Avaliando os Potenciais e as Limitações dos Algoritmos de Imputação em Bases de Dados Clínicos

1st Izadora Monken Ganem
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
izadoraganem@gmail.com; 0009-0005-3902-5402

2nd Marcos André Gonçalves
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
magoncalv@gmail.com; 0000-0002-2075-3363

3rd Leonardo Chaves Dutra da Rocha
dept. name of organization (of Aff.)
Universidade Federal de São João del-Rei
São João del-Rei, Brasil
lcrocha@ufsj.edu.br; 0009-0005-1481-8377

***Abstract*—Missing values are a central challenge in clinical datasets, directly affecting the reliability of predictive models. This work compares traditional imputation strategies (mean, KNN, MICE, and MissForest) with a foundation-model-based approach (TabPFN), evaluating both imputation accuracy and impact on mortality prediction. We introduce a quadrant-based structural analysis of the patient feature space to examine how local data organization influences performance. Results show that imputation effectiveness depends on clinical structure and that TabPFN demonstrates greater robustness in heterogeneous regions with stable computational cost. Overall, we show that imputation is not a neutral preprocessing step: its integration into the predictive pipeline significantly affects outcomes, highlighting the need for adaptive, task-oriented strategies in real-world health-care settings.**

I. Introdução

Aprendizado de Máquina (AM) têm se consolidado como uma ferramenta central para diagnóstico e prognóstico clínico, permitindo análises em larga escala e apoio à decisão médica. No entanto, seu desempenho depende diretamente da qualidade dos dados. Em cenários reais, valores ausentes são frequentes em bases clínicas e, quando tratados inadequadamente, podem introduzir vieses, reduzir a confiabilidade dos modelos e comprometer decisões em saúde [1]. Assim, a imputação deixa de ser um simples pré-processamento e passa a desempenhar papel estratégico no pipeline preditivo.

Neste trabalho, investigamos de forma sistemática o impacto de diferentes estratégias de imputação em uma grande coorte multicêntrica com mais de 16.000 pacientes hospitalizados com COVID-19, provenientes de 25 hospitais, contemplando variáveis de sinais vitais, comorbidades e exames laboratoriais. Inicialmente, comparamos cinco métodos amplamente utilizados — média, KNN, GAIN, MissForest e MICE — avaliando tanto erro de imputação quanto custo computacional. Observamos que MissForest e MICE apresentam melhor desempenho global, porém com elevado custo, o que limita sua aplicação em cenários operacionais.

A partir da hipótese de que a similaridade clínica entre os pacientes influencia diretamente a qualidade da imputação, propusemos o KNN++, uma versão otimizada do KNN adaptada à estrutura da base analisada. Embora não supere os métodos estado-da-arte em métricas globais, o KNN++ mostrou-se altamente eficiente e competitivo em regiões estruturalmente homogêneas. Demonstramos que seu desempenho é condicionado por dois fatores estruturais: a similaridade média entre vizinhos e a entropia local da variável imputada, evidenciando que a organização do espaço clínico é determinante.

Em uma etapa adicional, ampliamos a análise ao incorporar o TabPFN [2], modelo de fundação baseado em arquitetura Transformer para dados tabulares, explorado em cenário zero-shot. Avaliamos não apenas a acurácia da imputação, mas também seu impacto direto na predição de mortalidade hospitalar, integrando métodos tradicionais, análise estrutural por quadrantes e abordagens fundacionais.

Nossos objetivos foram: (i) comparar métodos tradicionais e baseados em modelos de fundação considerando erro de imputação e custo computacional; (ii) analisar como a estrutura local dos dados influencia o desempenho por meio de uma abordagem baseada em quadrantes; e (iii) avaliar o efeito da imputação na tarefa final de predição de mortalidade sob diferentes configurações de treino e teste.

De forma sintética, os resultados evidenciam que menor erro de imputação não implica, necessariamente, melhor desempenho clínico ou melhor relação custo-benefício. Mostramos que a estrutura local do espaço de pacientes — capturada por medidas de similaridade e entropia — condiciona fortemente o comportamento dos métodos, motivando a proposta do KNN++, que alcança desempenho competitivo em regiões homogêneas com custo significativamente reduzido. Ao in-

corporar o TabPFN na segunda etapa do trabalho, observamos maior robustez em cenários heterogêneos e estabilidade computacional. Além disso, demonstramos que a imputação impacta diretamente a predição de mortalidade, sendo mais eficaz quando integrada de forma sistêmica ao pipeline (com imputação conjunta de treino e teste). Esses achados reforçam a necessidade de uma abordagem orientada à tarefa, que considere simultaneamente estrutura dos dados, custo e impacto preditivo.

