INSPIRE INTELLIGENCE

Inspire Intelligence Journal
Vol. 1, No. 1, 2026
https://doi.org/10.65718/inspireAI.2026.1001

InspireQuill.org
Inspire Academic Press Inc. (Canada)

Research Paper

Toward Improved ICU Care - Phlebotomy Frequency Using Deep Learning

Rola Al Mallah [1, *], Alejandro Quintero [2]

[1] Computer Engineering department, Polytechnique Montreal, Montreal, QC, Canada
[*] Corresponding Author: rola.al-mallah@polymtl.ca

Abstract

Frequent phlebotomies in Intensive Care Units (ICUs) are crucial for patient monitoring but pose significant challenges, including complications such as iatrogenic anemia, which affects up to 70% of ICU patients. Traditional blood sampling methods do not sufficiently mitigate anemia or improve blood draw frequency. Recent advancements in Artificial Intelligence (AI) offer promising opportunities to enhance healthcare practices in the ICU, especially by analyzing diverse patient data, such as vital signs, blood test results, and demographics. However, existing AI models have predominantly focused on predicting isolated blood test results, rather than providing a holistic solution to forecast blood draw needs throughout a patient's ICU stay. This study addresses this gap by developing a Long Short-Term Memory (LSTM) model to predict blood draw frequency upon ICU admission, using key temporal, quantitative, and patient-specific features. Trained on data extracted from a real-world database, the model demonstrates promising performance, particularly by incorporating a novel loss function in the field to improve predictions of variable multi-output sequence lengths. This research introduces the first predictive model for ICU blood draw frequency, shifting the approach from reactive to proactive care. It underscores the potential of AI to personalize ICU practices, reduce unnecessary blood draws, and improve patient outcomes.

Keywords: Intensive Care Unit (ICU); Iatrogenic Anemia; Long-Short-Term Memory (LSTM); Phlebotomy Frequency Prediction; Variable Multi-Output Sequence

1. Introduction

Advancements in Artificial Intelligence (AI) are transforming patient care by improving decision-making, diagnostic accuracy, and treatment personalization using historical data and clinical insights [1]. Intensive Care Units (ICUs) have become a key research area for exploring AI's potential in enhancing critical care [2, 3]. One targeted focus is phlebotomy (blood draws) which raises concerns about complications like iatrogenic anemia due to its frequent use in ICUs. This condition affects over 70% of patients by day two [4] and nearly 95% have low hemoglobin level by day three [5]. Frequent blood draws also increase infection risk and patient discomfort, potentially hindering recovery and increasing mortality. Despite these issues, phlebotomy remains essential for real-time patient monitoring, helping healthcare providers assess treatment efficacy and detect complications early

Academic Editor: Ghazanfar Latif

Received: 16/06/2025
Revised: 28/08/2025
Accepted: 20/11/2025
Published: 29/12/2025

Citation
Al Mallah, R., & Quintero, A. (2026). Toward improved ICU care: Phlebotomy frequency using deep learning. *Inspire Intelligence Journal, 1*(1), 1-9.

[6]. Balancing the need for phlebotomy with the risks of anemia and other complications remains a challenge in ICU care [7].

To reduce blood loss, traditional strategies rely on clinical expertise and manual processes through blood management programs. These include using smaller blood collection tubes, closed blood sampling devices, point-of-care testing, staff education and a bundle of interventions [8, 9]. These interventions have reduced blood loss by approximately 25%, showing potential for improvement. However, the decision-making process of healthcare professionals remains poorly defined, as they may continue to order blood draws based on immediate clinical judgment. Standardized protocols also present challenges, factors such as the severity of a patient's condition, disease progression, and resource availability complicate their implementation [10]. If the patients' clinical characteristics are not the cause of increased blood sampling frequency and volume, the underlying decision-making by clinicians is unclear. It is plausible that blood sampling practice remains a matter of tradition, clinician preference or fear, in comparison to a reflection of the best available evidence [7]. This suggests that there may be a lack of evidence-based decision-making behind the frequency and volume of phlebotomy and thus underscores the need for a more personalized approach to blood sampling [11, 12]. AI-based research in this area predominantly and solely focuses on evaluating the necessity of blood draws by forecasting their values [13, 14]. They use machine learning techniques to predict blood test results at specific points in time, aiming to determine whether individual phlebotomies are necessary or not. However, this approach is limited by its focus on isolated time points, without considering the broader temporal context of phlebotomy during the patient's ICU stay, rendering it not significantly different from traditional methods. The contributions of this paper are:

