Federated Reinforcement Learning for Therapeutic Interventions over ICUs with Noisy Labels

Linxiao Cao †, Yifei Zhu‡, Haoquan Zhou†, Shilei Tan †, Wei Gong†*
† Centre for Leading Medicine and Advanced Technologies of IHM,
The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine,
University of Science and Technology of China, Hefei, China
‡ University of Michigan–Shanghai Jiao Tong University Joint Institute,
Shanghai Jiao Tong University, Shanghai, China
linxiaocao@mail.ustc.edu.cn, yifei.zhu@sjtu.edu.cn, zhouhq2005@qq.com,
tanshilei@ustc.edu.cn, weigong@ustc.edu.cn

***Abstract*—The proliferation of healthcare IoT devices and the resulting rich healthcare data sprout new possibilities for intelligent healthcare applications. Patients in intensive care units (ICUs) rely on various networked gadgets to continuously monitor their health and manage critical situations. Among the common therapeutic interventions in ICUs, invasive mechanical ventilation and injecting sedatives during ventilation play crucial roles in maintaining respiratory function and enhancing patient care. While existing therapeutic interventions largely depend on experience and intuition, we propose a federated inverse reinforcement learning framework, termed *FERRY*, which automatically and intelligently learns optimal therapeutic intervention policies across networked ICUs while keeping raw data local. Specifically, our federated approach overcomes limitations in medical data privacy and facilitates collaboration; our proposed inverse reinforcement learning framework learns the variational posterior distribution from historical trajectories to handle the unknown reward. Additionally, we enhance our framework with distributionally robust optimization to ensure worst-case performance and adaptively filter out noisy data through joint loss learning. Extensive experiments on the real-world dataset demonstrate that *FERRY* improves the overall ventilation and sedation decision-making accuracy by 36.75% compared to other state-of-the-art baselines.**

***Index Terms*—Federated learning, Inverse reinforcement learning, Distributionally robust optimization, Noisy label**

I. Introduction

The rise of IoT devices has led to an unprecedented increase in healthcare data in recent years. A global survey by Dell Technologies indicates that data in healthcare industries has surged by 878% over the past two years [1]. Intensive care units (ICUs) are crucial departments within hospitals, equipped with advanced devices that continuously monitor patients' vital signs and generate physiological data. This data-rich medical landscape creates an ideal environment for the development of automated and intelligent decision support tools, leveraging artificial intelligence and data analytics to uncover complex relationships within large datasets and enhance clinical decisions and treatment protocols, as illustrated in Fig. 1.

In ICUs, therapeutic interventions — such as invasive mechanical ventilation, thoracentesis, and sedative administration — are critical for helping patients survive acute life-threatening conditions. However, the timing and dosage of these interventions can greatly influence patient outcomes [2]. For instance, premature extubation during mechanical ventilation may hinder recovery, while prolonged ventilation can elevate the risk of infection [3]. Additionally, sedatives are frequently administered to ventilated patients to alleviate pain and maintain physiological stability. However, in traditional ICUs, physicians rely on their experience to interpret complex and diverse data, leading to variability in the effectiveness of these interventions [4]. This reliance on individual intuition and experience not only risks suboptimal treatment outcomes but also diminishes overall care quality [5]. To tackle these challenges, there is a growing interest in leveraging machine learning techniques to enhance clinical decision-making and improve patient outcomes based on data collected from medical devices.

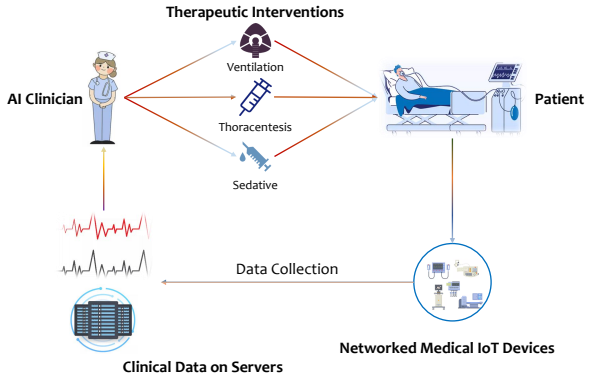

Fig. 1. The illustration of intelligent therapeutic interventions

Reinforcement learning (RL) has recently emerged as a powerful tool for assisting physicians in clinical decision-making due to its high adaptability to complex dynamic environments and strong capability to process uncertain information [6]. By retrieving essential information from clinical data, RL could provide accurate and personalized treatment policies that improve patient outcomes [7]. Notably, several studies

