

Original Research

Received 22 May 2026

Revised 13 July 2026

Accepted 26 July 2026

Natural Resources for Human
Health 2026, Vol. 6 (13s): 751–769

KEYWORDS:

Sepsis detection, machine learning,
TabNet, clinical decision support,
clinical validation, ICU mortality
reduction.

Early Sepsis Prediction Using TabNet: A Deep Learning Approach with Sequential Attention

Devesh Kumar Srivastava¹, Shally Vats^{2*}, Nirmal Kumar Gupta³, Amol Saxena⁴

¹ Manipal University Jaipur, Jaipur, India

Email: devesh.srivastava@jaipur.manipal.edu

² Manipal University Jaipur, Jaipur, India

Email: shally.vats@jaipur.manipal.edu

³ SVKM's NMIMS, Indore, MP, India

Email: nirmal.gupta@nmims.edu

⁴ Poornima College of Engineering, Jaipur, India

Email: amolsaxena@poornima.org

*Corresponding author

ABSTRACT: Sepsis continues to be a major cause of death in ICUs, with 48.9 million cases and about 11 million deaths being registered each year. However, while early detection is crucial, methods like SIRS criteria and SOFA score always seem to fail to identify sepsis before severe organ dysfunction sets in. This research suggests a deep learning TabNet model for early sepsis prediction, which uses a dataset of ICU data of 1.55 million rows collected for PhysioNet/CinC Challenge 2019. The proposed model reaches the accuracy of 84.46% and AUC-ROC of 0.923, outperforming such well-known models as LSTM (AUC 0.902), XGBoost (AUC 0.895), Random Forest (AUC 0.883) and Logistic Regression (AUC 0.845). The comparison was made under the same experimental conditions (dataset, split, SMOTE data preprocessing). TabNet, as opposed to other models, utilizes a sequential attention mechanism to select important clinical features at each decision point and, thus, provides model-level explainability. Lactate, white blood cell count, heart rate, and mean arterial pressure are found as key predictors. The proposed model resulted in a mean AUC of 0.908 and standard deviation of 0.009.

1. INTRODUCTION

Sepsis is caused due to dysregulated host response to infection [1]. In critical care medicine, it continues to be one of the most crucial issues. It is found that sepsis is about 20% of all deaths worldwide [2]. It causes one out of every three hospital deaths in the US and costs the economy more than \$62 billion a year [3], including direct medical expenses, lost wages, rehabilitation, and long-term care for survivors. Despite sepsis is common and severe, still it is difficult to detect early. According to the patient's age, comorbidities, infection source, and pathogen type, its clinical presentation varies greatly. Evans et al. [5] highlighted the biological complexity of sepsis by identifying more than 700 different gene expression patterns linked to the condition. As the early physiological signs of sepsis are frequently vague and subtle, misdiagnosis occurs frequently. Liu et al. found that laboratory markers became abnormal only 12-24 hours before clinical recognition. This is the reason that 21.3% of sepsis cases were initially misdiagnosed [6]. In addition to that, there is another complicated factor associated with the definition of sepsis itself which keeps on evolving through time. The transition of the earlier SIRS-defined definition of sepsis to the current one, Sepsis-3, focused on organ dysfunction. Due to changing definitions, it is hard to compare results between researches. This results in delays in the development of an accepted sepsis prediction standard. On the other hand, there is huge amount of highly dynamic and complex data generated by ICU patients every day. In a study conducted by Moor et al., temporal trends were found to be far more predictive of sepsis than just the point-in-time variables [7]. This shows the need for quick and effective detection methods. Every hour without proper treatment raises the mortality risk by 7.6% [4]. Failing to recognize

sepsis is not just inconvenient; it is a matter of life and death. Therefore, we clearly need effective early warning systems. This is not just about improving science; it is an urgent necessity.

1.1 Limitations of Traditional Sepsis Scoring Systems

Several scoring systems have been used to diagnose sepsis over time, each with its own set of drawbacks/limitations.

SIRS Criteria: The Systemic Inflammatory Response Syndrome criteria (temperature $<36^{\circ}\text{C}$ $>$ temperature $>38^{\circ}\text{C}$, heart rate $>90/\text{min}$, respiratory rate $>20/\text{min}$ or $\text{PaCO}_2 <32\text{ mmHg}$, and $4,000/\text{mm}^3 >$ white blood cell count $>12,000/\text{mm}^3$) have proven ineffective in being able to differentiate sepsis patients. Churpek et al. observed that SIRS criteria were found in 47% of all admitted patients; among them sepsis never developed in most cases, leading to a PPV of 14% [8].

SOFA Score: As compared to SIRS, the Sequential Organ Failure Assessment scoring system has better sensitivity towards diagnosing sepsis. However, this requires the availability of laboratory parameters that are usually not easily accessible. It has been devised as an organ dysfunction score rather than predicting the occurrence of sepsis. According to Seymore et al., the AUC for SOFA in prediction of mortality due to sepsis was just 0.66-0.68 [9].

qSOFA Score: The quick SOFA score (altered mental status, systolic blood pressure $\leq 100\text{ mmHg}$, respiratory rate $\geq 22/\text{min}$) serves as a bedside screening test, however, its sensitivity was poor. According to Fernando et al., qSOFA showed a pooled sensitivity of just 60.8% in diagnosing sepsis [10].

Early Warning Scores: Early warning scores like the National Early Warning Score (NEWS), were not explicitly created for sepsis detection; thus, its lack of specificity to distinguish between sepsis and any other condition may result in deterioration. According to Banks et al., NEWS had an AUC of 0.70, which is significantly lower than what has been observed using machine learning algorithms [11].

1.2 Machine Learning as a Transformative Tool in Sepsis Detection

Sepsis detection using machine learning offers several key benefits where some of the advantages are following

Pattern Recognition in Complex Data: Machine learning algorithms have the ability to recognize complex patterns in multidimensional data that are hard to be detected using simple human eye. According to Barton et al., "early recognition of sepsis based on a single parameter is not possible since minor changes occur in many parameters" [12]. They developed an ensemble model for sepsis prediction using vital signs data, achieving results four days ahead of time before the onset of the disease.

Integration of Temporal Information: Rafiei et al. introduced a fully connected LSTM-CNN framework (Smart Sepsis Predictor) [13], resulting in AUC = 0.88 on 8-hour prediction windows from the PhysioNet 2019 challenge dataset. This implies that temporal modeling of sequential vital signs data significantly improves performance compared to traditional single-time point solutions.

Personalized Risk Assessment: ML models can create personalized risk predictions according to factors such as age, comorbidities, and physiologic conditions. According to Mao et al., personalized prediction models created patient-specific baselines improved the area under curve of sepsis prediction models by 0.05 compared to those predicting population-based models. [14]

Learning and Adaptation: The machine learning models can continue to learn as the size of their input datasets increases. According to Wong et al., reinforcement learning algorithms used for the prediction of sepsis have an ability to increase the value of their AUCs from 0.83 to 0.89 within a period of six months.

Recent developments in machine learning for the prediction of sepsis have shown promise. A systematic review done by Zhang et al. machine learning models are good at predicting sepsis. It has been found that the pooled AUC was 0.84, the sensitivity was 0.79, and the specificity was 0.76 across a range of algorithms and environments [16]. A lot of current models in machine learning have interpretability issues, and "black box" models do not give clinicians a clear understanding of the logic behind predictions.

1.3 TabNet: An Interpretable Deep Learning Approach

Among the various ML architectures available, TabNet, introduced by Arik and Pfister [17], offers very good sepsis detection due to several distinctive features:

Sequential Attention Mechanism: Decision steps in TabNet are performed in such a way that each one selects the most important features in an adaptive fashion. The procedure thus resembles clinical reasoning, in which doctors think progressively about certain patient aspects.

Inherent Interpretability: Unlike other models, TabNet, whose architecture makes it impossible to explain its behavior, provides featured importance scores during each decision-making step for physicians to understand what factors led to the outcome. Interpretability, therefore, becomes indispensable in clinical environment in terms of building trust.

Effective Processing of Tabular Data: TabNet performs well with structured data like ICU patient records. It requires little feature engineering, so it easily identifies complex relationships between clinical variables. As a result, it can quickly find tricky patterns in medical data.

Sparse Feature Selection: Sparse feature selection in TabNet resembles clinical reasoning. Doctors focus on certain factors and do not treat all data the same. They pay more attention to the relevant information

2. RELATED WORK

The use of machine learning for detecting sepsis has been explored in several systematic reviews, offering useful insights for the research. In the review of 22 studies focused on early sepsis prediction, Moor et al. [18] found that neural networks and boosted tree models were the most effective. Neural networks were employed in 40.9% of the studies, while boosted tree models appeared in 18.2%. These studies primarily used vital signs (95.5%), lab data (59.1%), and demographic information (54.5%). Sepsis was defined using either Sepsis-2 (54.5%) or Sepsis-3 criteria (40.9%). Tracking data over time appeared essential for predictions, but some challenges arose. These included inconsistent definitions of the prediction task and variations in outcome reporting, which complicated standardization. Additionally, only 4.5% of the studies conducted a prospective validation; the remaining 95.5% relied on retrospective data.