- Proactive care planning: We introduce the first predictive model to estimate blood draw frequencies for ICU patients at the time of admission, enabling clinicians to plan care proactively. The model supports early identification of trends and potential complications, allowing preventive interventions.
- Data-driven personalization: Our approach leverages both established and novel features, including vital signs, laboratory values and demographics, using a realistic ICU database to enable more precise, personalized patient management.
- Broader research impact: The proposed model not only improves clinical decision-making and patient outcomes but also provides a foundation for causal inference and can be extended to other real-world healthcare applications.

The remainder of the paper is organized as follows: Section 2 provides a detailed explanation of the methodology used to construct the predictive model, Section 3 discusses the outcomes and findings, and Section 4 provides a discussion along with recommendations for future research, and the paper concludes in Section 5.

2. Materials and Methods

The prediction of phlebotomy frequency for ICU patients at admission time is determined by implementing a Long-Short-Term Memory (LSTM) network within a machine learning pipeline developed from scratch as shown in Figure 1.

2.1 Data Source and Feature Selection

This study uses the latest version of the Medical Information Mart for Intensive Care (MIMIC-IV, v2.2), an openly accessible database containing de-identified health data from patients admitted to critical care units at Beth Israel Deaconess Medical Center from 2012-2019 [15]. The database includes vital signs, blood test results, and various ICU data along with their exact timestamps, enhancing the credibility and reliability of the findings in real-world critical care scenarios. The study is limited to the context of adult ICU patients (> 18 years old) and does not extend to other medical settings or types of predictive models beyond the chosen LSTM and baseline approach.

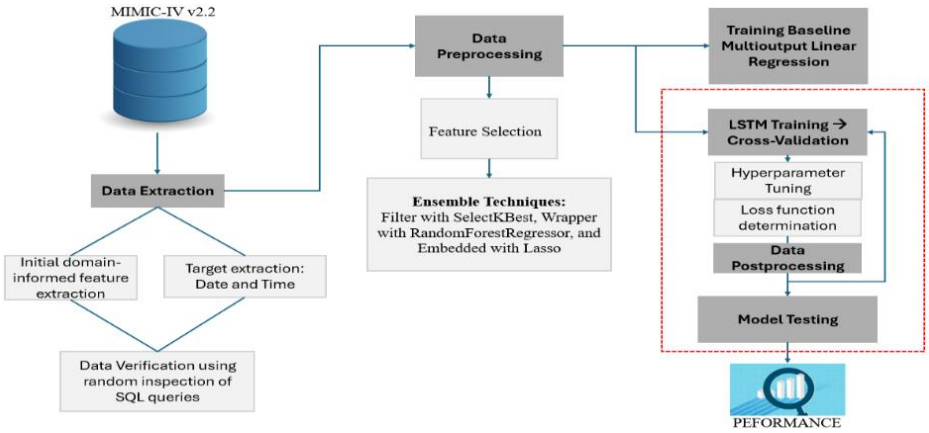

Figure 1. Machine learning pipeline for the prediction of variable multi-output time-series of phlebotomy events.

For feature selection, our model was informed by studies, which identified 12 key laboratory tests, 7 vital signs, age and gender as significant predictors of blood test results. These features were considered likely to serve as important determinants in the prediction of blood draw frequencies. We also considered additional features such as primary diagnosis, first ICU care unit, weight, insurance, marital status, and race. The data was extracted from Google Cloud's BigQuery. Quality verification was performed through result comparison and sample data inspection. To improve performance, an ensemble feature selection approach combining SelecKbest, RandomForestRegressor, and LassoCV from Scikit-learn was used, applying an intersection method similar to [16].

