**UNIVERSIDADE FEDERAL DE PELOTAS**
**Centro de Desenvolvimento Tecnológico**
**Programa de Pós-Graduação em Computação**

Dissertação

# Modelos Preditivos para Sepse: Um Estudo de Reprodutibilidade, Padronização e Confiabilidade em Aprendizado de Máquina

**Pedro Machado Wurzel**

Pelotas, 2025

**Pedro Machado Wurzel**

# Modelos Preditivos para Sepse: Um Estudo de Reprodutibilidade, Padronização e Confiabilidade em Aprendizado de Máquina

Dissertação apresentada ao Programa de Pós-Graduação em Computação do Centro de Desenvolvimento Tecnológico da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Ricardo Matsumura Araujo
Coorientador: Prof. Dr. Bruno Pereira Nunes

Pelotas, 2025

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação da Publicação

W972m Wurzel, Pedro Machado

Modelos preditivos para Sepse [recurso eletrônico] : um estudo de reprodutibilidade, padronização e confiabilidade em aprendizado de máquina / Pedro Machado Wurzel ; Ricardo Matsumura Araujo, orientador ; Bruno Pereira Nunes, coorientador. — Pelotas, 2025.
71 f. : il.

Dissertação (Mestrado) — Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2025.

1. Predição de Sepse. 2. Aprendizagem de máquina. 3. Predição conformal. 4. Avaliação de modelos clínicos. I. Araujo, Ricardo Matsumura, orient. II. Nunes, Bruno Pereira, coorient. III. Título.

CDD 005

Elaborada por Simone Godinho Maisonave CRB: 10/1733

**Pedro Machado Wurzel**

**Modelos Preditivos para Sepse: Um Estudo de Reprodutibilidade, Padronização e Confiabilidade em Aprendizado de Máquina**

Dissertação aprovada, como requisito parcial, para obtenção do grau de Mestre em Ciência da Computação, Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas.

**Data da Defesa:** 06 de agosto de 2025

**Banca Examinadora:**

Prof. Dr. Ricardo Matsumura Araujo (orientador)
Doutor em Computação pela Universidade Federal do Rio Grande do Sul.

Prof. Dr. Luiz Alexandre Chisini
Doutor em Odontologia pela Universidade Federal de Pelotas.

Prof. Dr. Ulisses Brisolara Corrêa
Doutor em Computação pela Universidade Federal de Pelotas

Dedico este trabalho aos meus pais, Márcia e Charles, pelo exemplo dado e pelo apoio incondicional em todas as etapas da minha jornada, e especialmente à minha parceira de vida, Vitória, que, mesmo enfrentando o período mais difícil de sua vida, esteve ao meu lado com força e sensibilidade. Sua presença foi essencial para que eu pudesse concluir esta dissertação.

# RESUMO

WURZEL, Pedro Machado. **Modelos Preditivos para Sepse: Um Estudo de Reprodutibilidade, Padronização e Confiabilidade em Aprendizado de Máquina**. Orientador: Ricardo Matsumura Araujo. 2025. 71 f. Dissertação (Mestrado em Ciência da Computação) – Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2025.

A sepse constitui uma condição médica grave caracterizada pela resposta inflamatória sistêmica do organismo a uma infecção e apresenta alta taxa de mortalidade, cuja detecção precoce é fundamental para o sucesso do tratamento. Nas últimas décadas, modelos baseados em aprendizagem de máquina têm sido amplamente explorados para predizer sua ocorrência, mas a literatura da área ainda apresenta sérios problemas de reprodutibilidade, comparabilidade e ausência de padronização metodológica. Esta dissertação propõe uma avaliação crítica do estado da arte na predição de sepse, aliada a uma análise criteriosa usando uma estrutura padronizado de diferentes métodos. Para isso, foi criado um conjunto de dados padronizado com base no MIMIC-IV v2.2, utilizando a definição Sepsis-3, com critérios clínicos reprodutíveis e janelas temporais bem definidas. Modelos do estado da arte foram reimplementados e avaliados tanto em seus contextos originais quanto nesse ambiente controlado. Além disso, esta pesquisa incorporou a técnica de Predição Conformal, em sua forma transdutiva, como forma de quantificar incertezas e aumentar a confiabilidade das predições. Os resultados mostraram que a reprodutibilidade ainda é um desafio, que o desempenho dos modelos varia significativamente sob diferentes janelas temporais e que a aplicação de predição conformal pode melhorar a segurança e a precisão das predições, embora sua eficácia dependa da calibragem e da natureza do modelo subjacente. As contribuições desta dissertação incluem a padronização de práticas experimentais, a análise crítica da literatura vigente, a demonstração das limitações dos modelos atuais e a proposta de caminhos mais robustos e confiáveis para o uso de IA em contextos clínicos sensíveis.

Palavras-chave: predição de sepse; aprendizagem de máquina; conformal prediction; avaliação de modelos clínicos.

# ABSTRACT

WURZEL, Pedro Machado. **Titulo do Trabalho em Ingles**. Advisor: Ricardo Matsumura Araujo. 2025. 71 f. Dissertation (Masters in Computer Science) – Technology Development Center, Federal University of Pelotas, Pelotas, 2025.

Sepsis is a serious medical condition characterized by the body's systemic inflammatory response to infection and is associated with a high mortality rate, making early detection essential for successful treatment. In recent decades, machine learning models have been widely explored to predict its occurrence, yet the literature still suffers from major issues related to reproducibility, comparability, and the lack of methodological standardization. This dissertation proposes a critical evaluation of the state of the art in sepsis prediction, alongside the development of a standardized framework for testing and comparing models. To this end, a standardized dataset was created based on MIMIC-IV v2.2, using the Sepsis-3 definition, with reproducible clinical criteria and well-defined time windows. State-of-the-art models were reimplemented and evaluated both in their original contexts and within this controlled environment. Additionally, this research incorporated the Conformal Prediction technique, in its transductive form, as a means of quantifying uncertainties and enhancing the reliability of predictions. The results showed that reproducibility remains a challenge, that model performance varies significantly across different time windows, and that the application of conformal prediction can improve the safety and precision of predictions, although its effectiveness depends on calibration and the nature of the underlying model. The contributions of this dissertation include the standardization of experimental practices, a critical analysis of the current literature, the demonstration of limitations in existing models, and the proposal of more robust and reliable pathways for the use of AI in sensitive clinical settings.

Keywords: sepsis prediction; machine learning; conformal prediction; clinical model evaluation.

# LISTA DE FIGURAS

# LISTA DE TABELAS

# LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| AM | Aprendizagem de Máquina |
| AUROC | Área sob a Curva Característica de Operação do Receptor |
| IC | Intervalo de Confiança |
| MIMIC | Medical Information Mart for Intensive Care |
| PC | Predição Conformal |
| PCI | Predição Conformal Indutiva |
| PCT | Predição Conformal Transdutiva |
| SOFA | Sequential Organ Failure Assessment |
| UTI | Unidade de Terapia Intensiva |

# SUMÁRIO

# 1 INTRODUÇÃO

A sepse constitui uma condição médica grave caracterizada pela resposta inflamatória sistêmica do organismo a uma infecção, culminando na disfunção ou falência de órgãos (Singer et al., 2016). Segundo dados da Organização Mundial da Saúde, a sepse impacta anualmente mais de 30 milhões de indivíduos globalmente, resultando em aproximadamente 6 milhões de óbitos (World Health Organization et al., 2018). A prontidão no diagnóstico e o imediato início do tratamento representam aspectos cruciais para aprimorar os desfechos em pacientes com sepse ou choque séptico (Evans et al., 2021; Seymour et al., 2016).

No contexto brasileiro, uma pesquisa conduzida por Machado et al. (2017) revela que cerca de um terço dos leitos de Unidades de Terapia Intensiva (UTIs) está ocupado por casos de sepse, apresentando uma letalidade global alarmante de 55%. Nos Estados Unidos, internações relacionadas à sepse ultrapassam as provocadas por infarto do miocárdio e acidente vascular cerebral, constituindo quase metade dos óbitos hospitalares (Fleischmann et al., 2016). Entre 2010 e 2019, o Brasil registrou aproximadamente 463 mil óbitos atribuídos à sepse, sendo a faixa etária acima de 60 anos a mais afetada, com uma taxa de 112,9 óbitos por 100 mil habitantes. Análises da tendência de mortalidade indicam um aumento contínuo ao longo desse período (Almeida et al., 2022).

A gravidade da sepse reside na sua capacidade de induzir disfunção em múltiplos órgãos e sistemas, acarretando elevadas taxas de mortalidade e encargos financeiros significativos para os sistemas de saúde (Jost et al., 2019). Diversas diretrizes foram concebidas ao longo dos anos para a identificação e tratamento da sepse, sendo a Sepsis-3, proposta por Singer et al. (2016), a mais recente. Apesar das melhorias introduzidas por essa atualização, ainda há oportunidades para alcançar resultados mais promissores. Nesse cenário, a aplicação de técnicas de aprendizagem de máquina (AM) tem ganhado destaque como alternativa promissora para a identificação precoce da sepse (Deng et al., 2022).

A predição de sepse representa um desafio relevante na área de AM, pois os sintomas são, muitas vezes, vagos e comuns, sobretudo nos estágios iniciais. Além disso,

não há um padrão-ouro bem estabelecido para o diagnóstico, e os métodos convencionais, como culturas microbiológicas, são lentos diante da gravidade da condição (Hunt, 2023; Kim; Choi, 2020).

Apesar do volume crescente de pesquisas na área, a literatura revela grande variação em metodologias, técnicas e definições adotadas. Há diferenças nas variáveis utilizadas, tipos de dados (quantitativos, textuais, demográficos), janelas de tempo para predição e estratégias para lidar com dados ausentes (Moor et al., 2021). Essa heterogeneidade compromete a reprodutibilidade dos estudos e dificulta a identificação de abordagens eficazes para aplicação clínica.

Neste trabalho, parte-se da hipótese de que a ausência de critérios padronizados compromete a validade das comparações entre modelos preditivos de sepse. O objetivo geral é propor uma abordagem padronizada para avaliação desses modelos com técnicas de AM, incorporando o conceito de predição conformal (PC) como forma de estimar incerteza nas predições realizadas, ampliando a confiabilidade dos modelos em contextos clínicos.

A metodologia completa será detalhada em capítulo específico. No entanto, de forma geral, ela inclui a reprodução de estudos representativos do estado da arte, a utilização de um conjunto de dados padronizado e a aplicação da técnica de PC. A Figura 1 ilustra de maneira esquemática as etapas do processo metodológico, desde a seleção dos estudos até a avaliação final.

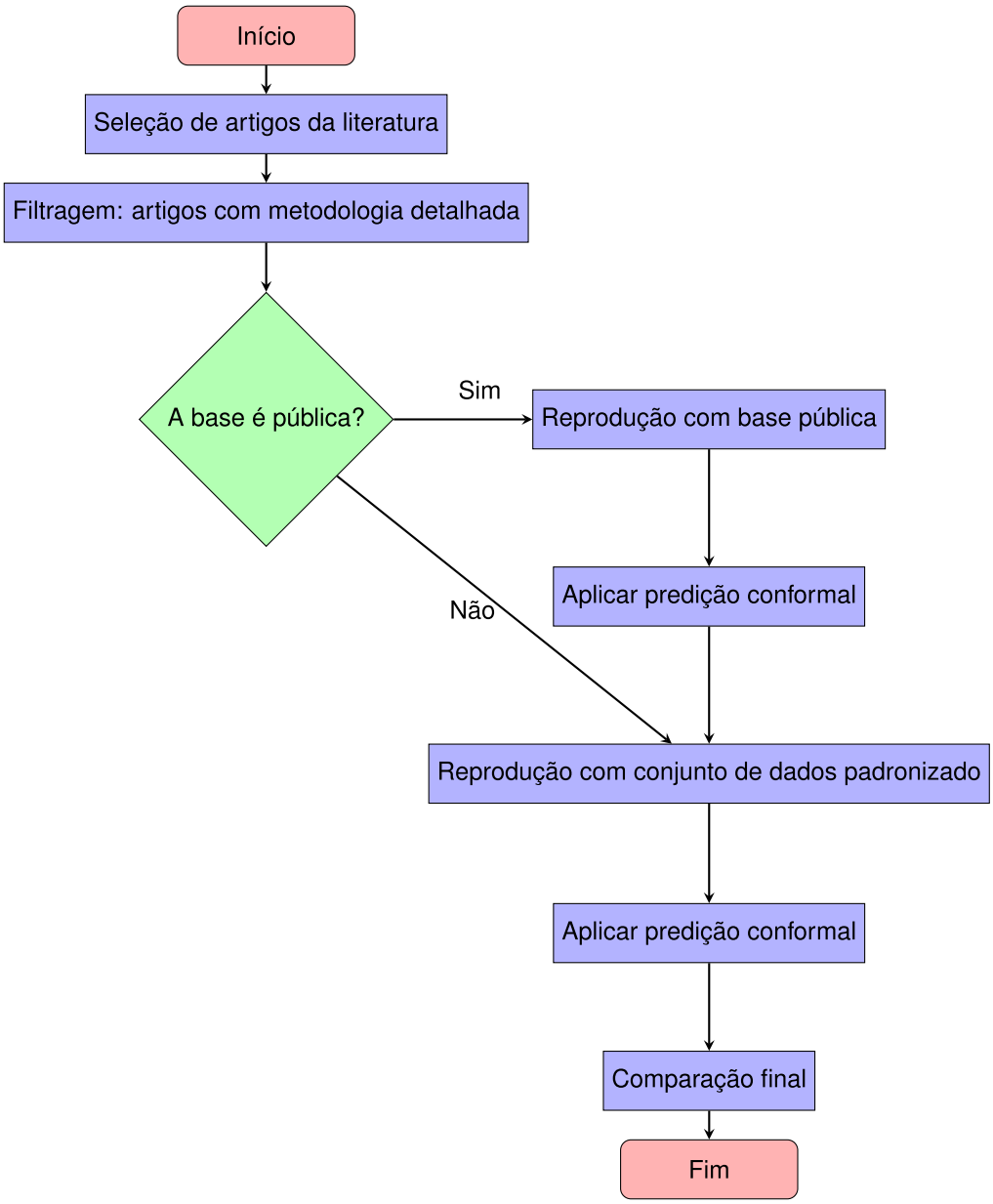


Figura 1 – Fluxograma da Metodologia Experimental: Reprodução, Padronização e Predição Conformal.

# 2 REVISÃO DA LITERATURA

## 2.1 Definição e Diagnóstico da Sepse

Ao longo do tempo, houveram evoluções nos critérios e definições relacionados à sepse, visando aprimorar a prestação de cuidados e a identificação precoce da condição. Notavelmente, as definições Sepsis-1, Sepse-2 e Sepse-3 emergiram como marcos fundamentais, recebendo amplo respaldo na comunidade científica.

A definição Sepsis-1, estabelecida em 1991, introduziu os critérios da Síndrome da Resposta Inflamatória Sistêmica (SIRS) para identificar pacientes com sepse. Os critérios SIRS incluíam: temperatura corporal $> 38°$C ou $< 36°$C, frequência cardíaca $> 90$ bpm, frequência respiratória $> 20$ irpm ou $PaCO_2 < 32$ mmHg, e contagem de leucócitos $> 12.000/mm^3$, $< 4.000/mm^3$ ou presença de $> 10\%$ de formas imaturas. A presença de dois ou mais desses critérios, juntamente com uma infecção suspeita ou confirmada, definia a sepse. A sepse grave era caracterizada pela presença de disfunção orgânica, e o choque séptico, por hipotensão persistente apesar da reposição volêmica adequada (Dugar; Choudhary; Duggal, 2020).

A definição Sepse-2, introduzida por Levy; Fink (2003), fundamenta-se na presença de uma Síndrome de Resposta Inflamatória Sistêmica (SIRS) desencadeada por uma infecção. Conforme estipulado por esta definição, a sepse é caracterizada pela presença de pelo menos dois critérios SIRS, abrangendo leucopenia ou leucocitose, temperatura corporal anormal (febre ou hipotermia), aumento da frequência cardíaca (taquicardia), aumento da frequência respiratória (taquipneia), e alterações na temperatura corporal. O choque séptico é definido como sepse grave acompanhada de hipotensão persistente, mesmo após ressuscitação com fluidos; enquanto a sepse grave se configura pela combinação de sepse com disfunção orgânica.

Já a definição Sepse-3, apresentada por Singer et al. (2016), foi concebida para aprimorar a precisão e utilidade clínica na detecção de sepse. Em contraste com a abordagem predominantemente baseada em critérios SIRS, a Sepse-3 enfatiza a falência de órgãos como característica definidora. O escore SOFA (*Sequential Organ Failure Assessment*), que avalia a função de sistemas como respiratório, cardiovas-

cular, hepático, renal, de coagulação e nervoso central, é empregado para medir a disfunção orgânica. Um aumento de dois ou mais pontos na pontuação SOFA, resultante de uma infecção, indica a presença de sepse. Adicionalmente, sinais como níveis elevados de lactato sanguíneo e hipotensão persistente, exigindo vasopressores para manter uma pressão arterial média de 65 mmHg ou superior, são considerados indicativos de choque séptico (Singer et al., 2016).

O choque séptico, segundo a Sepse-3, é definido como um subconjunto de sepse no qual anomalias circulatórias e celulares/metabólicas são suficientemente profundas para aumentar substancialmente a mortalidade. Clinicamente, isso é identificado por hipotensão que requer vasopressores para manter uma pressão arterial média $\geq 65$ mmHg e níveis de lactato sérico $> 2$ mmol/L, apesar de ressuscitação volêmica adequada (Devos, 2018).

Essas definições desempenham um papel crucial na padronização da identificação e gestão da sepse, facilitando diagnósticos oportunos e precisos, bem como intervenções eficazes para melhorar os desfechos dos pacientes. Destaca-se que a definição mais amplamente adotada na atualidade é a Sepse-3 (Seymour et al., 2016), que simplifica os critérios, focando na disfunção orgânica em detrimento dos critérios SIRS utilizados em definições anteriores. Essa simplificação não apenas reduz o potencial de erros de diagnóstico, mas também facilita a padronização do diagnóstico e pesquisa da sepse, promovendo uma comparação mais eficaz de estudos, compartilhamento de dados e o desenvolvimento de estratégias de gestão embasadas em evidências (Islam et al., 2023).

## 2.2 Predição de Sepse com Aprendizagem de Máquina

A aplicação de técnicas de AM na predição de sepse tem se mostrado promissora, permitindo a identificação precoce da condição e possibilitando intervenções clínicas mais eficazes. A seguir, são abordadas as principais técnicas utilizadas, bases de dados comuns na literatura e métricas empregadas para avaliação dos modelos.

### 2.2.1 Principais técnicas utilizadas

Diversas técnicas de AM têm sido empregadas na predição de sepse, com destaque para métodos supervisionados baseados em classificadores. Os algoritmos mais recorrentes na literatura são os baseados em árvores de decisão, como o *Random Forest*, e modelos de *ensemble*, incluindo o *Gradient Boosting*. Esses métodos são valorizados por sua capacidade de lidar com dados clínicos tabulares, com múltiplas variáveis heterogêneas e potenciais interações complexas entre os preditores. Modelos lineares, como a regressão logística, também continuam sendo amplamente utilizados, especialmente pela sua interpretabilidade e robustez frente a conjuntos de

dados clínicos com menor complexidade (Bomrah et al., 2024; Moor et al., 2021; Fleuren et al., 2020; Fascia, 2024).