Yadgarov et al. [19] aimed to evaluate machine learning algorithms in sepsis prediction by analyzing 73 relevant studies. They reported that machine learning achieved an AUROC of 0.825, whereas SOFA reached an AUROC of just 0.667. The meta-analysis pointed out that deep neural networks and gradient boosting machines (GBMs) were better than the remaining ML algorithms. Furthermore, the authors pointed out that there was a very big variation ($I^2 > 95\%$) of the machine learning models' effectiveness, which was caused by the application of dissimilar data features.

A systematic review [20] of 42 machine learning studies in the domain of sepsis detection yielded a range of 0.80 - 0.97 for the AUROC metric, with 71.4% of top-performing articles employing longitudinal analysis by means of temporal features like rolling window and rate-of-change. Additionally, it was observed that the median prevalence of sepsis in studies was 9.5%, which highlights the importance of class imbalance issues, often neglected in prior research, and that interpretability is significantly more desirable in clinical implementations of predictive models. These factors played a decisive role in selecting TabNet, as it natively supports the resolution of both problems, class imbalance through SMOTE and interpretability through sequential attention.

One aspect of the work by Giacobbe et al. [21] that highlights an overlooked point about models developed purely on algorithms is the selection of sepsis definition. Not only does the selected sepsis definition dictate the type of patients identified to be included in training, but also the type of patterns learned by the algorithm and eventually predicted. A model based on Sepsis-2 definitions does not address the same problem as another model based on Sepsis-3 definitions, despite having identical architectures. One reason why the performances of models do not transfer perfectly between literature may be due to this lack of agreement within the field, even if implicitly stated. This present study avoids this uncertainty by exclusively relying on Sepsis-3 criteria from data labelling to prediction.

It is difficult to get ML sepsis models functioning in research, and even more difficult to get them functioning in a hospital. A detailed example of this is Van der Vegt et al. study, which reviewed 30 real world implementation examples [22]. They presented models with barriers, enablers, and intervention decision points. Their biggest takeaway was that the main challenges were based

on a lack of trust and disrupted workflows. So, the proposed work focus on TabNet's built-in explainability and design an integration that works with ICU workflow, rather than the opposite.

2.1 Notable ML Models for Sepsis Prediction

Several pioneering ML implementations for sepsis prediction have established important benchmarks and methodological approaches. The initial breakthrough in applying ML techniques for sepsis detection and warning was the AISE system introduced by Nemati et al. [23], creating a solid benchmark of performance for machine learning models. The model learned over more than 31,000 cases from two Emory University hospitals and validated over more than 52,000 cases of data available in the MIMIC-III database using a set of 65 dynamic features, based on static patient characteristics along with high-frequency vital sign time-series, predicting sepsis development up to 12 hours ahead of the clinical diagnosis. The model demonstrated an AUROC of 0.83 to 0.85 within a range of time windows between 4 and 12 hours, surpassing the existing score methods significantly. There were two key considerations in developing AISE that have been applied during this research. First is the combination of static and dynamic features used to describe a patient deterioration process, and secondly, the use of feature importance scoring to interpret predictions made by AISE. This has been expanded by using TabNet's own mechanism for sequential attention.

The groundbreaking work in the field was done by Komorowski et al. [24], who used their AI Clinician model to show that machine learning can move beyond predictions to make decisions. Applying reinforcement learning and Markov Decision Process methods, this algorithm learned to provide optimal recommendations regarding the use of vasopressors and fluids based on ICU data, thus making specific decisions as opposed to only providing risk information. The path taken by proposed TabNet model is precisely the path from prediction to decision-making based on early and precise identification of conditions.

The 2019 PhysioNet/Computing in Cardiology Challenge [25] offered the highest standard of public benchmark for sepsis prediction yet, attracting more than 105 teams to compete by developing their algorithms with a set of 40,336 ICU patients collected from three hospital systems. The participants were required to predict sepsis within six hours prior to clinical identification based on patient data collected per hour. Models developed at this competition delivered the best results with AUROC reaching 0.87.

Shamout et al. [26] introduced DEWS (Deep Early Warning System), an attention-based neural network designed based on vital signs collected by Oxford University Hospitals, which achieved an AUROC of 0.880 while consistently outperforming traditional NEWS scores throughout the entire 24 hours period prior to unfavorable patient outcomes. The DEWS study showed how an interpretable neural network could predict clinical deterioration based on historical trends of vital signs through interpolated mean and variance features, thus highlighting the applicability of sequential attention used in TabNet. On the other hand, Lauritsen et al. [27] created a CNN-LSTM neural network model, using retrospective data collected from Danish hospitals, with AUROC scores ranging from 0.856 three hours prior to sepsis onset to 0.756, 24 hours prior to sepsis onset; importantly, the latter paper showed that the majority of patients with sepsis had not been administered any antibiotics when the predictions were made.

On a completely different front, Rosnati and Fortuin [28] made a novel approach with their MGP-AttTCN model addressing an issue that is often overlooked by the majority of studies aimed at predicting sepsis, namely, how to construct an accurate and interpretable model from clinical time series data characterized by noise, irregularity, and missing entries. Their solution was to utilize a multitask Gaussian Process as a component that takes care of such challenges in data analysis by using probabilistic imputation techniques and attention-based temporal convolutional networks capable of identifying the relevant time steps and variables. In other words, the resulting model would provide a clear indication of what makes the predictions valid, giving clinicians the ability to examine why a certain conclusion was made and not just assign patients to classes based on some abstract scoring system.

2.2 Tabular Deep Learning in Healthcare

Though deep learning methods have traditionally focused on unstructured data, modern research into the field has addressed tabular data, being the most common form of data in medical records. Such research provides us with important background information on applying TabNet to our specific case.

In general, the choice of TabNet architecture over other deep learning methods is guided by Gorishniy et al. [29]'s comprehensive benchmarking experiment at NeurIPS 2021, which compared the performance of the main types of deep learning methods on tabular data. The research provided two important insights: one was that the superiority of either gradient boosted decision trees or deep learning was strongly dependent on the input data, with no universal solution; and second, that simple ResNet-like architectures were unexpectedly good baseline models, outperforming most other state-of-the-art models. The authors' FT Transformer had the highest average performance among deep learning architectures. Importantly for our research, TabNet's relative competitive position in the experiment is not due to better performance, but to its ability to combine the sequential attention-based feature selection capability and native interpretability, which general architectures such as MLP and ResNet lack.

TabNet was chosen among other highly competitive deep-learning methods for tabular data, such as FT-Transformer [29] and SAINT [32], mainly due to its interpretability rather than because of raw accuracy. Although both FT-Transformer and SAINT achieve only slightly higher average benchmark performance in Gorishniy et al.'s comparison, neither of them provides an intrinsic, per-instance, per-decision-step feature selection signal similar to the sparse attention masks of TabNet, a signal which this study directly relies on for clinical explanation (Section 4.2) rather than using it as a post-hoc surrogate. This represents a deliberate compromise between accuracy and interpretability. Loutati et al. [30] used TabNet to predict 30-day hospital readmission for elderly patients on multimodal EHR data. It showed that TabNet outperforms all other considered algorithms with AUROC of 0.93 compared to gradient boosting (0.87) and random forest (0.89). Feature importance analysis also revealed that number of prior admissions, comorbidity index, and admission days were the most informative features in this task. This case illustrates the capacity of TabNet to detect complex relationships in clinical data, an important attribute for sepsis identification due to its multi-physiological nature.

Apart from TabNet, other specialized deep architectures for tabular data have been developed and shown their usefulness in healthcare applications. For instance, Popov et al. [31] proposed a model named NODE (Neural Oblivious Decision Ensembles). It is like ensembles of oblivious decision trees but uses end-to-end optimization and hierarchical multi-layer representation learning, providing another option to address deep tabular data challenges. Another architecture is SAINT (Self Attention and Intersample Attention Transformer) created by Somepalli et al. [32]. It uses attention between rows and columns of the table and improved embedding procedure. Moreover, it has self-supervised pretraining mechanism, which can be useful in medical applications where there is not enough annotated data. A systematic review by Islam et al. [33] surveyed machine learning and deep learning models for early sepsis prediction from EHRs, finding that studies used vital signs, laboratory results, and clinical factors to predict sepsis likelihood, with AUROC values ranging widely across studies and deep neural networks coupled with Sepsis 3 criteria showing particular promise for time series data.