This study employs a hybrid approach to handle missing data, combining Machine Learning (ML) Remove, ML-Unique and ML-Traditional methods to balance the integrity of the dataset and the sample size, a technique used by the authors in [17]. Categorical values are replaced with a distinct "Unknown" category when missing and encoded using One-Hot Encoding. Word Embeddings was implemented with a vector of 100 for primary diagnoses on both International Classification of Diseases (ICD) version 9 and 10. Continuous variables are imputed using population medians and binary indicators. Outliers are managed using Z-scores for normally distributed features and the Inter Quartile Range (IQR) for skewed data, with additional steps to retain clinically significant outliers and ensure meaningful data are preserved. They are preprocessed using the standardization scaling technique.

2.2 Model Development and Implementation

We initially used an LSTM model with a single hidden layer to establish a baseline, which was compared to a multi-output linear regression model as a theoretical reference. This comparison helped identify key features and informed further post-processing adjustments to handle variable output sequences. The model was initially trained using Mean Squared Error (MSE) loss. We then evaluated the impact of the Correntropy loss, Equation (1), which effectively addresses non-Gaussian noise often present in raw medical data. This additional testing aimed to assess whether Correntropy loss could improve the model's performance, as it did in the context of vehicle control flow prediction [18]. The kernel size (σ) was adjusted to identify the optimal value, using sizes derived from those in [19].

$$C(y, \hat{y}) = -\sum_{i=1}^{n} K(y_i - \hat{y}_i), \quad K(x) = e^{-\frac{x^2}{2\sigma^2}}, \tag{1}$$

where the parameter σ controls the bandwidth of the kernel.

The model's performance was evaluated on a separate test set (comprising 20% of the entire dataset) using a combination of metrics: average Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for both the predicted values (time intervals) and the Sequence Length (SL). These metrics were combined to provide a comprehensive evaluation of the predictive accuracy of the model. The architecture was progressively expanded by adding layers through Grid Search to improve performance, while early stopping was used to prevent overfitting. Finally, the model was provided with static patient data at the time of admission as inputs, representing a single timestamp with multiple features, and variable sequential blood draw data as outputs. The target values, which consist only of date and time with varying lengths, were transformed into intervals representing the elapsed time since the previous blood draw. The remaining portion of the sequence was padded with zeros.

3. Results

3.1 Outcomes of Data Preprocessing

For the laboratory test data, we focused on results taken within approximately six hours of ICU admission, using the *"charttime"* attribute as the reference for extraction in MIMIC-IV. The first recorded test for each patient, based on proximity to admission, was retained by assigning a unique rank to each test event. This ensured that duplicates were eliminated, preserving the integrity of the data. Similar queries were used to extract vital signs, demographic information, and medical details such as primary diagnoses, with the first recorded measurement at admission time being selected for analysis. Additionally, we carefully handled the timing of blood draw collection, using *"charttime"* as a proxy for blood draw times, as the exact time of blood draw was not provided. This approximation is a reasonable trade-off, as most blood draws in the ICU require rapid analysis, ensuring that the timing remains sufficiently accurate for our purposes. We grouped data by *"charttime"* to capture unique blood draw events and performed data cleaning by excluding intervals shorter than 5 minutes, which were likely due to human errors or duplications. The queries were verified by comparing the blood draws counts and confirming their accuracy through manual inspection of sample patient records. This process ensured the data's reliability, thus eliminating the need for synthetic data or simulations.

After preprocessing, the dataset contained 71,184 unique patients, though some may have been readmitted and counted as separate entries. Most patients underwent 50 or fewer blood draws (t1 to t50), with those exceeding this number categorized as outliers to streamline the analysis. Missing values were predominantly found in laboratory tests, particularly for HCO3, calcium, and phosphate. Feature selection revealed significant correlations among features like blood pressure measurements, hemoglobin, and hematocrit. Using ensemble of feature selection, we excluded demographic features like race, marital status, and insurance. The top 11 consistent features across methods included WBC, first care unit, Bpsystolic, calcium, respiratory rate, BUN, heart rate, primary diagnosis, RDW, phosphate, and hemoglobin, while the bottom four were insurance, marital status, race, and surprisingly, gender. Notably, gender, despite its initial exclusion, improved model performance when included. The analysis underscores the effectiveness of ensemble feature selection in enhancing model performance by reducing dimensionality and eliminating the need for time-consuming manual feature evaluation. The final set of selected features is presented in Once trained, the model struggled to accurately identify the appropriate stopping points at the timestamps, failing to predict the End-of-Sequence (EOS) token of zeros used for padding. Although we experimented with an alternative token, such as -1, this approach resulted in increased MAE and RMSE. To resolve this issue, we modified the heuristic and fine-tuned it to improve performance. The mechanism involved truncating the sequence when a pattern of five consecutive decreases in predictions was observed. Specifically, we monitored each prediction sequence and counted instances where the current value was less than the previous one. If this "decrease count" reached the threshold of five, tuned for optimal performance, we reset all values from the second-to-last detected decrease onward to zero, effectively truncating the sequence. We incorporated this