Todas as implementações e execuções dos experimentos foram realizadas pela aluna Izadora Ganem, sob a orientação dos professores Leonardo Rocha e Marcos Gonçalves. A concepção do projeto e as análises dos resultados foram desenvolvidas de forma conjunta entre a aluna e os professores. Este trabalho consolida as duas etapas do Projeto Orientado em Computação. A primeira etapa (POC 1) resultou na publicação de um artigo no SBBD 2025 (A4) [3]. A segunda etapa (POC 2) resultou na submissão de um artigo completo ao periódico BMC Medical Informatics and Decision Making (em avaliação).

II. Revisão da Literatura

A imputação de dados ausentes em bases clínicas é um tema bastante discutido na literatura [4], principalmente devido à complexidade e ao impacto prático que esses dados incompletos podem causar na área da saúde. Entre os métodos tradicionais baseados em aprendizado de máquina, destacam-se o MissForest e o Multiple Imputation by Chained Equations (MICE), frequentemente apontados como estado-da-arte na imputação de dados. O MissForest é um método não paramétrico que utiliza florestas aleatórias para estimar valores ausentes de forma iterativa [5]. Já o MICE realiza imputação múltipla por meio de uma sequência de modelos condicionais, em que cada variável com dados faltantes é modelada com base nas demais, refinando as estimativas até atingir convergência [6].

Além desses métodos, abordagens baseadas em aprendizado profundo também vêm sendo exploradas. O GAIN (Generative Adversarial Imputation Nets) [7], por exemplo, utiliza redes adversariais generativas compostas por um gerador responsável por imputar os valores ausentes e um discriminador, que tenta diferenciar valores reais de valores imputados. Esse tipo de abordagem permite capturar relações mais complexas nos dados, mas necessita de um ajuste fino de hiperparâmetros.

Apesar da robustez desses métodos mais sofisticados, o custo computacional pode ser um limitador em ambientes hospitalares, especialmente quando decisões precisam ser tomadas rapidamente. Nesse contexto, métodos mais simples como o k-Nearest Neighbors (KNN) [8] continuam sendo relevantes. O KNN realiza a imputação com base na média ou moda dos k vizinhos mais próximos, explorando padrões locais no espaço de atributos. Embora seja sensível à escolha de parâmetros, apresenta baixo custo computacional, é fácil de implementar e possui alta interpretabilidade.

Paralelamente à evolução das técnicas de imputação, modelos de fundação baseados em arquiteturas Transformer têm emergido como alternativas promissoras às abordagens tradicionais. Entre eles, destaca-se o Tabular Prior-Data Fitted Network (TabPFN) [9], que emprega in-context learning treinado em milhões de tarefas sintéticas para realizar inferência em uma única passagem, com mínima necessidade de ajuste de hiperparâmetros. Essa característica torna o TabPFN especialmente atrativo para a área da saúde, onde eficiência computacional e desempenho consistente são determinantes.

III. Metodologia Experimental

Para conduzir o estudo utilizamos uma coorte multicêntrica [10]–[12], composta por aproximadamente 16.000 pacientes hospitalizados com COVID-19 no Brasil, provenientes de 25 hospitais. A base contém 62 variáveis clínicas coletadas desde a admissão até os desfechos hospitalares, apresentando cerca de 16% de dados ausentes, o que impõe desafios relevantes para a construção de modelos preditivos confiáveis. Para garantir maior controle experimental, foi utilizado um recorte referente à primeira onda da COVID-19.

A. Avaliação Comparativa dos Métodos de Imputação e Análise Estrutural da Vizinhança

O primeiro objetivo deste trabalho foi comparar cinco métodos de imputação: Média, K-Nearest Neighbors (KNN), Generative Adversarial Imputation Networks (GAIN), Multiple Imputation by Chained Equations (MICE) e MissForest.