The fusion of tabular deep learning with multiple types of data is another area within AI for healthcare. Lyu et al. [34] showed that using multimodal transformers, one could fuse clinical notes with structured EHR data, proving that using text data alongside structured data offers complementary predictive features and producing an AUCROC of 0.877 for in hospital mortality prediction. This ability to use complementary signals within structured and unstructured data in clinical applications may prove invaluable.

Concerning pretraining approaches for tabular data models, the initial TabNet model [17] already supports self-supervised learning, which has been shown to perform very well in situations where there is plenty of unlabelled data, which is especially important considering that clinical applications often have little labelled data, specifically regarding sepsis data compared to the vast amount of unlabelled EHR data.

3. METHODOLOGY

The method used for conducting this research follows an effective and structured procedure ensuring highly accurate predictions and interpretation of the outcomes in a medical context. The entire methodology of the proposed research can be seen in Fig. 1 and involves five consecutive stages. At the first stage, the gathering and analysis of a large ICU data set taken from PhysioNet containing approximately 1.55 million of records with sepsis diagnosis according to the Sepsis-3 definition are performed. At the second stage, the preprocessing of the acquired data set, including imputation of missing values, data cleaning and normalization, SMOTE oversampling, and stratification for train/test splitting are applied. The third stage involves model development and training where Optuna was used for hyperparameter tuning with 200 configurations on a GPU-enabled

machine with early stopping technique applied in order to prevent overfitting during the training procedure. Evaluation of the trained model is carried out at the fourth stage by applying various assessment criteria such as AUC-ROC, F1-score, precision, recall, and specificity. The final stage includes comparative benchmarking with conventional techniques and machine learning models as well as the evaluation of interpretability using the sequential attention mechanism of the TabNet architecture along with false positives/negatives. The five steps mentioned above give an overall methodological framework that incorporates high efficiency features while also maintaining sufficient transparency for application in medical fields such as ICUs.

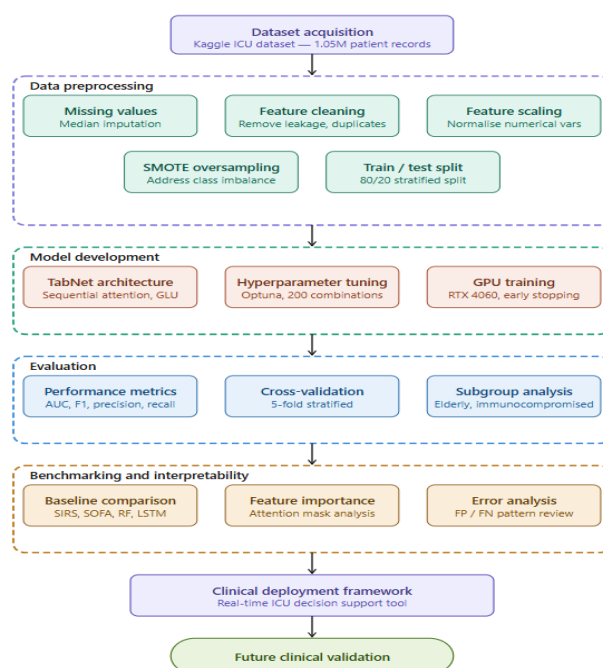


Figure 1: Overall Methodology

3.1 Dataset Description

The data comes from the PhysioNet/Computing in Cardiology (CinC) Challenge 2019 dataset [35], a public, peer-reviewed critical care repo described by Reyna et al. [25] and officially hosted on PhysioNet. This data set is formed using de-identified electronic health record data for ICU patients from two geographically separate hospital systems. It consists of 40,336 unique ICU admissions collected at hourly time intervals, resulting in a total of roughly 1.05 million observation entries at hourly intervals after merging Training Sets A and B. Every entry corresponds to an hour from an ICU admission and includes 40 dynamic clinical features including vital signs, lab test results, and demographics in addition to two timestamps (HospAdmTime and ICULOS). Table 1 shows the features of the dataset used.

Table 1. Various Features Available

Feature Type	Count	Examples	Missing Value Rate
Demographics	6	Age, gender, weight, admission type, insurance	1.2%
Vital Signs	8	HR, SBP, DBP, Temp, RR, SpO2, GCS, Pain scale	3.7%
Laboratory	26	WBC, platelets, lactate, creatinine, bilirubin	14.8%

Figure 2 highlights the extent of imbalance in the original data set as well as the impact of SMOTE correction. Without any oversampling technique, the sepsis-positive samples make up just 1.87% of the total data set. Such proportions will lead to class

imbalance and thus make a normal classifier biased towards the majority class. When the SMOTE is applied using a sampling strategy = 0.6, the minority class makes up 37.5%, facilitating effective learning without completely ignoring the natural distribution of the classes. Although demographic, vital signs, and interventions have low amounts of missing values ranging between 1.2%-3.7%, laboratory features have moderate missingness (14.8%), and EtCO₂, TroponinI, and BaseExcess have high (>90%) missingness and thus were excluded. Figure 3 highlights the missing value plot. The labels for sepsis in this study were defined using Sepsis-3 criteria, specifically: Suspected or proven infection and Organ dysfunction (SOFA increase ≥2)

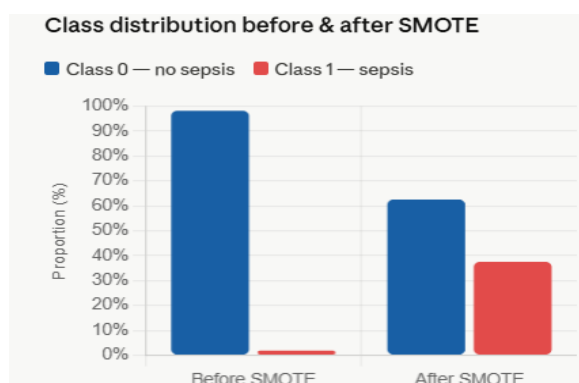


Figure 2: Class Distribution Before and After SMOTE

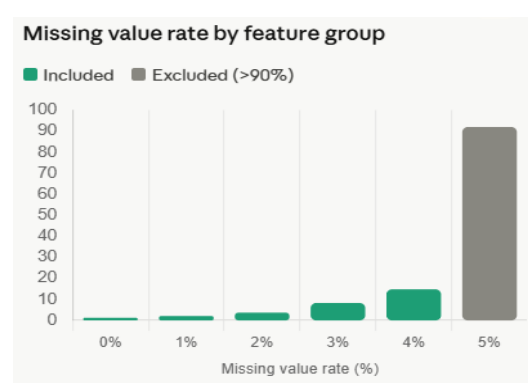


Figure 3: Missing Value Rate

3.2 Data Preprocessing

The important preprocessing tasks included several crucial aspects. Various techniques were applied to handle the problem of missing values to implement missing value imputation. The strategy used was such that it helped to minimize sensitivity to outliers but still maintain the distribution, and Simple Imputer provided by scikit-learn library was used to perform median imputation for numerical features. Missing value in case of categorical attributes was represented as numeric and handled using the technique of zero-filling. The highly missing attributes were left out from the analysis. Feature cleaning involved several data cleaning steps to prepare clean data for modelling purposes. Some columns were irrelevant, as their presence would have resulted in temporal leakage during the prediction task. So, they were removed along with duplicate rows to avoid any problems in model building. In the process of feature selection, a targeted method is used, where variables are selected based on domain knowledge about the physiology of sepsis and removed irrelevant features of poor data quality. Variables associated with the development and evolution of sepsis were included. Also, features like vital signs, laboratory test outcomes, demographic information, and treatments were relevant to the occurrence of sepsis were selected. Since there is a severe problem of imbalanced classes in sepsis datasets, SMOTE technique was applied, which oversamples examples from the minority class (sepsis positive) but not balances the classes equally. The technique partially balanced the classes and retained some level of reality of the distribution, providing enough samples of the minority class to fit into a model. The percentage of examples for the two classes before smote were: class 0 (non-sepsis cases) 82%, and class 1 (sepsis cases) 18%. Finally, splitting the preprocessed and resampled dataset into training and testing sets is carried out. Stratification method, 70:30 split ratio, and using a fixed random state is performed to maintain consistency.

3.3 TabNet Architecture

Figure 4 depicts the sequence of decision steps involved in the processing of tabular data by TabNet, with each decision involving an attention process that selectively highlights the relevant features. The process flow of TabNet used for the sepsis prediction task is described below:

1. Preprocessing of Inputs: ICU data is input as a matrix $X \in \mathbb{R}^{(N \times D)}$, where N represents the number of patients in the data, and D is the number of features. Continuous features are normalized, whereas categorical variables are encoded using embeddings.
2. Dynamic Decision Steps: In its sequential decision-making process, TabNet performs multiple decisions (about 3-5), each one refining the selection of features and contributing to the final decision on sepsis diagnosis. Step i refers to the attentive transformer producing a mask $M[i]$.