mechanism because we noticed that, rather than generating the EOS token, the LSTM was producing a downward trend after a certain timestamp, prompting us to truncate the sequence at that point (see Figure 2).

Once trained, the model struggled to accurately identify the appropriate stopping points at the timestamps, failing to predict the End-of-Sequence (EOS) token of zeros used for padding. Although we experimented with an alternative token, such as -1, this approach resulted in increased MAE and RMSE. To resolve this issue, we modified the heuristic and fine-tuned it to improve performance. The mechanism involved truncating the sequence when a pattern of five consecutive decreases in predictions was observed. Specifically, we monitored each prediction sequence and counted instances where the current value was less than the previous one. If this *"decrease count"* reached the threshold of five, tuned for optimal performance, we reset all values from the second-to-last detected decrease onward to zero, effectively truncating the sequence. We incorporated this mechanism because we noticed that, rather than generating the EOS token, the LSTM was producing a downward trend after a certain timestamp, prompting us to truncate the sequence at that point (see Figure **2**).

Table 1. Final set of patient features derived from both empirical studies, clinical settings and ensemble feature selection.

Category	**Selected features**
Demographics (total = 2)	Gender, Age
Vital signs (total = 8)	Weight, Heart Rate, Respiratory Rate, Oxygen, Saturation pulse (SPO2), Mean systolic blood pressure (Bpsystolic), Mean diastolic blood pressure (Bpdiastolic), Mean blood pressure (Bpmean), Temperature
Laboratory tests (total = 14)	Sodium (Na), Potassium (K), Bicarbonate (HCO3), Phosphate (PO4), Creatinine (Cr), Blood Urea Nitrogen (BUN), Platelet Count (Plt), Hemoglobin (hgb), White Blood Cells (WBC), Mean Corpuscular Hemoglobin Concentration (MCHC), Red Blood Cells (RBC), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Red Cell Distribution Width (RDW)
Clinical assessment (total = 2)	Primary Diagnosis, First Care Unit

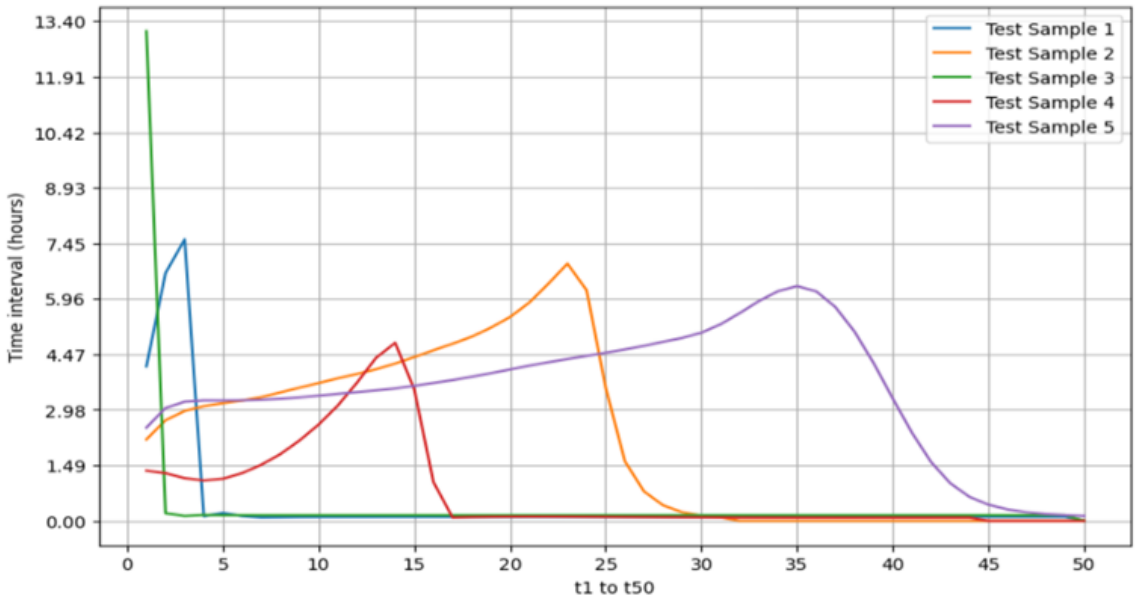

Figure 2. Trend of the predicted values at each timestamp using 5 samples from the test set.