Além disso, redes neurais artificiais, particularmente as redes *feedforward* e as redes recorrentes, têm sido exploradas com frequência crescente, sobretudo em contextos que envolvem séries temporais, como sinais vitais monitorados continuamente. As redes neurais recorrentes com unidades LSTM (*Long Short-Term Memory*) são particularmente úteis nesses cenários, por sua habilidade de capturar dependências temporais de longo prazo nos dados. Técnicas de regularização, seleção de variáveis e imputação de dados incompletos foram identificadas como estratégias comuns para melhorar a performance e a generalização dos modelos (Bomrah et al., 2024; Moor et al., 2021; Fascia, 2024).

Ainda existe grande variabilidade metodológica entre os estudos. Essa heterogeneidade afeta diretamente a capacidade de comparar os resultados de forma padronizada, mas reforça o interesse crescente na aplicação de modelos cada vez mais sofisticados no contexto da sepse (Fleuren et al., 2020).

### 2.2.2 Bases de dados comuns na literatura

A eficácia dos modelos de AM na predição de sepse está intrinsecamente relacionada à qualidade, abrangência e representatividade dos dados utilizados para o treinamento e validação desses algoritmos. Diversos bancos de dados clínicos têm sido amplamente empregados na literatura para esse propósito, destacando-se:

- **MIMIC-III (Medical Information Mart for Intensive Care III)**: Desenvolvido pelo MIT Lab for Computational Physiology, o MIMIC-III é um banco de dados público e amplamente utilizado que contém informações clínicas detalhadas de mais de 60.000 admissões hospitalares em UTIs do Beth Israel Deaconess Medical Center, entre 2001 e 2012. A base inclui variáveis como sinais vitais, resultados laboratoriais, medicamentos administrados, diagnósticos, anotações clínicas e dados demográficos. Devido à sua riqueza e granularidade, o MIMIC-III é considerado um recurso valioso para o desenvolvimento e validação de modelos preditivos de sepse (Johnson et al., 2016).

- **MIMIC-IV**: Lançado como sucessor do MIMIC-III, o MIMIC-IV traz dados mais atualizados de internações em UTIs e unidades hospitalares do mesmo centro médico, abrangendo o período de 2008 a 2022. Essa versão foi reestruturada para melhorar a organização e a consistência dos dados, e inclui não apenas informações clínicas de pacientes adultos, mas também dados administrativos, registros de prescrição, exames laboratoriais, sinais vitais e resultados de imagens diagnósticas (Johnson et al., 2020).

- **eICU Collaborative Research Database**: Desenvolvido em parceria entre o MIT e a empresa Philips Healthcare, o eICU Database contém informações de mais de 200.000 admissões em UTIs de 208 hospitais diferentes dos Estados Unidos, abrangendo um período entre 2014 e 2015. Este banco de dados multicêntrico é particularmente relevante para estudos de validação externa, por permitir a avaliação da generalização de modelos preditivos em diferentes contextos hospitalares. Os dados incluem sinais vitais, resultados laboratoriais, medicamentos, intervenções clínicas, escores de gravidade e dados demográficos dos pacientes (Pollard et al., 2018).
- **PhysioNet/Computing in Cardiology Challenge 2019**: Esta competição internacional teve como objetivo incentivar o desenvolvimento de modelos de predição precoce de sepse com base em dados clínicos temporais. O conjunto de dados fornecido no desafio inclui registros longitudinais de 40 variáveis clínicas, como frequência cardíaca, pressão arterial, temperatura corporal, níveis de oxigênio e parâmetros laboratoriais, coletadas ao longo do tempo para mais de 40.000 pacientes. Os dados são anonimizados e organizados em séries temporais, permitindo a modelagem de eventos dinâmicos e a detecção de padrões precoces associados à sepse. Essa base é amplamente utilizada como *benchmark* para comparação entre diferentes abordagens de AM (Reyna et al., 2020).

### 2.2.3 Métricas usadas para avaliação

A avaliação do desempenho de modelos de AM na predição de sepse é essencial para assegurar sua aplicabilidade clínica, sendo realizada por meio de diversas métricas estatísticas. A métrica mais amplamente empregada é a Área sob a Curva Característica de Operação do Receptor (AUROC), que quantifica a capacidade discriminativa do modelo em diferenciar corretamente entre pacientes com e sem sepse (Islam et al., 2023).

Entre as métricas fundamentais, destaca-se a sensibilidade (ou *recall*), que expressa a proporção de casos de sepse corretamente identificados. Essa medida é particularmente crítica em contextos clínicos, nos quais a falha em detectar precocemente a condição pode acarretar consequências severas para o paciente.

Complementarmente, a especificidade avalia a proporção de indivíduos saudáveis que são corretamente classificados como negativos, contribuindo para a redução de falsos positivos e, consequentemente, para a diminuição de alarmes desnecessários. A precisão, por sua vez, representa a proporção de predições positivas que são de fato corretas, sendo um indicador relevante da confiabilidade dos alertas emitidos pelo sistema de predição (Islam et al., 2023; Shashikumar et al., 2021).

Por fim, o *F1-score*, que também é comumente utilizado em trabalhos de predição de sepse, é definido como a média harmônica entre precisão e sensibilidade. Ele

constitui uma métrica especialmente útil em cenários com classes desbalanceadas, como é frequentemente o caso em bases de dados clínicas de sepse (Islam et al., 2023).

## 2.3 Desafios da Reprodutibilidade na Área

A predição de sepse por meio de técnicas de AM enfrenta desafios significativos relacionados à reprodutibilidade dos estudos. Tais obstáculos dificultam a validação independente dos resultados publicados, comprometem a confiabilidade das comparações entre abordagens e, por conseguinte, limitam tanto o avanço científico quanto a aplicação clínica segura dessas tecnologias. Entre os principais fatores que contribuem para esse cenário, destacam-se a heterogeneidade dos conjuntos de dados utilizados, a escassa descrição metodológica, as variações nos processos de pré-processamento e as dificuldades inerentes à comparação entre diferentes trabalhos.

Um dos entraves mais relevantes é a diversidade dos conjuntos de dados empregados nos estudos. Muitos trabalhos utilizam bases de dados privadas, de difícil acesso ou com restrições éticas e legais, o que inviabiliza a reprodução direta dos experimentos. Mesmo entre os estudos que utilizam bases públicas amplamente conhecidas, como o MIMIC-III ou o MIMIC-IV, observa-se grande variação nas estratégias de extração e organização dos dados, bem como nas definições clínicas adotadas para o desfecho de sepse. Essa falta de padronização resulta em diferenças substanciais na composição dos conjuntos de treinamento e teste, afetando diretamente as métricas de desempenho reportadas.

Adicionalmente, os métodos de pré-processamento variam amplamente entre os estudos. Tais variações incluem a seleção de variáveis, o tratamento de valores ausentes, a definição de janelas temporais de observação, a normalização dos dados e as estratégias para lidar com o desbalanceamento entre classes. Pequenas alterações nessas etapas podem gerar impactos significativos no comportamento dos modelos, dificultando a identificação de se um desempenho superior decorre, de fato, da técnica empregada ou de ajustes no tratamento dos dados. A ausência de padronização e, frequentemente, a escassez de documentação detalhada sobre esses procedimentos tornam a replicação dos resultados um desafio substancial.

Por fim, a comparação entre diferentes estudos também apresenta limitações importantes. A utilização de conjuntos de dados distintos, métricas variadas, diferentes períodos de avaliação e definições heterogêneas de sepse resulta em análises comparativas inconsistentes e, muitas vezes, imprecisas. Esse cenário prejudica a identificação de abordagens verdadeiramente robustas e eficazes em contextos clínicos reais. A ausência de um protocolo comum de avaliação contribui para um panorama fragmentado, no qual os avanços são difíceis de validar e consolidar.

Diante desses desafios, torna-se evidente a necessidade de uma abordagem padronizada para a avaliação de modelos de predição de sepse, baseada no uso de conjuntos de dados públicos, protocolos reprodutíveis e critérios consistentes de avaliação. Esta dissertação propõe-se a enfrentar essas limitações, oferecendo uma base metodológica sólida que possa servir de alicerce para futuras pesquisas na área.

# 3 OBJETIVOS E METODOLOGIA

## 3.1 Objetivos

Este trabalho tem como foco central a avaliação crítica do estado da arte na predição de sepse por meio de técnicas de AM, bem como a proposição de uma abordagem padronizada que permita comparações mais justas e reprodutíveis entre diferentes métodos. Considerando as lacunas observadas na literatura, como a falta de consenso sobre definições clínicas, a utilização de diferentes conjuntos de dados e a escassez de código e experimentos reproduzíveis, este estudo visa organizar, sistematizar e padronizar os elementos necessários para uma avaliação robusta e confiável dos modelos.

### 3.1.1 Objetivo principal

Avaliar criticamente o estado da arte em predição de sepse com AM, identificando os principais obstáculos à reprodutibilidade e propondo um protocolo padronizado de avaliação que permita a comparação justa entre diferentes metodologias, incluindo a aplicação de técnicas de PC.

### 3.1.2 Objetivos específicos e hipóteses

- **Investigar a reprodutibilidade de estudos existentes:** Selecionar e reproduzir experimentalmente estudos considerados estado da arte, avaliando os desafios práticos encontrados durante o processo e discutindo as limitações que dificultam a replicação dos resultados originais.
  **Hipótese 1:** A maioria dos estudos de predição de sepse com AM não pode ser reproduzida integralmente com as informações metodológicas disponíveis nas publicações originais.

- **Propor um conjunto de dados padronizado para avaliação:** Construir um conjunto de dados estruturado e documentado, com definição consistente do desfecho (sepse), seleção criteriosa de variáveis e estratégias de pré-processamento bem definidas, permitindo a aplicação de múltiplas técnicas sob

condições iguais.
**Hipótese 2:** A existência de um conjunto de dados padronizado com definições clínicas consistentes e fluxo de processamento de preparação documentado reduz a variabilidade nos resultados entre diferentes métodos de predição.

- **Investigar o impacto de diferentes configurações de predição:** Avaliar como variações em parâmetros como janelas temporais, tempo de antecedência da predição e balanceamento das classes afetam o desempenho dos modelos.
  **Hipótese 3:** Pequenas variações nas configurações de predição resultam em diferenças significativas no desempenho dos modelos, afetando sua comparabilidade.

- **Aplicar técnicas de PC:** Integrar PC aos modelos avaliados, a fim de incorporar estimativas de confiabilidade nas previsões realizadas, explorando seu impacto na interpretação clínica e na robustez dos resultados obtidos.
  **Hipótese 4:** Técnicas de PC podem tornar mais explícitas as incertezas do modelo, permitindo uma melhor interpretação dos resultados.

- **Propor diretrizes para avaliação padronizada na área:** Com base nos achados do estudo, sugerir boas práticas e recomendações para futuros trabalhos na área de predição de sepse com AM, promovendo maior reprodutibilidade e comparabilidade entre estudos.
  **Hipótese 5:** A adoção de diretrizes padronizadas pode melhorar substancialmente a reprodutibilidade e a comparabilidade de estudos futuros na área de predição de sepse com AM.

## 3.2 Descrição dos Experimentos

### 3.2.1 Reprodutibilidade dos estudos

A primeira etapa experimental consistiu na tentativa de reprodução de modelos propostos na literatura recente. Foram selecionados artigos representativos do estado da arte em predição de sepse, com base em critérios como uso de AM, detalhamento metodológico e alta performance preditiva. Para cada estudo, buscou-se replicar o fluxo de processamento descrito, incluindo etapas de pré-processamento, seleção de variáveis, configuração dos modelos e métricas de avaliação.

Nos casos em que os trabalhos utilizaram bases de dados públicas, a reprodução foi feita utilizando o mesmo conjunto de dados sempre que possível, além do padronizado. Para os estudos que utilizaram bases de dados privadas ou não disponíveis publicamente, os experimentos foram adaptados para utilizar apenas o conjunto de dados padronizado desenvolvido neste trabalho. Nestes casos, manteve-se ao máximo o

mesmo fluxo metodológico descrito pelos autores, incluindo arquitetura dos modelos, engenharia de atributos e abordagem de avaliação, de modo a preservar a lógica da proposta original e permitir uma análise comparativa sob condições padronizadas.

Durante a reprodução, foram registrados todos os obstáculos enfrentados, como ambiguidade na definição de variáveis, inconsistências nas métricas reportadas e diferenças relacionadas à definição de sepse utilizada. Os resultados obtidos foram comparados com os resultados originais, quando disponíveis, a fim de quantificar o grau de reprodutibilidade de cada estudo. Essa etapa teve como finalidade ilustrar os desafios práticos da replicação científica na área, servindo como motivação para a padronização proposta neste trabalho.

### 3.2.2 Avaliação dos métodos no conjunto de dados padronizado

Com base nas limitações identificadas na etapa anterior, foi desenvolvido um conjunto de dados padronizado a partir da base pública MIMIC-IV, versão 2.2 (Johnson et al., 2020). A escolha dessa base se deu por diversos motivos: (i) ampla adoção na literatura científica da área de saúde, especialmente em estudos de predição clínica; (ii) alta granularidade temporal dos dados, o que permite a modelagem temporal de eventos clínicos com precisão; (iii) presença de variáveis clínicas relevantes para a detecção e progressão de sepse, como sinais vitais, exames laboratoriais e anotações de prescrições; e (iv) caráter público e de livre acesso, o que favorece a reprodutibilidade e transparência dos experimentos.

A construção do conjunto de dados incluiu critérios de filtragem fundamentais, como a exclusão de pacientes com menos de 18 anos, visando manter a homogeneidade clínica e respeitar as diferenças entre populações pediátricas e adultas. Além disso, cada admissão à UTI foi considerada de forma independente, tratando múltiplas internações de um mesmo paciente como eventos distintos, o que reflete melhor a dinâmica de risco clínico real.

A padronização envolveu a definição clara do desfecho (sepse), com base nos critérios do Sepsis-3, e a aplicação de regras consistentes para extração, agregação e normalização das variáveis ao longo do tempo. Além disso, definiu-se uma estrutura experimental que permite testar diferentes janelas de observação e tempos de antecedência da predição, respeitando um cenário realista de aplicação clínica.

### 3.2.3 Testes com Predição Conformal

A etapa final dos experimentos consistiu na aplicação de técnicas de PC aos modelos treinados, com o objetivo de incorporar estimativas explícitas de confiabilidade às previsões realizadas. Neste trabalho, optou-se pela utilização da abordagem Predição Conformal Transdutiva (PCT), em detrimento de variantes indutivas, em razão de suas propriedades estatísticas mais rigorosas e da oferta de garantias teóricas válidas

sob suposições mínimas.

A escolha pelo PCT fundamenta-se, principalmente, em seu mecanismo de funcionamento: ele realiza a predição de cada exemplo de teste considerando diretamente o conjunto de calibração, sem a necessidade de treinar modelos auxiliares ou realizar particionamentos fixos dos dados. Essa característica permite a geração de conjuntos preditivos com validade estatística garantida para cada instância individual, mantendo controle formal sobre o erro condicional, aspecto particularmente relevante em contextos clínicos, nos quais a confiabilidade de cada predição é essencial. Conforme discutido por Vovk; Gammerman; Shafer (2005), essa abordagem oferece maior robustez teórica e rigor na quantificação da incerteza, reforçando sua adequação a aplicações sensíveis, como a predição de sepse.

Nos experimentos, o PCT foi aplicado sobre os modelos com melhor desempenho identificados na etapa anterior, respeitando o mesmo conjunto de dados e as mesmas configurações experimentais. Foram testados diferentes níveis de significância, permitindo avaliar o impacto da confiança desejada na largura dos conjuntos preditivos, na cobertura estatística e na utilidade clínica das predições.

A análise considerou tanto os aspectos quantitativos (como taxa de cobertura e eficiência) quanto qualitativos, buscando compreender como a inclusão de incerteza explícita pode influenciar a tomada de decisão médica baseada em modelos preditivos. Essa etapa representa um avanço em relação à simples classificação binária, adicionando uma camada de interpretabilidade e segurança às predições automatizadas de sepse (Shashikumar et al., 2021).

## 3.3 Conjuntos de Dados Utilizados

Durante a fase de reprodução de estudos do estado da arte, foram selecionados e implementados quatro trabalhos relevantes da literatura, cada um utilizando diferentes bases de dados para o treinamento e avaliação de modelos de predição de sepse. A diversidade das fontes de dados entre esses estudos evidencia a fragmentação existente na área, o que reforça a importância da padronização proposta neste trabalho.

O primeiro estudo reproduzido foi o de Zhao; Shen; Wang (2021), que utilizou os dados do PhysioNet/Computing in Cardiology Challenge 2019 (Reyna et al., 2020). Essa base é pública e voltada especificamente para tarefas de predição de sepse, contendo dados temporais multivariados de pacientes internados, com anotações específicas para o evento de sepse, o que possibilitou sua reprodução local em condições próximas às originais descritas no estudo.

O segundo trabalho, Kam; Kim (2017), fez uso do MIMIC-II (v3) (Saeed et al., 2002), um banco de dados público que contém registros clínicos de pacientes internados em UTIs do Beth Israel Deaconess Medical Center. Apesar de ser uma versão

anterior ao MIMIC-III e MIMIC-IV, os dados estavam disponíveis e puderam ser processados conforme descrito no artigo, permitindo a replicação do modelo e de sua avaliação.

Os dois estudos restantes, Delahanty et al. (2019) e Zhang et al. (2021), utilizaram bases de dados clínicas privadas, sem acesso público. Dado que as bases originais desses estudos não estavam disponíveis para download ou requisição por meio de protocolo de acesso, as reproduções foram feitas exclusivamente utilizando o conjunto de dados padronizado desenvolvido neste trabalho. As metodologias originais foram seguidas com o máximo de fidelidade, considerando o fluxo de processamento de pré-processamento, escolha de variáveis e arquitetura dos modelos, adaptados para os dados disponíveis.

Além desses bancos utilizados na fase de reprodução, a base MIMIC-IV (Johnson et al., 2020), versão 2.2, foi utilizada como fonte para a construção do conjunto de dados padronizado proposto nesta dissertação. Esse conjunto padronizado serviu como base comum para executar os modelos de todos os estudos selecionados, permitindo uma comparação justa e controlada sob um mesmo cenário experimental.

Essa abordagem de múltiplos conjuntos de dados, tanto os originais (quando disponíveis) quanto o padronizado, foi fundamental para avaliar a reprodutibilidade dos estudos e investigar o impacto que diferentes fontes e estruturas de dados têm sobre o desempenho dos modelos de predição de sepse.

## 3.4 Métricas de Avaliação

A avaliação dos modelos propostos neste trabalho foi realizada com base em métricas amplamente reconhecidas na literatura científica, considerando tanto o desempenho preditivo quanto a capacidade de quantificar incertezas por meio da PC. Dado o contexto clínico da predição de sepse, um problema sensível ao tempo e com impacto direto em decisões médicas, é fundamental utilizar métricas que reflitam não apenas a acurácia global, mas também o comportamento dos modelos frente à distribuição real dos dados.