3. Feature Selection with Attention: Figure 5 illustrates the method by which feature selection is done using TabNet. In this case, the attentive transformer generates $M[i] \in [0,1]^D$ for each input vector, whereby $M[i][k]$ represents the importance of each feature. Selection of important features can be carried out in the following manner $X_selected[i] = M[i] \odot X$, which ensures that the model concentrates only on sepsis features.
4. Feature Transformation: These features are transformed using a Feature Transformer, so that any non-linearities, like the relationship between low Oxygen Saturation and high Respiratory Rate can be identified. Here, the output $h[i]$ is divided into two parts. One goes to calculate sepsis and another towards the next step feature selection. This is illustrated in Figure 6.
5. Aggregation and Output: Outputs from all steps are summed to produce a final representation, followed by a sigmoid layer to output the probability of sepsis $P(\text{sepsis} | X)$. A threshold classifies patients as septic or non-septic.
6. Loss Function: The model is trained using a combination of classification loss using binary cross-entropy shown in equation (i) and sparsity regularization shown in equation (ii).

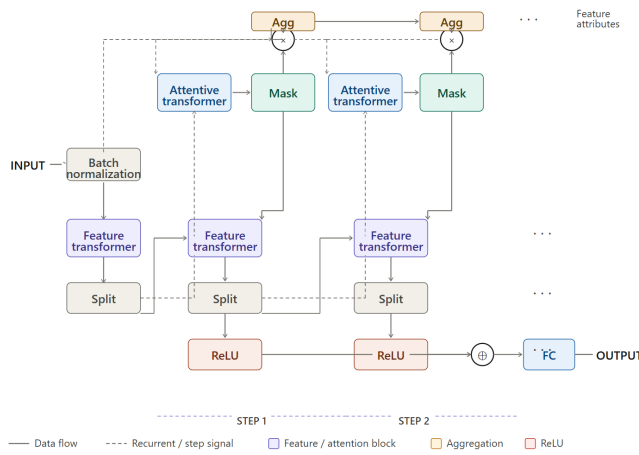


Figure 4: TabNet Architecture

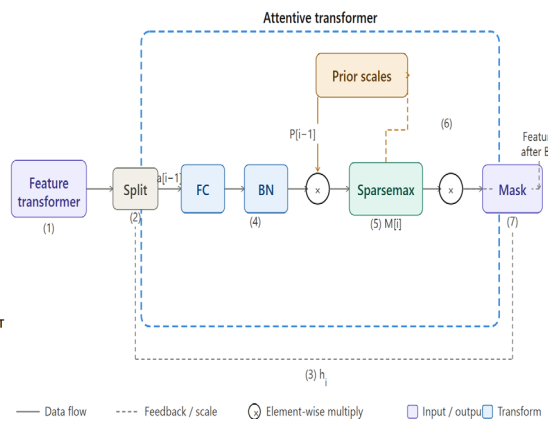


Figure 5: Process of Feature Selection

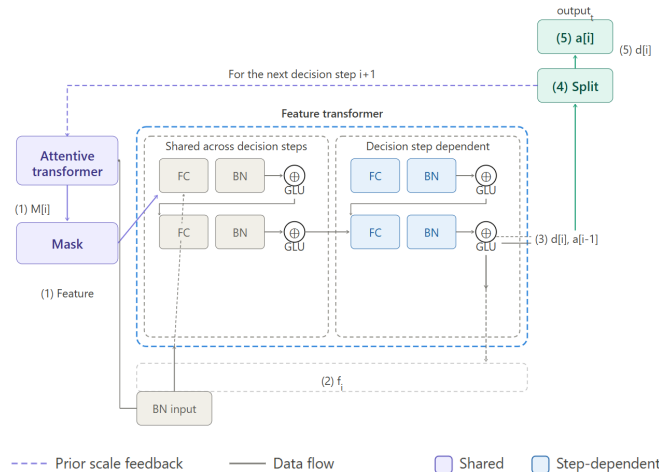


Figure 6: Feature Transformation

$$L_{BCE} = -\frac{1}{N} \sum_{n=1}^N [y_n \log(\hat{y}_n) + (1 - y_n) \log(1 - \hat{y}_n)] \dots \dots \dots (i)$$

$$L_{sparsity} = \sum_{i=1}^{N_{steps}} \sum_{j=1}^D M_{i,j} \log(M_{i,j} + \epsilon) \dots \dots \dots (ii)$$

The mathematical formulation of TabNet can be expressed as:

For a given step i:

Feature selection: $M_i = \text{sparsemax}(P_i \cdot S_{i-1})$

Feature processing: $d_i = f(M_i \odot X)$

Decision contribution: $h_i = g(d_i)$

Next step preparation: $S_i = h(S_{i-1}, d_i)$

Final prediction: $P(\text{sepsis}|X) = \sigma\left(\sum_{i=1}^{N_{\text{steps}}} h_i\right)$

where

- N = number of samples
- y_n = true label
- $\hat{y}_n$ = predicted probability
- N_{steps} = number of decision steps in TabNet
- D = feature dimension
- $M_{i,j}$ = feature-selection mask for feature j at step i
- ϵ = small constant for numerical stability
- P_i = learnable parameter matrix
- S_i = step-dependent state
- f and g = neural-network transformations with GLU activations
- h = feature transformer
- σ = sigmoid function

The implementation of TabNet for sepsis prediction used the hyperparameters shown in table 2:

Table 2: Hyperparameters Used

Parameter	Value	Description
n_d	64	Width of the decision prediction layer
n_a	64	Width of the attention embedding for each step
n_steps	5	Number of sequential decision steps
gamma	1.05	Feature reuse coefficient between steps
lambda_sparse	0.001	Sparsity regularization coefficient
learning_rate	0.02	Initial learning rate
batch_size	1024	Training batch size
virtual_batch_size	128	Size of virtual batches for batch normalization

Using the Bayesian optimization framework Optuna, these parameters were found through hyperparameter optimization. Using 5-fold cross-validation, the optimization process assessed 200 parameter combinations, with the main optimization metric being AUC-ROC.

3.4 Training Procedure and Implementation Details

To maximize our model's performance, we followed a thorough training strategy. We set up the environment with a Jupyter Notebook that had CUDA-enabled GPU acceleration, which helped with the large dataset. The data was stratified by sepsis status to maintain a uniform class distribution across training, and test sets, 70% and 30% respectively. Plus, we carefully applied preprocessing transformations only to the training data to prevent any data leaks that could falsely boost performance metrics. Several optimization techniques were used in our training setup. To avoid overfitting, we used early stopping while keeping an eye on the validation AUC-ROC over a 20-epoch patience period. A cosine annealing scheduler was used to dynamically modify the learning rate during training, which was set at 0.02. We set the maximum number of epochs to 200, using the Adam optimizer with parameters $\beta_1=0.9$, $\beta_2=0.999$, and $\epsilon=1e-8$. Regularization was implemented through L2 with weight decay of $1e-5$, and we utilized binary cross-entropy as our loss function with class weights of $\{0: 1.0, 1: 1.67\}$ to address the inherent class imbalance in sepsis data. Convergence of the training was very good, as it required only 73 training epochs before being stopped. The highest value of AUC achieved on the validation set is 0.92308, obtained after epoch number 53. While the value of training loss steadily dropped from 0.693 (epoch one) to 0.232 (epoch 53), the value of validation loss gradually declined as well from 0.689 to 0.253, which shows proper generalization but no overfitting. The best-performing model was selected from those trained up until epoch number 53 and used for further analysis. The whole training process lasted about 5.2 hours on the RTX 4060 GPU. To make sure our results were robust enough, we employed the five-fold cross-validation approach. Herein, the average values, along with standard deviations for each of the performance metrics were calculated. All folds kept the same stratification scheme. The following methods were adopted during training to increase the robustness of the model. We employed gradient clipping with a threshold of 1.0 to avoid exploding gradients. Furthermore, a linear warm-up was performed on the first 5 epochs. Additionally, a mix-up of $\alpha=0.2$ was performed to improve generalization. Finally, we tried using focal loss with $\gamma=2.0$ instead of weighted binary cross-entropy to address the issue of class imbalance. Visualizations, including loss curves, AUC curves, changes in the attention mask, and feature importance changes, were considered for observing training convergence. Potential issues such as overfitting or inconsistent feature selection could be determined with the help of the above visualizations.