Inicialmente, a avaliação foi centrada na variável Sat/FiO2 (razão entre saturação de oxigênio e fração inspirada de oxigênio), apontada como altamente relevante para predição de óbito em estudos anteriores [13]. Em seguida, foram removidas todas as observações com valores originalmente ausentes nessa variável, por não ter ground truth disponível para avaliar imputações sobre dados reais faltantes. A partir do subconjunto completo, foi introduzido valores faltantes artificiais em 50% das observações da variável de maneira aleatória utilizando o mecanismo Missing Completely at Random (MCAR) [14]. Essa estratégia permitiu comparar os valores imputados com os valores verdadeiros da variável.

Com o objetivo de ter validação estatística, realizamos k-fold cross validation com 10 partições. O conjunto com 50% de valores removidos foi dividido em dez subconjuntos aproximadamente iguais. Em cada iteração, nove partições foram utilizadas como treino (mantendo os valores reais conhecidos) e a décima foi utilizada para teste, onde os valores eram imputados. O processo foi repetido dez vezes, de modo que cada partição servisse uma única vez como teste. Para avaliar o desempenho dos métodos utilizamos o Erro Absoluto Médio (MAE) e do Erro Quadrático Médio (MSE) para medir a diferença entre os valores reais e os imputados.

Para investigar a hipótese central de que imputações baseadas em pacientes clinicamente semelhantes tendem a produzir estimativas mais precisas, realizamos ajustes no KNN personalizados para esta base. Inicialmente, foi realizada uma *permutation importance* [15] para verificar quais variáveis eram mais correlacionadas com a variável Sat/Fio2. Também

avaliamos diferentes métricas de distância (Manhattan, Euclidiana e similaridade de cosseno) e diferentes valores de k (3, 5 e 10 vizinhos). Os melhores resultados foram obtidos com k=3, utilizando distância de Manhattan e um subconjunto específico com as 10 variáveis clínicas mais relacionadas com a variável alvo da imputação: *vm* (ventilação mecânica), *po2/fio2* (pressão parcial de oxigênio/fração inspirada de oxigênio), *fr* (frequência respiratória), *óbito, HospitalGDP, plaquetas, fc* (frequência cardíaca), *bicarbonato, ph* (ph sanguíneo), *pcr* (proteína c reativa). Essa versão ajustada foi denominada KNN++ e avaliada no mesmo protocolo de validação cruzada, sendo comparada aos demais métodos.

Por fim, aprofundamos a análise investigando como o desempenho da imputação é influenciado pela estrutura local dos dados. Após as dez execuções da validação cruzada, cada paciente foi representado em um espaço bidimensional com o eixo X indicando a distância média aos k vizinhos mais próximos (indicador de similaridade local) e o eixo Y, a entropia da variável *Sat/Fio2* entre esses vizinhos, calculada conforme descrito por [16]. Esse espaço foi dividido em quatro quadrantes, com base na média desses indicadores, combinando alta ou baixa similaridade com alta ou baixa entropia local. O objetivo foi avaliar a relação entre erro de imputação e estrutura das vizinhanças.

A comparação entre o erro absoluto dos métodos KNN++, MissForest e MICE em cada quadrante foi realizada por meio do teste de Wilcoxon com correção de Bonferroni, permitindo verificar diferenças estatisticamente significativas entre os métodos em diferentes configurações estruturais.

B. Expansão da Metologia e Impacto na Predição de Mortalidade

Na segunda etapa do trabalho, a análise foi estendida para um conjunto de variáveis clinicamente relevantes em pacientes hospitalizados com COVID-19, além da saturação de oxigênio. Foram consideradas: creatinina sérica (função renal), proteína C reativa - CRP (inflamação sistêmica), contagem de plaquetas (estado hematológico) e frequência cardíaca (função cardiovascular). Cada variável foi avaliada individualmente quanto à qualidade da imputação, seguindo o mesmo protocolo experimental da Parte 1, com introdução de ausência artificial sob mecanismo MCAR (50%) e validação cruzada 10-fold.

A estrutura da análise local foi preservada, caracterizando cada observação por duas métricas locais derivadas de KNN, empregado tanto como imputador quanto como instrumento de particionamento do espaço: (i) similaridade média aos k vizinhos mais próximos e (ii) entropia local da variável analisada. Além dos métodos previamente avaliados, incorporou-se o TabPFN como novo método de imputação, comparado sob condições experimentais idênticas.