Inicialmente, todos os modelos foram comparados com base em quatro métricas principais: AUROC, *recall*, *F1-score* e acurácia. A acurácia fornece uma visão geral da proporção de acertos; a AUROC, amplamente utilizada na literatura, avalia a capacidade discriminativa dos modelos independentemente do limiar de decisão. O *F1-score* e o *recall* foram calculados tanto de forma macro (média entre as classes) quanto específica para a classe 1 (sepse), visando capturar o desempenho global e, especialmente, a eficácia dos modelos na identificação de casos positivos em um cenário clínico desbalanceado.

Além da avaliação tradicional, os modelos foram também analisados sob o pa-

radigma da PC, especificamente na sua forma transdutiva. Esse método possibilita estimar a incerteza associada a cada predição, gerando intervalos de confiança (IC) preditivos. Para cada nível de confiança (por exemplo, 90%, 95%), foram calculadas as mesmas métricas mencionadas anteriormente (AUROC, *recall*, *F1-score* e acurácia), restringindo a avaliação apenas às amostras para as quais o modelo gerou uma predição única (i.e., classe positiva ou negativa, excluindo os casos de indecisão).

Adicionalmente, foi analisado o número absoluto de amostras que ficaram fora do IC, ou seja, os casos em que o modelo se absteve de prever por não ter confiança suficiente. Essa métrica foi estratificada entre classes (casos de sepse e não-sepse), permitindo entender se o modelo apresenta maior incerteza em algum dos grupos e em que medida isso impacta a cobertura e a confiabilidade das predições. Também foram avaliadas as métricas de *coverage rate*, que indica a proporção de vezes em que o verdadeiro rótulo esteve contido no conjunto preditivo, e *set size*, que reflete o número médio de rótulos retornados por predição.

Essa abordagem combinada, que avalia tanto desempenho preditivo quanto a confiabilidade das decisões, oferece uma análise robusta e alinhada com as necessidades de aplicações em saúde, onde a confiança do modelo é tão importante quanto sua acurácia.

## 3.5 Ferramentas Utilizadas

Para a reprodução dos trabalhos e condução dos experimentos desta pesquisa, foi utilizado o ambiente *Jupyter Notebook*, com scripts escritos na linguagem `Python`, versão 3.12.7. A escolha dessa linguagem e ambiente se justifica por sua ampla aceitação na comunidade científica e pela flexibilidade no desenvolvimento de fluxos de processamento de AM.

A manipulação e o pré-processamento dos dados foram realizados utilizando as bibliotecas `pandas` e `numpy`, reconhecidas por sua eficiência em operações com dados tabulares e vetoriais. A construção e avaliação dos modelos foram conduzidas com o uso das bibliotecas `scikit-learn`, que oferece uma ampla gama de algoritmos e ferramentas de avaliação, e `PyTorch`, utilizada nos experimentos que envolveram redes neurais. As análises estatísticas básicas necessárias foram realizadas com recursos disponíveis na própria `scikit-learn`.

Para a visualização dos resultados e dados intermediários, foram utilizadas as bibliotecas `matplotlib` e `seaborn`, facilitando a geração de gráficos e representações gráficas dos modelos e métricas.

Os experimentos foram realizados em um computador pessoal com sistema operacional *Windows 11*, equipado com processador *Intel Core i7* de 10ª geração, pertencente à linha *Ideapad S145* da *Lenovo*. Esse ambiente foi suficiente para a execução

dos modelos propostos e reprodução dos estudos selecionados. O controle de versões foi mantido com `Git`, garantindo a rastreabilidade do código e dos experimentos ao longo do desenvolvimento.

Todo o código-fonte desenvolvido para esta pesquisa, incluindo scripts de pré-processamento, modelagem, avaliação e visualização, está disponível publicamente no seguinte repositório: `https://github.com/pedrowurzel/sepsis-prediction-framework`. A disponibilização visa assegurar a reprodutibilidade dos experimentos e facilitar o uso e extensão dos métodos aqui propostos por outros pesquisadores da área.

# 4 ANÁLISE CRÍTICA E EXPERIMENTOS DE REPRODUTIBILIDADE

## 4.1 Análise Crítica da Literatura

A literatura recente sobre o uso de AM e aprendizado profundo para a predição precoce de sepse apresenta avanços significativos, mas também revela lacunas metodológicas e desafios de padronização que dificultam a generalização dos resultados. A revisão sistemática conduzida por Islam et al. (2023) fornece uma visão consolidada desses desafios ao analisar 42 estudos publicados entre junho de 2016 e março de 2023, com foco exclusivo em pacientes adultos e no uso de dados provenientes de prontuários eletrônicos (EHRs).

### 4.1.1 Diversidade de Definições de Sepse e Janelas Temporais

Um dos achados mais relevantes da revisão de Islam et al. (2023) é a diversidade nas definições adotadas para o diagnóstico de sepse. Das 42 investigações incluídas, 22 utilizaram a definição Sepsis-3 (52,4%) e 14 adotaram a Sepsis-2 (33,3%), enquanto outras recorreram a códigos ICD-9, ICD-10 ou diagnósticos clínicos realizados por especialistas de UTI. Além disso, muitas dessas definições foram modificadas para se adequar à natureza dos dados disponíveis ou aos objetivos dos estudos, incluindo alterações nas janelas temporais utilizadas para caracterizar o início da sepse. Essa variabilidade metodológica, também observada nas revisões de Moor et al. (2021) e Deng et al. (2022), compromete a comparabilidade entre os estudos e dificulta a avaliação padronizada do desempenho dos modelos, uma vez que diferentes critérios diagnósticos influenciam diretamente a geração dos rótulos de saída.

Deng et al. destacam que a definição de sepse não foi considerada como critério de inclusão em muitos estudos, resultando em um cenário onde modelos distintos são treinados para identificar condições clínicas heterogêneas sob a mesma terminologia. Moor et al. corroboram essa constatação ao demonstrar que a prevalência de sepse variou de 3,3% a 63,6% entre os estudos analisados, refletindo populações-alvo e critérios diagnósticos bastante divergentes.

### 4.1.2 Falta de Padronização nas Bases de Dados, Variáveis e Avaliação

A análise conduzida por Islam et al. também evidencia a falta de padronização quanto às bases de dados utilizadas, às variáveis consideradas e às métricas de desempenho relatadas. Embora a base MIMIC-III tenha sido a mais utilizada (29,8%), diversos outros conjuntos de dados foram empregados, como o PhysioNet/CinC Challenge 2019, o Emory Healthcare System e bases institucionais específicas. Além disso, os dados foram extraídos de diferentes contextos hospitalares, como UTIs (81% dos estudos), departamentos de emergência e enfermarias gerais, o que contribui para uma grande heterogeneidade nos perfis dos pacientes.

A prevalência de sepse entre os estudos variou de 0,41% a 63,6%, com mediana de 9,5%, revelando o uso frequente de conjuntos de dados desequilibrados. Tal desequilíbrio exige o emprego de técnicas de balanceamento ou aumento de dados para garantir a robustez dos modelos preditivos. Essa realidade é compatível com as observações de Moor et al. (2021), que também relataram ampla variação nos critérios de inclusão, pré-processamento e proporção de casos positivos, mesmo em estudos que utilizaram a mesma base de dados.

No que diz respeito às variáveis utilizadas, os estudos revisados por Islam et al. (2023) utilizaram desde apenas dois até 168 atributos distintos, com mediana de 22. As variáveis mais frequentes foram sinais vitais, dados laboratoriais e características demográficas, sendo que apenas 21% dos estudos relataram utilizar modelos explicáveis (XAI). Apenas seis estudos disponibilizaram seus códigos ou fluxos de processamento de pré-processamento, e menos de 20% validaram seus modelos com dados externos, limitando significativamente a generalização dos resultados.

### 4.1.3 Métricas de Avaliação e Qualidade dos Estudos

Embora a maioria dos estudos tenha utilizado métricas como AUROC (29,8%), sensibilidade (19,8%) e especificidade (21,5%), a revisão de Islam et al. (2023) alerta para os riscos de se comparar diretamente os desempenhos relatados, dada a ampla variabilidade nos conjuntos de dados e metodologias adotadas. Os valores de AUROC variaram de 0,80 a 0,97, indicando desempenho potencialmente promissor, mas não conclusivo. Assim como argumentado por Moor et al. (2021), uma análise crítica dessas métricas só é válida se houver padronização prévia dos critérios diagnósticos e do momento de previsão da sepse.

A avaliação de qualidade realizada por Islam et al. (2023) reforça essa crítica: nenhum dos estudos cumpriu integralmente os 16 critérios de qualidade estabelecidos pelos autores, sendo que apenas cinco foram classificados como de alta qualidade. A maioria relatou o uso de modelos de AM, prevalência de sepse e técnicas de engenharia de atributos, mas poucos estudos discutiram a aplicabilidade clínica de suas abordagens ou realizaram validação em cenários reais. Esse achado é consistente

com os apontamentos de Deng et al. (2022), que destacam a falta de justificativa para as divisões entre treino, validação e teste, bem como a escassez de estudos prospectivos.

### 4.1.4 Considerações Finais

A literatura sobre predição precoce de sepse por meio de AM tem avançado significativamente nos últimos anos, mas continua marcada por lacunas metodológicas substanciais. A revisão de Islam et al. (2023), ao reunir evidências de 42 estudos recentes, destaca a necessidade urgente de padronização nos critérios diagnósticos, nas bases de dados, nas variáveis utilizadas e nas métricas de avaliação. As revisões anteriores de Moor et al. (2021) e Deng et al. (2022) reforçam esse diagnóstico, indicando que sem esforços coordenados para harmonizar esses aspectos, será difícil consolidar modelos realmente aplicáveis à prática clínica. Assim, futuras investigações devem priorizar transparência metodológica, validação externa e avaliação do impacto clínico real dos modelos desenvolvidos.

## 4.2 Experimentos de Reprodutibilidade

### 4.2.1 Critérios de Seleção dos Estudos

Para a realização dos experimentos de reprodutibilidade, foram selecionados trabalhos da área de AM aplicados à predição de sepse que atendessem aos seguintes critérios:

1. apresentassem resultados expressivos em termos de desempenho preditivo;
2. descrevessem com clareza os dados utilizados e as etapas de pré-processamento;
3. disponibilizassem detalhadamente os procedimentos metodológicos; e
4. tivessem relevância acadêmica reconhecida, considerando o número de citações, publicações em periódicos ou conferências relevantes, e/ou impacto da proposta metodológica.

### 4.2.2 Panorama Geral das Reproduções

Ao todo, foram analisados seis trabalhos distintos, abrangendo diferentes abordagens e bases de dados. Entre os estudos avaliados, três puderam ser reproduzidos com sucesso utilizando os dados originais ou padronizados desenvolvidos neste projeto. Dois tiveram suas reproduções descontinuadas devido a falhas descritivas ou indisponibilidade dos dados. Um estudo foi parcialmente reproduzido com base em

premissas alternativas. A seguir, são descritas individualmente as tentativas de reprodução, com ênfase nas dificuldades encontradas e nos motivos de sucesso ou insucesso.

### 4.2.3 Estudo de Zhao; Shen; Wang (2021)

O trabalho apresenta três abordagens distintas para predição de sepse: *mean processing*, *mean processing improved* e *feature generation*, todas implementadas com os algoritmos `XGBoost` e `LightGBM`. A descrição das etapas de pré-processamento é bastante detalhada, com uso da técnica `Miceforest` para imputação de valores ausentes e aplicação da base PhysioNet/Computing in Cardiology Challenge 2019.

Identificou-se uma inconsistência metodológica relevante: apesar de o objetivo do modelo ser a predição da sepse, os autores incluíram amostras do período de sepse (e não apenas do período pré-sepse) como rótulos positivos. Apesar dessa limitação conceitual, a reprodução do estudo foi possível com sucesso, tanto com a base original quanto com o conjunto de dados padronizado desenvolvido nesta dissertação.

### 4.2.4 Estudo de Zhang et al. (2021)

O estudo utiliza um banco de dados não público (2019 DII Challenge), o que inviabilizou o uso da base original. No entanto, foi possível realizar a reprodução com o conjunto de dados padronizado, dado que este apresentava variáveis fisiológicas e laboratoriais semelhantes, bem como frequência de aquisição compatível.

O trabalho é bem descrito metodologicamente, e a reprodução com a base alternativa foi bem-sucedida.

### 4.2.5 Estudo de Rafiei et al. (2021)

Apesar de utilizar a base pública PhysioNet/CinC 2019, a reprodução foi descontinuada devido a diversas ambiguidades metodológicas. A descrição do uso de janelas temporais de 4h, 8h e 12h é inconsistente, sem esclarecimento se tais janelas representam períodos fixos de coleta ou janelas retroativas à ocorrência da sepse.

Além disso, técnicas como *window slicing* e *noise injection* são mencionadas, mas não há qualquer detalhamento sobre sua parametrização, implementação ou impacto na composição das amostras. Tais lacunas impossibilitaram a reprodução fiel do experimento.

### 4.2.6 Estudo de Kamaleswaran et al. (2021)

Este trabalho destaca-se por sua ênfase na interpretabilidade dos modelos. No entanto, a base de dados utilizada é de acesso restrito, coletada em ambiente hospitalar pelos próprios autores. Embora tenha sido considerada a possibilidade de reprodução com a base padronizada, diferenças significativas em termos de frequência de

coleta e densidade dos dados inviabilizaram uma comparação justa. Por esse motivo, a reprodução foi interrompida.

### 4.2.7 Estudo de Kam; Kim (2017)

Utilizando a base MIMIC-II, este trabalho apresenta inconsistências na descrição dos critérios de inclusão, especialmente em relação ao uso dos critérios *SIRS* nas primeiras horas de admissão. Divergências nos números de pacientes identificados sugerem possível erro ou omissão na descrição dos filtros utilizados. Além disso, não são incluídos códigos importantes, como choque séptico ou infecções septicêmicas, na seleção de casos.

Diante disso, optou-se por seguir o protocolo descrito por Calvert et al. (2016), citado pelos próprios autores como referência. Com essa abordagem, a reprodução foi realizada com sucesso, utilizando tanto a base MIMIC-II quanto a base padronizada.

### 4.2.8 Estudo de Delahanty et al. (2019)

A base utilizada pelos autores pertence à Tenet Healthcare e não está disponível publicamente. Ainda assim, foi possível realizar a reprodução com a base padronizada, cujas variáveis e frequência de coleta são comparáveis.

O trabalho apresenta excelente descrição metodológica, o que possibilitou sua reprodução com sucesso, mesmo sem o acesso à base original.

## 4.3 Impacto na Comparação Entre Métodos

A avaliação comparativa entre diferentes técnicas de predição de sepse é essencial para o avanço científico na área, especialmente diante da complexidade clínica envolvida. No entanto, os problemas encontrados durante a tentativa de reprodução dos trabalhos selecionados revelam obstáculos significativos à condução de comparações justas e objetivas.

Primeiramente, a indisponibilidade de dados utilizados em diversos estudos compromete diretamente a possibilidade de replicar experimentos com fidelidade. A utilização de bases de dados proprietárias ou não públicas, como observado nos trabalhos de Kamaleswaran et al. (2021) e Delahanty et al. (2019), limita a reprodutibilidade e impede que resultados obtidos possam ser validados ou confrontados por terceiros. Mesmo nos casos em que se utilizou uma base alternativa, como o conjunto de dados padronizado desenvolvido nesta dissertação, permanece uma incerteza quanto à equivalência real das distribuições de dados, à densidade temporal das variáveis e à natureza das populações estudadas. Essa disparidade afeta diretamente os resultados dos modelos e compromete a comparabilidade.

Além disso, muitos trabalhos apresentam lacunas na descrição de seus métodos,

como critérios de inclusão de pacientes, definições de janelas temporais e parâmetros utilizados em técnicas de aumento de dados (data augmentation). Tais imprecisões foram particularmente evidentes nos estudos de Rafiei et al. (2021) e Kam; Kim (2017). Sem um entendimento completo e preciso dessas etapas, é impossível assegurar que os modelos reproduzidos operam sob as mesmas condições dos originais, comprometendo a validade da comparação entre abordagens.

Outro aspecto crítico diz respeito ao próprio objetivo de predição. Como observado no trabalho de Zhao; Shen; Wang (2021), a utilização de amostras do período de sepse (e não apenas do período pré-séptico) como rótulos positivos introduz viés na tarefa preditiva, pois o modelo pode estar, de fato, aprendendo a reconhecer pacientes já diagnosticados, e não antecipando a ocorrência de sepse. Esse tipo de incongruência metodológica afeta diretamente os resultados obtidos e distorce a comparação com outros métodos que seguem uma definição mais estrita do problema.

Portanto, as falhas na documentação, os critérios inconsistentes e as limitações de acesso aos dados dificultam a replicação fiel dos experimentos e introduzem variáveis de confusão na comparação de desempenho entre diferentes técnicas. Em um cenário ideal, todos os estudos seriam reproduzidos sob condições padronizadas e com acesso total às bases de dados e configurações experimentais. A ausência desses elementos compromete a confiabilidade das análises comparativas e evidencia a necessidade de maior rigor metodológico e transparência na publicação de estudos científicos na área de predição clínica.

# 5 PROPOSTA E METODOLOGIA PARA UMA AVALIAÇÃO PADRONIZADA

A análise crítica realizada nos capítulos anteriores revelou uma série de desafios que comprometem a reprodutibilidade, a comparabilidade e a confiabilidade dos estudos existentes sobre predição de sepse com técnicas de AM. A ausência de dados acessíveis, a descrição incompleta de métodos e pré-processamentos, e as inconsistências nas definições clínicas e critérios de seleção de amostras dificultam sobremaneira qualquer tentativa de replicação ou comparação objetiva entre os modelos propostos na literatura. Mesmo os estudos mais promissores, ao serem submetidos à tentativa de reprodução, demonstraram fragilidades metodológicas que afetam a validade de seus resultados.

Nesse contexto, este capítulo surge como uma resposta direta às lacunas identificadas. Dando continuidade ao objetivo principal deste trabalho, que é promover uma avaliação crítica do estado da arte e propor uma abordagem padronizada para a predição de sepse, apresenta-se aqui uma estrutura sistemática e coerente de padronização. A proposta visa estabelecer diretrizes claras quanto à seleção de pacientes, definição de sepse, manipulação de dados, métricas de avaliação e outros aspectos fundamentais do fluxo de processamento de modelagem.

Mais do que uma simples recomendação metodológica, a padronização proposta neste capítulo busca oferecer uma base sólida para experimentos futuros, permitindo que diferentes abordagens possam ser avaliadas sob condições equivalentes. Com isso, pretende-se não apenas facilitar a reprodutibilidade dos estudos, mas também viabilizar comparações justas entre métodos, contribuindo de forma significativa para o avanço científico da área.

## 5.1 Criação de um Conjunto de Dados Padronizado

Com base nas limitações identificadas na literatura e na necessidade de um ambiente experimental unificado para a comparação justa entre modelos de predição de sepse, foi desenvolvido um conjunto de dados padronizado a partir do banco de dados

MIMIC-IV, versão 2.2. O MIMIC-IV é atualmente a versão mais recente do MIMIC e é amplamente reconhecido como a principal fonte de dados clínicos abertos na área de pesquisa em predição de eventos críticos, como sepse. Sua ampla adoção na comunidade científica, como evidenciado por Islam et al. (2023), aliada à riqueza e granularidade dos dados disponíveis, torna-o uma escolha apropriada para construção de um conjunto de dados robusto e representativo.