have highlighted the effectiveness of RL in this domain. For instance, Komorowski et al. [8] employ a model-based policy iteration algorithm to develop treatment decisions for patients with sepsis, achieving better outcomes than those made by human clinicians. Similarly, Raghu et al. [9] utilize the Double Deep Q Network to learn clinically interpretable treatment policies for sepsis within a continuous state space. Prasad et al. [10] focus on developing mechanical ventilation weaning protocols using the Fitted Q Iteration algorithm, demonstrating a promising reduction in reintubation rates. Additionally, Yu et al. [11] apply bayesian inverse reinforcement learning (IRL) to optimize mechanical ventilation policies, resulting in personalized reward functions that not only reduce ventilation-associated lung injuries but also improve overall clinical outcomes.

Despite offering some benefits, these approaches suffer from several challenges. First, *privacy*. Many methods rely on data from a single hospital due to privacy issues [12], [13], which can lead to inadequate datasets and, consequently, reduced model performance. How to fully exploit the scattered data across different hospitals to train a model without exposing local hospital data is non-trivial. Second, *unknown reward*. The reward functions in these schemes are designed solely based on the physicians' clinical experience, as quantifying complex physiological changes into a logical reward system proves difficult. This resilience can lead to subjective and inconsistent rewards, introducing bias into the model [7]. How to develop policies that are close to physicians from existing expert trajectories is challenging. Last but not least, *noisy data*. These schemes often overlook the impact of data noise, which can stem from incorrect inputs, inconsistent naming conventions, and coding errors within clinical datasets [14]. How to mitigate the impact of data noise on model training and generate a functional model is not easy.

To overcome the aforementioned issues, we propose a novel framework called *FERRY*, which seamlessly integrates federated learning (FL) and IRL. Specifically, *FERRY* leverages FL to overcome medical data privacy limitations and enable collaboration across hospitals. The integration of IRL addresses the challenge of formulating reward functions by learning from expert demonstrations. Furthermore, *FERRY* incorporates distributionally robust optimization (DRO) to mitigate the negative effects of data heterogeneity and ensure strong model performance even in worst-case scenarios. However, naively combining existing IRL approaches with FL struggles to handle the noisy data, leading to significant performance degradation. In response, *FERRY* applies a joint loss learning to reduce the impact of noisy data, thereby enhancing both accuracy and robustness. The detailed methodology is presented in Section II.

The main contributions of the paper are summarized as follows:

- We propose a novel framework to jointly combine IRL and FL in the ICU domain to solve medical decision-making problems and apply DRO to ensure that the model performs well, even under worst-case distribution shifts within the data.
- We design a pseudo-Siamese paradigm that utilizes both local and global models to perform joint loss clipping, specifically targeting clients with noisy data. This approach effectively mitigates the negative effects of data noise during model training.
- Extensive experiments on the real-world dataset show that *FERRY* achieves central-competitive performance without aggregating the sensitive raw data.

II. Methodology

The detailed architecture of *FERRY* is shown in Fig. 2. The process starts with the cloud server initializing a global model, which is distributed to all participating hospitals. Each hospital then updates this global model using its local clinical data (i.e., demonstrations) to obtain a local model. For hospitals dealing with noisy data, the trained local and global models are processed through a joint loss training approach, producing an updated local model for the current round. Afterward, each hospital uploads its updated local model to the cloud server. The server aggregates these models into a new global model, which is redistributed to all hospitals for the next training round. This iterative process continues until the global model converges. Throughout the training, only model parameters are exchanged, ensuring that raw clinical data remains local and patient privacy is preserved. Additionally, employing the IRL framework handles the design of the reward function, leading to improved policy model accuracy. Moreover, joint loss learning helps mitigate the effects of noisy data, further enhancing the model's efficiency. The specifics regarding local update are elaborated in subsection II-A, the joint loss learning approach is detailed in subsection II-C, and the aggregation process is introduced in subsection II-D.