Em seguida, para investigar o impacto da imputação em tarefas preditivas, conduzimos experimentos de predição de mortalidade hospitalar utilizando três algoritmos de gradient boosting: LightGBM, XGBoost e CatBoost. A base original foi particionada em 10 folds de treino e de teste por validação cruzada, preservando o padrão real de ausência dos dados. Todas as variáveis selecionadas foram imputadas com o método de melhor desempenho global na etapa anterior: o TabPFN.

A imputação foi avaliada em três configurações experimentais, tendo como linha de base o conjunto de treino e de teste originais, sem imputação: (i) treino com dados imputados e teste com dados originais; (ii) treino com dados originais e teste com dados imputados; e (iii) treino e teste com dados imputados. Em todas as configurações, foi empregado um procedimento de validação cruzada estratificada com 10 partições de treino e de teste, garantindo a preservação da proporção entre as classes de óbito e não óbito em cada partição.

O desempenho dos classificadores foi avaliado por meio das métricas Micro-F1, Macro-F1, Precisão, Recall e F1-Score por classe. Para garantir maior rigor estatístico na comparação dos resultados, foram reportados intervalos de confiança de 95%, estimados a partir do erro padrão da média e da aproximação normal ($z = 1{,}96$).

IV. Resultados Experimentais

Nessa seção apresentamos os resultados obtidos a partir da aplicação da metodologia descrita na seção anterior. Dividimos nossas analises em duas linhas: (*i*) Avaliação Comparativa de Métodos de Imputação; e (*ii*) Extensão da Avaliação e Impacto Preditivo da Imputação.

A. Avaliação Comparativa de Métodos de Imputação

Foram comparados cinco métodos de imputação para a variável Sat/FiO: Média, KNN, GAIN, MICE e MissForest. A variável apresenta média igual a 369,83, utilizada como referência para contextualização dos erros observados.

A análise global dos resultados, considerando MAE, MSE e tempo médio de execução (Tabela I), indicou desempenho superior dos métodos MICE e MissForest em termos de acurácia. Ambos apresentaram os menores valores de erro, confirmando sua robustez na imputação da variável analisada. Em contraste, os métodos Média e GAIN apresentaram desempenho inferior. O baixo desempenho da imputação pela média era esperado devido à ausência de modelagem das relações entre variáveis. Já o desempenho limitado do GAIN sugere possível necessidade de calibração adicional de hiperparâmetros, embora tal ajuste implique maior custo computacional.

O KNN apresentou desempenho intermediário, ligeiramente superior ao método Média, porém significativamente inferior a MICE e MissForest. Entretanto, destacou-se pelo baixo tempo de execução, tornando-se atrativo sob a perspectiva da eficiência computacional e da simplicidade de implementação. Diante dessas características, o KNN foi selecionado para uma análise aprofundada com ajuste fino de seus parâmetros.

Após o processo de otimização, foi obtida a versão denominada KNN++. Essa versão apresentou melhora em relação ao KNN original, reduzindo o erro médio absoluto para. Ainda assim, permaneceu abaixo dos resultados de MissForest. Em contrapartida, o ganho em eficiência computacional foi substancial: o KNN++ apresentou tempo de execução

aproximadamente 70 vezes menor que o MICE e 2.850 vezes menor que o MissForest, evidenciando um trade-off claro entre acurácia e custo computacional.

Para investigar o comportamento do KNN++ em diferentes estruturas locais do espaço de dados, os pacientes foram distribuídos em quatro quadrantes definidos pela combinação entre a similaridade média aos k vizinhos (distância média) e a entropia local da variável Sat/FiO2 (Figura 2). Essa análise permitiu avaliar como a estrutura local influencia o erro de imputação.

Métricas	Média	KNN	KNN++	MissForest	MICE	GAIN
MAE	95.74	94.47	87.32	52.17	64.88	94.51
MSE	19061	25731	20101	11281	15742	18013
Tempo (seg.)	0.014	0.16	0.082	228.95	5.15	9.39

TABLE I: Efetividade e Custo Computacional (tempo) dos Métodos de Imputação.