A definição de sepse adotada neste trabalho seguiu os critérios estabelecidos pela Sepsis-3, a definição atualmente aceita e recomendada por especialistas na área (**?**). Para a identificação dos pacientes com sepse segundo essa definição, foi utilizada a implementação oficial disponibilizada pela equipe mantenedora do MIMIC, através de scripts públicos acessíveis no repositório do GitHub[1]. Esses scripts realizam a detecção dos casos a partir da combinação entre sinais de infecção suspeita e disfunção orgânica, operacionalizados pelo escore SOFA.

O momento de início (*onset*) da sepse foi definido como o menor tempo entre os registros de `suspected_infection_time` e `sofa_time`. Essa escolha se justifica pela intenção de garantir maior sensibilidade na identificação do momento crítico de transição clínica. Considerando que tanto a suspeita de infecção quanto a disfunção orgânica podem ser detectadas em momentos distintos da internação, adotar o menor tempo entre os dois garante que o modelo aprenda a prever a sepse com base em informações disponíveis antes que a condição esteja plenamente estabelecida.

Para a definição dos pacientes sem sepse, foram considerados todos os indivíduos que não apresentaram nenhum código diagnóstico associado à condição séptica em seus registros. Os códigos utilizados para essa filtragem abrangeram três versões da Classificação Internacional de Doenças (CID), a fim de capturar possíveis variações entre diferentes sistemas de codificação. Os códigos empregados foram:

- **CID-9:**
  - 038: Septicemia
  - 995.9: Sepse não especificada
  - 785.52: Choque séptico
- **CID-10:**
  - A40: Septicemia estreptocócica
  - A41: Outras septicemias
  - P36: Septicemia neonatal
  - R65.2: Sepse grave e choque séptico

[1] `https://github.com/MIT-LCP/mimic-code/tree/main/mimic-iv/concepts_postgres`

- **CID-11:**
  - 1G4: Sepse
  - 293771399: Códigos relacionados à sepse (mapeamento específico da versão CID-11)

A justificativa para esse filtro está em assegurar que os pacientes identificados como “sem sepse” não apresentem registros diagnósticos que possam comprometer a clareza da separação entre as classes positiva e negativa do problema de classificação. Essa separação é fundamental para garantir a qualidade dos dados de treinamento e evitar a introdução de ruído nas análises.

Além disso, foram considerados apenas pacientes com idade igual ou superior a 18 anos, de modo a restringir a análise ao público adulto e evitar a heterogeneidade clínica entre faixas etárias. Cada internação hospitalar foi tratada como uma instância independente, respeitando o caráter episódico dos dados clínicos.

Com esse processo de filtragem, foram identificados 32.970 pacientes com sepse e 35.259 pacientes sem sepse.

A construção do conjunto de dados seguiu princípios consistentes com o objetivo de prever a ocorrência de sepse antes de sua manifestação clínica. Para os pacientes com sepse, foram extraídos dados de sinais vitais, exames laboratoriais e outras variáveis relevantes desde o momento da admissão até o instante do *onset* da sepse, excluindo-se explicitamente qualquer dado posterior a esse ponto. Para os pacientes sem sepse, foi gerado aleatoriamente um ponto de corte dentro da internação hospitalar e os dados considerados foram aqueles entre a admissão e esse ponto. Essa abordagem visa simular realisticamente o cenário clínico de tomada de decisão, onde a sepse ainda não se manifestou e a tarefa do modelo é identificar seu risco iminente.

Demais aspectos do fluxo de processamento de modelagem, como as janelas temporais utilizadas, métodos de imputação de dados nulos, seleção de variáveis, engenharia de atributos e normalização, foram mantidos conforme descritos em cada um dos trabalhos analisados durante as tentativas de reprodução, de modo a preservar a comparabilidade e permitir avaliações controladas.

## 5.2 Comparação de Métodos em Condições Controladas

Após a construção do conjunto de dados padronizado com base no MIMIC-IV v2.2 e na definição clínica de sepse segundo os critérios da Sepsis-3, foram conduzidos experimentos controlados com o objetivo de avaliar, de forma justa e reproduzível, diferentes técnicas de predição identificadas na literatura. A padronização dos dados e do fluxo de processamento experimental possibilitou a eliminação de variáveis metodológicas externas, permitindo focar na avaliação comparativa real do desempenho

das técnicas.

### 5.2.1 Aplicação das Técnicas do Estado da Arte

Foram reimplementadas sobre o conjunto de dados padronizado as principais abordagens de AM utilizadas nos estudos reproduzidos, incluindo modelos como XGBoost e LightGBM, e redes neurais profundas, como LSTM. Cada modelo foi treinado e avaliado sob as mesmas condições de entrada, utilizando o mesmo conjunto de amostras.

Para assegurar uma comparação objetiva entre os modelos avaliados, foram utilizadas quatro métricas principais de desempenho: acurácia, AUROC, *recall* e *F1-score*. Cada uma dessas métricas oferece uma perspectiva complementar sobre a performance dos modelos, especialmente em cenários clínicos marcados por desbalanceamento entre as classes:

- **AUROC**: avalia a capacidade do modelo de distinguir entre pacientes com e sem sepse ao longo de todos os thresholds possíveis;
- ***Recall***: mede a proporção de casos de sepse corretamente identificados (verdadeiros positivos em relação ao total de positivos reais). Essa métrica é crítica em contextos de predição precoce, pois quantifica a sensibilidade do modelo à detecção da condição-alvo. No presente trabalho, o *recall* será avaliado tanto de forma macro (média entre as classes) quanto especificamente para a classe 1 (sepse), refletindo a prioridade clínica em minimizar falsos negativos;
- ***F1-score***: representa o equilíbrio entre precisão e *recall*, sendo particularmente útil quando há desbalanceamento entre as classes. Assim como o *recall*, será avaliado em sua versão macro e focada na classe positiva, permitindo entender tanto o desempenho global quanto a efetividade do modelo na detecção da sepse;
- **Acurácia**: indica a proporção total de predições corretas entre todas as instâncias avaliadas. Embora amplamente utilizada, essa métrica pode ser enganosa em problemas com classes desbalanceadas, motivo pelo qual é usada aqui de forma complementar às demais.

Essas métricas foram escolhidas por refletirem não apenas o desempenho global do modelo, mas também sua eficácia prática em um ambiente clínico, onde a sensibilidade e o equilíbrio entre erro tipo I e tipo II são cruciais para decisões médicas seguras.

### 5.2.2 Variação de Parâmetros e Condições Experimentais

Além da comparação direta entre modelos, também foram conduzidos experimentos com variação da antecedência da predição, com o objetivo de avaliar a robustez

e a adaptabilidade das abordagens em diferentes contextos clínicos. Foram testadas predições realizadas com 24 horas, 12 horas, 6 horas e no momento do *onset* (0h), simulando cenários com diferentes níveis de antecipação na tomada de decisão médica.

Essa variação permite observar como o desempenho dos modelos se comporta à medida que a janela temporal para previsão se aproxima do evento crítico, refletindo diretamente na sua utilidade clínica em contextos de maior ou menor urgência. Com isso, tornou-se possível avaliar não apenas o desempenho absoluto dos modelos, mas também sua consistência temporal, identificando quais técnicas mantêm desempenho estável mesmo quando a tarefa de predição se torna mais desafiadora.

Os resultados desses experimentos, detalhados no capítulo seguinte, fornecem subsídios concretos para a seleção de modelos que não apenas apresentem altos índices de desempenho, mas que também sejam viáveis, robustos e confiáveis para aplicação prática em ambientes hospitalares reais.

## 5.3 Aplicação de PC

### 5.3.1 Definição e Justificativa para o Uso

A última etapa dos experimentos consistiu na aplicação de técnicas de PC, com o objetivo de incorporar estimativas explícitas de confiabilidade às predições realizadas pelos modelos de AM. Diferentemente das abordagens tradicionais que retornam uma única classe com uma probabilidade associada, a PC fornece conjuntos de predição com níveis de confiança estatística controláveis, permitindo interpretar com maior segurança os resultados do modelo, aspecto crucial em contextos clínicos de alta sensibilidade, como a predição de sepse.

Neste trabalho, foi adotada a abordagem PCT, em detrimento de variantes como o Predição Conformal Indutiva (PCI), devido à sua robustez estatística e às garantias teóricas válidas mesmo sob suposições fracas, como *exchangeability* (Vovk; Gammerman; Shafer, 2005). O PCT realiza a predição individualmente para cada instância de teste, utilizando diretamente o conjunto de calibração, o que permite gerar ICs (ou conjuntos preditivos) adaptados a cada exemplo. Essa característica fornece controle formal sobre o erro condicional e garante que, com uma probabilidade de pelo menos $1 - \varepsilon$, o verdadeiro rótulo estará contido no conjunto predito, independentemente da complexidade do modelo ou da distribuição dos dados.

A escolha pelo PCT se justifica também pelo fato de que, em ambientes clínicos, o risco associado a decisões automatizadas é significativo, e a transparência quanto à incerteza pode apoiar decisões mais seguras por parte dos profissionais de saúde. Como reforçado por Shafer e Vovk (Vovk; Gammerman; Shafer, 2005), a PC é generalizável a qualquer modelo subjacente e não depende de premissas como independên-

cia ou normalidade, o que o torna especialmente adequado para dados clínicos reais, frequentemente ruidosos, incompletos e heterogêneos.

Nos experimentos deste trabalho, o PCT foi aplicado sobre os modelos com melhor desempenho identificados anteriormente, mantendo-se o mesmo conjunto de dados padronizado e configurações experimentais. Foram testados diferentes níveis de significância ($\varepsilon = 0{,}15$, $0{,}1$ e $0{,}05$), permitindo avaliar a relação entre confiança estatística, taxa de cobertura e largura dos conjuntos preditivos. Essa avaliação incluiu tanto métricas quantitativas (como *coverage rate* e *set size*) quanto implicações qualitativas para o uso clínico.

A escolha dos níveis de significância foi orientada por critérios estatísticos e pela necessidade de avaliar diferentes graus de confiança nos conjuntos preditivos gerados. Esses valores correspondem, respectivamente, a níveis de confiança de 85%, 90% e 95%, amplamente utilizados em aplicações clínicas e estatísticas. Essa variação permitiu investigar o *trade-off* entre a largura dos conjuntos preditivos e a taxa de cobertura, fator essencial em cenários de decisão clínica sensível, como o diagnóstico de sepse. Conforme argumentado por Vovk e Shafer (Vovk; Gammerman; Shafer, 2005), o parâmetro $\varepsilon$ é definido pelo usuário e determina diretamente a taxa de erro esperada do preditor conformal, sendo que valores menores garantem maior confiabilidade estatística, ao custo de conjuntos possivelmente mais amplos.

De forma complementar, Angelopoulos e Bates (Angelopoulos; Bates, 2021) destacam que valores como 0,05 e 0,10 são comumente empregados na prática para estabelecer limites superiores de incerteza em aplicações sensíveis, recomendando ainda a experimentação com múltiplos níveis de significância como uma forma de avaliar o impacto da confiança na utilidade dos conjuntos preditivos. Assim, a adoção de diferentes valores de $\varepsilon$ neste trabalho visa fornecer uma análise abrangente do desempenho do PCT, considerando tanto métricas quantitativas quanto implicações qualitativas para o apoio à decisão médica.

### 5.3.2 Comparação com Outras Abordagens

A adoção do PC, especialmente em sua variante transdutiva, representa um avanço significativo em relação aos classificadores probabilísticos tradicionais, que, embora forneçam estimativas de confiança na forma de scores ou probabilidades, não oferecem garantias estatísticas formais sobre sua acurácia. Diversos estudos demonstram que esses modelos, como redes neurais profundas, árvores de decisão e algoritmos de *boosting*, frequentemente produzem predições mal calibradas, isto é, as probabilidades atribuídas não refletem de forma confiável a frequência real dos eventos (Guo et al., 2017; Niculescu-mizil; Caruana, 2005).

Em contrapartida, o PCT fornece conjuntos preditivos com validade estatística garantida, mesmo sob condições mínimas, como a suposição de *exchangeability*, e man-

tém essa propriedade em cenários *online* ou em tempo real. Tal capacidade é particularmente valiosa em contextos clínicos, como a predição de sepse, nos quais a confiabilidade de cada decisão automatizada pode ter implicações diretas na segurança do paciente e na condução do tratamento médico.

O trabalho de Yang et al. (2024) ilustra claramente o potencial da PC em cenários clínicos reais. Ao desenvolver o modelo CPMORS para predição do risco de mortalidade por sepse, os autores demonstraram que o uso do PC não apenas reduz significativamente a taxa de erro, como também sinaliza automaticamente previsões incertas. Esses casos foram associados a maior gravidade clínica e maior tempo de internação, indicando que a PC pode funcionar como um mecanismo de alerta adicional. Ademais, ao integrar valores de SHAP para interpretação das variáveis, o modelo tornou-se mais transparente e confiável para os profissionais de saúde.

O estudo de Shashikumar et al. (2021) reforça o potencial da PC como ferramenta essencial para aumentar a segurança e a confiabilidade de sistemas de inteligência artificial em ambientes clínicos. Nesse trabalho, a técnica foi empregada para detectar amostras fora da distribuição esperada, ou seja, casos cuja estrutura de dados divergia substancialmente do conjunto de treinamento, situações que podem ocorrer devido a *data drift*, registros incompletos ou variações no perfil populacional. Nessas condições, a abordagem conformal adotada permitiu que o sistema classificasse tais instâncias como “indeterminadas”, evitando a emissão de predições potencialmente errôneas. Essa funcionalidade teve impacto direto na prática clínica: ao se abster de prever quando a confiança não era estatisticamente garantida, o modelo contribuiu para uma redução expressiva de alarmes falsos e possibilitou a integração segura com fluxos de decisão médica.

A capacidade de sinalizar quando o modelo não deve se pronunciar destaca um dos principais diferenciais da PC frente a classificadores tradicionais, que, mesmo diante de incertezas, ainda retornam uma única decisão. Assim, o estudo demonstra como a PC pode ser decisiva em contextos críticos, como a sepse, oferecendo uma camada adicional de controle, interpretabilidade e segurança no uso clínico de algoritmos preditivos.

Esse tipo de interpretação, onde a incerteza é explicitamente representada e associada a sinais clínicos objetivos, constitui um diferencial importante em relação às abordagens binárias tradicionais. Ao permitir que o modelo retorne conjuntos como sepse, não sepse (previsão ambígua) ou conjuntos vazios (modelo não confiável), o PCT sinaliza os limites de sua própria competência, facultando que decisões críticas sejam delegadas ao julgamento clínico humano quando necessário.

Por outro lado, conforme discutido por Papadopoulos; Vovk; Gammerman (2007), abordagens como o PCI surgem como alternativas viáveis quando há restrições computacionais, especialmente ao empregar redes neurais profundas. No entanto, a flexi-

bilidade computacional do PCI vem ao custo de uma menor sensibilidade à estrutura específica do conjunto de teste, pois ele não é reavaliado ponto a ponto, como ocorre no PCT. Neste estudo, considerando que os modelos base subjacentes são relativamente leves e que a prioridade é a validade estatística individual, o PCT mostrou-se mais adequado.

Portanto, a incorporação da PC na variante PCT agrega uma camada de interpretabilidade e segurança operacional essencial para o uso de modelos preditivos em ambientes hospitalares. Essa abordagem promove uma transição da simples acurácia para a confiança quantificada, fornecendo aos profissionais de saúde não apenas o "quê" prever, mas também o "quão confiável" é essa previsão, permitindo um uso mais ético, transparente e eficaz da inteligência artificial na prática clínica.

# 6 RESULTADOS E DISCUSSÃO

Este capítulo apresenta e discute os principais resultados obtidos ao longo dos experimentos, organizados em três frentes complementares: (i) a reprodutibilidade de estudos do estado da arte, (ii) a comparação entre modelos em um ambiente padronizado e (iii) a avaliação da PC como forma de quantificar a incerteza. Cada uma dessas seções explora diferentes aspectos da pesquisa, relacionando os resultados aos objetivos propostos no trabalho.

A análise é feita de forma crítica, evidenciando tanto os desempenhos alcançados quanto as limitações metodológicas dos estudos reproduzidos. Também são discutidos os efeitos da padronização experimental na comparação entre modelos e o valor prático de previsões probabilísticas acompanhadas de medidas de confiança.

## 6.1 Resultados da Reprodutibilidade dos Estudos

Esta seção apresenta os resultados obtidos a partir da tentativa de reprodução dos principais trabalhos do estado da arte selecionados para este estudo. O objetivo foi verificar, na prática, o quanto os modelos publicados na literatura são efetivamente reprodutíveis quando reimplementados a partir das descrições fornecidas nos artigos originais, utilizando os mesmos bancos de dados.

A Tabela 1 apresenta a comparação entre os valores reportados nos artigos originais e os resultados obtidos durante as reproduções. Dois trabalhos atenderam aos critérios de reprodução completos: Zhao; Shen; Wang (2021) e Kam; Kim (2017). Ambos utilizaram bases públicas e forneceram detalhes suficientes para replicação dos experimentos com razoável fidelidade.

Tabela 1 – Comparação entre os resultados originais e reproduzidos com o mesmo conjunto de dados

| Métrica | Zhao; Shen; Wang (2021) | Kam; Kim (2017) |
|---|---|---|
| AUROC (Orig.) | 0.98 | 0.93 |
| AUROC (Reprod.) | 0.97 | 0.83 |
| Recall Classe 1 (Orig.) | 0.65 | 0.91 |
| Recall Classe 1 (Reprod.) | 0.56 | 0.68 |
| F1 Classe 1 (Orig.) | 0.72 | - |
| F1 Classe 1 (Reprod.) | 0.69 | 0.71 |
| Acurácia (Orig.) | 0.93 | 0.93 |
| Acurácia (Reprod.) | 0.92 | 0.72 |

*Nota:* Resultados com fundo cinza referem-se aos valores originalmente reportados nos artigos, enquanto os demais correspondem aos obtidos nas reproduções utilizando o mesmo conjunto de dados. A métrica “Classe 1” refere-se à classe positiva (sepse). Valores ausentes indicados por “–”.

O modelo de Zhao; Shen; Wang (2021) foi originalmente treinado para realizar a predição de sepse com 6 horas de antecedência, e seus resultados originais reportaram um AUROC de 0,98 e *F1-score* de 0,72 para a classe positiva (sepse). Na reprodução, os resultados se mantiveram bastante próximos, com um AUROC de 0,97 e *F1-score* de 0,69 para a mesma classe. O *recall* da classe 1 foi levemente inferior (0,56 contra 0,65), mas a consistência geral dos resultados reforça a boa qualidade da descrição metodológica do trabalho, além de demonstrar que o modelo possui desempenho robusto mesmo quando executado em ambiente externo ao estudo original. A acurácia geral permaneceu alta em ambas as execuções, superior a 0,91.

