Identification of Sepsis Subphenotypes via ICU Data Clustering

Giovana Assis[1], Victoria F. Mello[1], Haniel B. Ribeiro[1], Alexandre G. Barros[2] , Gisele L. Pappa[1], Wagner Meira Jr.[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Belo Horizonte – MG – Brasil

[2]Faculdade de Medicina – Universidade Federal de Minas Gerais
Belo Horizonte – MG – Brasil

{giovana.assis,victoriaflores,haniel.botelho,glpappa,meira}@dcc.ufmg.br

{xandebarros}@gmail.com

Abstract. *This study aims to identify sepsis subphenotypes in ICU patients using clustering techniques on MIMIC-IV clinical data. Missing data were imputed using MissForest. We applied UMAP (trustworthiness = 0.97) for dimensionality reduction and K-means (K = 5, silhouette score = 0.30) for clustering, revealing five distinct subphenotypes with unique clinical characteristics. Our findings suggest that these subphenotypes could guide efficient diagnosis and personalized treatments, surpassing the current SOFA score's limitations. The results highlight the need for further exploration of additional variables to improve sepsis diagnosis and treatment.*

Resumo. *Este estudo tem como objetivo identificar subfenótipos de sepse em pacientes de UTI usando técnicas de clusterização em dados clínicos do MIMIC-IV. Dados ausentes foram imputados usando MissForest. Aplicamos UMAP (trustworthiness = 0,97) para redução de dimensionalidade e K-means (K = 5, silhouette score = 0,30) para clusterização, revelando cinco subfenótipos distintos com características clínicas únicas. Nossos achados sugerem que esses subfenótipos podem orientar diagnósticos mais eficientes e tratamentos personalizados, superando as limitações do escore SOFA atual. Os resultados destacam a necessidade de explorar variáveis adicionais para melhorar o diagnóstico e tratamento da sepse.*

1. Introdução

A sepse é uma resposta descontrolada do corpo a infecções, sendo uma condição clínica grave e frequentemente letal [Seymour et al. 2019]. Essa síndrome afeta cerca de 48,9 milhões de pessoas anualmente em todo o mundo, com taxas de mortalidade que podem chegar a 35,3% em ambientes hospitalares, conforme observado em uma auditoria que incluiu 10.069 pacientes de diversos continentes, como Europa, Ásia e América [Sakr et al. 2018]. A diversidade nas manifestações de sepse e as respostas imunes variadas entre os pacientes constituem desafios significativos para o desenvolvimento de tratamentos eficazes.

A variabilidade na resposta à sepse entre os pacientes dificulta a criação de tratamentos padronizados. Nesse contexto, a identificação de subfenótipos de sepse pode oferecer uma estratégia promissora para personalizar as intervenções terapêuticas. Estudos recentes sugerem que subgrupos de pacientes com sepse podem ser delineados com base em características demográficas, laboratoriais e clínicas semelhantes, possibilitando tratamentos mais direcionados e eficazes [Seymour et al. 2019].

Embora modelos de aprendizado de máquina tenham demonstrado potencial ao utilizar grandes volumes de dados para melhorar a predição de desfechos clínicos em estudos anteriores, os resultados variáveis desses trabalhos refletem a complexidade e a heterogeneidade da sepse. Isso ressalta a necessidade de desenvolver abordagens que possam capturar essa variabilidade, especialmente nas fases iniciais da internação em unidades de terapia intensiva.

Este estudo aborda o desafio de identificar subfenótipos de sepse em pacientes de UTIs (Unidades de Terapia Intensiva), uma condição altamente heterogênea e letal, utilizando dados clínicos disponíveis na base de dados MIMIC-IV [Johnson et al. 2023], reconhecida por sua riqueza em informações clínicas. O objetivo é não apenas identificar subfenótipos, mas também examinar como diferentes técnicas de aprendizado de máquina podem ajudar a superar as limitações do SOFA no diagnóstico e tratamento da sepse.

O trabalho também enfrenta desafios relacionados à presença de dados faltantes, especialmente em exames laboratoriais, o que exigiu a aplicação de técnicas de imputação para manter a integridade da análise. Para identificar grupos homogêneos de pacientes sépticos, utilizamos o UMAP (*Uniform Manifold Approximation and Projection for Dimension Reduction*) [McInnes et al. 2018] para a redução da dimensionalidade e o K-means para o agrupamento. Os resultados revelaram a existência de cinco subfenótipos de sepse com características distintas. Assim, este trabalho busca não apenas contribuir para uma melhor compreensão desses subfenótipos, visando o desenvolvimento de ferramentas diagnósticas mais eficazes, como também demonstrar como a combinação de diferentes técnicas de aprendizado de máquina pode ser uma abordagem eficaz para lidar com desafios dessa natureza.