A. Local IRL Update

Typically, applying RL to solve clinical decision problems requires modeling the entire treatment process as a Markov decision process (MDP), consisting of the tuple (S, A, P, R), where S is the state space, A is the action space, P is the transition function, and R is the reward function [15]. However, in our work, we cannot directly inform about the potential reward or transitions due to the lack of knowledge about R and P. Fortunately, we have access to treatment trajectories from expert clinicians, which allows us to infer R based on their observed behavior. With that in mind, we define the two remaining elements as follows:

State: $s_t \in S$ denotes the physiological state of the patient at time t. In designing the state space, we incorporate observed physiological data to assess the patient's condition. The state at time t is a high-dimensional feature vector that includes essential patient information such as age, weight, and other key physiological indicators.

Action: $a_t \in A$ represents the action taken by the agent we trained based on the patient's state at time t. The action space is defined separately for ventilator status and sedative dosage.

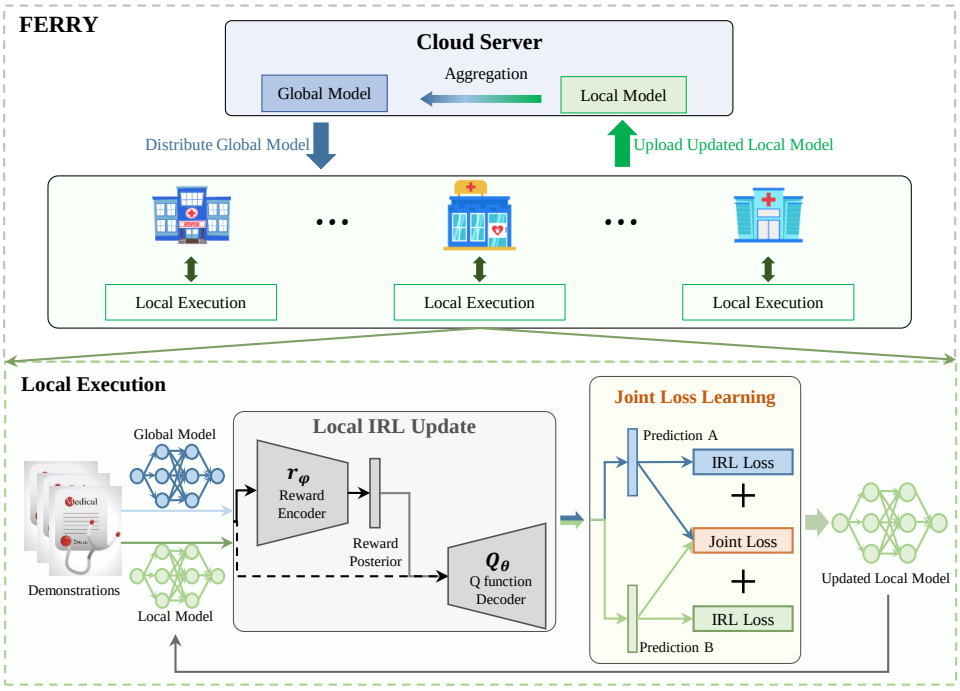

Fig. 2. The illustration of the proposed framework of *FERRY*.