No caso do trabalho de Kam; Kim (2017), a predição foi realizada com antecedência de 3 horas, e o estudo original reportou desempenho bastante elevado, com AUROC de 0,93 e *recall* de 0,91 para a classe de sepse. No entanto, como discutido na seção 4.2.7, a reprodução enfrentou inconsistências nos critérios de seleção de pacientes e na aplicação dos filtros descritos pelos autores, o que levou à formação de um conjunto de teste muito reduzido, com apenas 50 amostras. Essa limitação pode ter comprometido a validade estatística da reprodução, tornando os resultados mais sensíveis a variações pontuais e ruído. Ainda assim, a reprodução obteve AUROC de 0,83 e *recall* de 0,68 para a classe positiva, valores razoáveis, porém consideravelmente inferiores aos relatados originalmente. A diferença pode indicar que parte do bom desempenho reportado no artigo original pode depender de escolhas específicas de segmentação dos dados que não foram suficientemente detalhadas para reprodução fidedigna.

A comparação entre os dois estudos também ilustra o impacto da antecedência da predição no desempenho dos modelos. Enquanto Zhao; Shen; Wang (2021) obtiveram bons resultados com uma janela de 6 horas, Kam; Kim (2017) trabalharam com apenas 3 horas de antecedência, o que, em tese, deveria facilitar a tarefa preditiva (já que mais próximos do evento-alvo). No entanto, o desempenho inferior na reprodução pode sugerir que o modelo de Kam; Kim (2017) seja mais sensível às configurações específicas de amostragem ou à definição operacional de sepse utilizada.

É importante destacar a diferença observada entre os valores de *recall* e *F1-score* avaliados de forma macro e especificamente para a classe 1 (sepse). Em ambos os estudos reproduzidos, os valores macro foram superiores, o que pode ser atribuído ao desbalanceamento entre as classes: como a maioria dos pacientes não apresenta sepse, os modelos tendem a ter desempenho melhor na classe majoritária, elevando artificialmente as médias.

Em síntese, os resultados indicam que, embora seja possível reproduzir parcialmente os experimentos originais, pequenas diferenças nos dados e falta de padronização na descrição dos métodos podem comprometer a comparabilidade entre estudos. O trabalho de Zhao; Shen; Wang (2021) destaca-se positivamente por sua transparência metodológica e reprodutibilidade, ao passo que o de Kam; Kim (2017) exemplifica os desafios práticos enfrentados ao tentar replicar estudos que dependem de filtros clínicos complexos e pouco detalhados.

## 6.2 Comparação de Modelos no Conjunto de Dados Padronizado

O objetivo desta etapa foi avaliar a robustez dos modelos quando aplicados fora do contexto original, em um ambiente controlado e comparável, o que possibilita uma análise mais justa entre técnicas distintas. A Tabela 2 apresenta os resultados da reprodução dos estudos selecionados utilizando o conjunto de dados padronizado construído com base no MIMIC-IV v2.2, sob condições experimentais homogêneas.

De forma geral, observa-se que todos os modelos mantiveram desempenho competitivo mesmo com a mudança de base de dados, embora com variações relevantes em algumas métricas, principalmente nas associadas à classe positiva (sepse). Zhao; Shen; Wang (2021), originalmente treinado com janelas de 6 horas antes do *onset*, a reprodução com o conjunto de dados padronizado resultou em queda no AUROC (de 0,98 para 0,92), mas um aumento significativo no *recall* da classe 1 (de 0,65 para 0,73), sugerindo um modelo mais sensível, ainda que com leve perda na discriminação. O *F1-score* da classe 1 também melhorou (de 0,72 para 0,79), indicando um desempenho mais equilibrado entre precisão e sensibilidade no novo contexto.

O modelo de Kam; Kim (2017), com predição 3 horas antes do *onset*, também manteve bons resultados após a adaptação. O AUROC caiu de 0,93 para 0,89, mas

Tabela 2 – Comparação entre resultados originais e reprodução padronizada

| Trabalho | AUC | R-M | R-1 | F1-M | F1-1 | Acc. |
|---|---|---|---|---|---|---|
| Zhao; Shen; Wang (2021) (Orig.) | 0.98 | - | 0.65 | - | 0.72 | 0.93 |
| Zhao; Shen; Wang (2021) (Repr.) | 0.92 | 0.85 | 0.73 | 0.87 | 0.79 | 0.91 |
| Kam; Kim (2017) (Orig.) | 0.93 | - | 0.91 | - | - | 0.93 |
| Kam; Kim (2017) (Repr.) | 0.89 | 0.83 | 0.72 | 0.83 | 0.81 | 0.83 |
| Zhang et al. (2021) (Orig.) | 0.94 | - | - | - | - | - |
| Zhang et al. (2021) (Repr.) | 0.90 | 0.79 | 0.60 | 0.82 | 0.69 | 0.91 |
| Delahanty et al. (2019) (1h) (Orig.) | 0.93 | - | 0.68 | - | - | - |
| Delahanty et al. (2019) (1h) (Repr.) | 0.93 | 0.85 | 0.74 | 0.87 | 0.81 | 0.90 |
| Delahanty et al. (2019) (3h) (Orig.) | 0.95 | - | 0.72 | - | - | - |
| Delahanty et al. (2019) (3h) (Repr.) | 0.94 | 0.87 | 0.76 | 0.90 | 0.84 | 0.93 |
| Delahanty et al. (2019) (6h) (Orig.) | 0.96 | - | 0.75 | - | - | - |
| Delahanty et al. (2019) (6h) (Repr.) | 0.93 | 0.87 | 0.76 | 0.89 | 0.83 | 0.93 |
| Delahanty et al. (2019) (12h) (Orig.) | 0.97 | - | 0.79 | - | - | - |
| Delahanty et al. (2019) (12h) (Repr.) | 0.93 | 0.86 | 0.73 | 0.89 | 0.82 | 0.94 |
| Delahanty et al. (2019) (24h) (Orig.) | 0.97 | - | 0.85 | - | - | - |
| Delahanty et al. (2019) (24h) (Repr.) | 0.93 | 0.86 | 0.73 | 0.89 | 0.81 | 0.95 |

Notas: AUC = AUROC, R-M = *Recall* Macro, R-1 = *Recall* Classe 1, F1-M = *F1-score* Macro, F1-1 = *F1-score* Classe 1, Acc. = Acurácia. Valores originalmente em porcentagem foram convertidos para formato decimal com duas casas decimais. “Orig.” refere-se aos valores dos artigos originais, “Repr.” à reprodução padronizada.

o *recall* da classe 1 permaneceu elevado (0,72), e o *F1-score* da classe 1 chegou a 0,81, valor bastante expressivo. Considerando os desafios metodológicos mencionados anteriormente e o fato de o conjunto de dados padronizado seguir uma estrutura distinta da original, esses resultados reforçam a relativa capacidade de generalização do modelo.

No caso de Zhang et al. (2021), cuja predição ocorre 4 horas antes do *onset*, os resultados reproduzidos mostram um desempenho inferior ao reportado originalmente (AUROC de 0,90 contra 0,94), especialmente no *recall* da classe 1 (0,60), mas ainda assim aceitável no contexto clínico. Isso pode estar associado à ausência do banco de dados original, o que obrigou a reprodução a depender exclusivamente da base padronizada, com ajustes aproximados em variáveis e frequência de coleta. Apesar disso, o modelo demonstrou comportamento estável, com *F1-score* macro e acurácia semelhantes aos demais trabalhos reproduzidos.

O estudo de Delahanty et al. (2019) realiza um experimento sistemático de predição de sepse com diferentes janelas de antecedência (1h, 3h, 6h, 12h e 24h antes do *onset*), o que permite avaliar como o tempo até o evento influencia a performance dos modelos. Curiosamente, no artigo original, o desempenho aumenta à medida que a janela de predição se afasta do *onset*: o AUROC vai de 0,93 (1h antes) até 0,97

(12h e 24h antes), e o *recall* da classe positiva cresce de 67% (1h) para 84% (24h). Esse comportamento contraria a expectativa comum de que, quanto mais próximo do evento, mais sinais clínicos relevantes estejam disponíveis, facilitando a predição. Uma possível explicação para esse fenômeno é o uso, no estudo original, de critérios que podem ter introduzido *leakage* de informações clínicas nos dados mais próximos do *onset*, gerando viés na separação entre classes ao longo do tempo. Também é possível que o modelo original, por ser mais sensível a padrões precoces, tenha capturado sinais sistêmicos de risco generalizado (como infecção ou instabilidade hemodinâmica) mais presentes em registros distantes do diagnóstico formal de sepse.

Na reprodução com o conjunto de dados padronizado, observou-se um comportamento mais estável e esperado: os valores de AUROC oscilaram levemente entre 0,93 e 0,94, com diferenças pouco significativas entre as janelas. O *recall* da classe 1 e o *F1-score* apresentaram uma leve queda conforme a predição se afastava do *onset*, indo de 0,74 (1h) a 0,73 (24h), o que está mais alinhado com a hipótese de que predições mais próximas do evento têm acesso a informações mais diretas sobre a deterioração clínica do paciente. Além disso, os valores de *F1-score* da classe positiva variaram entre 0,81 e 0,84, mantendo-se em um patamar elevado e consistente, demonstrando robustez e generalização do modelo mesmo fora do ambiente original. Esses resultados sugerem que, sob uma padronização mais rigorosa dos dados e dos critérios de definição de sepse, a relação entre tempo e performance segue um comportamento mais lógico e clínico, e que os ganhos observados no estudo original com maior antecedência podem ter sido influenciados por fatores específicos de modelagem ou seleção de dados não replicáveis em ambientes externos.

Em resumo, os resultados obtidos com o conjunto de dados padronizado revelam que, mesmo com variações na base de dados e ajustes nos critérios de inclusão, os modelos reimplementados conseguiram manter desempenho robusto, com pequenas perdas em AUROC, mas ganhos consideráveis em *recall* e *F1-score* da classe 1 em alguns casos. Esse comportamento reforça a importância da padronização na comparação entre técnicas e evidencia que certos modelos conseguem generalizar bem para outros contextos clínicos, enquanto outros demonstram maior sensibilidade às características dos dados originais.

### 6.2.1 Impacto da Variação da Antecedência da Predição

As Tabelas 3, 4, 5 e 6 apresentam os resultados da aplicação dos quatro modelos em diferentes janelas de antecedência à sepse (0h, 6h, 12h e 24h), simulando cenários clínicos com distintos níveis de urgência e disponibilidade de informações prévias. De maneira geral, observa-se que todos os modelos apresentaram melhor desempenho quanto mais próximos do momento do *onset*, o que está de acordo com a expectativa clínica: à medida que a sepse se aproxima, os sinais fisiológicos tornam-

se mais pronunciados e informativos.

Tabela 3 – Resultados de reprodução com conjunto de dados padronizado em diferentes intervalos de tempo — 0h

| Métrica | Zhao; Shen; Wang (2021) | Kam; Kim (2017) | Zhang et al. (2021) | Delahanty et al. (2019) |
|---|---|---|---|---|
| AUROC | 0,94 | 0,91 | 0,94 | **0,95** |
| *Recall* macro | 0,86 | 0,84 | **0,88** | 0,87 |
| *Recall* (classe 1) | 0,74 | 0,78 | 0,80 | **0,82** |
| *F1-score* macro | 0,88 | 0,84 | **0,90** | 0,87 |
| *F1-score* (classe 1) | 0,81 | 0,83 | **0,85** | **0,85** |
| Acurácia | **0,93** | 0,84 | 0,92 | 0,87 |

O modelo de Zhao; Shen; Wang (2021), originalmente ajustado para uma janela de 6h, apresentou seu melhor desempenho na janela de 0h, com AUROC de 0,94 e *F1-score* da classe 1 de 0,81 (Figura 2, 4). No entanto, os resultados para 6h permaneceram bastante consistentes (AUROC de 0,92 e *F1-score* de 0,79), demonstrando que o modelo preserva boa capacidade preditiva mesmo com menor proximidade do evento. A partir de 12h e, especialmente, em 24h, as métricas apresentaram declínio acentuado, com *recall* da classe 1 caindo para 0,50 e *F1-score* para 0,56 (Figura 3), o que indica perda substancial de sensibilidade em cenários de predição mais antecipada. Essa degradação visualiza-se claramente nos gráficos, revelando uma sensibilidade progressivamente limitada à medida que o modelo se afasta do momento do onset.

Tabela 4 – Resultados de reprodução com conjunto de dados padronizado em diferentes intervalos de tempo — 6h

| Métrica | Zhao; Shen; Wang (2021) | Kam; Kim (2017) | Zhang et al. (2021) | Delahanty et al. (2019) |
|---|---|---|---|---|
| AUROC | 0,92 | 0,86 | 0,90 | **0,93** |
| *Recall* macro | 0,85 | 0,78 | 0,76 | **0,87** |
| *Recall* (classe 1) | 0,73 | 0,67 | 0,55 | **0,76** |
| *F1-score* macro | 0,87 | 0,78 | 0,79 | **0,89** |
| *F1-score* (classe 1) | 0,79 | 0,76 | 0,65 | **0,83** |
| Acurácia | 0,91 | 0,78 | 0,90 | **0,93** |

Para Kam; Kim (2017), cujo modelo foi originalmente ajustado para 3h de antecedência, o comportamento foi semelhante: o desempenho máximo foi atingido em 0h (*F1-score* da classe 1 de 0,83), com degradação progressiva nas janelas mais longas. Em 24h, o *recall* da classe 1 caiu para 0,59 e o *F1-score* para 0,70. Esses valores, destacados nas Figuras 3 e 4, evidenciam maior dificuldade do modelo em identificar corretamente os pacientes com sepse a partir de dados mais distantes do evento.

O modelo de Zhang et al. (2021) também confirmou esse padrão de desempenho decrescente com o tempo. Em 0h, os resultados foram notavelmente altos (AUROC de 0,94, *F1-score* da classe 1 de 0,85), mas houve queda contínua nas janelas seguintes. O *F1-score* da classe 1 caiu para 0,64 em 24h, enquanto o *recall* caiu de 0,80 para 0,53. As Figuras 3 e 4 evidenciam essa tendência de forma clara, apontando que o

Tabela 5 – Resultados de reprodução com conjunto de dados padronizado em diferentes intervalos de tempo — 12h

| Métrica | Zhao; Shen; Wang (2021) | Kam; Kim (2017) | Zhang et al. (2021) | Delahanty et al. (2019) |
|---|---|---|---|---|
| AUROC | 0,88 | 0,81 | 0,89 | **0,93** |
| *Recall* macro | 0,77 | 0,74 | 0,76 | **0,86** |
| *Recall* (classe 1) | 0,63 | 0,60 | 0,53 | **0,73** |
| *F1-score* macro | 0,79 | 0,74 | 0,80 | **0,89** |
| *F1-score* (classe 1) | 0,68 | 0,70 | 0,65 | **0,82** |
| Acurácia | 0,84 | 0,74 | 0,91 | **0,94** |

modelo é fortemente dependente da proximidade temporal com o *onset* para manter desempenho elevado.

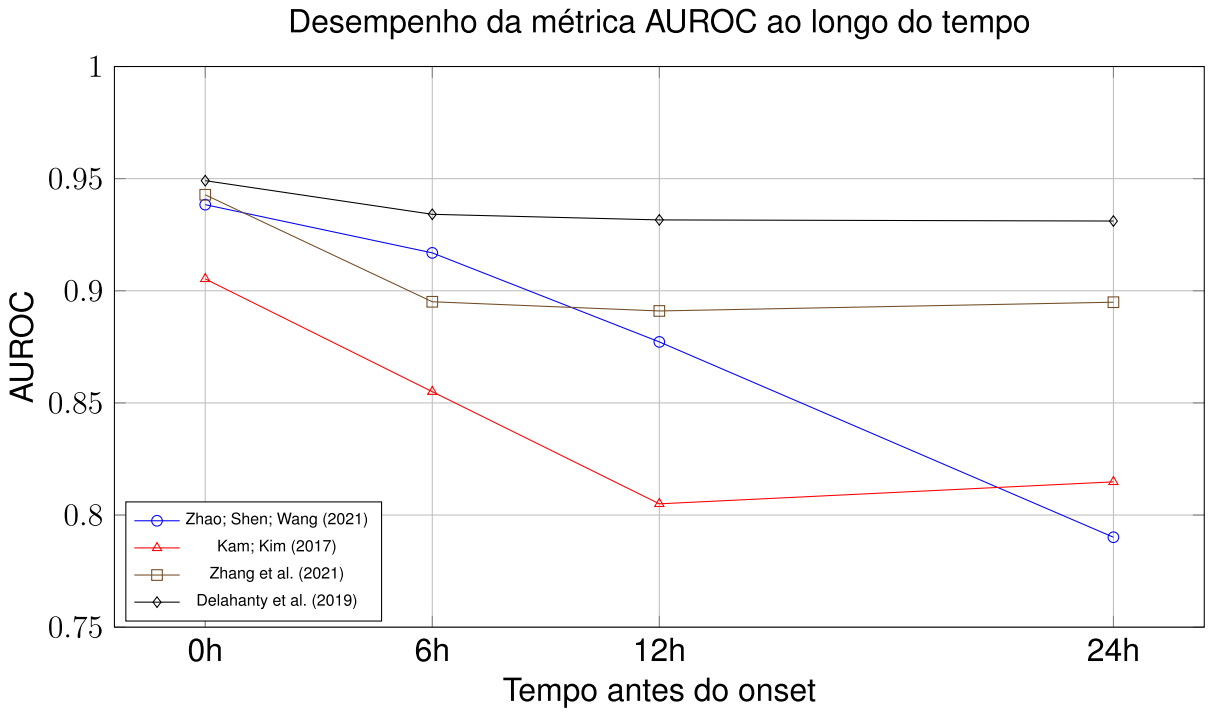


Figura 2 – Evolução da métrica AUROC ao longo das janelas de tempo.

Por outro lado, o modelo de Delahanty et al. (2019) demonstrou o comportamento mais estável ao longo das diferentes janelas. Embora o melhor desempenho tenha sido registrado em 0h (*F1-score* da classe 1 de 0,85), a degradação nos pontos seguintes foi muito mais suave. Em 24h, o *F1-score* da classe 1 ainda se manteve elevado (0,81), e o AUROC variou de forma mínima (de 0,95 para 0,93), como demonstram as Figuras 2 e 4. Essa estabilidade sugere maior capacidade de generalização temporal do modelo, o que pode estar relacionado à sua arquitetura ou ao uso de características clínicas mais robustas e estáveis ao longo do tempo.