3.2 Model Evaluation and Improvement

Initially, a multi-output linear regression model was evaluated, yielding an average MAE of 1.33 hours and RMSE of 3.58 hours. The performance of SL prediction was also lacking, with an average MAE of 12.06 blood draws and an RMSE of 18.98 blood draws. Feature selection and the introduction of a timing-related feature, representing the time elapsed since the first blood draw after admission, further improved performance on both target values and SL. This enhancement resulted in an MAE of 1.31 hours, an RMSE of 3.58 hours, and a 45.85% within ±3 blood draws improvement in SL prediction. Building on these improvements, the LSTM model was trained with a single train-test split and evaluated on a separate test dataset. However, further evaluation using methods like k-fold cross-validation is recommended for more robust and generalized results.

Experimenting with Correntropy loss and optimizing the kernel size to 1 yielded the most accurate SL predictions, achieving 55.28% within ±3 blood draws. Though its impact on time intervals was minimal, as slight timing deviations persisted, model optimization was achieved through hyperparameter tuning with GridSearch, adjusting hidden layers, units, dropout rates, and batch sizes. The optimal configuration consisted of four hidden layers with dropout rates of 0.5 (except for the last layer) and 512, 256, 128, and 50 units, respectively. Introducing a novel feature, the count of blood draws per patient, led to reductions in both MAE with 0.98 hours and RMSE with 2.81 hours. The tuned model achieved an MAE of 1.22 hours, RMSE of 3.62 hours for time intervals. This resulted in 88.26% ±3 blood draws in performance for SL predictions, highlighting the importance of temporal and quantitative data. Final performance results are shown in Figure 3.,
Figure ***4***. and
Figure **5**., with true and predicted sequence lengths nearly overlapping.