2. Trabalhos Relacionados

O estudo de subfenótipos de sepse tem recebido atenção significativa devido à sua relevância na melhoria dos tratamentos. O trabalho de [Seymour et al. 2019] é um estudo importante nessa área, destacando a identificação de fenótipos clínicos distintos para aprimorar a terapia e os cuidados. Com dados de 20.189 pacientes, o estudo aplicou a técnica de clusterização K-means, derivando quatro fenótipos clínicos. Os fenótipos α, β, γ e δ, baseados em variáveis clínicas, correlacionaram-se com padrões de biomarcadores e desfechos clínicos, revelando diferenças em mortalidade e respostas ao tratamento. Este estudo demonstra a utilidade da clusterização na compreensão da heterogeneidade da sepse.

Por outro lado, [Fohner et al. 2019] utilizaram a modelagem de tópicos não supervisionada para investigar a variabilidade nos padrões de tratamento da sepse. Com dados de 29.253 pacientes e a técnica *Latent Dirichlet Allocation* (LDA), identificaram 42 tópicos de tratamento clínico. A análise indicou uma variabilidade significativa nos padrões de tratamento e sugere que a modelagem de tópicos pode fornecer uma caracterização clínica detalhada, contrastando com a abordagem centrada em fenótipos

clínicos de [Seymour et al. 2019].

Adicionalmente, [Ibrahim et al. 2020] analisaram a heterogeneidade na sepse ao estratificar os tipos de disfunção orgânica em pacientes com sepse na UTI. O estudo concluiu que a estratificação de subpopulações de sepse melhora a previsão em modelos de aprendizado de máquina, destacando a importância de considerar a heterogeneidade clínica para aprimorar a personalização dos modelos. Embora compartilhe a ênfase na heterogeneidade com o estudo de [Seymour et al. 2019], foca na estratificação das disfunções orgânicas.

A escolha da técnica de agrupamento é crucial, pois pode impactar os resultados. Nesse contexto, [Koutroulis et al. 2022] compararam a Análise de Classe Latente (LCA) com K-means para derivar fenótipos clínicos na sepse pediátrica. O estudo encontrou que a LCA ofereceu uma segmentação mais robusta em comparação ao K-means, indicando a eficácia superior da LCA na análise de coortes heterogêneas e na orientação de terapias direcionadas. Essa comparação ilustra a eficácia de abordagens alternativas em relação ao K-means.

Apesar dos achados de [Koutroulis et al. 2022], que consideraram a LCA como modelo mais robusto, a pesquisa de [Hu et al. 2022] focou na identificação de subfenótipos de sepse utilizando K-means na base de dados MIMIC-IV, semelhante ao que foi feito em nosso trabalho. O estudo revelou dois subfenótipos com perfis clínicos distintos, mostrando que a análise de grupos baseada em dados clínicos rotineiros pode identificar rapidamente subfenótipos com desfechos diferentes.

Este trabalho visa expandir essas abordagens ao explorar uma variedade de técnicas de aprendizado de máquina, incluindo UMAP e K-means, além de métodos de imputação de dados como KNN e MissForest. Embora compartilhemos algumas técnicas com estudos anteriores, nosso projeto se destaca pela integração de múltiplas abordagens para identificar não apenas subfenótipos, mas também novas variáveis que possam aprimorar a identificação da sepse. Devido às características dos dados utilizados, que não provêm de um estudo longitudinal, os resultados não são diretamente comparáveis aos de pesquisas anteriores, mas as métricas encontradas são relevantes para entender as dinâmicas da sepse e possibilitar o desenvolvimento de estratégias de tratamento mais adequadas às necessidades dos pacientes e de um diagnóstico mais eficaz.

3. Caracterização dos Dados

O conjunto de dados utilizado neste estudo é oriundo do MIMIC-IV, um banco de dados público amplamente reconhecido, desenvolvido em colaboração entre o Beth Israel Deaconess Medical Center (BIDMC) e o Massachusetts Institute of Technology (MIT). O MIMIC-IV compreende registros eletrônicos de saúde (EHR) de pacientes internados em Unidades de Tratamento Intensivo (UTI), coletados entre 2008 e 2019. O MIMIC-IV é notável pela sua riqueza e heterogeneidade, fornecendo uma base de dados robusta para estudos que exigem informações detalhadas e variadas sobre pacientes críticos como descrito em [Johnson et al. 2023].