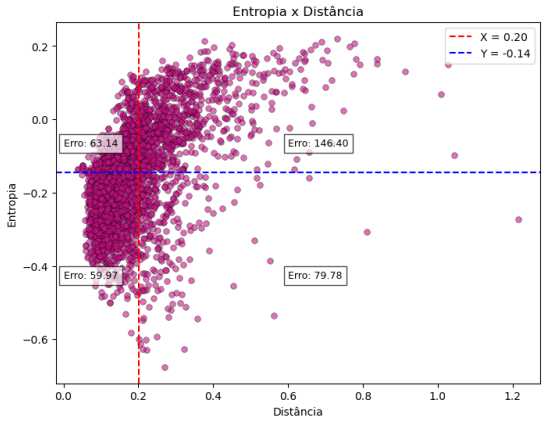

Fig. 1: Entropia x Distância média.

Fig. 2: Avaliação do Impacto da Estrutura Local dos Dados na Imputação.

Os resultados (Tabela II) mostraram que o desempenho do KNN++ varia significativamente conforme a estrutura da vizinhança. Em regiões caracterizadas por alta similaridade e baixa entropia (quadrantes Inferior Esquerdo e Superior Esquerdo), o KNN++ apresentou erros médios próximos aos obtidos com o MICE, sem diferença estatisticamente significativa. É importante observar que o quadrante Inferior Esquerdo, onde o KNN++ obteve seu menor erro médio, concentra aproximadamente 42% dos pacientes, indicando relevância prática dessa configuração estrutural.

Por outro lado, em regiões mais heterogêneas, especialmente no quadrante Superior Direito (alta distância e alta entropia), o desempenho do KNN++ foi significativamente inferior, o que contribuiu para o aumento de seu erro global. Nesses cenários, o MissForest demonstrou maior robustez, apresentando as menores medianas de erro em todos os quadrantes.

De forma geral, os resultados indicam que, embora o KNN++ não alcance o desempenho global dos métodos estado-da-arte, sua performance é competitiva em regiões estruturalmente homogêneas e densas, especialmente considerando seu custo computacional inferior. Esses achados reforçam a hipótese de que a similaridade clínica local desempenha papel determinante na qualidade da imputação baseada em vizinhança.

Quadrante	KNN++	MissForest	MICE	Porcentagem da Coorte
Direita Inferior	79.78 (↓,*)	65.61	79.03	9%
Direita Superior	146.40 (↓,↓)	54.94	79.32	28%
Esquerda Inferior	59.97 (↓,↓)	47.25	52.59	42%
Esquerda Superior	63.14 (↓,*)	50.44	60.62	21%

TABLE II: Média de erros por quadrante com significância estatística do KNN++ em relação aos demais métodos. Os símbolos ↓ e * indicam, respectivamente, perda e empate estatístico.

B. Extensão da Avaliação e Impacto Preditivo da Imputação

Com a adição de modelos de fundação para dados tabulares no pipeline de imputação, a Tabela III compara o desempenho dos cinco métodos de imputação (Média, KNN, MICE, MissForest e TabPFN) nas variáveis clínicas da primeira onda, considerando MAE, MSE e tempo de execução. O TabPFN apresentou, de forma consistente, os menores valores de MAE e MSE em todas as variáveis, tanto laboratoriais quanto clínicas, mantendo ainda tempo de execução estável e reduzido (entre 1,13 e 1,29 segundos). Esse resultado evidencia uma combinação superior entre qualidade de imputação e eficiência computacional e que o TabPFN é capaz de romper o trade-off entre acurácia e eficiência encontrada da etapa anterior.

O MissForest obteve desempenho intermediário, com erros menores que os métodos mais simples, porém superiores ao TabPFN, além de apresentar o maior custo computacional (cerca de 69 a 79 segundos). O MICE também mostrou desempenho intermediário, com tempo moderado (aproximadamente 1,3 a 3,2 segundos). Já o KNN e a imputação pela Média registraram os maiores erros, embora com tempos de execução bastante reduzidos, especialmente no caso da Média.

De forma geral, os resultados indicam vantagem clara do TabPFN, que alcança a melhor relação entre acurácia e custo computacional em todas as variáveis analisadas de maneira consistente.

A Tabela IV apresenta os valores de MAE estratificados em quadrantes definidos com base na similaridade média com os vizinhos mais próximos e na entropia local. Observa-se um padrão consistente entre variáveis e métodos: os maiores erros concentram-se no quadrante superior direito (baixa similaridade e alta entropia), enquanto os menores valores de MAE predominam nos quadrantes inferior esquerdo e superior esquerdo, associados a maior similaridade e menor heterogeneidade local.

