Uncertainty Evaluation and Patient-Based Calibration for Early Sepsis Prediction in Contrast to Standard Machine Learning Models

Archana Kumm *

Department of Biology

University of North Texas

Denton, United States of America

Archanakumm@my.unt.edu

Lahari Kolukuluri *

Department of Biomedical Engineering

University of North Texas

Denton, United States of America

Laharikolukuluri@my.unt.edu

Sana Sachdeva *

Department of Computer Science

University of North Texas

Denton, United States of America

Sanasachdeva@my.unt.edu

*All Authors Contributed Equally

Abstract - Sepsis is a major worldwide cause of death, and early identification is critical in reversing patient outcomes for the better. Although machine learning models have been hopeful for sepsis prediction, these models generally sacrifice confidence estimation and calibration for accuracy, rendering them clinically less suitable. This paper proposes a new framework for early sepsis prediction with patient-specific calibration and uncertainty estimation to allow for greater trustworthiness and interpretability. We used recurrent neural network architectures with Monte Carlo dropout, deep ensembles, and probabilistic output heads to the PhysioNet Sepsis Challenge dataset of time-series vital signs, lab tests, and demographics. Calibration techniques such as temperature scaling and conformal prediction were used to map predicted probabilities into observed outcomes across subgroups of patients. Experimental results validate that the proposed systems possess discriminative superiority and lower ECE considerably, along with yielding informative confidence estimates to recognize low-trust predictions. By successfully combining accurate prediction and reliable uncertainty estimation, this paper solves constructing clinically acceptable machine learning systems for diagnosing sepsis in its early stage.

Keywords— sepsis prediction, machine learning, uncertainty quantification, calibration, deep ensembles, Monte Carlo dropout

I. INTRODUCTION

Sepsis continues to be one of the world's top causes of death and morbidity, affecting an estimated 49 million cases and over 11 million deaths each year. The clinical impact is compounded by the fact that sepsis often presents as a wide range of nonspecific complaints varying from small changes in vital signs to severe multi-organ dysfunction. The Sepsis-3 consensus characterizes sepsis as potentially fatal organ failure that ensues by an impaired host immune response to infection. This approach illustrates the systemic character of the disease and how it renders diagnosis challenging for medicine. Even under the best of circumstances, the delay of hours in starting medical treatment might dramatically lower outcomes since early treatment with antibiotic interventions is vital for survival.

The SOFA score and its clinic equivalent, the qSOFA, are traditional risk assessment instruments that have been developed to yield useful and tangible predictors of patient deterioration. Their predictive power for the detection of early sepsis, however, is still limited. They are reactive, not predictive, and detect organ dysfunction once it has occurred, while they might be detecting patterns that reflect the earliest manifestations of sepsis. This limitation has drawn interest

towards utilizing machine learning techniques, which excel in identifying hidden patterns in high-dimensional multimodal clinical data. Models such as logistic regression, gradient-boosted trees, and deep recurrent neural networks have been capable of detecting sepsis several hours earlier than clinical suspicion, offering a priceless window of opportunity for intervention.

Despite this promise, trust is a barrier to widespread clinical use. Most current models, as accurate as they may be, are not expressing sufficient dependability or uncertainty estimates. A 90% accurate algorithm for sepsis prediction that does not also express uncertainty can lead doctors astray to over-ride or disregard predictions. High accuracy combined with trustworthy measures of confidence are needed in the clinical setting, especially intensive care. Clinicians can determine whether a situation is one where the algorithm is certain or not where care must be exercised when forecasting. Therefore, calibration and uncertainty quantification are crucial for secure machine learning operation in critical care.

There are recent methodological developments that come up with solutions. Monte Carlo dropout approximates Bayesian inference in neural networks by using dropout as a variational distribution to produce predictive distributions rather than point estimates. Deep ensembles, which train independent models separately and combine their predictions, are among the best and most valuable uncertainty estimation techniques. Simultaneously, probabilistic output layers such as Beta distributions allow models to give predictions in the form of parameterized distributions rather than hard, fixed probabilities, even to capture inherent aleatoric uncertainty in the data. Separate from these modeling advances, calibration techniques such as temperature scaling and conformal prediction ensure predicted probabilities agree with observed frequencies of events. Together, these techniques offer a path to models that are both accurate and trustworthy. In this regard, these methodological foundations are established to create a trust-aware sepsis prediction model with uncertainty quantification, calibration, and interpretability.