Inicialmente, a base de dados MIMIC-IV conta com um total de 431.231 admissões hospitalares, das quais 73.181 referem-se a admissões em UTI [Johnson et al. 2023]. Para o presente estudo, aplicamos uma série de critérios de

filtragem para selecionar um subconjunto de 36.581 internações. Os critérios de inclusão foram: (i) pacientes que estiveram internados em um leito de UTI, sendo que cada entrada na UTI foi considerada como um caso distinto, mesmo que se trate do mesmo paciente; (ii) pacientes diagnosticados com sepse, englobando tanto aqueles que foram admitidos na UTI já com o diagnóstico de sepse quanto aqueles que desenvolveram a condição durante a estadia na UTI, conforme a definição de sepse-3 apresentada em [Singer et al. 2016].

A partir desse conjunto de pacientes, extraímos um conjunto de variáveis clínicas para análise. Essas 29 variáveis foram selecionadas com base no estudo de [Seymour et al. 2019], que utilizou como critério de escolha a associação das variáveis com o início ou desfecho da sepse, sua incorporação em modelos conceituais de fisiopatologia da sepse e tolerância do hospedeiro, e sua disponibilidade no prontuário eletrônico do hospital – complementadas pela variável que representa a pontuação no escore SOFA, além de três novas variáveis demográficas, totalizando 33 variáveis. As variáveis demográficas incluíram idade no momento da internação, gênero, raça, estado civil e tipo de admissão. As variáveis de sinais vitais abrangeram frequência cardíaca, pressão arterial sistólica, temperatura corporal, frequência respiratória e saturação de oxigênio. As variáveis laboratoriais selecionadas foram ALT (alanina aminotransferase), AST (aspartato aminotransferase), albumina, contagem de neutrófilos imaturos (bands), bilirrubina total, troponina, nitrogênio ureico no sangue (BUN), proteína C-reativa, cloreto, creatinina, creatinina máxima, taxa de sedimentação de eritrócitos (ESR), glicose, relação normalizada internacional (INR), lactato, bicarbonato, hemoglobina, contagem de leucócitos (WBC), relação PaO2/FiO2 com e sem ventilação, e contagem de plaquetas. Além disso, incluímos variáveis relacionadas ao escore de gravidade, como o escore SOFA e a pontuação na Escala de Coma de Glasgow (GCS).

Para cada uma dessas variáveis, extraímos o valor mais anormal (que foge o padrão de normalidade) registrado nas primeiras 24 horas após a admissão na UTI. Os dados faltantes foram considerados como sendo faltantes aleatórios, e variáveis com mais de 80% de valores nulos foram excluídas da análise. Foram excluídas também variáveis com alta correlação, de maneira que selecionamos apenas uma variável de cada par com correlação elevada, baseando nossa escolha na variável com a menor porcentagem de valores nulos.

As variáveis obtidas ao final podem ser observadas na Tabela 1. Os modelos selecionados para o tratamento dos valores nulos nas variáveis selecionadas serão detalhados na seção seguinte.

4. Metodologia

Para a identificação de subfenótipos de sepse, adotamos uma abordagem sistemática para o tratamento dos dados, conforme descrito a seguir. Após a preparação dos dados, detalhada na seção anterior, realizamos a imputação de valores nulos. Embora variáveis com mais de 80% de valores ausentes tenham sido removidas, ainda assim restaram dados faltantes em nosso conjunto. Para lidar com esses dados, testamos dois métodos de imputação: KNN (*K-Nearest Neighbors*) e MissForest. O KNN é um algoritmo que preenche valores ausentes com base na média dos valores de seus k vizinhos mais próximos. Em contrapartida, o MissForest utiliza florestas aleatórias para prever valores ausentes,

Tabela 1. Tabela contendo o conjunto final de variáveis utilizadas.

Laboratoriais	**Demográficas**
ALT (U/L)	Idade
Albumina (g/dL)	Tipo de admissão
Bilirrubina (mg/dL)	Estado civil
BUN (mg/dL)	Raça
Cloreto (mEq/L)	Gênero
Glicose (mg/dL)	**Sinais Vitais**
INR	Frequência cardíaca (bpm)
Lactato (mmol/L)	Saturação de oxigênio (%)
Leucócitos (WBC) ($\times 10^3/\mu L$)	Frequência respiratória (respirações/min)
Bicarbonato (mEq/L)	Temperatura (°C)
Hemoglobina (g/dL)	**Escores**
Relação PaO2/FiO2 com ventilação	SOFA
Contagem de plaquetas ($\times 10^3/\mu L$)	Escala de Coma de Glasgow

avaliando relações simultâneas entre as diferentes variáveis do conjunto de dados para capturar padrões e dependências entre elas.