Algorithm 1 The procedures of local IRL Update

Input: Clinical demonstrations D, state space S, action space A, learning rate η, mini-batch size ξ

Output: Parameter of reward distribution ϕ, parameter of policy function θ

Initialise ϕ, θ
while *not converged* **do**
  Sample D_{mini} from D;
  $f(\phi, \theta, D) = \mathrm{E}[\frac{n}{\xi} f(\phi, \theta, D_{mini})]$;
  $(\phi^{'}, \theta^{'}) \longleftarrow (\phi, \theta) + \eta \nabla_{\phi,\theta} f(\phi, \theta, D)$;
  $\phi, \theta \longleftarrow \phi^{'}, \theta^{'}$
end while
Return ϕ, θ

Specifically, we approximate six commonly used sedatives by mapping them onto a dose scale, which is then discretized into four distinct levels. The action $a_t \in A$ reflects the treatment at time step t, where $a_t[0] \in [0, 1]$ indicates whether the ventilator is on or off, and $a_t[1] \in [0, 1, 2, 3]$ specifies the sedative dose level. The full action space is as follows:

$$A = \left\{ \begin{bmatrix}0\\0\end{bmatrix} \begin{bmatrix}0\\1\end{bmatrix} \begin{bmatrix}0\\2\end{bmatrix} \begin{bmatrix}0\\3\end{bmatrix} \begin{bmatrix}1\\0\end{bmatrix} \begin{bmatrix}1\\1\end{bmatrix} \begin{bmatrix}1\\2\end{bmatrix} \begin{bmatrix}1\\3\end{bmatrix} \right\}$$

For each time step t, the patient state s_t and the action a_t taken by the physician together form the trajectory for that time step. The complete set of trajectories from individual patients constitutes the dataset $D = \{(s_1^i, a_2^i, \cdots, s_\tau^i, a_{\tau+1}^i)\}_{i=1}^m$, where τ denotes the maximum time step for the i_{th} patient and m is the total number of patients. Given that we do not have direct access to the reward function R or the transition probability P, we adopt IRL to infer R from the expert trajectories. IRL refers to a class of algorithms designed to derive an unknown reward function from demonstrations, enabling the reward function to be informed by observations of expert behavior [16]–[19].

To achieve this, *FERRY* defines a surrogate distribution $r_\phi(R)$, parameterized by ϕ, along with a policy network Q_θ, parameterized by θ. Next, *FERRY* minimizes the Kullback-Leibler (KL) divergence between r_ϕ and the posterior distribution $p(R|D)$, forming an optimization objective:

$$\min_{\phi} D_{KL}(r_\phi(R)) || p(R|D) \tag{1}$$

Then, *FERRY* updates both ϕ and θ simultaneously using stochastic gradient descent (SGD). The optimization continues until convergence, yielding a reward function that explains the expert's demonstrations and a clinical policy optimized for performance under that reward function. The detailed procedures are outlined in pseudo-code in Algorithm 1.

B. DRO Design

As previously mentioned, we want to develop a cross-institutional clinical decision tool within a distributed setting. However, it is crucial to ensure the reliability of the policies across multiple hospitals, particularly given the conservative and rigorous nature of healthcare. Local data stored in different hospitals can vary significantly in both quality and quantity, which may impact model performance. To tackle this challenge, *FERRY* incorporates DRO, a robust learning paradigm designed to enhance the model's resilience by minimizing worst-case empirical risk. Inspired by previous work [20]

that utilized DRO to address the challenges posed by data heterogeneity, we combine various local loss functions with learnable weights and sample clients into the joint learning process based on their performance in each iteration. Our optimization objective can be formulated as:

$$\min_{\theta\in\Theta}\max_{\lambda\in\Lambda}=\sum_{i=1}^{N}\lambda_i f_i(\theta) \tag{2}$$

where $\lambda \in \Lambda$ is the global weight for each local loss function; N is the number of participating clients. A more detailed description of this approach is provided in Algorithm 2.

Algorithm 2 *FERRY* (DRO)

Input: N edge servers, total number of iterations K, local number of steps T, initial model θ_0
Output: Final model θ_K

for k = 0 to K **do**
  Cloud server samples $t^{'} \in [0, T]$ randomly
  Cloud server broadcasts θ_0 and $t^{'}$ to all edge servers.
  for client $i \in N$ in parallel **do**
    $\theta_k^i, \theta_{t'}^i \longleftarrow$ **Joint Loss Training**$(i, \theta_k, t^{'}, j)$
  end for
  $\overline{\theta}_k$=$\sum_{i\in N}\omega_i\theta_k^i$
  $\overline{\theta}_{k,t'} = \sum_{i\in N}\omega_i\theta_{t'}^i$
  Cloud server broadcast $\overline{\theta}_{k,t'}$ to all clients and receive $f_i(\overline{\theta}_{k,t'}, \xi)$ from N clients
  Identify noisy clients based on loss and mark as j
end for
Return θ_K