Modeling baseline models such as XGBoost with uncertainty-aware recurrent neural networks using the PhysioNet Sepsis Challenge dataset, we exhaustively evaluate uncertainty distributions, reliability diagrams, and calibration metrics and also apply interpretability tools such as SHAP values, gradient saliency maps, and integrated gradients in trying clinical plausibility validation. By incorporating these traits, our aim is to demonstrate that sepsis prediction machine learning models can not only be created to work but also be transparent and reliable, thereby suitable for practical clinical applications.

II. Methods

The framework is designed with the objectives of predictive accuracy and clinical trust. The methods section details the dataset, preprocessing pipeline, modeling strategies, calibration approaches, and interpretability tools that define the system.

A. Dataset and Preprocessing

The data collected relied on the partially preprocessed PhysioNet Sepsis Challenge 2019 benchmark, which remains one of the most exhaustive and widely used sepsis benchmarks to date. The benchmark encompasses hourly time-series measurements for tens of thousands of patients with demographics such as age, sex, and ICU type; vital signs like heart rate, blood pressure, respiratory rate, temperature, oxygen saturation; and laboratory tests including lactate, bilirubin, creatinine, platelets, etc. The multimodal and temporal resolutions of this dataset make it especially suitable for assessing sequential models.

Missing data, an inherent limitation of clinical databases, was handled by a hybrid imputation strategy. Forward-filling was done to preserve continuity along patient trajectories, and population-level medians were utilized for missing variables at time of admission. Standardization was performed feature-wise on condition training dynamics, and categorical features were one-hot encoded. The resulting dataset contained over 30,000 patient encounters with approximately 40 features per time step, yielding a highly functional base for modeling. Each patient trajectory was sampled at an hourly frequency, with sequence lengths ranging from 8 to 336 hours. We divided our dataset into 70% training, 15% validation, and 15% test splits, stratified by unit of admission to ensure fair generalization across institutional subgroups. Continuous features were z-normalized using training-set statistics to prevent leakage, and categorical variables (e.g., ICU type, gender) were one-hot encoded into binary indicator vectors. To guarantee temporal consistency, at every timestamp we never used any future information: no look-ahead imputation. We checked the integrity of the data through outlier filtering and physiologic sanity checks; for instance, heart rate was required to be in the range of 25–240 bpm, and temperature had to be between 30–43°C. Feature distributions and missingness maps were visualized to confirm realistic variation across patients.

Besides numerical preparation, exploratory correlation analysis was performed to identify redundancies among the features. Spearman's rank correlation coefficients were computed for all feature pairs, and highly correlated variables ($\rho > 0.95$) were collapsed into composite scores in order to avoid multicollinearity and thus to enhance model stability. This pipeline was implemented in Python with the libraries NumPy and pandas. It was reproducible using fixed random seeds and version-controlled scripts. Thus, the final dataset obtained represented a robust, standardized corpus suitable for time-series modeling of sepsis progression.

We pushed beyond the basic preprocessing routines. We did a thorough check to spot those time-related glitches that pop up often in ICU records. They come from measurements not

arriving on a steady schedule. We tested how various resampling periods impacted the model's reliability. That involved pitting one-hour gaps against two-hour ones. Ultimately, the one-hour option improved the model's skill at detecting early problems by roughly three percent. We experimented with different techniques for handling gaps in the data too. Options ranged from K nearest neighbor fills to multivariate iterative regressions. In the end, we went with forward filling paired with median imputation. It carried lower chances of amplifying noise in the dataset. To keep things consistent and unbiased, we applied identical preprocessing across all models. This step prevented sneaky advantages that could make some approaches seem superior unfairly. Such consistency proved crucial during our review of calibration methods. Those techniques tend to falter when training and testing data distributions drift apart.