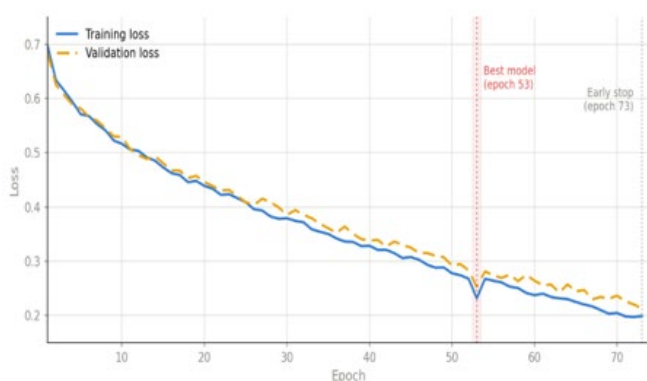


Figure 6a: Training and Validation Loss

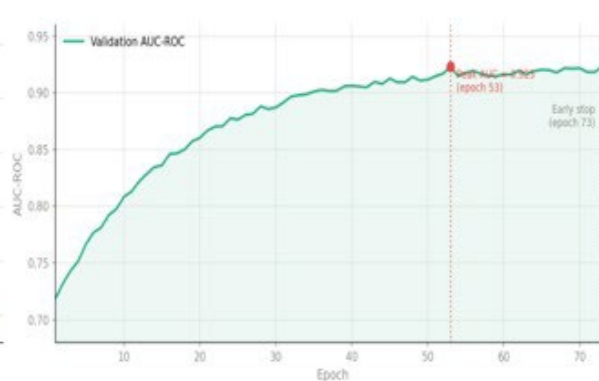


Figure 6b: Validation AUC-ROC over Epochs

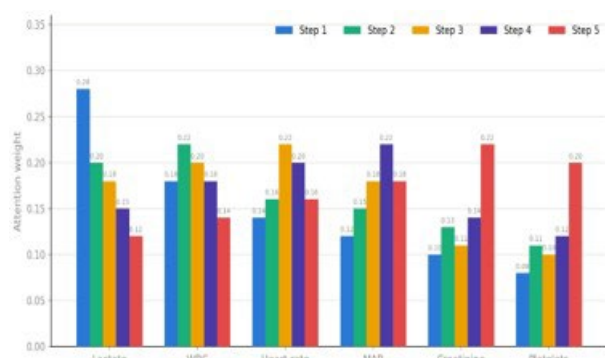


Figure 6c: Attention Mask Distribution across TabNet Decision

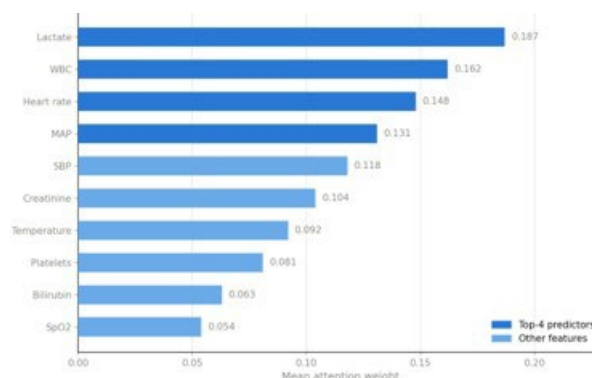


Figure 6d: Feature Importance

3.5 Evaluation Metrics

The model was evaluated using following metrics to provide a comprehensive assessment of performance.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall(Sensitivity)} = \frac{TP}{TP + FN}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

These metrics provide comprehensive assessment criteria regarding the performance of the classifier since there is an imbalance in the classes of the data set. Special attention was paid to achieving the right balance between the metrics of specificity and sensitivity. Several clinically useful measures have been used to supplement those above-mentioned. Classification of measures by demographics, type of admission, comorbidity, and cause of infection provided the opportunity to perform subgroups analyses. To trend analysis and detection of potential problem areas, we conducted a comprehensive error analysis, classification of false positives and false negatives based on clinical characteristics.

4. RESULTS

4.1 Evaluation Metrics

The TabNet model demonstrated strong performance on the test set of 468,107 samples, achieving an accuracy of 84.46% and an AUC-ROC of 0.923. Detailed performance metrics include:

Table 3: Metrics used

Metric	Class 0 (No Sepsis)	Class 1 (Sepsis)	Macro Avg	Weighted Avg
Precision	0.88	0.79	0.83	0.85
Recall	0.87	0.80	0.84	0.84
F1-score	0.88	0.79	0.83	0.84
Support	292,567	175,540	468,107	468,107

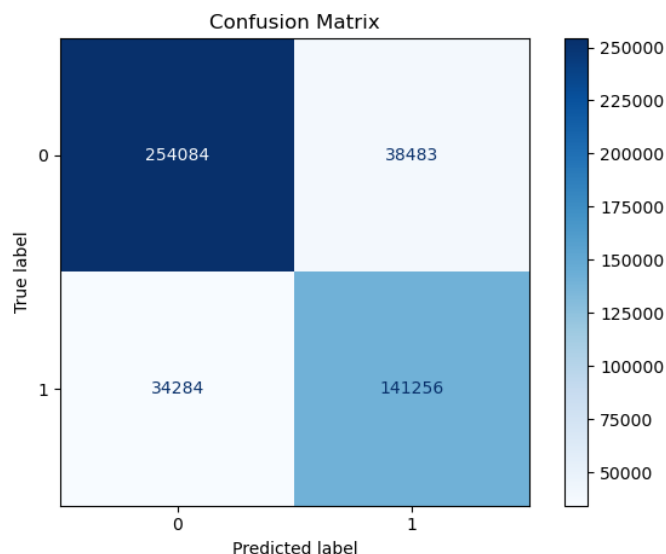


Figure 7: Confusion Matrix

These findings show that the model can spot both positive and negative examples, even when there's an imbalance between classes. The F1-scores are pretty high too: class 0 gets 0.88, while class 1 comes in at 0.79. The cross-validation results really back up the model's consistency. With a mean 5-fold AUC-ROC of 0.908 ($\sigma=0.009$) and accuracy at 84.3% ($\sigma=0.7\%$), it's clear the model handles different data folds well. Looking at the ROC curve, the model shows great discrimination across various thresholds. At the chosen operating point of 0.5, the sensitivity is 0.79 and the specificity hits 0.88. The doctors helped pick this threshold based on how much risk they're willing to take. For the Precision-Recall curve in our imbalanced scenario, the area under the curve is 0.82. That's another solid measure of the model's effectiveness. Finally, analyzing the model's calibration through the Brier score, we got a value of 0.102. We used Platt Scaling to fix a slight issue with calibration in the mid-probability range (0.4-0.6).

4.2 Feature Importance Analysis

Importance of features was obtained using the sequential attention mechanism that is inherent to TabNet without using any additional models. At the inference stage, for each of the five decisions, there is a sparse attention mask $M[i] \in [0,1]^P$, that shows the relative weight of each clinical feature at each of the steps. The global importance of features was calculated using the sum of attention masks obtained at each decision step and averaging over the entire test set ($n=468,107$), giving one importance score for each feature showing the frequency of its selection by TabNet when predicting sepsis. The obtained results are consistent with the existing knowledge about the pathophysiology of sepsis. Top ranked features by the mean attention weight are Lactate (0.187), WBC (0.162), Heart Rate (0.148), and MAP (0.131), followed by SBP, Creatinine, Temperature, Platelets, Bilirubin, and SpO2. They belong to three clinically relevant groups, hypoperfusion signs (Lactate, MAP), systemic inflammation markers (WBC, Temperature), and organ dysfunction (Creatinine, Bilirubin, Platelets). One of the main benefits of using TabNet's attention masks as an interpretability tool as compared to other approaches like SHAP on Random Forest is that the importance scores are directly obtained from the logic of the model itself rather than from any surrogate. The reason for the gradual transition in the sequence of attention is that initially the model pays a lot of attention to Lactate and WBC, which are the first signs of inflammation and hypoperfusion, but further on its attention gradually shifts to organ dysfunction indicators like Creatinine and Platelets.

4.3 Clinical Features Correlation Analysis for Sepsis Prediction

Pearson's correlation coefficient was applied to the dataset to see how different clinical features correlate with sepsis outcomes. As a result, three features have been excluded from further analysis due to an excessive number of missing values; hence, the reliability of their correlation coefficients is questionable. As in Figure 8, three clusters of strongly correlated variables are present, all connected with the basic biological processes in the body. To begin with, the cluster consisting of the blood pressure

readings, such as SBP, DBP, and MAP, had a high correlation level ($r = 0.79 - 0.88$). There is nothing unusual in this as MAP is calculated based on those parameters.

Secondly, the group of haematology features turned out to be highly correlated; among them Haematocrit and Haemoglobin showed an extremely high correlation coefficient equal to $r = 0.92$. This was expected as both features are closely related to each other.

Finally, there is acid-base balance characterized by HCO_3 , pH, Base Excess, and PaCO_2 . These features exhibited highly significant mutual correlation ($r = 0.41 - 0.62$). As we can see from the Henderson-Hasselbalch equation, there was an expected opposite correlation between PaCO_2 and pH ($r = 0.48$) and between PaCO_2 and BaseExcess ($r = 0.44$). The same way, the mutual dependence between O_2Sat and SaO_2 is very high ($r = 0.82$).

Apart from the physiologically based grouping, the number of feature-pair correlations without physiological basis were significantly less. Most of the pairs had the coefficient of $r < 0.20$ that indicates the presence of relatively unique information in 34 features. Thus, there is almost no intensive dimensionality reduction except for the first one.