C. Joint Loss Learning

After applying DRO, *FERRY* obtains the softmax layer outputs P_1 and P_2 of the global and local models. To enhance the policy's robustness against noisy data, *FERRY* employs joint loss learning through a pseudo-siamese paradigm, which integrates both global and local models. This approach simultaneously trains two networks with distinct parameters, enabling the model to leverage shared knowledge across hospitals while accommodating the unique characteristics of each hospital's data [21]. The losses of the two models are computed jointly, based on the premise that both models should agree on samples with correct labels while diverging on incorrect ones. To implement this, the local side establishes two models with identical structures but different initialization states: one model is derived from the cloud server, while the other is based on the local historical model from the previous training round. *FERRY* updates both models using a joint loss function that incorporates the IRL loss from the previous step along with a co-regularization loss. We define the joint loss function as follows:

$$\mathcal{L}(\xi) = \mathcal{L}_{IRL}(\xi) + \mathcal{L}_{Co-Reg}(\xi) \tag{3}$$

where ξ is the inputs, and $\mathcal{L}_{IRL}$ refers to the IRL loss, while $\mathcal{L}_{Co-Reg}$ indicates the contrast loss between the predicted distributions of the two networks. Co-regularization assists the model in selecting data with correct labels, as a smaller co-regularization loss suggests that the two networks have reached an agreement in their predictions. We quantify the co-regularization term using the KL divergence between the prediction distributions as follows:

$$\mathcal{L}_{Co-Reg} = D_{KL}(P_1||P_2) + D_{KL}(P_2||P_1) \tag{4}$$

where P_1 and P_2 denote the prediction distributions generated by the two models for the same inputs.

Algorithm 3 The procedures of joint loss learning

Input: Dataset D, global model parameters α, local model β
Output: Final model α'

Initialise α, β
while *not converged* **do**
  Sample mini-batch ξ from local data set D
  P_1=$f(\alpha, \xi)$; P_2=$f(\beta, \xi)$
  Calculate the joint loss using P_1, P_2
    $L = \frac{1}{D}\sum_{\xi\in D} l(\xi)$ ▷ $\tilde{\xi}$ denotes the small-loss sets
  Update $\alpha' = \alpha - \eta\nabla L$, $\beta' = \beta - \eta\nabla L$
end while

Intuitively, samples with smaller losses are more likely to be correctly labeled. Consequently, those samples with relatively low loss values are prioritized for updating the policy. The details of the joint loss learning process are outlined in the Algorithm 3.

D. Aggregation Process

FERRY adopts the aggregating weighted loss functions approach [22] to implement the aggregation process. This approach can be mathematically defined as follows:

$$\min_{\theta} f(\theta) = \sum_{i=1}^{N} p_i f_i(\theta) \tag{5}$$

where N is the number of clients each with m_i training samples. The total number of samples is $M = \sum_{i=1}^{N} m_i$, and the sample proportion $p_i = m_i/M$. After the global model parameters are consolidated, the server scatters them down to each hospital again to start a new round of communication.

III. Experiment

A. Experimental Setup

Data Pre-processing. Experiments are conducted on the Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC-III) database [23], [24], which comprises nearly 60,000 ICU inpatient records. To ensure a more objective evaluation of our weaning policy, *FERRY* first filters out data from patients who were ventilated for less than 24 hours or who failed to be discharged at the end of their admission. This resulted in 3545 patients' hourly clinical data. Next, the remaining dataset is then divided into a training set, comprising 2,836 patients and a total of 260,559 trajectories, and a test set, consisting of 709 patients with a total of 67,720