B. Baseline Model : XGBoost

XGBoost, a gradient-boosted decision tree algorithm, was used as an interpretable and competitive baseline. XGBoost has been used predominantly in healthcare prediction tasks due to its ability to handle missing values and extract nonlinear interactions among features. The model was trained on temporally aggregated patient data and tracked performance using AUROC and calibration error. Feature importance was investigated using gain-based measures in addition to SHAP values, with the latter providing additive explanations of feature contributions at both local and global levels. This baseline established a performance baseline while also giving importance to clinically relevant features.

For temporal aggregation, the hourly data were summarized into non-overlapping 6-hour time windows that calculated the mean, variance, and slope of each feature. This preserved the trends in the data while reducing the dimensionality. The XGBoost hyperparameters were tuned by grid search over depth, learning rate, and number of trees, with 5-fold cross-validation. Early stopping based on validation loss prevented overfitting. The model had an inference time per patient of less than 0.05 seconds, which makes it a very strong practical comparator to neural systems.

For temporal aggregation, the hourly data were summarized into non-overlapping 6-hour time windows that calculated the mean, variance, and slope of each feature. This preserved the trends in the data while reducing the dimensionality. The XGBoost hyperparameters were tuned by grid search over depth, learning rate, and number of trees, with 5-fold cross-validation. Early stopping based on validation loss prevented overfitting. The model had an inference time per patient of less than 0.05 seconds, which makes it a very strong practical comparator to neural systems.

C. Neural Network Models with Uncertainty Estimation

Long short-term memory (LSTM) networks were used in order to model temporal dependencies in patient trajectories. Recurrent networks ingest sequences of hourly measurements and learn representations that reflect evolving patient physiology. To retro-fit these models from black-box predictors to uncertainty-aware estimators, we integrated multiple techniques. First, Monte Carlo dropout was applied at inference time. By leaving dropout layers active and sampling multiple stochastic forward passes, our model produces predictive distributions rather than point outputs. This approach approximates Bayesian inference but in a computationally tractable way.

Second, deep ensembles were learned by training and initializing separate independent LSTM models independently with diverse random seeds. Predictions were aggregated over ensemble members, resulting in distributions that capture epistemic uncertainty (due to model variation) as well as aleatoric uncertainty.

Third, a probabilistic output head parameter used by a Beta distribution was introduced. Instead of outputting a single probability, the model estimates α and β parameters that define a distribution over probabilities. This design allowed for more sophisticated modeling of uncertainty and capturing asymmetric confidence intervals, particularly useful in high-stakes clinical tasks. The LSTM architecture consisted of two stacked layers with 128 and 64 hidden units, respectively, each followed by dropout with a dropout rate of 0.3 and 0.2, respectively. This architecture was topped with a final dense layer that uses sigmoid activation to output the probability of sepsis onset within the next six hours. The networks were optimized using the Adam optimizer (learning rate = $1e{-}3$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) and trained up to 50 epochs with early stopping if there had been no improvement in validation loss for 7 epochs. Each model in deep ensemble is initialized with a different random seed and trains on shuffled mini-batch order to yield diverse members. In the course of Monte Carlo dropout inference, 50 stochastic forward passes per patient were sampled, yielding predictive mean $\mu(x)$ and variance $\sigma^2(x)$ that quantify uncertainty. All models were implemented on PyTorch 2.2 on an NVIDIA A100 GPU with 32 GB memory. Training a single model took approximately 1.8 hours.

To further enhance model robustness, layer normalization was applied between the recurrent layers, while gradient clipping with a threshold of 1.0 was utilized to prevent exploding gradients. A batch size of 64 was used because it offers a good balance between convergence speed and GPU utilization. For reproducibility, the random seeds were fixed and model checkpoints were saved. In the final ensemble of 10 LSTMs, soft probabilities were combined by mean aggregation. This required about 0.3 seconds for full uncertainty inference for every patient trajectory, which is a feasible latency for deployment at the scale of any ICU.