O segundo passo da nossa metodologia foi a redução da dimensionalidade dos dados, uma etapa que visa simplificar o conjunto de dados e melhorar a eficiência computacional. Para essa finalidade, avaliamos dois métodos populares de redução de dimensionalidade: UMAP e t-SNE (*t-Distributed Stochastic Neighbor Embedding*). O UMAP preserva tanto a estrutura local quanto a global dos dados, tornando-o eficaz em manter as relações inerentes mesmo após a redução dimensional. Além disso, o UMAP se destaca por sua eficiência computacional em conjuntos de dados complexos. Por outro lado, o t-SNE é conhecido por sua capacidade de separar claramente clusters em espaços de baixa dimensionalidade, facilitando a visualização dos dados. No entanto, uma limitação do t-SNE é sua incapacidade de gerar uma função de mapeamento explícita que pode ser aplicada a novos dados, o que restringe sua utilidade em cenários de aprendizado supervisionado onde a generalização é necessária.

Após a imputação dos dados e a redução da dimensionalidade, avançamos para o agrupamento. Nesta etapa, comparamos dois modelos: K-means e *Gaussian Mixture Model* (GMM). O K-means é um dos métodos de agrupamento mais utilizados, que busca minimizar a variância intra-cluster, assumindo que os grupos são esféricos e de tamanho similar. O GMM, por sua vez, é um modelo probabilístico que presume que os dados são gerados a partir de uma mistura de distribuições gaussianas, permitindo a formação de grupos com formas e tamanhos variados. O K-means foi escolhido inicialmente devido à sua aplicação frequente em estudos semelhantes, enquanto o GMM foi considerado por sua flexibilidade em capturar a complexidade das distribuições dos dados. Ambos os modelos foram treinados e avaliados em relação à sua capacidade de criar grupos que refletem a estrutura subjacente dos dados.

5. Experimentos e Resultados

Esta seção apresenta os experimentos e resultados obtidos ao realizarmos inicialmente uma análise quantitativa seguida de uma análise qualitativa dos dados, feita por um especialista.

5.1. Análise Quantitativa

Nesta seção, apresentamos os experimentos realizados, organizados em três partes: (i) seleção dos melhores métodos de imputação de valores nulos, (ii) redução de dimensionalidade e (iii) agrupamento para identificar subfenótipos de sepse no nosso conjunto de dados. Cada experimento foi estruturado para investigar a eficácia dos métodos escolhidos, priorizando a identificação de padrões significativos nos dados.

O primeiro experimento focou na comparação entre os métodos KNN e MissForest para a imputação de valores nulos. O KNN foi testado com k = 5, enquanto o MissForest foi utilizado com sua configuração padrão. A análise foi baseada no desempenho dos métodos em termos do escore de silhueta, um indicador da coesão e separação dos grupos formados após a imputação. O MissForest apresentou desempenho superior, alcançando uma silhueta de 0,303 para 5 grupos, enquanto o KNN obteve como melhor resultado uma silhueta de 0,284 para 6 grupos. Apesar de a escolha do método de imputação ser uma etapa importante, optamos por não incluir visualizações desses experimentos devido a limitações de espaço.

O segundo experimento visou a escolha do modelo mais adequado para a redução da dimensionalidade dos dados, no qual avaliamos os métodos UMAP e t-SNE. O UMAP foi testado variando seu número de componentes de 2 a 10, enquanto o t-SNE, devido a sua limitação em cenários com número alto de dimensões, foi testado com 2 a 3 componentes.

Para validar a eficácia da redução da dimensionalidade, utilizamos a métrica de *trustworthiness*, que avalia quão bem a estrutura local dos dados originais é preservada na projeção de baixa dimensionalidade. Essa métrica varia de 0 a 1, com valores próximos de 1 indicando maior preservação das relações de vizinhança. O *trustworthiness* é calculado considerando o número total de amostras n e os k vizinhos mais próximos de cada amostra i. O ranking $r(i, j)$ de cada vizinho j de i no espaço original é comparado com sua posição na projeção de baixa dimensionalidade. Se a posição de j muda significativamente na projeção, a diferença é penalizada, reduzindo o valor do *trustworthiness*. A fórmula utilizada para cálculo do *trustworthiness* é apresentada na Equação 1.

$$T(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in N_i \; k} \max\left(0, (r(i, j) - k)\right) \tag{1}$$

Os resultados obtidos no experimento mostraram que o índice de *trustworthiness* aumentou de 0,84 para 0,97 à medida que o número de componentes aumentava de 2 para 6, estabilizando em torno deste valor para componentes adicionais (Figura 1). O t-SNE, por sua vez, apresentou bons resultados ao reduzir os dados para 3 componentes, evidenciando grupos bem definidos. No entanto, devido à sua limitação em gerar uma função de mapeamento explícita para novos dados de validação, decidimos que o UMAP seria a

escolha mais apropriada, especialmente porque o UMAP, configurado para 6 componentes, apresentou o melhor índice de *trustworthiness* – 0,97, preservando de forma eficaz as relações de proximidade entre os pontos de dados originais no espaço reduzido.