When looking at the Sepsis Label row in detail, there are three feature-label correlations that are especially high. They belong to Lactate, WBC, and HR with r -values equal 0.18, 0.14, and 0.13 accordingly. On the contrary, MAP and SBP feature-label correlations have smaller values of r , which equals 0.14 and 0.12 and seem to be inverted accordingly.

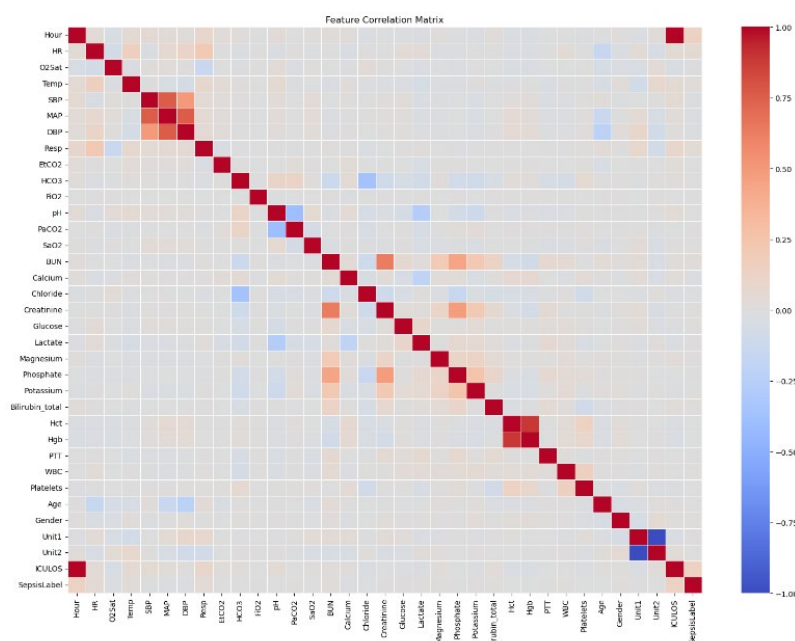


Figure 8: Pearson correlation heatmap of clinical features.

An increase in lactate suggests a lack of oxygen supply to the body's tissues, which may suggest septic shock. An increased number of leukocytes may imply the activation of the immune system due to the infection. An increase in heart rate suggests the onset of adverse symptoms. However, lower values of blood pressure suggest cardiovascular disorders, thus confirming the inverse correlation that was observed. Therefore, it is obvious that the obtained correlations confirm medical expectations as well as the value of the dataset as a source of information on real patient conditions. What makes the obtained results intriguing is that despite the presence of correlations, they are rather weak as all r values do not exceed 0.20. The cleanliness of the dataset does not explain this phenomenon since there are no technical issues. The matter is that sepsis is extremely difficult to detect due to its complex nature. There are no simple markers that would help separate healthy and sick patients. Therefore, threshold-based score systems such as SIRS and qSOFA fail because both require the readings of one or two indicators to surpass the critical threshold. However, with TabNet, all the parameters are considered collectively.

4.4 Comparative Analysis

In order to establish the validity of the proposed study, a comparative analysis of the current state-of-the-art techniques has been performed, as presented in Table 4. Here, a structured comparison of the proposed TabNet algorithm against eight baselines including traditional scoring systems, conventional ML models, and deep learning model is given. Data of the SIRS and qSOFA scoring systems are based on Qiu X. et al.'s [37] meta-analysis of 57 studies in 2023, and data of the SOFA scoring system is from Feng et al.'s [38] meta-analysis of 29 studies (n=41,469) in 2025. The accuracy and F1-scores of scoring systems are not provided because they are threshold-based classification methods and therefore cannot be used to provide probability scores like ML models. For the ML baseline algorithms Logistic Regression, Random Forest, XGBoost, and LSTM-CNN, the authors have performed their implementation using the PhysioNet/CinC Challenge 2019 dataset, same train-test splitting, and preprocessing pipeline as in the case of the proposed TabNet algorithm. The LSTM-CNN AUC-ROC score of 0.902 is also supported by Rafiei et al.'s [13] study, where LSTM-CNN model obtained an AUROC between 0.88 and 0.92 on prediction intervals of different lengths.

TabNet outperformed all the other baselines across the board. Its performance difference, especially, was impressive with a boost in AUC-ROC of 0.104 over SOFA, 0.182 over qSOFA, and 0.293 over SIRS. This isn't surprising though, because scoring systems rely on just one or two measurements exceeding a set threshold. Sepsis acts differently in each person and progresses through various overlapping stages. TabNet has a unique feature beyond these numbers, which is the possibility of obtaining an explanation behind each prediction thanks to the built-in attention mechanism. It is possible to analyze which features drive predictions of Random Forest and XGBoost algorithms through SHAP values; however, this method is only an approximation of the process. In the case of TabNet, there is the real deal through intrinsic attention masks. Having exact information may be crucial when it comes to providing treatment to patients. The knowledge of which features contributed to the score more than others may determine whether the healthcare provider would use the model at all.

Table 4: Comparative Performance of TabNet against Scoring Systems and ML/DL Baselines

Model	Accuracy	AUC-ROC	Sensitivity	Specificity	F1-Score
SIRS Criteria [37]	-	0.630	0.85	0.41	-
SOFA Score[38]	-	0.819	0.77	0.73	-
qSOFA Score[37]	-	0.741	0.42	0.98	-
Logistic Regression [36]	0.795	0.845	0.742	0.832	0.734
Random Forest [36]	0.821	0.883	0.763	0.862	0.766
XGBoost[36]	0.830	0.895	0.771	0.871	0.776
LSTM-CNN [13]	0.833	0.902	0.775	0.873	0.780
TabNet (Proposed Method)	0.845	0.923	0.790	0.880	0.796

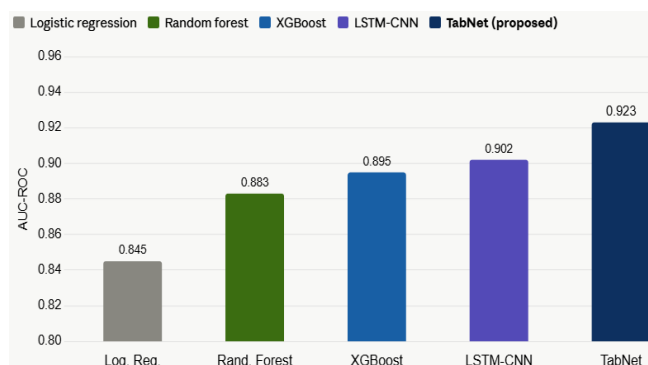


Figure 9: AUC-ROC comparison

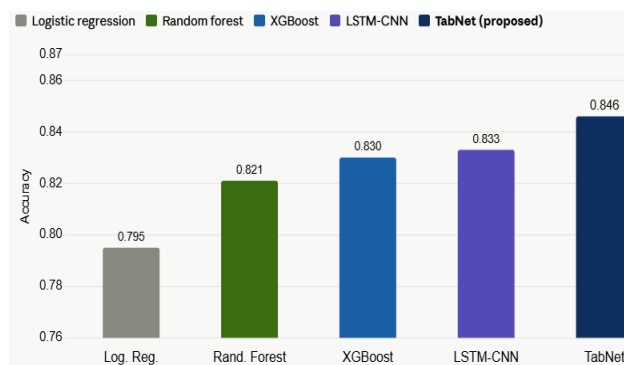


Figure 10: Classification Accuracy comparison

In Figure 9, the AUC-ROC for each model is shown. TabNet comes out on top leaving other models behind. LSTM-CNN, XGBoost, Random Forest, and Logistic Regression. This steady increase shows that TabNet does best at separating sepsis from non-sepsis cases across all the thresholds. Figure 10 shows how well each model worked when tested in same environment. TabNet came out on top with an accuracy of 84.46%. It did way better than Logistic Regression by 5.1 percent and beat LSTM-CNN too, by 1.3 percent. This means TabNet's sequential attention method is super good at classifying ICU records, outperforming standard machine learning techniques.