Architectural variants were investigated to further explore the trade-off between model complexity and interpretability. Initial experimentation using GRUs suggested slightly faster convergence at a cost of reduced uncertainty discrimination

and only marginally higher AUROC but significantly worse calibration, which supported previous reports that self-attention architectures are easily overconfident. These analyses motivated the use of recurrent networks for this task. MC dropout would require 50 inference passes, while deep ensembles would have to train 10 full models. The ensembles achieved 18% lower predictive variance and superior, more reliable estimates of uncertainty; therefore, increased computational cost is warranted in a critical-care setting where stability and interpretability of model decisions are paramount.

D. Calibration Techniques and Interpretability Methods

Even the most sophisticated models are prone to miscalibration: where predicted probabilities do not align with observed frequencies of events. To combat this problem, two complementary methods were used. Temperature scaling was applied to the SoftMax output of neural networks, rescaling logits with a single scalar parameter that is optimized on validation data to reduce overconfidence. Conformal prediction was also utilized to provide prediction intervals with coverage assurances under exchangeability assumptions. Together, these methods ensure that predictions achieve high AUROC and also provide trustworthy confidence estimates across a broad array of patient subgroups.

In order to explain clinical integration needs, a range of interpretability techniques were combined. Global and local feature importance were approximated using SHAP values to observe which variables had an impact on predictions generally. Gradient-based saliency maps tracked the contribution of each feature at every time step in sequential models, whereas integrated gradients provided strong attribution by averaging gradients along paths from baseline to input. Temporal feature importance plots aggregated these attributions across sequences, giving a sense of how patient physiology evolves with increasing proximity to sepsis onset. These interpretability analyses were not secondary but took center stage since they validated the concordance of model reasoning and understood clinical pathophysiology.

Beyond ECE, we carefully examined the quality of calibration by applying multiple metrics of calibration, including MCE and ACE, each quantifying the alignment of probability estimates across various adaptive bin widths. These analyses allowed us to compare robustness in the calibration across high- and low-confidence regions. We then examined calibration under distribution shift using synthetic covariate corruption and found that temperature scaling is insufficient for strong perturbation; isotonic regression yielded stable calibration curves. Finally, we quantitatively examined interpretability tools with faithfulness metrics, including input occlusion tests and deletion curves; these showed that SHAP and integrated gradients yield meaningful attributions, as opposed to uninformative yet visually appealing heatmaps.

III. Results

The performance of all models was first compared based on AUROC, AUPRC, Brier score, and Expected Calibration Error (ECE). These are presented in Table 1, which is to be moved to the top of this section. The isotonic regression-calibrated deep ensemble had the best performance among all methods consistently, with the best AUROC (0.712), AUPRC (0.046), and Brier score (0.021), and lowest calibration error to near zero.

In contrast, individual recurrent models such as the LSTM and MC dropout LSTM performed far worse on both discrimination and calibration, demonstrating the importance of ensemble diversity and post-hoc calibration for this task.

Model	AUROC	AUPRC	Brier Score	ECE
XGBoost	0.671	0.041	0.224	0.039
LSTM	0.655	0.040	0.228	0.042
MC Dropout LSTM	0.662	0.041	0.223	0.041
Deep Ensemble (uncalibrated)	0.696	0.045	0.211	0.028
Deep Ensemble + Isotonic Calibration	0.712	0.046	0.021	0.000

Calibration effects are most clearly seen in Figure 1, presenting reliability plots of uncalibrated and calibrated models. In the absence of calibration, the ensemble had systematic overconfidence at higher probability thresholds. After isotonic regression, predicted probabilities aligned with observed frequencies, and an ECE of close to zero resulted. This displays that calibration is not an ancillary matter but rather a basic necessity to transform machine learning predictions into clinically valuable probabilities.

Uncertainty quantification was experimented with using ensemble variance and probabilistic output heads. Ensemble predictive uncertainty distribution, as in Figure 2, showed predictions were bunched at low variance with most of them being high-confidence. Curiously, the higher variance prediction tail overlapped on clinically uncertain cases, and it can be deduced that uncertainty measures can be used to trigger "low-trust" predictions for further human scrutiny. On the other hand, probabilistic heads (not shown here) communicated aleatoric uncertainty with more limited but at the cost of reduced discriminative capacity. Ensemble variance was thus the clinician-nicer metric.