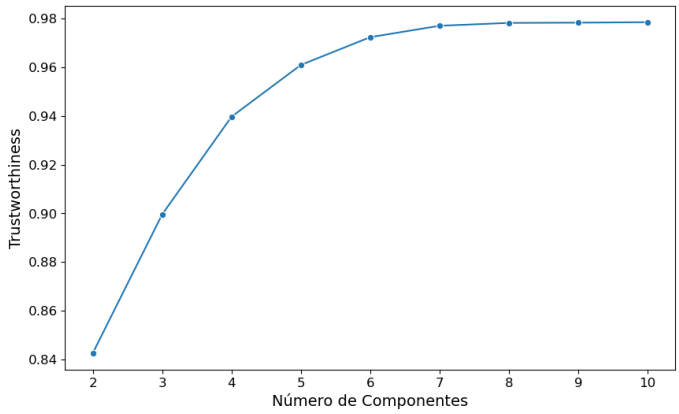

Figura 1. Gráfico de *trustworthiness* por Número de Componentes no UMAP.

O terceiro experimento teve como objetivo comparar os modelos de clusterização K-means e GMM. Ambos os modelos foram aplicados aos dados pré-processados utilizando MissForest para imputação e UMAP com 6 componentes para redução de dimensionalidade. O K-means foi testado com o número de grupos variando de 2 a 10, enquanto o GMM foi avaliado sob as mesmas condições, utilizando covariância do tipo esférica, que, em testes preliminares, se mostrou a mais eficaz. Os resultados indicaram que, embora ambos os métodos apresentassem desempenho semelhante, o K-means obteve valores de silhueta superiores, atingindo 0,30 para 5 grupos, enquanto o GMM alcançou 0,29 (Figura 2). Com base nos resultados apresentados na Tabela 2, que detalha as métricas de avaliação do K-means – incluindo a menor distância intra-cluster, maior distância inter-cluster e o índice de Davies-Bouldin, uma métrica que avalia a compacidade e separação dos grupos, de modo que um valor mais baixo do DBI indica uma melhor qualidade de clusterização, significando que os clusters são compactos e bem separados – concluímos que o K-means executado com o valor de k igual a 5 era a escolha mais apropriada para a nossa análise de subfenótipos.

Tabela 2. Métricas de Avaliação por Número de Grupos usando K-means.

k	Silhueta	Dist. Intra-Cluster	Dist. Inter-Cluster	Davies-Bouldin Ind.
2	0.268864	2.40978	3.32456	1.48618
3	0.271941	2.1407	3.26371	1.43886
4	0.249301	1.93756	3.23571	1.33285
5	0.303211	1.7026	3.20945	1.15952
6	0.27048	1.64664	3.12592	1.32761
7	0.269163	1.59189	3.05623	1.33131
8	0.290649	1.42281	3.25714	1.21797
9	0.28548	1.3768	3.21006	1.15969
10	0.280816	1.34303	3.13898	1.19137

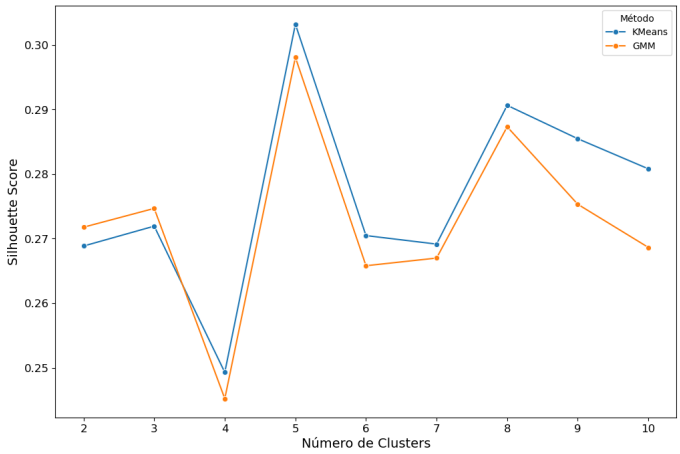

Figura 2. Gráfico comparativo da silhueta para diferentes valores de K obtidos com K-means e GMM.