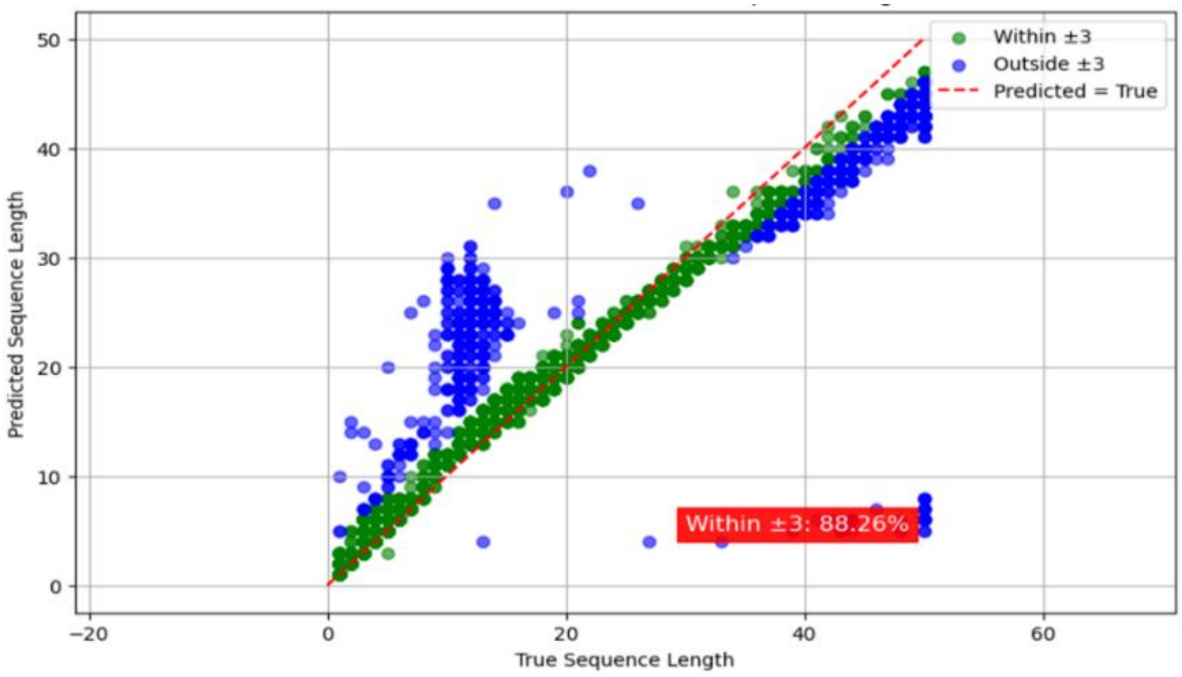

Figure 3. Scatter plot of the true vs. predicted sequence length with tuned LSTM using selected features, time of first blood draw, and total number of blood draws trained with Correntropy Loss with.