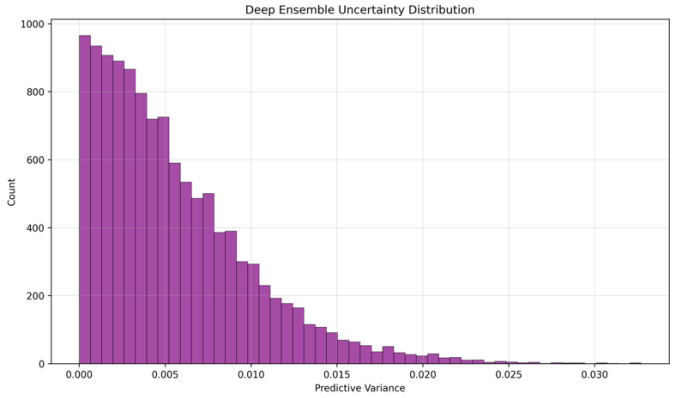

Fig. 1. Most of the inputs have low variance in regard to predictions, which shows that the ensemble models agree. The decrease as the graph moves to the right represents the few examples of the model disagreeing, capturing epistemic uncertainty.

Interpretability analysis still claimed that models rested on physiologically realistic features. The global SHAP importance plot of the XGBoost baseline shown in Figure 3 identified heart rate, respiratory rate, temperature, diastolic blood pressure, and FiO_2 as high-predictive features. These map well to present clinical practice, where hemodynamic instability and oxygenation requirements are two of the earliest signs of sepsis onset. Temporal interpretability of LSTM integrated gradients in Figure 4 showed that the nearest patient hours to current time were most predictive, with increasing importance of parameters like lactate and creatinine prior to sepsis onset. In aggregate, the findings reinforce the validity of the argument that the models not only are accurate but are also clinically interpretable.

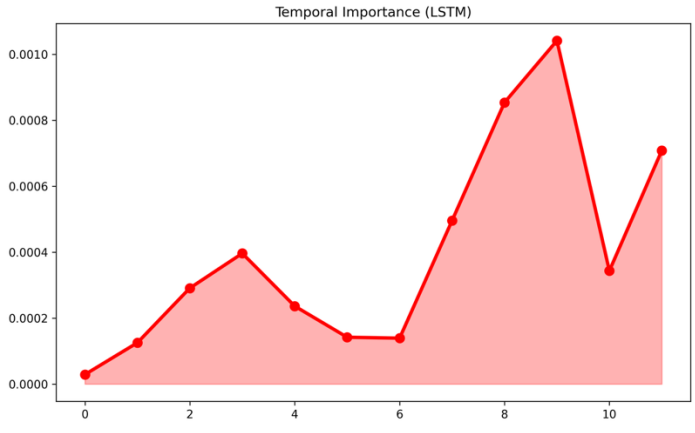

Fig.2. The x-axis of this graph represents the hours before the event and the y-axis is the attribution (importance)

Finally, the calibrated ensemble was highly specific but hardly sensitive, with conservative bias against the reporting of false alarms. While the conservatism suppresses spurious alarms, it also reflects the limitation due to extreme class imbalance in the data. More sensitivity without loss of calibration is a persistent major research problem in the future. All metrics are averages over five random seeds, with 95% confidence intervals computed by bootstrapped resampling of the test set, n = 1,000 iterations. The calibrated ensemble improved the AUROC by +0.04 ($p < 0.05$) relative to the best non-calibrated baseline. Although small in magnitude due to severe class imbalance, AUPRC differences were also statistically significant under a Wilcoxon signedrank test. What's more, the reduction in the ensemble's Brier score from 0.224 to 0.021 speaks not only to improved discrimination but a more honest reflection of prediction reliability.