Os experimentos e a análise das métricas utilizadas, juntamente com suas visualizações, possibilitaram a seleção dos métodos mais adequados para cada etapa, resultando em uma segmentação mais clara dos subfenótipos de sepse. Os resultados e as características dos grupos identificados serão discutidos na próxima seção, onde exploraremos suas implicações.

5.2. Análise Qualitativa

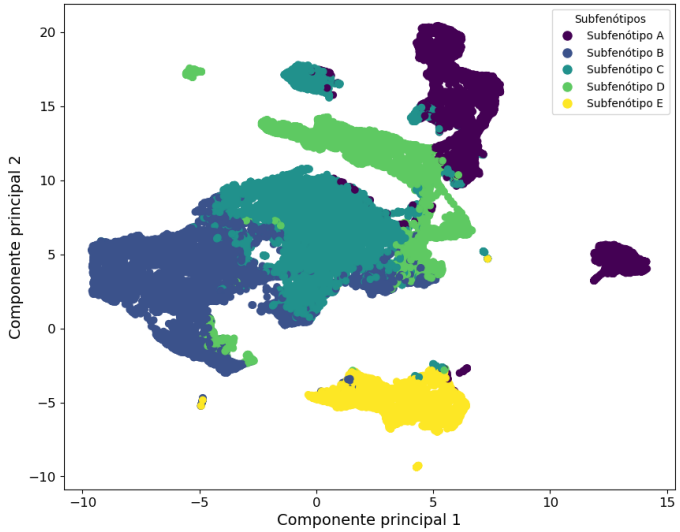

Figura 3. Visualização dos grupos.

Antes de analisarmos as variáveis mais relevantes de cada grupo, podemos observar a Figura 3, que ilustra, segundo as duas componentes mais relevantes do UMAP, os grupos encontrados em um espaço 2D. Embora nem todos os grupos ocupem uma região única do espaço devido à alta redução de dimensionalidade utilizada para construir o gráfico, no geral, eles estão bem separados.

A Tabela 3 apresenta os intervalos de algumas características dos grupos identificados, que, segundo especialistas, são frequentemente utilizadas para avaliar as condições

dos pacientes. As duas primeiras linhas da tabela listam o número de pacientes em cada grupo e a taxa de mortalidade, seguidas da mediana (Q1 - Q3) de outras variáveis relevantes. Já a Figura 4 apresenta os boxplots de 16 variáveis – selecionadas por um especialista dentre as 24 disponíveis pelo papel que desempenham na definição do quadro clínico do paciente – para cada um dos 5 subfenótipos identificados, permitindo observar um panorama mais detalhado das características clínicas de cada grupo.

Observando o subfenótipo A, por exemplo, notamos que ele abriga pacientes com os maiores valores de SOFA, menores contagens de plaquetas — indicando um sistema imunológico debilitado — e PaO2/FiO2 entre os mais baixos, sugerindo possíveis problemas pulmonares. Os boxplots dessa categoria, observados na Figura 4, exibem variáveis como ALT, BUN, INR e lactato, revelam valores extremos, corroborando a presença de disfunções multissistêmicas graves e o alto índice de mortalidade observado.

Em contraste, o subfenótipo B apresenta valores de SOFA mais baixos e contagens de plaquetas mais altas, caracterizando-o como o grupo com menor risco. Entretanto, é importante ressaltar que os níveis de creatinina são os mais altos entre todos os grupos, sugerindo problemas na função renal. Os boxplots refletem essa menor variabilidade e valores medianos reduzidos em variáveis críticas, como frequência respiratória e lactato, reforçando a ideia de um quadro clínico menos severo.

Os subfenótipos C e D, que englobam mais de 50% dos pacientes sépticos, compartilham características semelhantes, mas uma diferença significativa se destaca: os pacientes do subfenótipo D estão, em média, um dia mais tempo na UTI que os do subfenótipo C. Essa maior duração de internação se reflete em menores contagens de plaquetas e função pulmonar mais debilitada. As sutis diferenças observadas nos boxplots, como em frequência respiratória e saturação de oxigênio, sugerem uma piora no estado clínico dos pacientes do grupo D, possivelmente devido ao tempo mais prolongado na UTI.

Por fim, o subfenótipo E, que inclui pacientes com um tempo de internação médio de 4 dias e um SOFA score de 7, apresenta padrões de variabilidade nas variáveis clínicas que indicam uma complexidade moderada, mas ainda assim com um risco de mortalidade significativo, conforme evidenciado pelos seus boxplots.

