y, Chen Q, He B. Federated Learning on Non-IID Data Silos: An Experimental Study. 2022 IEEE 38th Int. Conf. Data Eng. ICDE, 2022, p. 965–78. doi:10.1109/ICDE53745.2022.00077.

[17] Mondrejevski L, Miliou I, Montanino A, Pitts D, Hollmén J, Papapetrou P. FLICU: A Federated Learning Workflow for Intensive Care Unit Mortality Prediction. 2022 IEEE 35th Int. Symp. Comput.-Based Med. Syst. CBMS, 2022, p. 32–7. doi:10.1109/CBMS55023.2022.00013.