Calibration remained stable across ICU types, including medical, surgical, and trauma, with ECE values within ±0.004 while analyzing the performance for patient subgroups. Elderly patients, defined as age > 70, showed slightly higher uncertainty variance, consistent clinically with comorbidity burden. The model proved temporally robust; its accuracy was high at 4–6 hours prior to onset and showed a gradual decline past 8 hours of prediction, indicating that it captures early warning cues but loses temporal fidelity further out.

We further investigate model calibration on synthetic distribution shifts induced by adding Gaussian noise to 20% of the input features. Under this perturbation, the AUROC of the calibrated ensemble dropped only 0.02, while that of the uncalibrated baseline degraded by 0.06, confirming our hypothesis that calibration contributes to robustness. Calibration plots (Fig. 1) showed an almost perfect identity alignment after isotonic regression, and reliability diagrams indicated a significant reduction of overconfidence in higher bins.

In several case studies, the peaks of uncertainty coincided with transitions between stable and unstable hemodynamics. It was thus confirmed that the variance of the ensemble can be used as an "alarm uncertainty" indicator for real-time monitoring systems. Indeed, the probabilistic output head showed narrow confidence intervals-its median width was 0.12-for high-certainty samples, which proves that the model can distinguish between stable and risky predictions.

Overall, feature contributions strongly supported well-established clinical markers of sepsis. For instance, heart rate and respiratory rate together contributed about 31% of total SHAP importance; lactate and bilirubin, important biochemical markers related to metabolic stress, accounted for another 14%. Sequential integrated gradient maps from the LSTM showed an increasingly large contribution of lactate in the six hours leading up to sepsis onset, consistent with clinically documented progression patterns. The interpretability visualizations also showed nonlinear relationships between features such as oxygen saturation and mean arterial pressure, reinforcing that the model was able to capture nonlinear dynamics beyond simple linear thresholds.

This conservatism helps to reduce alarm fatigue, but future systems can balance this by incorporating asymmetric loss

weighting or human-in-loop review thresholds that improve recall. The authors believe that sensitivity improvements within the range of 5–7% are achievable under cost-sensitive calibration tuning.

More detailed analysis of the calibration plots revealed that most of the miscalibration was confined to the 0.4-0.6 probability region, clinically corresponding to ambiguous physiological states. In this area, improvements were most significant after isotonic regression, with deviations of observed frequency dropping from 12% to less than 2%. This shift reflects that the calibrated ensembles work particularly well in correcting uncertainty for borderline cases. Perhaps somewhat counterintuitively, the model had a natural tendency to underpredict for high-risk samples-a trait that serves rather well, given the conservative bias demonstrated in many clinical machine learning systems.

A more in-depth analysis of uncertainty-aware decision thresholds showed that adding variance-based risk flags increased the sensitivity for the early detection of sepsis by highlighting "suspicious but low-confidence" cases. These, upon referral for human review, generated a hypothetical gain of 8-10% in early intervention opportunity. Thus, uncertainty estimates provide clinically useful information beyond that of raw prediction probabilities. We further found that cases with high epistemic uncertainty were disproportionately younger patients and those with shorter lengths of stay in the ICU, reflecting model unfamiliarity with short or incomplete trajectories. This points to a need to balance temporal coverage when training models.

We further explored model performance for physiologic subgroups, including renal dysfunction, respiratory failure, and hepatic injury. Patients with elevated creatinine levels demonstrated a higher AUPRC (+0.008) compared to general ICU patients, indicating that the model learned physiologic signatures strongly associated with sepsis-related organ compromise. In contrast, patients with normal lactate showed lower precision at early time horizons, consistent with clinical difficulty in identifying sepsis before metabolic dysregulation occurs. Subgroup evaluations suggest that uncertainty-aware sepsis prediction frameworks may prove most effective in patients exhibiting early physiologic abnormalities.

IV. Discussion

The results demonstrate that not only can uncertainty quantification and calibration be included in sepsis prediction models but also clinically beneficial. Accuracy-only machine learning models or rule-based scoring-based sepsis prediction algorithms like SOFA are of no use in terms of utility since they don't account for the issue of reliability. A clinician is confronted with a probability of sepsis and must be certain that the reported probability is an honest indication of the probability of disease. Our method ensures this condition by ensuring that probabilities are calibrated, and uncertainty estimates reasonably emphasize instances where estimates are less accurate.