As diferenças observadas entre os subfenótipos, evidenciadas na análise integrada das características da Tabela 3 e dos boxplots da Figura 4, ressaltam a complexidade intrínseca da sepse e expõem limitações do SOFA escore. Observamos que grupos com SOFA semelhantes foram alocados em subfenótipos diferentes, sugerindo que esse escore, isoladamente, pode não capturar nuances importantes da condição clínica dos pacientes. Por exemplo, embora as médias do SOFA entre os subfenótipos A e E sejam próximas, esses grupos apresentam comportamentos clínicos significativamente distintos. Esses achados sublinham a necessidade de uma abordagem multifatorial no diagnóstico e tratamento, que vá além dos indicadores tradicionais. A inclusão de um maior número de variáveis de análise se apresenta como um caminho promissor para aprimorar tanto a modelagem preditiva quanto a compreensão das manifestações da sepse. Com uma base de atributos mais ampla, seria possível alcançar agrupamentos mais precisos e obter insights mais profundos, possibilitando intervenções mais direcionadas e eficazes, além de facilitar a personalização do tratamento.

Tabela 3. Valores da mediana (Q1-Q3) de diferentes características para cada subfenótipo.

Características	Subfenótipo A	Subfenótipo B	Subfenótipo C	Subfenótipo D	Subfenótipo E
# Pacientes	5725	10401	11008	5681	3766
Mortalidade, n (%)	1404 (24.52)	694 (6.67)	1346 (12.23)	661 (11.64)	765 (20.31)
Tempo na UTI (h)	74.59 (38.87 - 164.65)	45.77 (17.24 - 113.80)	50.80 (21.15 - 113.37)	71.14 (33.61 - 158.71)	99.49 (47.67 - 220.67)
Bilirrubina	3.1 (1.10 - 6.50)	0.6 (0.40 - 1.00)	0.6 (0.40 - 1.20)	0.7 (0.40 - 1.20)	0.6 (0.40 - 1.20)
Creatinina	1.6 (1.00 - 2.80)	0.9 (0.70 - 1.30)	1.4 (0.90 - 2.50)	1.1 (0.80 - 1.50)	1.1 (0.80 - 1.70)
SOFA score	8.0 (6.00 - 11.00)	3.0 (2.00 - 5.00)	4.0 (3.00 - 6.00)	5.0 (4.00 - 8.00)	7.0 (5.00 - 10.00)
Plaquetas	116 (64.00 - 179.00)	176 (129.00 - 233.00)	194 (129.00 - 284.00)	147 (109.00 - 197.00)	169 (119.00 - 226.00)
PaO2/FiO2 com vent.	195.5 (128.00 - 272.50)	222.5 (154.00 - 320.00)	209.5 (152.00 - 272.00)	188.0 (123.33 - 272.50)	210.0 (136.67 - 306.92)

6. Conclusões

Neste estudo, abordamos a identificação de subfenótipos de sepse em pacientes de UTI, utilizando técnicas de imputação de dados faltantes e redução de dimensionalidade combinadas com métodos de clusterização. Através da análise dos dados clínicos disponíveis no MIMIC-IV, conseguimos identificar cinco subfenótipos distintos de sepse, sugerindo que a variabilidade dessa condição não é completamente capturada por métricas tradicionais como o escore SOFA.

Os subfenótipos identificados apontam para a necessidade de uma análise mais detalhada da sepse, dado que os padrões encontrados refletem a heterogeneidade da resposta clínica dos pacientes. Isso reforça a importância de se considerar um conjunto mais amplo de variáveis e abordagens metodológicas mais refinadas ao investigar condições complexas como a sepse. A metodologia aplicada neste trabalho, incluindo a seleção e tratamento dos dados, demonstrou a viabilidade de integrar múltiplas técnicas analíticas para lidar com a natureza não supervisionada do problema.

Embora os resultados ofereçam uma nova perspectiva sobre a estratificação de pacientes sépticos, ainda há limitações a serem enfrentadas. A inclusão de variáveis adicionais e a validação dos subfenótipos em diferentes bases de dados são caminhos que podem refinar a análise e proporcionar um entendimento mais robusto das manifestações da sepse.