O TabPFN apresentou os menores valores de MAE em todos os quadrantes e para todas as variáveis analisadas. Esses achados indicam que a estrutura local do espaço de dados exerce influência direta sobre a qualidade da imputação, impactando todos os métodos, ainda que em magnitudes distintas, e evidenciam a superioridade consistente do TabPFN em comparação aos demais modelos. Diante desse desempenho, o TabPFN foi selecionado para a etapa subsequente de avaliação

TABLE III: Comparação dos métodos de imputação segundo MAE, MSE e tempo de execução.

Variável	Métrica	Média	KNN	MICE	MissForest	TabPFN
$SatO_2/FiO_2$	MAE	95.29	86.54	64.84	51.10	41.34
	MSE	19064.33	19529.65	14951.00	11178.08	9097.37
	Tempo (s)	0.00	0.07	2.35	78.06	1.29
PCR (mg/L)	MAE	64.35	66.02	58.38	58.72	50.96
	MSE	7108.68	6896.44	6305.48	6223.47	5028.68
	Tempo (s)	0.00	0.08	2.20	69.52	1.13
Creatinina (mg/L)	MAE	0.62	0.72	0.41	0.40	0.32
	MSE	1.74	1.86	0.77	0.92	0.60
	Tempo (s)	0.00	0.07	2.45	75.31	1.19
Frequência cardíaca (bpm)	MAE	13.05	13.00	13.37	12.68	11.84
	MSE	278.33	312.40	337.22	262.98	233.52
	Tempo (s)	0.00	0.09	3.17	78.93	1.26
Plaquetas	MAE	66250.40	63606.53	70946.14	62089.64	55758.61
	MSE	7908080000.00	7483011000.00	215942700000.00	7159054000.00	5887709000.00
	Tempo (s)	0.00	0.08	1.31	76.81	1.27

dos efeitos da imputação nas tarefas preditivas de mortalidade hospitalar.

A Tabela V apresenta os resultados da predição de mortalidade hospitalar com o LightGBM em quatro cenários: baseline (sem imputação), apenas treino imputado, apenas teste imputado e imputação simultânea em treino e teste.

A imputação aplicada a ambos os conjuntos produziu o melhor desempenho global, com ganhos consistentes em Macro-F1, Micro-F1, precisão, recall e F1 da classe 1 (óbito). Em contraste, imputar apenas o treino levou, na maioria das métricas, a desempenho inferior ao baseline. Quando aplicada somente ao teste, os resultados permaneceram próximos ao baseline, sem melhorias relevantes. A imputação aplicada a ambos os conjuntos produziu o melhor desempenho global, com ganhos consistentes em Macro-F1, Micro-F1, precisão, recall e F1 da classe 1 (óbito). Em contraste, imputar apenas o treino levou, na maioria das métricas, a desempenho inferior ao baseline. Quando aplicada somente ao teste, os resultados permaneceram próximos ao baseline, sem melhorias relevantes. O efeito é ainda mais expressivo quando a imputação é aplicada simultaneamente em todas as variáveis clínicas relevantes nos conjuntos de treino e de teste: nessa configuração, o Macro-F1 da predição de mortalidade hospitalar sobe de 74.07 para 88.67, um ganho relativo de aproximadamente 20%, e o ganho é particularmente acentuado na classe 1 (óbito), cujo F1 passa de 55.36 para 80.85, um aumento de cerca de 46%. Como a classe 1 é minoritária e clinicamente crítica, essa melhora indica que a imputação integrada beneficia sobretudo a identificação dos casos de maior risco. Esses resultados indicam que a imputação melhora a performance preditiva apenas quando integrada de forma consistente às etapas de treinamento e avaliação.