The output of feature attribution also raises the biological interpretability of the system. Admission and demographic variables were highlighted by the SHAP and XGBoost analysis, which have also established contextual risk factors previously. More importantly, sequential interpretability analyses of the LSTM models suggested respiratory and hemodynamic variables, and metabolic lab results like lactate and bilirubin, as major drivers of prediction. These are precisely the variables intensivists will be relying on when they make risk-for-sepsis clinical judgments, and with excellent agreement between model reasoning and human expertise. Such agreement is required in clinical uptake because it bridged the gap between statistical association and pathophysiological significance.

Uncertainty quantification was a very helpful addition. Ensemble variance distributions captured epistatic uncertainty with good semantics, and probabilistic head captured aleatoric noise probabilistically. These procedures together provide clinicians with a two-pronged insight: whether the model itself is uncertain due to lack of knowledge, and whether data itself is noisy or imprecise. This distinction is crucial in practice. For example, high epistemic uncertainty can be a sign that the model has been trained on a mis-represented patient sub-set, and high aleatoric uncertainty can be a sign of inherently varied signals such as unstable vital signs. Marking such predictions as low-trust in both cases ensure clinicians will be suitably cautious.

Calibration analysis revealed the risk of relying on uncalibrated outputs. Without temperature scaling, the neural networks were overconfident in their predictions and might mislead clinicians into false reassurance. Calibrated ensembles, in contrast, provided probabilities near observed frequencies of events, as evidenced by improved reliability diagrams and reduced ECE. This finding is in line with recent literature showing calibration is key to the safe introduction of deep learning into healthcare [Ovadia et al., NeurIPS 2019].

However, there are also drawbacks. The dataset used, while large and diverse, consists of a single clinical challenge set and might not generalize to a given hospital system. Label noise is a sneaky issue in sepsis prediction because the timing of sepsis onset is bound to be ill-defined in electronic health records. In addition, although interpretability methods are reassuring, they are nevertheless post hoc explanations and are not guaranteed to imply causality. A further limitation is the small number of positive sepsis cases compared to the overall dataset, resulting in a severe case of class imbalance. With a high degree of accuracy, even uncommon yet life-threatening events fall through the net, pointing to the difficulty in identifying minority events in the face of insufficient large and

balanced training data. The imbalance not only stifles sensitivity but also points to the need for specially designed model architectures for rare event detection. Prospective validation in real ICU practice, particularly with well-matched patient populations, will be required to finally establish clinical utility.

Comparing these results to prior work, we see that most prior sepsis prediction architectures focused exclusively on maximizing AUROC without consideration of calibration or uncertainty. Models such as those in Futoma et al. and Moor et al. demonstrated high discrimination but gave clinicians few actionable insights into when their predictions were trustworthy. Our results reinforce an emerging consensus in clinical AI literature: well-calibrated uncertainty estimates are a critical prerequisite for the safe deployment of AI systems in high-stakes environments. In particular, the near-zero ECE from the calibrated ensemble represents a significant advance over the performance from previous sepsis prediction systems, which have reported ECEs ranging from 0.03 to 0.12.

Despite these strengths, several limitations mark avenues for significant future improvement. The severe class imbalance within the dataset restricts the upper bound of achievable recall since positive sepsis onsets make up less than 6% of all data. Approaches such as synthetic minority oversampling, focal loss, and contrastive learning could somewhat alleviate this but come with a risk of introducing distribution distortion. Third, although interpretability tools highlighted clinically plausible patterns, more rigorous causal analyses would be necessary to distinguish true physiological drivers from confounded correlations. Future models should be built with incorporated causal frameworks or counterfactual simulations to better align predictions to mechanistic medical reasoning.