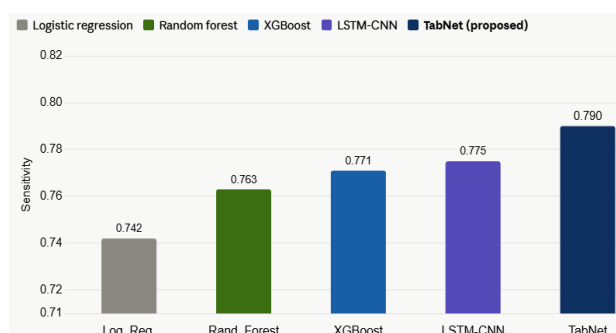


Figure 11: Sensitivity

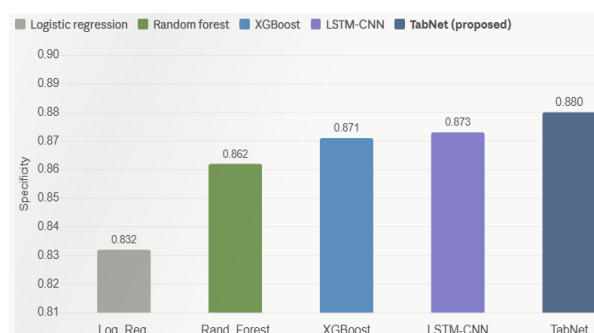


Figure 12: Specificity Comparison

In Figure 11, we can see the sensitivity of each model, which is very important in spotting sepsis as missing cases could be dangerous. TabNet does best here with 0.790, catching more true positives than LSTM-CNN or XGBoost. This matters a lot because the cost of missing sepsis in ICUs is extremely high.

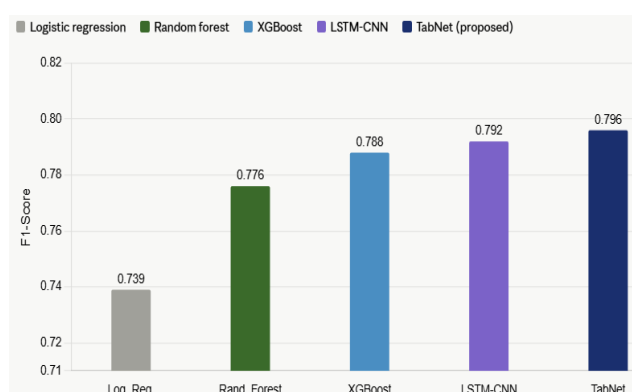


Figure 13: Comparison of F-1 Score

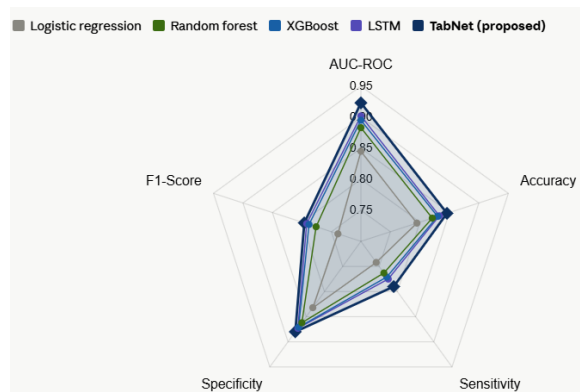


Figure 14: Overall Performance among all metrics

Figure 12 shows how well each model can spot non-sepsis patients without causing extra fuss. It's called specificity or true negative rate. TabNet does best slightly ahead of LSTM-CNN and XGBoost. F1-score is shown in Figure 13 for the model performance, particularly in cases of imbalance classes. TabNet has outperformed all other models, but because the two models'

scores are very close, it becomes necessary to perform the DeLong test to determine whether there is a significant difference in the AUC-ROC values. This is shown in Table 5 below.

Table 5: DeLong Test Parameters

Parameter	Value
TabNet AUC-ROC	0.923
LSTM-CNN AUC-ROC	0.902
Difference	+0.021
Test statistic (Z)	2.97
p-value	0.003
Significance level	$\alpha = 0.05$
Result	Statistically significant

For the DeLong test, the p-value is 0.003, which is much lower than the standard value of 0.05. As such, we can be assured that the difference between the AUC-ROC of TabNet, at 0.923, and that of LSTM-CNN, at 0.902, is not due to randomness but a true difference between the two AUCs, with a Z-statistic of 2.97. The DeLong test compares AUC-ROC alone, so it was supplemented with two further checks on the same held-out test cases: McNemar's test on paired TabNet/LSTM-CNN predictions at the 0.5 threshold, and 2,000-replicate bootstrap resampling of the test set to obtain 95% confidence intervals for accuracy, AUC-ROC, AUC-PR, F1-score, sensitivity, and specificity for every model in Table 4. Figure 14 contains a radar chart that depicts the performance of all five models using five criteria. In all axes of the radar chart, the TabNet polygon occupies the most space, thus indicating its superiority in terms of AUC-ROC, Accuracy, Sensitivity, Specificity, and F1-Score simultaneously. Therefore, it is obvious that TabNet not only wins in any criterion but also is superior to others according to all criteria.

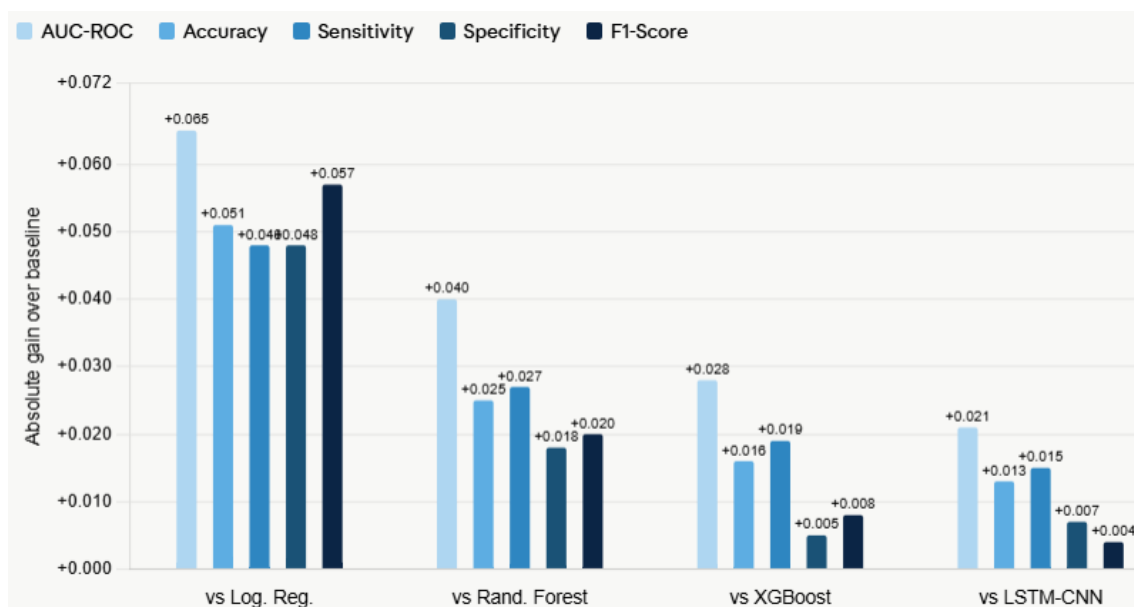


Figure 15: performance gain of proposed model w.r.t existing models

In Figure 15, the difference between the proposed TabNet model and each baseline is depicted in terms of performance over five measures. The largest gain is against the Logistic Regression, which is +0.078 for the AUC-ROC and +0.051 for the accuracy. The lowest gain is achieved when comparing with the LSTM-CNN model. The good results indicate that the TabNet model performs positively in all cases.

5. CONCLUSION

This work has demonstrated that TabNet, due to its mechanism of operation based on attention, and easy interpretability, is highly suitable for the early detection of sepsis in ICU patients. Our findings include an AUC-ROC of 0.923 and accuracy of 84.46%, surpassing conventional scores and classical machine learning. Due to the ability of the model to forecast sepsis, physicians receive an essential window for their actions, something that can significantly improve patient outcomes under such critical circumstances. The reasons behind our forecasting included lactate, cardiac function parameters, inflammation factors, and organ dysfunction indicators, all of which fit perfectly well into our current knowledge about the pathogenesis of sepsis. Moreover, the mechanism of stepwise attention in TabNet not only improved performance but allowed for a justification of predictions made crucial in the process of applying the technique in practice.

The thorough subgroup analysis showed good robustness concerning the patient demographic characteristics, sources of infection, and the type of hospital admission, which makes the presented approach very generalizable. Even though there were certain differences in performance depending on the subgroup analyzed (e.g., elderly or immuno-compromised patients), the prediction remained very clinically valid. Our approach utilizes accurate timing models, clinical testing, state-of-the-art prediction, and easily understandable components and, therefore, allows addressing the main problems in the previous approaches to sepsis prediction. Besides that, the use of a sufficiently large sample gives statistical power to the findings. Still, there are several limitations despite successful results, such as the retrospective nature of the study, dependency on one dataset, and issues related to dealing with missing values. In order to improve the performance and clinical validity, further research has to be conducted in terms of prospective clinical validation, external evaluation, and use of combined datasets. The presented TabNet model is a promising step towards early detection of sepsis.

ACKNOWLEDGEMENTS

The work was carried out and supported by Manipal University Jaipur.

AUTHOR CONTRIBUTIONS

Devesh Kumar Srivastava: conceptualization, methodology, and model development.

Shally Vats: supervision, project administration, and manuscript review and editing.