TABLE IV: Erro Absoluto Médio por Quadrante

Variável	**Quadrante**	**Média**	**KNN**	**MICE**	**MissForest**	**TabPFN**
$SatO_2/FiO_2$	Direito Inferior	86.55	74.32	73.40	54.64	41.17
	Direito Superior	126.02	122.33	73.40	59.00	46.45
	Esquerdo Inferior	75.32	62.96	55.52	44.22	37.40
	Esquerdo Superior	80.36	71.26	61.70	47.25	39.34
PCT (mg/L)	Direito Inferior	73.39	74.04	68.06	69.15	59.08
	Direito Superior	69.98	69.84	65.58	67.35	56.93
	Esquerdo Inferior	58.81	62.09	50.49	50.32	45.14
	Esquerdo Superior	58.54	61.18	53.80	51.85	45.66
Creatinina (mg/L)	Direito Inferior	0.57	0.68	0.42	0.38	0.30
	Direito Superior	0.91	0.98	0.63	0.68	0.52
	Esquerdo Inferior	0.42	0.55	0.24	0.20	0.17
	Esquerdo Superior	0.46	0.54	0.29	0.23	0.22
Frequência Cardíaca	Direito Inferior	14.18	14.52	15.79	13.31	13.02
	Direito Superior	13.71	14.00	15.50	13.32	12.48
	Esquerdo Inferior	12.35	13.43	11.30	12.18	11.08
	Esquerdo Superior	12.36	13.18	11.67	11.94	11.32
Plaquetas	Direito Inferior	72968.775	69530.048	144726.036	69508.305	62249.771
	Direito Superior	70926.970	67668.490	68172.434	67852.048	58277.714
	Esquerdo Inferior	61545.299	60643.601	57054.830	56534.113	52952.643
	Esquerdo Superior	61611.022	56534.543	57315.986	56410.858	51735.208

TABLE V: Desempenho do Método LightGBM em Diferentes Configurações de Imputação (IC 95%).