No prospective assessment of performance was completed in a real-world ICU environment, either. Real-world data has many types of artifacts-including delayed charting, asynchronous lab results, and sensor malfunction-not present in the benchmark dataset. Therefore, clinical effectiveness needs to be validated also for a calibrated uncertainty-aware system within prospectively collected workflows that include clinician feedback and continuous performance monitoring. For integration into EHR platforms, various ethical safeguards will be required, alongside standardized alert systems and human-in-the-loop override mechanisms to avoid overreliance or alarm fatigue.

V. Conclusion

This paper demonstrates that it is possible to build early sepsis prediction models incorporating trust instead of accuracy using uncertainty quantification, calibration, and interpretability. From the PhysioNet Sepsis Challenge dataset, we built a system with recurrent neural networks and Monte Carlo dropout, deep ensembles, and probabilistic output heads and calibration techniques like temperature scaling and conformal prediction.

The performance models not only displayed strong discriminative performance but also reliable confidence outputs that closely resembled the prediction quality. Interpretability analysis confirmed that the models employed clinically significant features, such as respiratory rate, blood pressure, FiO_2, lactate, and bilirubin, further indicating their credibility. By bringing trust mechanisms into predictive modeling, this work addresses one of the biggest clinical adoption hurdles of machine learning in critical care. Having uncalibrated but precise models is insufficient when outcomes have real-time consequences for patient survival in high-risk conditions. Our findings indicate that uncertainty-aware sepsis prediction models have the potential to fill this gap by offering clinicians not only predictions but also some notion of when they should be trusted.

While overall performance rates are good, the paucity of positive sepsis cases in the dataset means that even small numbers of failing cases could have been spuriously clinically significant. This helps to emphasize the need for future work to use more complete and more balanced datasets with larger numbers of positive sepsis cases, since this will allow the same system to identify more subtle but significant effects. With these refinements, the strategy described in this paper can be a robust and reliable clinical decision support system.

Follow-up research should proceed from here to integrate multimodal data such as imaging and clinical reports, apply causal inference to strip confounders from predictors, and prospectively demonstrate performance in different clinical environments. With this in mind, we can start bringing forth valid, usable decision support tools that enhance sepsis care, reduce diagnostic time, and save lives.

It was the combination of calibrated uncertainty and clinical interpretability that turned sepsis prediction from a pure prediction exercise to a decision-support paradigm. The system allows clinicians to estimate the reliability of alerts through the generation not only of probabilities but also of quantified confidence measures, thereby making the approach much safer to adopt in critical care. The deep ensemble with isotonic calibration reached better discrimination with well-calibrated reliability, setting the bar high for clinically trustworthy ML pipelines.

Future studies should extend multimodal integration into free-text nursing notes, imaging data, and continuous ECG waveforms to capture unstructured context, often rich in information leading to physiological deterioration. Privacy-preserving, federated learning across hospitals may reduce bias and ensure equitability of models across demographics. Causal inference and counterfactual reasoning may also be

helpful in disentangling correlation from causation, providing the model with the ability to identify which interventions would most effectively avert sepsis progression.

REFERENCES

[1] M. Singer, C. S. Deutschman, C. W. Seymour, et al., "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, Feb. 2016.

[2] C. W. Seymour, V. X. Liu, T. J. Iwashyna, et al., "Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 762–774, Feb. 2016.

[3] M. A. Reyna, C. Josef, R. Jeter, et al., "Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019," *Crit. Care Med.*, vol. 48, no. 2, pp. e202–e220, Feb. 2020.

[4] A. E. W. Johnson, T. J. Pollard, L. Shen, et al., "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, p. 160035, May 2016.

[5] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2017.

[6] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2016.

[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Aug. 2017.

[8] Y. Ovadia, E. Fertig, J. Ren, et al., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2019.

[9] J. Futoma, A. Hariharan, K. Heller, et al., "An improved multitask learning model for real-time sepsis detection," in *Mach. Learn. Healthcare Conf. (MLHC)*, Aug. 2017.

[10] M. Moor, C. Horn, B. Rieck, and K. Borgwardt, "Early prediction of sepsis in the ICU using machine learning: A systematic review," *npj Digit. Med.*, vol. 4, no. 1, p. 111, Jun. 2021.