De forma geral, a análise evidencia que a antecedência da predição exerce forte impacto sobre a sensibilidade e a efetividade dos modelos, sendo mais crítica nos algoritmos que dependem fortemente de padrões imediatos à deterioração clínica. Modelos mais robustos, como o de Delahanty et al. (2019), apresentam melhor desempenho mesmo em cenários com menos informações recentes, o que é desejável

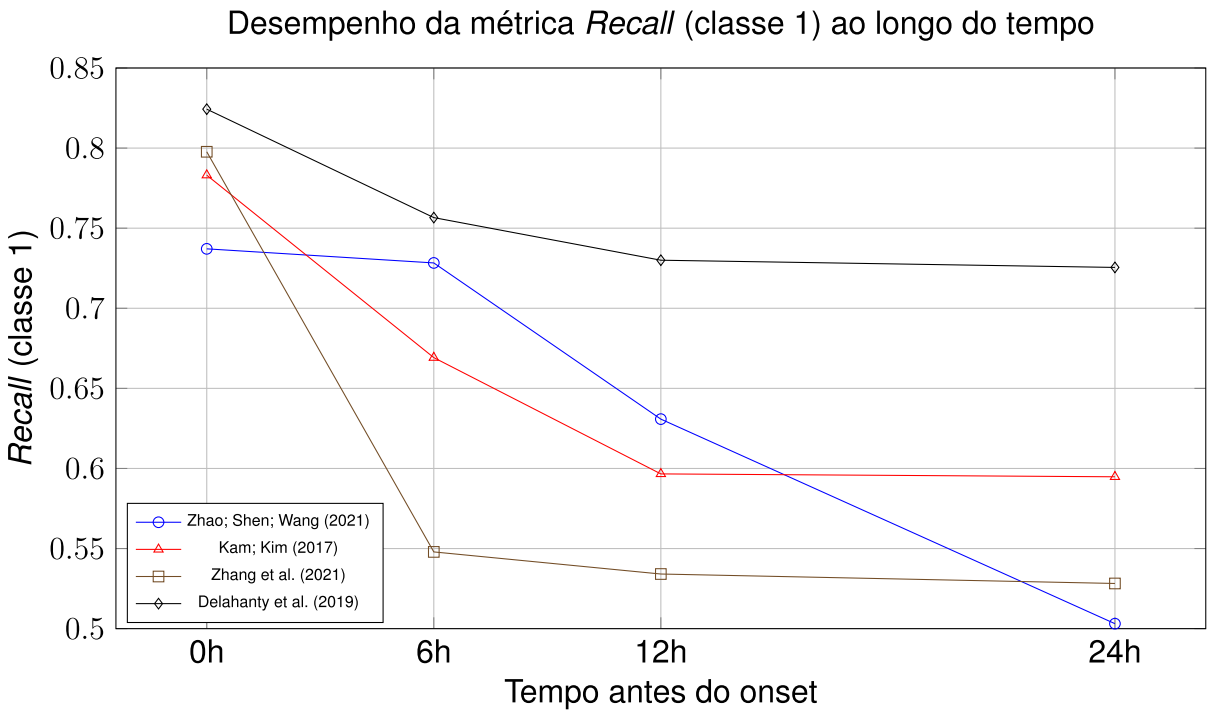


Figura 3 – Evolução da métrica *Recall* (classe 1) ao longo das janelas de tempo.

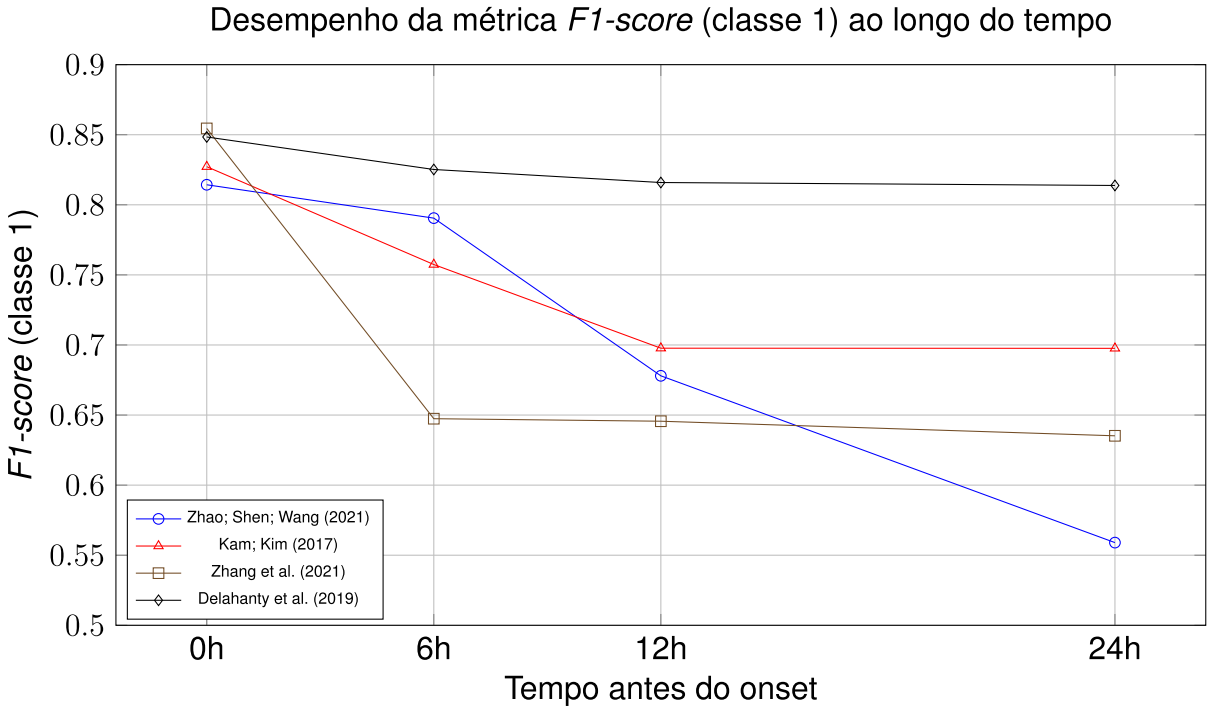


Figura 4 – Evolução da métrica *F1-score* (classe 1) ao longo das janelas de tempo.

Tabela 6 – Resultados de reprodução com conjunto de dados padronizado em diferentes intervalos de tempo — 24h

| Métrica | Zhao; Shen; Wang (2021) | Kam; Kim (2017) | Zhang et al. (2021) | Delahanty et al. (2019) |
|---|---|---|---|---|
| AUROC | 0,79 | 0,81 | 0,89 | **0,93** |
| *Recall* macro | 0,68 | 0,75 | 0,75 | **0,86** |
| *Recall* (classe 1) | 0,50 | 0,59 | 0,53 | **0,73** |
| *F1-score* macro | 0,69 | 0,74 | 0,79 | **0,89** |
| *F1-score* (classe 1) | 0,56 | 0,70 | 0,64 | **0,81** |
| Acurácia | 0,75 | 0,75 | 0,91 | **0,95** |

em contextos reais onde intervenções precoces são ideais. Essa variação reforça a importância de considerar janelas de tempo como variável experimental essencial em qualquer *benchmark* de predição de sepse.

## 6.3 Avaliação da PC

Nesta seção, são apresentados os resultados da aplicação da técnica de PC aos modelos reimplementados neste trabalho. O objetivo é investigar como a introdução de estimativas formais de incerteza afeta o desempenho, a confiabilidade e a utilidade clínica das predições. Foram adotados diferentes níveis de significância para simular cenários com distintas exigências de confiança, avaliando-se o impacto sobre métricas clássicas (como AUROC, *recall* e *F1-score*), bem como métricas específicas da PC, como *coverage rate* e *set size*.

Os experimentos foram realizados tanto nos conjuntos de dados originais utilizados nos artigos quanto no conjunto de dados padronizado proposto neste estudo, permitindo uma análise abrangente do comportamento dos modelos sob condições variadas de risco e ambiguidade.

### 6.3.1 Resultados com o Conjunto de Dados Original

A aplicação de PC ao modelo de Zhao; Shen; Wang (2021) evidencia um comportamento típico e desejável dessa abordagem: à medida que se aumenta o nível de significância ($\epsilon$) de $0{,}05$ para $0{,}15$, a *coverage rate* diminui gradualmente (de 1 para 0,88), enquanto o número de amostras removidas (ou predições abstidas) cresce consideravelmente, passando de 124 no total ($\epsilon = 0.05$) para 3.074 ($\epsilon = 0.15$). Essa dinâmica mostra que o modelo passa a rejeitar mais exemplos à medida que se exige menor confiabilidade estatística, o que é consistente com as garantias teóricas da técnica. Os resultados explorados nesta subseção estão detalhados na Tabela 7, onde é possível observar a evolução dos principais indicadores conforme a variação do nível de significância.

Ao mesmo tempo, nota-se um aumento progressivo do desempenho nas métricas de *F1-score* e *recall* da classe 1, que passam de 0,71 e 0,57 ($\epsilon = 0.05$) para 0,76 e 0,63 ($\epsilon = 0.15$), respectivamente. Isso ocorre porque o modelo passa a concentrar

Tabela 7 – Reprodução com conjunto de dados original — PC (níveis de significância $0{,}05$, $0{,}1$ e $0{,}15$)

| $\epsilon$ (nível de significância) | Métrica | Zhao; Shen; Wang (2021) | Kam; Kim (2017) |
|---|---|---|---|
| 0,05 | AUROC | **0,96** ($-1\%$) | 0,71 ($-14\%$) |
| | *Recall* (classe 1) | **0,57** ($+3\%$) | 0,52 ($-24\%$) |
| | *F1-score* (classe 1) | **0,71** ($+3\%$) | 0,60 ($-15\%$) |
| | Acurácia | **0,92** ($+0\%$) | 0,66 ($-8\%$) |
| | Coverage rate | **1** | **1** |
| | Set size | **1,08** | 1,82 |
| | Sem sepse removidos | 23 | 0 |
| | Com sepse removidos | 101 | 0 |
| 0,10 | AUROC | **0,96** ($-1\%$) | 0,71 ($-14\%$) |
| | *Recall* (classe 1) | **0,60** ($+7\%$) | 0,52 ($-24\%$) |
| | *F1-score* (classe 1) | **0,73** ($+6\%$) | 0,60 ($-15\%$) |
| | Acurácia | **0,94** ($+2\%$) | 0,66 ($-8\%$) |
| | Coverage rate | 0,95 | **1,00** |
| | Set size | **0,96** | 1,72 |
| | Sem sepse removidos | 297 | 0 |
| | Com sepse removidos | 1111 | 0 |
| 0,15 | AUROC | **0,95** ($-2\%$) | 0,71 ($-14\%$) |
| | *Recall* (classe 1) | **0,63** ($+13\%$) | 0,52 ($-24\%$) |
| | *F1-score* (classe 1) | **0,76** ($+10\%$) | 0,60 ($-15\%$) |
| | Acurácia | **0,96** ($+4\%$) | 0,66 ($-8\%$) |
| | Coverage rate | 0,88 | **1** |
| | Set size | **0,89** | 1,66 |
| | Sem sepse removidos | 836 | 0 |
| | Com sepse removidos | 2238 | 0 |

**Nota:** Os valores entre parênteses indicam a variação percentual em relação à reprodução sem aplicação de predição conformal (ver Tabela 1).

suas predições em instâncias mais confiáveis, o que reduz erros, especialmente falsos positivos. Essa seletividade também se reflete no tamanho médio dos conjuntos preditivos (*set size*), que diminui de 1,08 para 0,89, indicando um maior número de predições unárias e, portanto, mais informativas. Do ponto de vista clínico, esse comportamento é extremamente relevante: ao rejeitar casos ambíguos, o sistema protege o processo de decisão contra classificações imprecisas, o que é crucial em contextos sensíveis como a detecção de sepse.

Já no modelo de Kam; Kim (2017), o efeito da aplicação do PCT foi nulo ou limitado. Em todas as configurações de significância ($0{,}05$, $0{,}1$ e $0{,}15$), não houve nenhuma amostra removida, e a *coverage rate* se manteve constante em 1, o que significa que o modelo não identificou nenhuma predição como suficientemente incerta a ponto de ser rejeitada. O *set size* permaneceu ligeiramente acima de 1 (variando entre 1,82 e 1,66), o que indica predições ambíguas em média, mas sem que isso se traduzisse em abstinência de decisão. Isso pode estar relacionado a dois fatores principais: (i) o tamanho extremamente reduzido do conjunto de teste usado na reprodução (apenas 50 amostras), e (ii) a baixa variabilidade no escore de não conformidade das predições, possivelmente causada por um modelo com baixa sensibilidade à incerteza. O fato de os resultados de AUROC e F1 permanecerem absolutamente inalterados em todos os níveis de $\epsilon$ reforça a hipótese de que o modelo não se beneficiou da introdução de incerteza calibrada, seja por limitação técnica ou pelo baixo volume de dados.

Em síntese, a aplicação do PCT ao modelo de Zhao; Shen; Wang (2021) demonstrou na prática o que é teorizado na literatura: aumento da precisão e da confiabilidade das decisões ao custo de abstinência controlada. Já no caso de Kam; Kim (2017), a ausência de abstinência e de variação nas métricas sugere que o modelo ou o cenário reimplementado não foi sensível o suficiente à técnica de PC, ilustrando um caso em que a aplicação da metodologia não agrega valor. Esses resultados reforçam a importância de avaliar a sensibilidade do modelo ao risco estatístico, bem como a adequação do volume e diversidade de dados ao uso de métodos confiáveis de quantificação de incerteza.

### 6.3.2 Resultados com o Conjunto de Dados Padronizado

Os resultados descritos nesta seção estão apresentados nas tabelas 8 e 9.

O modelo de Zhao; Shen; Wang (2021) mostrou-se altamente responsivo à aplicação da PC. Com $\varepsilon = 0{,}05$, a *coverage rate* foi de 99%, com apenas 50 amostras removidas (25 positivas e 25 negativas). À medida que a significância aumenta para 0,1 e 0,15, a cobertura cai (para 95% e 90%, respectivamente), e o número de amostras abstidas cresce significativamente: mais de 1000 exemplos são removidos em $\varepsilon = 0{,}15$.

Do ponto de vista de desempenho, o modelo mostrou uma melhoria progressiva

Tabela 8 – Reprodução com conjunto de dados padronizado — PC (níveis de significância $0{,}05$, $0{,}1$ e $0{,}15$) — Modelos ZH21, KAM17 e ZG21

| $\epsilon$ (nível de significância) | Métrica | ZH21 | KAM17 | ZG21 |
|---|---|---|---|---|
| 0,05 | AUROC | 0,92 ($+0\%$) | 0,87 ($-3\%$) | 0,75 ($-16\%$) |
| | *Recall* macro | 0,86 ($+2\%$) | 0,81 ($-2\%$) | 0,69 ($-13\%$) |
| | *Recall* (classe 1) | 0,75 ($+3\%$) | 0,68 ($-5\%$) | 0,43 ($-28\%$) |
| | *F1-score* macro | 0,89 ($+2\%$) | 0,80 ($-3\%$) | 0,71 ($-14\%$) |
| | *F1-score* (classe 1) | 0,82 ($+3\%$) | 0,78 ($-3\%$) | 0,50 ($-28\%$) |
| | Acurácia | 0,93 ($+2\%$) | 0,81 ($-2\%$) | 0,85 ($-6\%$) |
| | Coverage rate | **1** | **1** | **1** |
| | Set size | **1,13** | 1,38 | 1,13 |
| | Sem sepse removidos | 25 | 0 | 0 |
| | Com sepse removidos | 25 | 0 | 0 |
| 0,10 | AUROC | 0,92 ($+0\%$) | 0,87 ($-3\%$) | 0,75 ($-17\%$) |
| | *Recall* macro | 0,87 ($+3\%$) | 0,81 ($-2\%$) | 0,69 ($-13\%$) |
| | *Recall* (classe 1) | 0,76 ($+4\%$) | 0,68 ($-5\%$) | 0,43 ($-28\%$) |
| | *F1-score* macro | 0,89 ($+3\%$) | 0,81 ($-3\%$) | 0,71 ($-13\%$) |
| | *F1-score* (classe 1) | 0,83 ($+5\%$) | 0,78 ($-3\%$) | 0,50 ($-27\%$) |
| | Acurácia | 0,94 ($+4\%$) | 0,81 ($-2\%$) | 0,86 ($-5\%$) |
| | Coverage rate | 0,95 | **1** | 0,99 |
| | Set size | 0,98 | 1,18 | **1** |
| | Sem sepse removidos | 274 | 0 | 25 |
| | Com sepse removidos | 203 | 0 | 19 |
| 0,15 | AUROC | 0,92 ($+0\%$) | 0,87 ($-3\%$) | 0,74 ($-18\%$) |
| | *Recall* macro | 0,88 ($+3\%$) | 0,81 ($-2\%$) | 0,68 ($-14\%$) |
| | *Recall* (classe 1) | 0,77 ($+6\%$) | 0,69 ($-5\%$) | 0,40 ($-33\%$) |
| | *F1-score* macro | 0,91 ($+4\%$) | 0,81 ($-3\%$) | 0,71 ($-14\%$) |
| | *F1-score* (classe 1) | 0,85 ($+7\%$) | 0,78 ($-3\%$) | 0,49 ($-29\%$) |
| | Acurácia | 0,95 ($+4\%$) | 0,81 ($-2\%$) | 0,87 ($-4\%$) |
| | Coverage rate | 0,90 | **1** | 0,91 |
| | Set size | 0,91 | **1,06** | 0,91 |
| | Sem sepse removidos | 629 | 1 | 170 |
| | Com sepse removidos | 406 | 5 | 92 |

**Identificadores:** ZH21 — Zhao; Shen; Wang (2021); KAM17 — Kam; Kim (2017); ZG21 — Zhang et al. (2021).
**Nota:** Os valores entre parênteses indicam a variação percentual em relação à reprodução sem aplicação de predição conformal (ver Tabela 2).