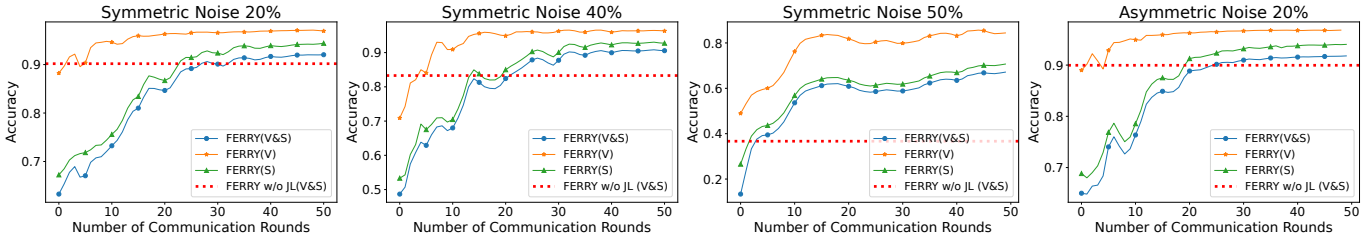

Fig. 3. Comparison of *FERRY* performance under different noise levels

trajectories. Additionally, the training set is split into eight subsets[1], which are distributed to eight hospitals, ensuring that each hospital receives only one subset of data.

Parameter Setting. We set the number of communication rounds between the cloud server and hospitals to 100, with 3,000 iterations of the local model occurring between every two communication rounds. The cloud server uses the Adam optimizer with a learning rate of 10^{-2}, while each hospital employs a 3-layer MLP network architecture with 64 units and ELU activation, training the local model with the Adam optimizer at a learning rate of 10^{-4}. The model parameters are randomly initialized on the cloud server at the start of training.

Benchmarks. We compare *FERRY* against the following methods:

- FQI [10]: A classic offline RL algorithm used to develop clinical decision-making tools for ventilation and sedative management on a centralized dataset.
- AVRIL [25]: A scalable and robust IRL algorithm designed to learn reward functions from expert demonstrations on a centralized dataset.
- Fed-NFQI [10], [22]: A method that integrates FL with neural FQI to continuously optimize policies in a distributed setting.
- *FERRY* w/o JL: A variant of *FERRY* without joint loss learning, used to assess the contribution of this component.

Metrics. We evaluate *FERRY* with *Ventilation Accuracy (V)*, *Sedative Accuracy (S)*, *Joint Accuracy (V&S)*, and *Rounds to Target Accuracy (RA)*. The term *Accuracy* denotes the ratio of predictions that match the doctor's actions to the totals. As an example, *V&S* indicates whether the prediction of the ventilation and sedative is consistent with the doctor's actions simultaneously. *RA* is the number of communications between the cloud server and clients required for the model to reach a target accuracy of 85%.

B. Experimental Results

Accuracy. Table I compares the performance of various methods on the MIMIC-III dataset, focusing on five metrics. *FERRY* achieves superior performance across all metrics, with *V&S* at 91.75% and reaching the target accuracy in 34 communication rounds, demonstrating a 27.66% improvement over its variant without joint loss learning (*FERRY* w/o JL). *FERRY*'s performance is notable for its privacy-preserving distributed approach, achieving similar results to centralized methods like AVRIL while protecting patient data.

TABLE I
RESULTS OF DIFFERENT METHODS ON MIMIC-III.

Method	*V*	*S*	*V&S*	*RA*	Improvement(%)
CNN	86.07%	51.3%	47.52%	Not applicable	
FQI	85%	58%	55%	Not applicable	
AVRIL	96.38%	96.71%	93.43%	Not applicable	
Fed-NFQI	71.89%	16.89%	14.38%	Can't reach	Not applicable
FERRY w/o JL	96.21%	91.69%	89.16%	47	0
FERRY	**96.76%**	**94.31%**	**91.75%**	**34**	**27.66**

Impact of DRO. We evaluate the robustness of the final trained model by evaluating its performance across different hospitals, which exhibit varying data distributions. Our focus is on the metric *V&S*, allowing us to assess the model's generalization and adaptability effectively. The results, illustrated in Figure 4, compare the performance of *FERRY*, *FERRY* without DRO, and Fed-NFQI on the hospital with the worst data distributions under different noise environments. *FERRY* consistently outperforms the other methods, with the performance gap widening as the noise in training labels increases. This highlights *FERRY*'s ability to better handle noisy and heterogeneous data through its selective co-training of the worse-performing hospital in each iteration of DRO, ultimately enhancing robustness.