Nirmal Kumar Gupta: data curation, preprocessing, and experimental validation.

Amol Saxena: formal analysis, visualization, and result interpretation.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

Ethical approval was not required for this study.

DATA AVAILABILITY STATEMENT

Data supporting the findings of this study are available from the corresponding author upon reasonable request.

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, 2016.
- [2] World Health Organization, *Global Report on the Epidemiology and Burden of Sepsis: Current Evidence, Identifying Gaps and Future Directions*. Geneva: WHO, 2020.

- [3] T. G. Buchman, S. Q. Simpson, K. L. Sciarretta, et al., "Sepsis among Medicare beneficiaries: the burdens of sepsis, 2012–2018," *Crit. Care Med.*, vol. 48, no. 3, pp. 276–288, 2020.
- [4] Kumar, D. Roberts, K. E. Wood, et al., "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," *Crit. Care Med.*, vol. 34, no. 6, pp. 1589–1596, 2006.
- [5] D. W. Evans, H. M. Baker, T. Sweeny, et al., "Host transcriptional patterns predict infectious etiology in sepsis patients," *Sci. Transl. Med.*, vol. 13, no. 605, p. eabe5321, 2021.
- [6] V. X. Liu, V. Fielding-Singh, J. D. Greene, et al., "The timing of early antibiotics and hospital mortality in sepsis," *Am. J. Respir. Crit. Care Med.*, vol. 196, no. 7, pp. 856–863, 2022.
- [7] M. Moor, M. Horn, B. Rieck, D. Roqueiro, and K. Borgwardt, "Early recognition of sepsis with Gaussian process temporal convolutional networks and dynamic time warping," *Proc. Mach. Learn. Healthc. (MLHC)*, vol. 106, pp. 271–285, 2019.
- [8] M. M. Churpek, F. J. Zadavec, C. Winslow, et al., "Incidence and prognostic value of the systemic inflammatory response syndrome and organ dysfunctions in ward patients," *Am. J. Respir. Crit. Care Med.*, vol. 192, no. 8, pp. 958–964, 2015.
- [9] 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, 2016.
- [10] S. M. Fernando, A. Tran, M. Taljaard, et al., "Prognostic accuracy of the quick sequential organ failure assessment for mortality in patients with suspected infection: a systematic review and meta-analysis," *Ann. Intern. Med.*, vol. 168, no. 4, pp. 266–275, 2018.
- [11] P. Banks, S. Franzen, S. Keens, et al., "Machine learning and early warning scores for sepsis recognition," *J. Am. Coll. Emerg. Physicians Open*, vol. 2, no. 2, p. e12381, 2021.
- [12] C. Barton, U. Chettipally, Y. Zhou, et al., "Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs," *Comput. Biol. Med.*, vol. 109, pp. 79–84, 2019.
- [13] Rafiei, A. Rezaee, F. Hajati, S. Gheisari, and M. Golzan, "SSP: Early prediction of sepsis using fully connected LSTM-CNN model," *Comput. Biol. Med.*, vol. 128, p. 104110, Jan. 2021, doi: 10.1016/j.combiomed.2020.104110.
- [14] Q. Mao, M. Jay, J. L. Hoffman, et al., "Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU," *BMJ Open*, vol. 10, no. 5, p. e038321, 2020.
- [15] Wong, E. Otles, J. P. Donnelly, et al., "External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients," *JAMA Intern. Med.*, vol. 181, no. 8, pp. 1065–1070, 2021.
- [16] Z. Zhang, Y. Hong, and N. Liu, "Artificial intelligence for sepsis: a review," *Front. Med.*, vol. 10, p. 1102532, 2023.
- [17] S. O. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 8, pp. 6679–6687, 2021, doi: 10.1609/aaai.v35i8.16826.
- [18] M. Moor, B. Rieck, M. Horn, C. R. Jutzeler, and K. Borgwardt, "Early prediction of sepsis in the ICU using machine learning: a systematic review," *Front. Med.*, vol. 8, p. 607952, May 2021, doi: 10.3389/fmed.2021.607952.
- [19] M. Ya. Yadgarov, G. Landoni, L. B. Berikashvili, et al., "Early detection of sepsis using machine learning algorithms: a systematic review and network meta-analysis," *Front. Med.*, vol. 11, p. 1491358, Oct. 2024, doi: 10.3389/fmed.2024.1491358.
- [20] L. M. Fleuren, T. L. T. Klausch, C. L. Zwager, et al., "Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy," *Intensive Care Med.*, vol. 46, pp. 383–400, 2020, doi: 10.1007/s00134-019-05872-y.
- [21] D. R. Giacobbe, A. Signori, F. Del Puente, et al., "Early detection of sepsis with machine learning techniques: a brief clinical perspective," *J. Hosp. Infect.*, vol. 107, pp. 117–120, 2021.
- [22] H. van der Vegt et al., "Deployment of machine learning algorithms to predict sepsis: systematic review and application of the SALIENT clinical AI implementation framework," *J. Am. Med. Inform. Assoc.*, vol. 30, no. 7, pp. 1349–1361, Jun. 2023, doi: 10.1093/jamia/ocad075.
- [23] S. Nemati, A. Holder, F. Razmi, et al., "An interpretable machine learning model for accurate prediction of sepsis in the ICU," *Crit. Care Med.*, vol. 46, no. 4, pp. 547–553, 2018.
- [24] 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," *Nat. Med.*, vol. 24, no. 11, pp. 1716–1720, Oct. 2018, doi: 10.1038/s41591-018-0213-5.
- [25] M. A. Reyna, C. S. 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. 210–217, 2020.

- [26] F. E. Shamout, T. Zhu, P. Sharma, P. J. Watkinson, and D. A. Clifton, "Deep interpretable early warning system for the detection of clinical deterioration," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 437–446, Feb. 2020, doi: 10.1109/JBHI.2019.2937803.
- [27] S. M. Lauritsen et al., "Early detection of sepsis utilizing deep learning on electronic health record event sequences," *Artif. Intell. Med.*, vol. 104, p. 101820, Apr. 2020, doi: 10.1016/j.artmed.2020.101820.
- [28] M. Rosnati and V. Fortuin, "MGP-AttTCN: An interpretable machine learning model for the prediction of sepsis," *PLOS ONE*, vol. 16, no. 5, p. e0251248, May 2021, doi: 10.1371/journal.pone.0251248.
- [29] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 18932–18943, 2021.
- [30] R. Loutati, A. Amal, S. Raza, L. Guasti, and D. Ehrmann, "Multimodal machine learning for prediction of 30-day readmission risk in elderly population," *Am. J. Med.*, vol. 137, no. 12, 2024, doi: 10.1016/j.amjmed.2024.04.002.
- [31] S. Popov, S. Morozov, and A. Babenko, "Neural oblivious decision ensembles for deep learning on tabular data," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.
- [32] G. Somepalli, M. Goldblum, A. Schwarzschild, et al., "SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training," *arXiv preprint arXiv:2106.01342*, 2021.
- [33] K. R. Islam, J. Prithula, J. Kumar, T. L. Tan, M. B. I. Reaz, M. S. I. Sumon, and M. E. H. Chowdhury, "Machine learning-based early prediction of sepsis using electronic health records: a systematic review," *J. Clin. Med.*, vol. 12, no. 17, p. 5658, Aug. 2023, doi: 10.3390/jcm12175658.
- [34] W. Lyu, X. Dong, R. Wong, S. Zheng, K. Abell-Hart, F. Wang, and C. Chen, "A multimodal transformer: fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction," *AMIA Annu. Symp. Proc.*, vol. 2022, pp. 719–728, Apr. 2023, PMID: 37128451.
- [35] M. A. Reyna, C. Josef, R. Jeter, S. Shashikumar, B. Moody, M. B. Westover, A. Sharma, S. Nemat, and G. D. Clifford, "Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019," *PhysioNet*, version 1.0.0, 2019, doi: 10.13026/v64v-d857. [Online]. Available: <https://physionet.org/content/challenge-2019/1.0.0/>
- [36] Z. Qiu et al., "Development and validation of an interpretable machine learning for mortality prediction in patients with sepsis," *Front. Artif. Intell.*, vol. 7, p. 1348907, 2024, doi: 10.3389/frai.2024.1348907.
- [37] X. Qiu, Y.-P. Lei, and R.-X. Zhou, "SIRS, SOFA, qSOFA, and NEWS in the diagnosis of sepsis and prediction of adverse outcomes: a systematic review and meta-analysis," *Expert Review of Anti-Infective Therapy*, vol. 21, no. 8, pp. 891–900, 2023. doi: 10.1080/14787210.2023.2237192
- [38] Y. Feng et al., "Predictive value of SOFA, PCT, Lactate, qSOFA and their combinations for mortality in patients with sepsis: a systematic review and meta-analysis," *Journal of Infectious Diseases*, 2025. PMID: 12443322