Tabela 9 – Reprodução com conjunto de dados padronizado — PC (níveis de significância $0{,}05$, $0{,}1$ e $0{,}15$) — Modelos DL19

| $\epsilon$ | Métrica | DL19-1h | DL19-3h | DL19-6h | DL19-12h | DL19-24h |
|---|---|---|---|---|---|---|
| 0,05 | AUROC | 0,93 (+0%) | **0,94** (+0%) | **0,94** (+1%) | 0,93 (+0%) | 0,93 (+0%) |
| | *Recall* macro | 0,88 (+3%) | **0,90** (+3%) | 0,89 (+3%) | 0,89 (+4%) | 0,89 (+4%) |
| | *Recall* (classe 1) | 0,79 (+6%) | **0,81** (+6%) | 0,80 (+5%) | 0,79 (+8%) | 0,79 (+8%) |
| | *F1-score* macro | 0,90 (+3%) | **0,92** (+2%) | 0,92 (+4%) | 0,92 (+4%) | 0,92 (+4%) |
| | *F1-score* (classe 1) | 0,85 (+5%) | **0,88** (+4%) | 0,87 (+5%) | 0,87 (+6%) | 0,86 (+7%) |
| | Acurácia | 0,92 (+2%) | 0,95 (+2%) | 0,96 (+3%) | 0,96 (+2%) | **0,97** (+2%) |
| | Coverage rate | **1** | **1** | **1** | **1** | 0,99 |
| | Set size | 1,23 | 1,11 | 1,07 | 1,04 | 1,03 |
| | Sem sepse removidos | 0 | 0 | 0 | 18 | 40 |
| | Com sepse removidos | 0 | 0 | 1 | 9 | 33 |
| 0,10 | AUROC | 0,93 (+0%) | **0,94** (+0%) | 0,93 (+0%) | 0,93 (+0%) | 0,93 (+0%) |
| | *Recall* macro | 0,86 (+1%) | **0,88** (+2%) | **0,88** (+1%) | **0,88** (+2%) | **0,88** (+2%) |
| | *Recall* (classe 1) | 0,75 (+1%) | **0,78** (+3%) | 0,77 (+2%) | 0,76 (+4%) | 0,76 (+4%) |
| | *F1-score* macro | 0,88 (+1%) | **0,91** (+1%) | **0,91** (+2%) | **0,91** (+2%) | **0,91** (+2%) |
| | *F1-score* (classe 1) | 0,82 (+2%) | **0,86** (+2%) | 0,85 (+2%) | 0,84 (+3%) | 0,85 (+4%) |
| | Acurácia | 0,91 (+1%) | 0,94 (+1%) | 0,95 (+2%) | 0,95 (+1%) | **0,96** (+1%) |
| | Coverage rate | 0,98 | 0,96 | 0,95 | 0,95 | 0,94 |
| | Set size | 1,01 | 0,96 | 0,95 | 0,95 | 0,94 |
| | Sem sepse removidos | 177 | 473 | 467 | 551 | 600 |
| | Com sepse removidos | 119 | 202 | 268 | 220 | 259 |
| 0,15 | AUROC | 0,93 (+0%) | **0,94** (+0%) | **0,94** (+1%) | 0,93 (+0%) | 0,93 (+0%) |
| | *Recall* macro | 0,88 (+3%) | **0,90** (+3%) | 0,89 (+3%) | 0,89 (+4%) | 0,89 (+4%) |
| | *Recall* (classe 1) | 0,79 (+6%) | **0,81** (+6%) | 0,80 (+5%) | 0,79 (+8%) | 0,79 (+8%) |
| | *F1-score* macro | 0,90 (+3%) | **0,92** (+2%) | **0,92** (+4%) | **0,92** (+4%) | **0,92** (+4%) |
| | *F1-score* (classe 1) | 0,85 (+5%) | **0,88** (+4%) | 0,87 (+5%) | 0,87 (+6%) | 0,86 (+7%) |
| | Acurácia | 0,92 (+1%) | 0,95 (+2%) | 0,96 (+3%) | 0,96 (+2%) | **0,97** (+2%) |
| | Coverage rate | 0,92 | 0,90 | 0,89 | 0,89 | 0,88 |
| | Set size | 0,93 | 0,90 | 0,89 | 0,89 | 0,88 |
| | Sem sepse removidos | 911 | 1093 | 1115 | 1181 | 1238 |
| | Com sepse removidos | 442 | 468 | 503 | 424 | 442 |

**Identificadores:** DL19 — Delahanty et al. (2019), com antecedência de predição de 1h, 3h, 6h, 12h e 24h.
**Nota:** Os valores entre parênteses indicam a variação percentual em relação à reprodução sem aplicação de predição conformal (ver Tabela 2).

nas métricas de *recall* e F1 da classe 1. O *F1-score* para casos de sepse cresceu de 0,82 ($\varepsilon = 0{,}05$) para 0,85 ($\varepsilon = 0{,}15$), e o *recall* de 0,75 para 0,77, refletindo que o modelo passou a focar em predições mais confiáveis, mesmo ao custo de rejeitar uma fração maior dos dados. O *set size* diminui de 1,13 para 0,91, indicando uma elevação na quantidade de predições unárias, o que torna o modelo mais assertivo e interpretável clinicamente. Esses resultados ilustram bem a capacidade do PCT de modular o risco da decisão automática, tornando o modelo mais seletivo em contextos incertos.

A aplicação de PC ao modelo de Kam; Kim (2017) demonstrou comportamento bastante estável e, até certo ponto, conservador. Em todas as configurações de significância ($\varepsilon = 0{,}05$, $0{,}1$, $0{,}15$), as métricas de desempenho permaneceram praticamente inalteradas, com AUROC fixado em 0,87 e *F1-score* da classe 1 em torno de 0,78. A *coverage rate* foi consistentemente alta, se mantendo em 1, e o número de amostras rejeitadas foi mínimo, com apenas cinco exemplos excluídos na configuração mais permissiva ($\varepsilon = 0{,}15$).

Esse resultado sugere que o modelo opera com alta autoconfiança e baixa sensibilidade à incerteza estatística, o que pode ter origens múltiplas. Primeiramente, a quantidade reduzida de amostras e a própria natureza do classificador podem não ter proporcionado variação suficiente nos escores de não conformidade, fazendo com que o PCT identificasse quase todas as predições como confiáveis. Em segundo lugar, é possível que o modelo apresente uma calibragem subótima de probabilidade, com escores extremos que não refletem bem a incerteza subjacente, cenário em que a PC tem dificuldade de atuação. Ainda assim, o desempenho do modelo foi sólido, mas a técnica de PC não agregou benefícios mensuráveis neste caso, funcionando mais como uma verificação de robustez do que como um mecanismo de refinamento.

O modelo de Zhang et al. (2021) apresentou um comportamento atípico e único entre os avaliados: a aplicação do PCT resultou em degradação de desempenho, especialmente visível nas métricas voltadas à classe positiva (sepse). O *F1-score* da classe 1 caiu de 0,69 (sem PC) para 0,50 ($\varepsilon = 0{,}1$) e 0,49 ($\varepsilon = 0{,}15$), com o *recall* da classe 1 também reduzido para 0,40 na configuração mais permissiva. Este foi o único modelo entre os avaliados que apresentou piora sistemática com a introdução da quantificação de incerteza.

Algumas hipóteses podem ser levantadas para explicar esse fenômeno. Em primeiro lugar, o desempenho original do modelo já se mostrava inferior aos demais, com F1 e *recall* relativamente baixos, o que sugere uma estrutura de decisão mais instável ou um maior grau de sobreajuste a padrões específicos do conjunto de dados. Ao aplicar PC, que tende a rejeitar casos de baixa conformidade, o modelo provavelmente passou a descartar amostras em que seu próprio julgamento era impreciso, o que, paradoxalmente, reduziu a cobertura sem ganhos compensatórios nas predições

mantidas. Além disso, o *set size* permaneceu relativamente alto, mesmo com níveis de significância elevados (ex: 0,91 com $\varepsilon = 0{,}15$), o que indica predições frequentemente ambíguas ou com baixa assertividade estatística.

Outra possibilidade é que o comportamento do modelo esteja associado ao volume reduzido de dados utilizados (3.000 amostras), o que pode ter gerado instabilidade na calibração dos escores de não conformidade. Em modelos mais sensíveis a variações de dados ou com estruturas complexas, como redes neurais profundas no caso do modelo avaliado, a aplicação de PCT com conjuntos pequenos pode não refletir bem o comportamento esperado em escala maior.

Já o modelo de Delahanty et al. (2019), avaliado em diferentes janelas de antecedência (1h a 24h), apresentou desempenho robusto e comportamento muito consistente com a aplicação de PC. Em todas as janelas, observou-se a mesma tendência de melhoria progressiva das métricas à medida que se aceitava maior significância estatística.

Por exemplo, na janela de 3h, o F1 da classe 1 aumentou de 0,83 ($\varepsilon = 0{,}05$) para 0,88 ($\varepsilon = 0{,}15$), e o *recall* subiu de 0,76 para 0,81. Esse padrão se repetiu em todas as janelas, com a cobertura caindo de valores próximos a 100% para cerca de 88–89%, e o número de amostras removidas chegando a mais de 1.500 em alguns casos. Ainda assim, os valores de AUROC se mantiveram estáveis, e a acurácia geral subiu em todos os cenários.

O *set size* também diminuiu sistematicamente com o aumento de $\varepsilon$ (ex: de 1,23 para 0,92 em 1h), o que demonstra que o modelo passou a emitir predições mais específicas, sem comprometer sua capacidade discriminativa. Esse comportamento reforça o potencial da PC em ambientes clínicos, nos quais a confiabilidade individual da decisão é tão importante quanto sua acurácia média. A resposta positiva do modelo de Delahanty et al. (2019) à técnica também sugere que ele possui uma estrutura probabilística estável e bem calibrada, mesmo em janelas mais distantes do *onset* da sepse.

### 6.3.3 Resultados comparativos com abordagens tradicionais

Os resultados obtidos com a aplicação de PC, especialmente na variante transdutiva (PCT), evidenciam diferenças importantes em relação às abordagens tradicionais de classificação binária. Quando comparados aos modelos originais reimplementados sem PC, os modelos com PC demonstraram, em geral, melhor equilíbrio entre sensibilidade e precisão, refletido no aumento consistente dos valores de *F1-score* da classe positiva (sepse), mesmo sob condições de incerteza. No entanto, os experimentos também revelaram que os benefícios da PC não são universais, e seu efeito depende fortemente das características do modelo base e da estabilidade dos escores de predição.

No caso do modelo de Zhao; Shen; Wang (2021), por exemplo, o *F1-score* da classe 1 saltou de 0,69 (sem PC) para até 0,85 com PC em $\varepsilon = 0{,}15$, ao custo de uma cobertura ligeiramente reduzida (de 100% para aproximadamente 90%). A mesma tendência foi observada com o modelo de Delahanty et al. (2019), que manteve um AUROC elevado, acima de 0,93, enquanto aumentava o *F1-score* e o *recall* da classe 1 em todas as janelas de predição com a aplicação do PC. Esses ganhos reforçam o valor prático da técnica, especialmente por permitir que o modelo abstenha-se de prever em casos incertos, elevando assim a confiabilidade das instâncias classificadas.

Contudo, no caso do modelo de Zhang et al. (2021), a aplicação do PCT resultou em degradação de desempenho, com queda no *F1-score* da classe 1 (de 0,69 para 0,49) e no *recall*. Esse foi o único modelo avaliado em que a introdução da PC comprometeu a qualidade das predições mantidas. Esse comportamento pode estar relacionado à instabilidade dos escores de predição do modelo ou à má calibração probabilística, o que torna o mecanismo de conformidade menos eficaz na diferenciação entre amostras confiáveis e incertas. Além disso, o uso de um subconjunto de 3.000 amostras para os testes pode ter amplificado esse efeito em modelos mais sensíveis a variações amostrais.

O modelo de Kam; Kim (2017), por outro lado, apresentou um comportamento neutro: todas as métricas permaneceram inalteradas com a aplicação da PC, e a *coverage rate* se manteve próxima de 100% em todos os níveis de significância. Isso sugere que o modelo operava com decisões de alta confiança, mas possivelmente mal calibradas do ponto de vista estatístico, o que impediu o PCT de atuar como um mecanismo de rejeição eficaz.

Diferentemente das abordagens tradicionais que fornecem apenas *scores* contínuos ou probabilidades, o PCT entrega garantias estatísticas formais sobre as previsões individuais, algo particularmente útil em contextos clínicos. Essa diferença metodológica se reflete diretamente em indicadores de qualidade como a *coverage rate* e o *set size*, que não apenas complementam as métricas clássicas, mas oferecem meios objetivos de medir a confiabilidade das decisões.

Esses achados reforçam que o uso de PC tem alto potencial em modelos bem calibrados e estáveis, mas sua adoção deve ser acompanhada de uma análise cuidadosa da natureza do modelo base, da distribuição dos dados e da robustez dos escores de confiança.

### 6.3.4 Benefícios e limitações identificados

A aplicação de PC, especialmente na sua forma transdutiva (PCT), demonstrou benefícios substanciais no contexto da predição de sepse, particularmente em modelos com boa calibração e estabilidade preditiva. O principal diferencial da técnica é sua capacidade de quantificar formalmente a incerteza, permitindo que o sistema

se abstenha de prever em situações ambíguas. Esse mecanismo de rejeição seletiva é especialmente valioso em ambientes clínicos de alta criticidade, como UTIs, onde uma decisão incorreta pode levar a consequências graves. Em dois modelos testados, os de Zhao; Shen; Wang (2021) e Delahanty et al. (2019), a introdução do PCT não apenas preservou métricas tradicionais como AUROC e acurácia, mas também elevou substancialmente os valores de *F1-score* e *recall* da classe positiva, melhorando a sensibilidade da predição em contextos de desbalanceamento, característica típica dos casos de sepse.

Outro benefício notável é a robustez estatística da técnica, que requer apenas a suposição de *exchangeability* para garantir validade preditiva. Além disso, a PC é algoritmicamente flexível: pode ser incorporado a qualquer modelo base, seja uma árvore de decisão, rede neural ou *ensemble*, com a simples definição de uma medida de não conformidade. Essa adaptabilidade permite que a técnica seja usada em diferentes cenários clínicos, ajustando o nível de seletividade de acordo com o risco operacional e a tolerância ao erro da instituição.

No entanto, os resultados também mostraram que o uso de PC não é universalmente benéfico. O caso do modelo de Zhang et al. (2021) evidenciou que a aplicação do PCT pode, em certos contextos, piorar o desempenho nas instâncias mantidas, especialmente quando o modelo base apresenta baixa estabilidade, má calibração ou alto grau de incerteza estrutural. Nessa situação, o PCT tende a rejeitar predições com maior frequência, mas sem ganhos claros nas instâncias aceitas, o que leva à redução de *recall* e *F1-score*, comportamento contrário ao observado nos demais modelos. Esse resultado indica que a PC, embora poderoso, não corrige deficiências fundamentais do modelo subjacente, podendo até amplificá-las sob determinados parâmetros.

Adicionalmente, o modelo de Kam; Kim (2017) demonstrou que modelos mal calibrados ou com decisões excessivamente confiantes podem resultar em *coverage* artificialmente alto (próximo de 100%) e ausência de rejeições, o que anula os benefícios esperados da PC. Nestes casos, a técnica se torna ineficaz como ferramenta de controle de confiança, atuando apenas como um verificador passivo.

Outro desafio importante é o custo computacional elevado do PCT, especialmente em grandes conjuntos de dados. Como a abordagem exige recalcular os *scores* de não conformidade para cada instância de teste, sua aplicação em escala hospitalar ou em tempo real requer otimizações adicionais ou a substituição por variantes mais eficientes, como o PCI. Por fim, a interpretação prática de métricas como *set size* e das instâncias rejeitadas ainda carece de maturidade no ambiente clínico, podendo gerar confusão em contextos nos quais decisões precisam ser rápidas, objetivas e transparentes.

Em síntese, PC se mostra uma ferramenta promissora e relevante para tornar mo-

delos preditivos mais confiáveis, auditáveis e seguros, especialmente em áreas críticas como a medicina intensiva. Contudo, sua aplicação exige cuidados metodológicos rigorosos, desde a seleção do modelo base até a interpretação clínica dos resultados. Quando bem implementada, a técnica contribui significativamente para a elevação da maturidade técnica e ética da inteligência artificial aplicada à saúde.

## 6.4 Análise Crítica

Os experimentos conduzidos ao longo deste trabalho revelam um panorama complexo do estado da arte na predição de sepse com técnicas de AM. A análise comparativa de modelos reproduzidos, tanto em seus contextos originais quanto em ambiente padronizado, evidencia avanços significativos na área, mas também expõe limitações estruturais que comprometem a solidez das conclusões encontradas na literatura.

Em primeiro lugar, a variabilidade nos resultados obtidos ao longo das reproduções indica que a reprodutibilidade ainda é um problema central. Apesar de alguns trabalhos, como o de Zhao; Shen; Wang (2021), apresentarem descrições metodológicas detalhadas e passíveis de replicação, outros, mesmo quando bem-intencionados, carecem de informações fundamentais, como critérios exatos de seleção de pacientes, detalhes de pré-processamento ou configurações de parâmetros. Essa falta de transparência metodológica compromete a comparação entre estudos e reforça a necessidade de padrões mais rigorosos de documentação e disponibilização de código e dados.

A aplicação dos modelos em um conjunto de dados padronizado permitiu uma avaliação em condições controladas, revelando diferenças expressivas entre algoritmos que, à primeira vista, apresentavam desempenhos semelhantes. A sensibilidade às janelas de tempo, por exemplo, variou substancialmente entre os modelos. Enquanto alguns demonstraram robustez mesmo com antecedência elevada (como Delahanty et al. (2019)), outros dependem fortemente de dados muito próximos ao *onset* da sepse para manter desempenho adequado. Isso sugere que nem todos os modelos do estado da arte são igualmente generalizáveis ou clinicamente úteis em cenários com tempo limitado para intervenção, um aspecto crítico quando se busca aplicações reais em ambientes hospitalares.

Além das limitações metodológicas, os experimentos também permitiram confrontar as promessas dos estudos originais com os resultados reproduzidos. No caso de Delahanty et al. (2019), o modelo original apresentou desempenho crescente à medida que a janela de predição se afastava do evento de sepse, um resultado contraintuitivo do ponto de vista clínico. Já nas reproduções conduzidas nesta dissertação, observou-se o comportamento inverso: os modelos obtiveram melhor desempenho quanto mais próximos do *onset*, o que está mais alinhado com a dinâmica real da

deterioração clínica. No estudo de Kam; Kim (2017), embora os autores tenham reportado AUROC superior a 0,93 mesmo com antecedência de 3 horas, a reprodução apresentou valores mais modestos e sujeitos a flutuação, agravada pelo tamanho reduzido do conjunto de testes. Por fim, o modelo de Zhang et al. (2021), que afirma superioridade por meio de arquiteturas LSTM e técnicas de atenção, não apenas teve desempenho inferior nas reproduções, como foi o único entre os estudados em que a aplicação de PC resultou em piora sistemática, sugerindo fragilidade na calibragem estatística e instabilidade frente à incerteza. Tais contrastes entre afirmações e resultados evidenciam a fragilidade de conclusões derivadas de bases não padronizadas e a importância de reavaliar o estado da arte sob condições controladas.

Outro ponto relevante diz respeito ao uso de métricas inadequadas ou incompletas na avaliação dos modelos. Muitos trabalhos priorizam AUROC como métrica principal, mas os experimentos deste estudo mostram que essa medida pode mascarar deficiências importantes na detecção da classe positiva (sepse). Métricas como *F1-score* e *recall* da classe 1, especialmente em contextos desbalanceados, revelaram-se mais alinhadas com os objetivos clínicos de detecção precoce. Essa constatação reforça a importância de adotar métricas sensíveis ao contexto e às consequências clínicas de erros de predição.

Por fim, a aplicação de PC revelou-se uma contribuição importante para a confiabilidade das predições. A capacidade de rejeitar instâncias incertas e oferecer garantias formais de cobertura representa um avanço em relação às abordagens determinísticas predominantes na literatura. Os resultados mostraram que, mesmo modelos com bom desempenho, como os de Zhao; Shen; Wang (2021) e Delahanty et al. (2019), se beneficiam substancialmente da introdução de quantificação explícita de incerteza, o que reforça o potencial da técnica para aplicações clínicas reais.

Os resultados também evidenciaram que a incorporação de técnicas como PC, embora promissora, não é isenta de limitações práticas. O caso do modelo de Zhang et al. (2021), que apresentou piora consistente no desempenho após a aplicação do PCT, demonstra que a técnica depende fortemente da estabilidade e calibragem do modelo base. Em vez de melhorar a qualidade das predições aceitas, como ocorreu com outros modelos, o PCT resultou em degradação do *recall* e do *F1-score* da classe positiva, mesmo em níveis elevados de cobertura. Isso indica que, em modelos mais sensíveis a ruído ou com maior instabilidade estrutural, a PC pode atuar de forma contraproducente. O modelo de Kam; Kim (2017), por sua vez, exibiu cobertura total e nenhuma instância rejeitada, o que sugere que decisões excessivamente confiantes (ou mal calibradas) impedem o PCT de exercer seu papel seletivo, anulando seus benefícios potenciais. Esses achados reforçam que, embora a PC ofereça vantagens teóricas robustas, sua eficácia na prática clínica depende de pré-condições metodológicas que nem todos os modelos atendem.