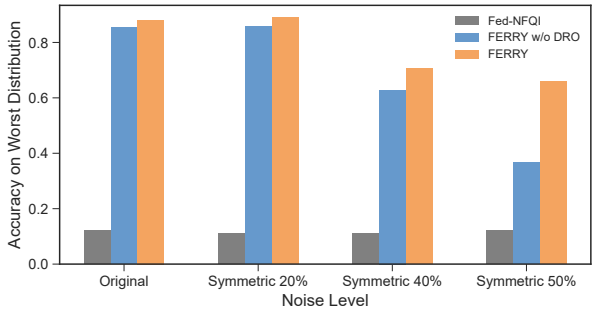

Fig. 4. The accuracy of *FERRY* and benchmarks on the worst distribution

Impact of joint loss learning. We further experiment to verify

[1]In this paper, the number of hospitals participating in the model training process is set to 8.

the ability of *FERRY* to handle noisy data. We create synthetic symmetric noisy labels[2] and asymmetric noisy labels[3] by flipping the original labels. Figure 3 presents the accuracy trends of *FERRY* and *FERRY* w/o JL under different noise conditions[4]. (symmetric 20%, 40%, 50%, and asymmetric 20%) over 50 communication rounds. It shows that *FERRY* consistently achieves higher accuracy and converges faster, especially under higher noise levels. In contrast, *FERRY* w/o JL struggles with much lower performance, particularly under symmetric 50% noise, where it barely improves after a few rounds. Table II complements this by providing the final accuracy values for both methods. *FERRY* outperforms *FERRY* w/o JL across all noise levels, with significant gains under higher noise conditions, reaffirming the robustness of the joint loss learning mechanism in handling noisy data.

TABLE II
THE TERMINAL ACCURACY OF *FERRY* & *FERRY* W/O JL UNDER DIFFERENT NOISE LEVEL

Noise Level	Method					
	FERRY w/o JL			*FERRY*		
	V(%)	*S*(%)	*V&S*(%)	*V*(%)	*S*(%)	*V&S*(%)
Symmetric 20%	96.23	92.48	90.18	97.09	94.19	92.01
Symmetric 40%	90.75	85.53	83.3	96.47	93.05	90.82
Symmetric 50%	53.89	40.05	36.68	84.41	70.69	67.13
Aymmetric 20%	96.56%	92.59%	89.96%	96.83%	94.06%	91.83%

IV. CONCLUSIONS

In this paper, we introduce *FERRY*, a distributed medical decision-making tool designed to provide timely therapeutic recommendations based on patient data while addressing critical issues such as privacy concerns, suboptimal reward function design, and the detrimental effects of noisy data. We evaluate *FERRY* using a real-world database and demonstrate its effectiveness in facilitating intelligent management of ventilation and sedation. Our results show that *FERRY* not only outperforms traditional FL methods in terms of accuracy and noise resilience but also achieves collaborative learning while effectively safeguarding patient privacy.

V. ACKNOWLEDGEMENTS

This work is supported in part by the Research Funds of Centre for Leading Medicine and Advanced Technologies of IHM under Grant 2023IHM01081 and 2023IHM01085 and National Natural Science Foundation of China under Grant 61932017. The authors would like to thank the Information Science Laboratory Center of USTC for the hardware and software services.

[2]Symmetric noise labeling refers to either ventilation or sedative, where the current action is changed to its opposite action, e.g., intubation is changed to extubation

[3]Asymmetric noise labeling refers to the random modification of the selected action to any other action.

[4]The noise level is defined as the ratio of the number of flipped labels to total samples.

REFERENCES

[1] D. Technologies, "Dell emc global data protection index survey." https://www.dellemc.com/en-us/data-protection/gdpi/index.htm.