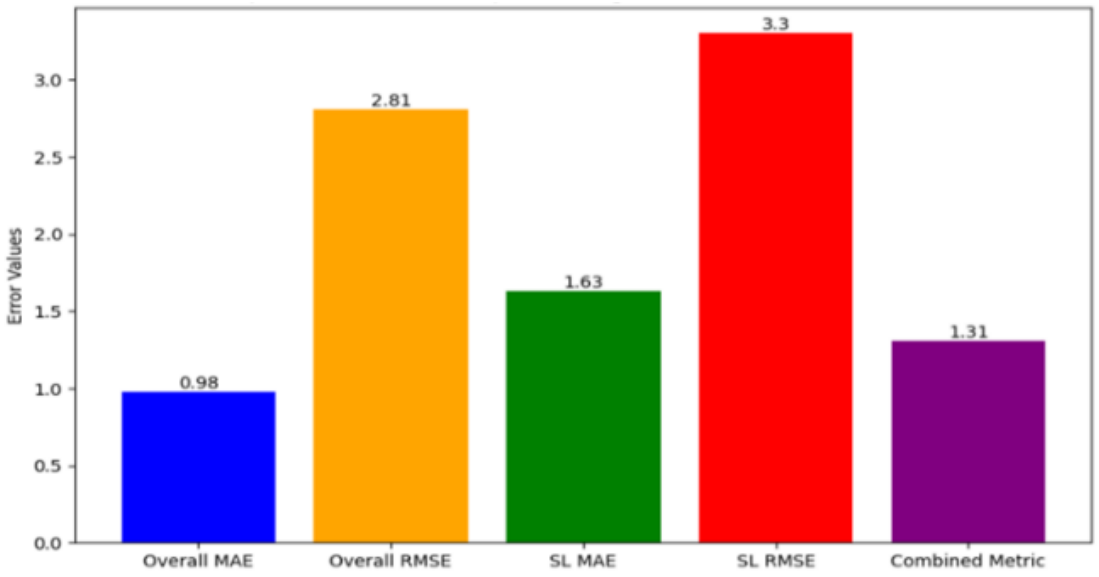

Figure 4. Model Performance Metrics

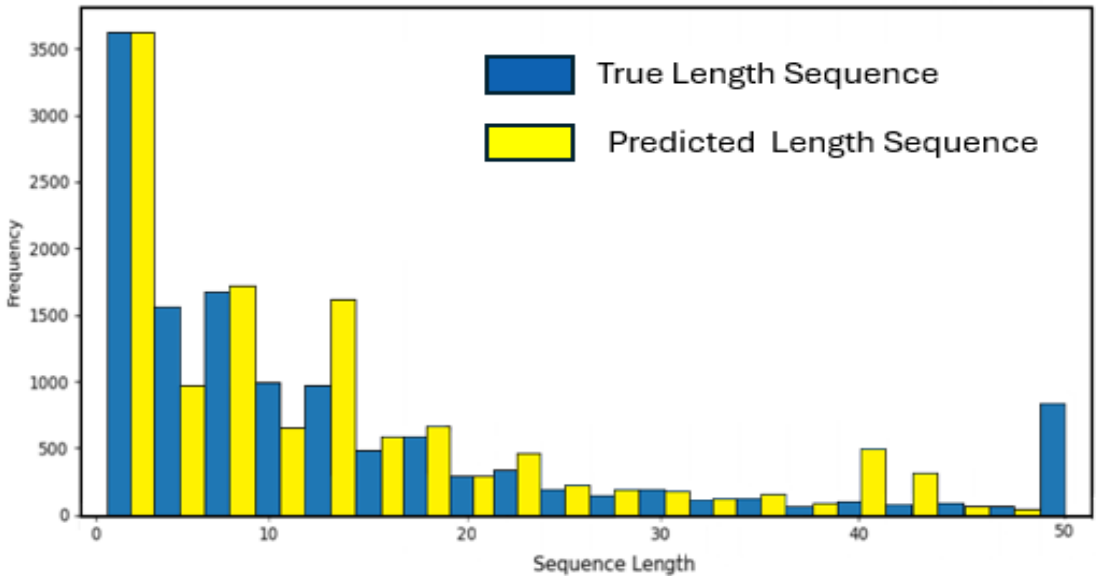

Figure 5. Distribution of true vs. predicted SL with final LSTM using selected features and Correntropy loss.

4. Discussion

This research highlights the potential of AI in predictive medicine. The key contribution of the study is its focus on forecasting the entire sequence of phlebotomy at the time of ICU admission using LSTM networks to improve resource allocation, minimize unnecessary phlebotomies, and potentially reduce iatrogenic anemia. Leveraging a realistic database, MIMIC-IV v2.2, the research effectively identified the most important features, and the LSTM

proved to be effective in modeling variable output sequence lengths. A detailed evaluation of the key features revealed that vital signs, blood test results, demographics, and temporal and quantitative data (e.g., the timing of the first blood draw and total blood draws per patient) were crucial in predicting blood draw frequency. Features such as calcium and chloride were excluded due to high correlation and missing data, while vital signs like heart rate and respiratory rate, along with lab results like phosphate and white blood cell count, played a significant role in improving model accuracy. The model faced challenges in predicting accurate sequence lengths, particularly with LSTM's difficulty in detecting the EOS token, suggesting the need for additional solutions such as decomposing the task or exploring alternative architectures like transformers or attention mechanisms. However, dataset size should be carefully considered, as these architectures are particularly effective when trained on larger datasets.

The study found that MSE Loss was effective in predicting interval timings for each blood draw, while Correntropy loss performed better in capturing the variability in sequence lengths. The study suggests a potential approach of using Correntropy loss for predicting sequence length, followed by MSE for predicting phlebotomy time intervals, or developing a combined loss function. These findings set the stage for future improvements, particularly through exploring more advanced models, refining feature selection, and incorporating additional contextual data to enhance the robustness of predictions and model generalization.

5. Conclusion

This study addresses key challenges in ICU blood management by developing a framework to predict blood test frequencies upon ICU admission, potentially minimizing unnecessary procedures, reducing patient discomfort, and most importantly preventing iatrogenic anemia. By leveraging demographics, vital signs, blood draws, temporal and quantitative data from the MIMIC-IV database, the study demonstrates how machine learning, specifically LSTM models, can forecast phlebotomy frequencies. The findings open the door for integrating this approach with multi-task frameworks, ultimately enhancing patient care and clinical workflows in Clinical Decision Support Systems (CDSS) for monitoring phlebotomy in ICUs.

Data Availability Statement
Not applicable.

Funding
This work was supported without any funding.

Conflicts of Interest
The author declares no conflicts of interest.

Ethical Approval and Consent to Participate
Not applicable.

References

[1]. Mueller, B., Kinoshita, T., Peebles, A., Graber, M. A., Lee, S.: Artificial intelligence and machine learning in emergency medicine: a narrative review. *Acute Med. Surg.* **9**(1), e740 (2022)

[2]. Alghatani, K., Ammar, N., Rezgui, A., Shaban-Nejad, A.: Predicting intensive care unit length of stay and mortality using patient vital signs: machine learning model development and validation. *JMIR Med. Inform.* **9**(5), e21347 (2021). https://doi.org/10.2196/21347

[3]. Gutierrez, G.: Artificial intelligence in the intensive care unit. *Crit. Care* **24**, 101 (2020). https://doi.org/10.1186/s13054-020-2785-y