Além das limitações individuais observadas, chama atenção o fato de que os dois únicos casos em que a aplicação de PC não trouxe benefícios, Kam; Kim (2017) e Zhang et al. (2021), envolvem arquiteturas baseadas em redes neurais. Esse padrão sugere que modelos neurais, apesar de sua capacidade expressiva, podem gerar distribuições de saída mal calibradas ou altamente concentradas, dificultando a atuação eficaz da PC. Como o PCT depende de *scores* de não conformidade sensíveis à variação estatística entre instâncias, modelos que produzem predições excessivamente confiantes ou pouco diferenciadas podem comprometer a utilidade do mecanismo de rejeição. Essa observação indica que a combinação entre redes neurais e PC requer atenção especial à calibragem dos *outputs*, e reforça a importância de considerar a natureza do classificador ao aplicar métodos de quantificação de incerteza em contextos clínicos.

Em síntese, os experimentos conduzidos nesta dissertação revelam que, embora o campo da predição de sepse com AM tenha avançado tecnicamente, ainda enfrenta desafios fundamentais. A análise crítica mostrou que o desempenho dos modelos é influenciado não apenas por aspectos como janelas temporais ou características dos dados, mas também por elementos muitas vezes negligenciados, como estabilidade estatística e calibração probabilística. A inconsistência dos resultados ao aplicar metodologias mais rigorosas, como a PC, evidencia que muitos modelos não foram concebidos com robustez metodológica em mente. Esse cenário revela um desequilíbrio: enquanto a área avança em complexidade algorítmica, carece de maturidade científica no que diz respeito à avaliação crítica, confiabilidade e reprodutibilidade dos modelos desenvolvidos. A consolidação da predição clínica baseada em IA requer, portanto, não apenas inovação técnica, mas também o estabelecimento de práticas científicas mais rigorosas e transparentes.

# 7 CONCLUSÃO

Esta seção final apresenta uma síntese dos principais achados obtidos ao longo do estudo, destacando as lições extraídas sobre reprodutibilidade, padronização e desempenho de modelos de predição de sepse com AM. Também são discutidas as contribuições diretas deste trabalho para a área, tanto em nível metodológico quanto aplicado, e apontadas as limitações enfrentadas, juntamente com sugestões para futuros desdobramentos e aprofundamentos da pesquisa.

## 7.1 Principais Achados

Este trabalho investigou criticamente o estado da arte na predição de sepse com técnicas de AM, com foco especial na reprodutibilidade e padronização dos experimentos. Os resultados obtidos ao longo das etapas experimentais indicam que a literatura da área ainda enfrenta desafios importantes relacionados à replicação de estudos: dificuldades de acesso a dados, ausência de documentação adequada sobre pré-processamento e critérios clínicos heterogêneos comprometem a comparação entre modelos e a validação independente de resultados. Esses achados confirmam a **Hipótese 1**, que previa que a maioria dos estudos da área não pode ser reproduzida integralmente com as informações metodológicas disponíveis nas publicações originais.

A reprodução de quatro estudos com diferentes metodologias revelou disparidades relevantes entre os resultados reportados nos artigos originais e os obtidos neste trabalho. Por exemplo, o modelo de Delahanty et al. (2019) manteve desempenho elevado nas reproduções, com *AUROC* acima de 0,93 em todas as janelas temporais, o que corrobora as afirmações dos autores sobre sua robustez em diferentes horizontes de predição. O estudo original já relatava crescimento progressivo da discriminatividade do modelo conforme o evento de sepse se aproximava (de AUROC 0,93 em 1h até 0,97 em 24h), e os experimentos reproduzidos confirmaram essa tendência. Tal consistência reforça o potencial clínico do modelo, especialmente em contextos de predição com antecedência.

Por outro lado, o modelo de Kam; Kim (2017) demonstrou reprodutibilidade limitada. Apesar dos autores afirmarem desempenho próximo a 0,93 em AUROC com 3h de antecedência, os experimentos aqui realizados alcançaram valores notavelmente inferiores, mesmo sob tentativas de replicar a arquitetura e janela temporal descritas. A discrepância é agravada pela limitação do conjjunto de dados original, que impôs um conjunto de teste com apenas 50 amostras. Esse fator compromete a confiabilidade das métricas reportadas e revela fragilidades na metodologia do estudo, especialmente na avaliação estatística de desempenho.

O modelo de Zhang et al. (2021), embora reportasse desempenho competitivo no artigo original (com AUROC de 0,94), apresentou dificuldades de reprodutibilidade tanto em sua implementação quanto na generalização para o conjunto de dados padronizado. Em particular, os experimentos indicaram degradação expressiva em métricas de *recall* e *F1-score* da classe positiva, sobretudo em janelas mais amplas de antecedência. A confiança depositada pelos autores na capacidade do modelo de aprender representações complexas ao longo do tempo não se refletiu nos resultados reproduzidos, indicando possível sobreajuste às condições originais do estudo.

Ao aplicar todos os modelos em um conjunto de dados padronizado com critérios clínicos unificados (MIMIC-IV, Sepsis-3), foi possível realizar uma comparação justa em condições controladas. Nessa configuração, evidenciou-se que nem todos os algoritmos do estado da arte mantêm desempenho estável: modelos baseados em técnicas tradicionais, como o de Delahanty et al. (2019), mostraram maior resiliência, enquanto redes neurais mais complexas apresentaram maior sensibilidade à variação de contexto. Esses achados dão suporte à **Hipótese 2**, ao demonstrar que a padronização dos dados permite revelar com mais clareza as diferenças reais entre modelos, reduzindo a influência de fatores externos e melhorando a comparabilidade dos resultados.

Além disso, os experimentos mostraram que pequenas variações na janela de predição ou na configuração temporal influenciam significativamente o desempenho dos modelos, especialmente em termos de sensibilidade e equilíbrio de classes. Tais variações revelaram-se críticas para a interpretação e a utilidade clínica dos algoritmos, confirmando a **Hipótese 3**, que previa impacto relevante de mudanças sutis nas configurações experimentais.

Por fim, a aplicação da técnica de PC demonstrou o potencial de incorporar estimativas formais de incerteza aos modelos preditivos, agregando uma camada de confiabilidade especialmente relevante para contextos clínicos. A utilização da variante transdutiva (PCT) permitiu, em grande parte dos casos, melhorar métricas como *F1-score* e *recall* da classe positiva, ao mesmo tempo em que elevou a segurança das predições por meio da rejeição de instâncias ambíguas. No entanto, os experimentos também revelaram que os ganhos com a técnica não são universais: em modelos ins-

táveis ou mal calibrados, como o de Zhang et al. (2021), a aplicação do PCT resultou em degradação de desempenho, enquanto em outros, como o de Kam; Kim (2017), não houve impacto mensurável. Esses achados validam parcialmente a **Hipótese 4**, ao mostrarem que a técnica torna as incertezas mais explícitas, mas também que sua eficácia depende da calibragem e estabilidade dos modelos subjacentes.

Em resumo, os resultados desta dissertação sustentam também a **Hipótese 5**: ao adotar práticas rigorosas de documentação, padronização de dados e aplicação de métricas apropriadas, foi possível identificar de forma mais clara as limitações e os méritos de diferentes abordagens. Essa constatação reforça a necessidade de diretrizes bem definidas para aumentar a reprodutibilidade e promover avanços mais consistentes na aplicação de IA para suporte clínico na predição de sepse.

## 7.2 Contribuições para a Área

Esta dissertação apresenta um conjunto de contribuições significativas para o campo da predição clínica com AM, com ênfase no desafio persistente da predição de sepse. Em um cenário científico marcado pela heterogeneidade metodológica e pela baixa reprodutibilidade dos estudos, este trabalho se destaca por propor, executar e validar uma estrutura padronizada e rigorosa para a avaliação comparativa de modelos preditivos, oferecendo um caminho claro para o avanço técnico e científico da área.

A primeira e mais concreta contribuição é a criação de um conjunto de dados padronizado baseado no MIMIC-IV v2.2, com extração de pacientes utilizando a definição clínica Sepsis-3, amplamente reconhecida como o padrão atual de diagnóstico. Esta base de dados, construída com filtros clínicos transparentes, critérios reprodutíveis e janelas de tempo bem definidas, serve como um marco de referência para futuras pesquisas, permitindo que diferentes grupos comparem seus modelos sob as mesmas condições experimentais. Isso ataca diretamente um dos maiores gargalos da literatura atual: a impossibilidade de comparação justa entre estudos que utilizam bases distintas, métricas diferentes e definições clínicas não homogêneas.

Em segundo lugar, o trabalho realiza uma análise crítica e empírica da reprodutibilidade de modelos do estado da arte, demonstrando na prática que muitos dos resultados publicados não podem ser replicados sem acesso a dados específicos ou sem informações detalhadas sobre etapas fundamentais do fluxo de processamento de predição. A reimplementação de quatro estudos, com diferentes níveis de sucesso, evidencia que a comunidade ainda falha em adotar boas práticas de ciência aberta. Ao tornar essas dificuldades explícitas e documentadas, esta dissertação contribui para o debate sobre a necessidade de transparência e padronização na pesquisa em IA médica, fortalecendo a cultura da replicabilidade.

Outra contribuição central e inovadora deste trabalho é a incorporação da técnica de PC, com foco na sua variante transdutiva, aplicada a modelos clínicos para predição de sepse. Ao introduzir uma abordagem formal para quantificar a incerteza nas predições, esta dissertação transcende os limites das classificações determinísticas tradicionais e propõe um modelo que *sabe quando deve se abster de prever*, algo particularmente valioso em ambientes de alto risco, como unidades de terapia intensiva. Ainda que os resultados tenham mostrado benefícios consistentes em diversos modelos, também foi observado que a técnica não é universalmente eficaz, especialmente em modelos instáveis ou mal calibrados. Essa descoberta acrescenta uma dimensão crítica à literatura, alertando para os limites práticos da PC e para a importância de pré-condições estatísticas para seu sucesso.

Do ponto de vista metodológico, a dissertação propõe uma estrutura completa de avaliação, que integra reprodutibilidade, padronização, análise em diferentes janelas de tempo, métricas sensíveis ao desbalanceamento de classes (como *F1-score* e *recall* da classe positiva), e medidas específicas de confiabilidade (como *coverage rate* e *set size*). Esse conjunto articulado de elementos oferece uma base metodológica sólida e generalizável, que pode ser adaptada para outras condições clínicas além da sepse, como falência renal, choque ou deterioração aguda.

Por fim, esta dissertação também tem um impacto prático claro: ela aproxima a pesquisa de inteligência artificial médica das necessidades reais da prática clínica, ao enfatizar confiabilidade, reprodutibilidade, interpretabilidade e segurança, elementos frequentemente negligenciados em estudos puramente técnicos. Ao integrar fundamentos estatísticos rigorosos com aplicações clínicas sensíveis, o trabalho contribui para elevar o nível de maturidade da área, aproximando-a de um cenário em que modelos de IA possam ser não apenas eficazes, mas também confiáveis, auditáveis e implementáveis em hospitais reais.

Em conjunto, estas contribuições posicionam esta dissertação como um trabalho de referência para a consolidação de boas práticas em predição clínica com AM, promovendo avanços não apenas técnicos, mas também epistemológicos e éticos no desenvolvimento de tecnologias aplicadas à saúde.

## 7.3 Limitações e Trabalhos Futuros

Embora este estudo tenha avançado na sistematização e avaliação de modelos preditivos de sepse, algumas limitações devem ser reconhecidas. Primeiramente, a reprodução dos trabalhos foi limitada ao que estava descrito nos artigos e disponível publicamente. Em alguns casos, a falta de código ou de acesso ao banco de dados original impediu a reprodução completa ou confiável de determinadas etapas, o que restringe a abrangência da análise.

Além disso, a aplicação de PC foi restrita a alguns modelos selecionados e a determinadas janelas de tempo. O custo computacional do PCT também impôs restrições à escala dos experimentos, especialmente em conjuntos maiores. Futuramente, seria interessante explorar a aplicação de variantes mais eficientes da técnica, como o PCI.

Outra direção relevante seria ampliar o escopo clínico dos experimentos, avaliando modelos em múltiplos conjuntos de dados institucionais para verificar sua generalização entre hospitais. Adicionalmente, o desenvolvimento de interfaces interpretáveis para uso clínico em tempo real, associadas a mecanismos de abstinência baseados em incerteza, representa um passo necessário para a transição desses modelos da pesquisa para a prática médica.

# REFERÊNCIAS


ALMEIDA, N. R. C. d. et al. Análise de tendência de mortalidade por sepse no Brasil e por regiões de 2010 a 2019. **Revista de Saúde Pública**, [S.l.], v.56, p.25, 2022.

ANGELOPOULOS, A. N.; BATES, S. **A gentle introduction to conformal prediction and distribution-free uncertainty quantification**. 2021.

BOMRAH, S. et al. A scoping review of machine learning for sepsis prediction-feature engineering strategies and model performance: a step towards explainability. **Critical Care**, [S.l.], v.28, n.1, p.180, 2024.

CALVERT, J. S. et al. A computational approach to early sepsis detection. **Computers in Biology and Medicine**, [S.l.], v.74, p.69–73, 2016.

DELAHANTY, R. J. et al. Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. **Annals of Emergency Medicine**, [S.l.], v.73, n.4, p.334–344, 2019.

DENG, H.-F. et al. Evaluating machine learning models for sepsis prediction: A systematic review of methodologies. **iScience**, [S.l.], v.25, n.1, 2022.

DEVOS, E. L. **Updates and Controversies in the Early Management of Sepsis and Septic Shock**. 2018.

DUGAR, S.; CHOUDHARY, C.; DUGGAL, A. Sepsis and septic shock: Guideline-based management. **Cleveland Clinic Journal of Medicine**, [S.l.], v.87, n.1, p.53–64, 2020.

EVANS, L. et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. **Intensive Care Medicine**, [S.l.], v.47, n.11, p.1181–1247, 2021.

FASCIA, M. **Machine learning applications in medical prognostics**: a comprehensive review. 2024.

FLEISCHMANN, C. et al. Assessment of global incidence and mortality of hospital-treated sepsis. Current estimates and limitations. **American Journal of Respiratory and Critical Care Medicine**, [S.l.], v.193, n.3, p.259–272, 2016.

FLEUREN, L. M. et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. **Intensive Care Medicine**, [S.l.], v.46, p.383–400, 2020.

GUO, C. et al. On calibration of modern neural networks. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 2017. **Anais...** PMLR, 2017. p.1321–1330.

HUNT, A. Sepsis: an overview of the signs, symptoms, diagnosis, treatment and pathophysiology. **Emergency Nurse**, [S.l.], v.31, n.6, 2023.

ISLAM, K. R. et al. Machine Learning-Based Early Prediction of Sepsis Using Electronic Health Records: A Systematic Review. **Journal of Clinical Medicine**, [S.l.], v.12, n.17, p.5658, 2023.

JOHNSON, A. E. et al. MIMIC-III, a freely accessible critical care database. **Scientific Data**, [S.l.], v.3, n.1, p.1–9, 2016.

JOHNSON, A. et al. **MIMIC-IV**. 2020. 49–55p. Accessed: 2021-08-23, `https://physionet.org/content/mimiciv/1.0/`.

JOST, M. T. et al. Morbimortalidade e custo por internação dos pacientes com sepse no Brasil, Rio Grande do Sul e Porto Alegre. **Revista de Epidemiologia e Controle de Infecção**, [S.l.], v.9, n.2, p.149–154, 2019.

KAM, H. J.; KIM, H. Y. Learning representations for the early detection of sepsis with deep neural networks. **Computers in Biology and Medicine**, [S.l.], v.89, p.248–255, 2017.

KAMALESWARAN, R. et al. Artificial intelligence may predict early sepsis after liver transplantation. **Frontiers in Physiology**, [S.l.], v.12, p.692667, 2021.

KIM, M.-H.; CHOI, J.-H. An update on sepsis biomarkers. **Infection & Chemotherapy**, [S.l.], v.52, n.1, p.1, 2020.

LEVY, M. M.; FINK, M. P. SCCM/ESICM/ACCP/ATs/SIS International Sepsis Definitions Conference. In: PMLR, 2003. **Anais...** [S.l.: s.n.], 2003.

MACHADO, F. R. et al. The epidemiology of sepsis in Brazilian intensive care units (the Sepsis PREvalence Assessment Database, SPREAD): an observational study. **The Lancet Infectious Diseases**, [S.l.], v.17, n.11, p.1180–1189, 2017.

MOOR, M. et al. Early prediction of sepsis in the ICU using machine learning: a systematic review. **Frontiers in Medicine**, [S.l.], v.8, p.607952, 2021.

NICULESCU-MIZIL, A.; CARUANA, R. Predicting good probabilities with supervised learning. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 22., 2005. **Proceedings...** [S.l.: s.n.], 2005. p.625–632.

PAPADOPOULOS, H.; VOVK, V.; GAMMERMAN, A. Conformal prediction with neural networks. In: IEEE INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE (ICTAI 2007), 19., 2007. **Anais...** IEEE, 2007. p.388–395.

POLLARD, T. J. et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. **Scientific Data**, [S.l.], v.5, n.1, p.1–13, 2018.

RAFIEI, A. et al. SSP: Early prediction of sepsis using fully connected LSTM-CNN model. **Computers in Biology and Medicine**, [S.l.], v.128, p.104110, 2021.

REYNA, M. A. et al. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. **Critical Care Medicine**, [S.l.], v.48, n.2, p.210–217, 2020.

SAEED, M. et al. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. In: COMPUTERS IN CARDIOLOGY, 2002. **Anais...** IEEE, 2002. p.641–644.

SEYMOUR, C. W. et al. Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). **JAMA**, [S.l.], v.315, n.8, p.762–774, 2016.

SHASHIKUMAR, S. P. et al. Artificial intelligence sepsis prediction algorithm learns to say “I don’t know”. **NPJ Digital Medicine**, [S.l.], v.4, n.1, p.134, 2021.

SINGER, M. et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). **JAMA**, [S.l.], v.315, n.8, p.801–810, 2016.

VOVK, V.; GAMMERMAN, A.; SHAFER, G. **Algorithmic Learning in a Random World**. New York: Springer, 2005.

World Health Organization et al. **WHO sepsis technical expert meeting, 16–17 January 2018**. 2018. [S.l.]: World Health Organization, 2018.

YANG, M. et al. Development and Validation of an Interpretable Conformal Predictor to Predict Sepsis Mortality Risk: Retrospective Cohort Study. **Journal of Medical Internet Research**, [S.l.], v.26, p.e50369, 2024.

ZHANG, D. et al. An interpretable deep-learning model for early prediction of sepsis in the emergency department. **Patterns**, [S.l.], v.2, n.2, 2021.

ZHAO, X.; SHEN, W.; WANG, G. Early prediction of sepsis based on machine learning algorithm. **Computational Intelligence and Neuroscience**, [S.l.], v.2021, n.1, p.6522633, 2021.