Em resumo, os resultados deste estudo revelam subfenótipos distintos de sepse, oferecendo uma visão mais aprofundada sobre a diversidade de manifestações da doença. Essa identificação aponta para a necessidade de aprimorar métricas como o escore SOFA, com o objetivo de potencialmente melhorar a eficácia das intervenções clínicas. Contudo, é imprescindível validar essas descobertas e aprofundar a análise do impacto dessas diferenças no tratamento e prognóstico dos pacientes. As conclusões apresentadas, portanto, devem ser vistas como um ponto de partida para investigações futuras, visando não apenas refinar as metodologias empregadas, mas também ampliar a compreensão das

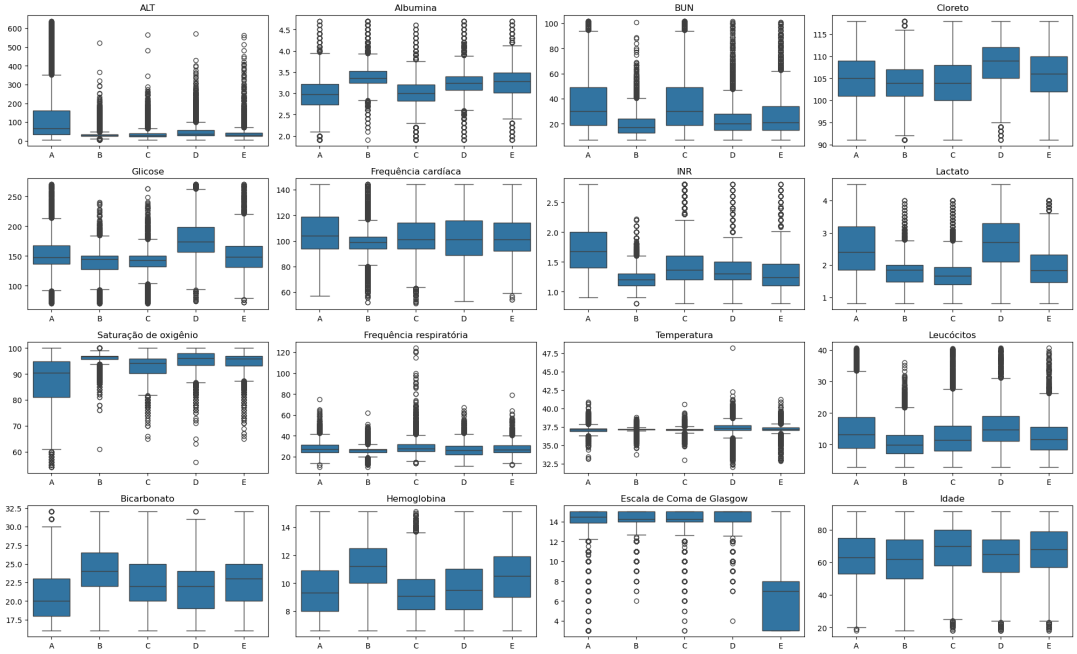

Figura 4. Distribuição em Boxplot dos 5 grupos para 16 variáveis analisadas.

complexidades que envolvem a sepse.

Referências

Fohner, A. E., Greene, J. D., Lawson, B. L., Chen, J. H., Kipnis, P., Escobar, G. J., and Liu, V. X. (2019). Assessing clinical heterogeneity in sepsis through treatment patterns and machine learning. *Journal of the American Medical Informatics Association*, 26(12):1466–1477.

Hu, C., Li, Y., Wang, F., et al. (2022). Application of machine learning for clinical subphenotype identification in sepsis. *Infectious Disease Therapy*, 11:1949–1964.

Ibrahim, Z. M., Wu, H., Hamoud, A., Stappen, L., Dobson, R. J. B., and Agarossi, A. (2020). On classifying sepsis heterogeneity in the icu: Insight using machine learning. *Journal of the American Medical Informatics Association*, 27(3):437–443.

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L. H., Celi, L. A., and Mark, R. G. (2023). Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Koutroulis, I., Velez, T., Wang, T., Yohannes, S., Galarraga, J. E., Morales, J. A., Freishtat, R. J., and Chamberlain, J. M. (2022). Pediatric sepsis phenotypes for enhanced therapeutics: An application of clustering to electronic health records. *Journal of the American College of Emergency Physicians Open*, 3(1):e12660.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Sakr, Y., Jaschinski, U., Wittebole, X., Szakmany, T., Lipman, J., Ñamendys Silva, S. A., Martin-Loeches, I., Leone, M., Lupu, M.-N., Vincent, J.-L., and Investigators, I.

(2018). Sepsis in intensive care unit patients: Worldwide data from the intensive care over nations audit. *Open Forum Infectious Diseases*, 5(12).

Seymour, C. W., Kennedy, J. N., Wang, S., et al. (2019). Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA*, 321(20):2003–2017.

Singer, M., Deutschman, C., Seymour, C., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G., Chiche, J., Coopersmith, C., Hotchkiss, R., Levy, M., Marshall, J., Martin, G., Opal, S., Rubenfeld, G., van der Poll, T., Vincent, J., and Angus, D. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810.