[4]. Whitehead, N.S., et al.: Interventions to prevent iatrogenic anemia: a Laboratory Medicine Best Practices systematic review. *Crit. Care* **23**(1), 278 (2019). https://doi.org/10.1186/s13054-019-2511-9

[5]. Ullman, A.J., Keogh, S., Coyer, F., Long, D.A., New, K., Rickard, C.M.: 'True Blood' the critical care story: an audit of blood sampling practice across three adult, paediatric and neonatal intensive care settings. *Aust. Crit. Care* **29**(2), 90–95 (2016). https://doi.org/10.1016/j.aucc.2015.06.002

[6]. Olver, P., Bohn, M.K., Adeli, K.: Central role of laboratory medicine in public health and patient care. *Clin. Chem. Lab. Med.* **61**(4), 666–673 (2023). https://doi.org/10.1515/cclm-2022-1075

[7]. Matzek, L.J., et al.: A contemporary analysis of phlebotomy and iatrogenic anemia development throughout hospitalization in critically ill adults. *Anesth. Analg.* **135**(3), 501–510 (2022). https://doi.org/10.1213/ANE.0000000000006127

[8]. Neef, V., et al.: Effect of using smaller blood volume tubes and closed blood collection devices on total blood loss in patients undergoing major cardiac and vascular surgery. *Can. J. Anesth.* (2024). https://doi.org/10.1007/s12630-023-02643-8

[9]. François, T., et al.: Strategies to reduce diagnostic blood loss and anemia in hospitalized patients: a scoping review. *Pediatr. Crit. Care Med.* **24**(1), e44–e53 (2023). https://doi.org/10.1097/PCC.0000000000003094

[10]. Yang, Z., Cui, X., Song, Z.: Predicting sepsis onset in ICU using machine learning models: a systematic review and meta-analysis. *BMC Infect. Dis.* **23**(1), 635 (2023). https://doi.org/10.1186/s12879-023-08614-0

[11]. Devis, L., et al.: Interventions to improve appropriateness of laboratory testing in the intensive care unit: a narrative review. *Ann. Intensive Care* **14**(1), 9 (2024). https://doi.org/10.1186/s13613-024-01244-y

[12]. Lennox, S., Bench, S.: Blood sampling in adult critical care: a mixed methods study. *Int. J. Orthop. Trauma Nurs.* **45**, 100923 (2022). https://doi.org/10.1016/j.ijotn.2022.100923

[13]. Huang, T., Li, L.T., Bernstam, E.V., Jiang, X.: Confidence-based laboratory test reduction recommendation algorithm. *BMC Med. Inform. Decis. Mak.* **23**(1), 93 (2023). https://doi.org/10.1186/s12911-023-02187-3

[14]. Yu, L., Zhang, Q., Bernstam, E.V., Jiang, X.: Predict or draw blood: an integrated method to reduce lab tests. *J. Biomed. Inform.* **104**, 103394 (2020). https://doi.org/10.1016/j.jbi.2020.103394

[15]. Johnson, A.E.W., et al.: MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023). https://doi.org/10.1038/s41597-022-01899-x

[16]. Chen, C., Tsai, Y., Chang, F., Lin, W.: Ensemble feature selection in medical datasets: combining filter, wrapper, and embedded feature selection results. *Expert Syst.* **37**(5), e12553 (2020). https://doi.org/10.1111/exsy.12553

[17]. Rios, R., et al.: Handling missing values in machine learning to predict patient-specific risk of adverse cardiac events: insights from REFINE SPECT registry. *Comput. Biol. Med.* **145**, 105449 (2022). https://doi.org/10.1016/j.compbiomed.2022.105449

[18]. Ye, B.-L., Zhang, M., Li, L., Liu, C., Wu, W.: A survey of traffic flow prediction methods based on long short-term memory networks. *IEEE Intell. Transp. Syst. Mag.* **-**, 2–27 (2024). https://doi.org/10.1109/MITS.2024.3400679

[19]. Cai, L., Lei, M., Zhang, S., Yu, Y., Zhou, T., Qin, J.: A noise-immune LSTM network for short-term traffic flow forecasting. *Chaos* **30**(2), 023135 (2020). https://doi.org/10.1063/1.5120502