[2] M. Raza, M. Awais, N. Singh, M. Imran, and S. Hussain, "Intelligent iot framework for indoor healthcare monitoring of parkinson's disease patient," *IEEE JSAC*, vol. 39, pp. 593–602, 2020.

[3] H. Wunsch, J. Wagner, M. Herlim, D. Chong, A. Kramer, and S. D. Halpern, "Icu occupancy and mechanical ventilator use in the united states," *Critical care medicine*, vol. 41, 2013.

[4] S. B. Patel and J. P. Kress, "Sedation and analgesia in the mechanically ventilated patient," *American journal of respiratory and critical care medicine*, vol. 185, pp. 486–497, 2012.

[5] G. Conti, J. Mantz, D. Longrois, and P. Tonner, "Sedation and weaning from mechanical ventilation: time for 'best practice'to catch up with new realities?," *Multidisciplinary respiratory medicine*, vol. 9, pp. 1–5, 2014.

[6] W. Gong, L. Cao, Y. Zhu, F. Zuo, X. He, and H. Zhou, "Federated inverse reinforcement learning for smart icus with differential privacy," *IEEE IoTJ*, vol. 10, no. 21, pp. 19117–19124, 2023.

[7] C. Yu, J. Liu, S. Nemati, and G. Yin, "Reinforcement learning in healthcare: A survey," *ACM Computing Surveys (CSUR)*, vol. 55, pp. 1–36, 2021.

[8] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care," *Nature medicine*, vol. 24, pp. 1716–1720, 2018.

[9] A. Raghu, M. Komorowski, I. Ahmed, L. Celi, P. Szolovits, and M. Ghassemi, "Deep reinforcement learning for sepsis treatment," *arXiv preprint arXiv:1711.09602*, 2017.

[10] N. Prasad, L.-F. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt, "A reinforcement learning approach to weaning of mechanical ventilation in intensive care units," *UAI*, 2018.

[11] C. Yu, J. Liu, and H. Zhao, "Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units," *BMC medical informatics and decision making*, vol. 19, pp. 111–120, 2019.

[12] G. S. Aujla and A. Jindal, "A decoupled blockchain approach for edge-envisioned iot-based healthcare monitoring," *IEEE JSAC*, vol. 39, pp. 491–499, 2020.

[13] Y. Zhang, D. He, M. S. Obaidat, P. Vijayakumar, and K.-F. Hsiao, "Efficient identity-based distributed decryption scheme for electronic personal health record sharing system," *IEEE JSAC*, vol. 39, pp. 384–395, 2020.

[14] C. Lee, Z. Luo, K. Y. Ngiam, M. Zhang, K. Zheng, G. Chen, B. C. Ooi, and W. L. J. Yip, "Big healthcare data analytics: Challenges and applications," *Handbook of large-scale distributed computing in smart healthcare*, pp. 11–41, 2017.

[15] A. G. Barto and R. S. Sutton, "Reinforcement learning," *Handbook of brain theory and neural networks*, pp. 804–809, 1995.

[16] A. Y. Ng, S. Russell, *et al.*, "Algorithms for inverse reinforcement learning.," in *Proc. of ICML*, 2000.

[17] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proc. of ICML*, 2004.

[18] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey, *et al.*, "Maximum entropy inverse reinforcement learning.," in *Proc. of AAAI*, 2008.

[19] S. Levine, Z. Popovic, and V. Koltun, "Nonlinear inverse reinforcement learning with gaussian processes," *Proc. of NeurIPS*, 2011.

[20] Y. Deng, M. M. Kamani, and M. Mahdavi, "Distributionally robust federated averaging," *Proc. of NeurIPS*, 2020.

[21] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proc. of IEEE/CVF CVPR*, 2020.

[22] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. of AISTATS*, 2017.

[23] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, pp. 1–9, 2016.

[24] A. Johnson, T. Pollard, and R. Mark, "MIMIC-III clinical database," 2020.

[25] A. J. Chan and M. van der Schaar, "Scalable bayesian inverse reinforcement learning," *Proc. of ICLR*, 2022.