Variável	Métrica	Baseline	Treino Imp. + Teste Orig.	Treino Orig. + Teste Imp.	Treino Imp. + Teste Imp.
$SatO_2/FiO_2$	Macro-F1	74.07 [72.85–75.71]	71.03 [69.33–72.73]	73.34 [71.71–74.97]	78.66 [77.28–80.04]
	Micro-F1	87.55 [86.75–88.35]	83.59 [82.65–84.53]	87.35 [86.69–88.01]	89.42 [88.76–90.08]
	Prec. C0	90.03 [89.53–90.53]	90.64 [90.00–91.28]	89.75 [89.19–90.31]	91.58 [91.10–92.06]
	Rec. C0	95.67 [95.67–96.45]	89.56 [88.71–90.41]	95.77 [95.26–96.28]	96.20 [95.69–96.71]
	F1 C0	92.77 [92.30–93.24]	90.11 [89.53–90.69]	92.67 [92.29–93.05]	93.81 [93.43–94.19]
	Prec. C1	68.59 [64.58–72.60]	50.51 [47.52–53.50]	68.04 [65.23–70.85]	74.46 [71.58–77.34]
	Rec. C1	46.66 [43.66–49.66]	53.53 [50.08–56.98]	45.01 [41.65–48.37]	55.45 [52.79–58.11]
	F1 C1	55.36 [52.49–58.23]	51.94 [49.05–54.83]	54.03 [51.11–56.95]	63.49 [61.10–65.66]
PCR (mg/L)	Macro-F1	74.07 [72.85–75.71]	65.67 [64.08–67.26]	74.47 [72.98–75.96]	79.28 [77.46–81.10]
	Micro-F1	87.55 [86.75–88.35]	80.94 [79.96–81.92]	87.82 [87.13–88.51]	89.78 [89.08–90.48]
	Prec. C0	90.03 [89.53–90.53]	88.66 [88.12–89.20]	90.09 [89.62–90.56]	91.78 [91.04–92.52]
	Rec. C0	95.67 [95.67–96.45]	88.49 [87.53–89.45]	96.00 [95.35–96.65]	96.36 [95.84–96.88]
	F1 C0	92.77 [92.30–93.24]	88.56 [87.94–89.18]	92.94 [92.53–93.35]	94.02 [93.63–94.41]
	Prec. C1	68.59 [64.58–72.60]	42.75 [39.88–45.62]	70.17 [66.80–73.54]	75.71 [73.18–78.24]
	Rec. C1	46.66 [43.66–49.66]	43.01 [40.04–45.98]	46.76 [43.94–49.58]	56.52 [52.18–60.86]
	F1 C1	55.36 [52.49–58.23]	42.78 [40.11–45.45]	55.99 [53.37–58.61]	64.53 [61.25–67.81]
Creatinina (mg/L)	Macro-F1	74.07 [72.85–75.71]	73.48 [71.68–75.28]	74.60 [73.28–75.92]	74.53 [72.75–76.31]
	Micro-F1	87.55 [86.75–88.35]	87.46 [86.43–88.49]	87.79 [86.58–89.00]	87.74 [86.59–88.89]
	Prec. C0	90.03 [89.53–90.53]	89.75 [89.14–90.36]	90.16 [89.51–90.81]	90.17 [89.59–90.75]
	Rec. C0	95.67 [95.67–96.45]	95.92 [95.07–96.77]	95.81 [94.27–97.35]	95.73 [94.62–96.84]
	F1 C0	92.77 [92.30–93.24]	92.74 [92.11–93.37]	92.90 [92.01–93.79]	92.86 [92.13–93.59]
	Prec. C1	68.59 [64.58–72.60]	69.05 [65.74–72.36]	69.69 [67.91–71.47]	69.24 [65.84–72.64]
	Rec. C1	46.66 [43.66–49.66]	44.83 [41.65–48.01]	47.45 [44.04–50.86]	47.53 [44.53–50.53]
	F1 C1	55.36 [52.49–58.23]	54.21 [51.19–57.23]	56.30 [54.26–58.34]	56.23 [53.32–59.14]
Frequência cardíaca (bpm)	Macro-F1	74.07 [72.85–75.71]	70.08 [68.24–71.92]	74.14 [72.18–76.10]	75.93 [74.38–77.48]
	Micro-F1	87.55 [86.75–88.35]	85.03 [84.38–85.68]	87.60 [86.65–88.55]	88.33 [87.65–89.01]
	Prec. C0	90.03 [89.53–90.53]	89.23 [88.64–89.82]	90.05 [89.45–90.65]	90.62 [90.14–91.10]
	Rec. C0	95.67 [95.67–96.45]	93.31 [92.96–93.66]	95.75 [94.96–96.54]	95.94 [95.37–96.51]
	F1 C0	92.77 [92.30–93.24]	91.22 [90.86–91.58]	92.81 [92.25–93.37]	93.20 [92.81–93.59]
	Prec. C1	68.59 [64.58–72.60]	56.90 [54.22–59.58]	68.87 [64.40–73.34]	71.23 [67.86–74.60]
	Rec. C1	46.66 [43.66–49.66]	43.37 [39.80–46.94]	46.76 [43.35–50.17]	50.04 [47.21–52.87]
	F1 C1	55.36 [52.49–58.23]	48.95 [45.62–52.28]	55.48 [52.06–58.90]	58.64 [55.89–61.39]
Plaquetas	Macro-F1	74.07 [72.85–75.71]	60.23 [58.45–62.01]	73.60 [72.32–74.88]	83.42 [82.01–84.83]
	Micro-F1	87.55 [86.75–88.35]	70.07 [69.10–71.04]	87.61 [87.12–88.10]	91.65 [90.97–92.33]
	Prec. C0	90.03 [89.53–90.53]	90.26 [89.78–90.74]	89.73 [89.28–90.18]	93.09 [92.56–93.62]
	Rec. C0	95.67 [95.67–96.45]	71.86 [70.91–72.81]	96.16 [95.59–96.73]	97.23 [96.65–97.81]
	F1 C0	92.77 [92.30–93.24]	79.97 [79.39–80.55]	92.84 [92.55–93.13]	95.10 [94.70–95.50]
	Prec. C1	68.59 [64.58–72.60]	30.30 [28.28–32.32]	70.05 [67.37–72.73]	82.35 [79.39–85.31]
	Rec. C1	46.66 [43.66–49.66]	61.06 [58.28–63.84]	44.52 [41.60–47.44]	63.68 [60.68–66.68]
	F1 C1	55.36 [52.49–58.23]	40.45 [37.40–43.50]	54.38 [52.03–56.73]	71.70 [69.24–74.16]
Todas as Variáveis	Macro-F1	74.07 [72.85–75.71]	65.61 [63.97–67.25]	73.56 [72.26–74.86]	88.67 [87.24–90.10]
	Micro-F1	87.55 [86.75–88.35]	77.70 [76.90–78.50]	87.93 [87.46–88.40]	94.10 [93.37–94.83]
	Prec. C0	90.03 [89.53–90.53]	90.25 [89.75–90.75]	89.53 [89.08–89.98]	95.18 [94.70–95.66]
	Rec. C0	95.67 [95.67–96.45]	82.13 [81.35–82.91]	96.86 [96.47–97.25]	97.89 [97.32–98.46]
	F1 C0	92.77 [92.30–93.24]	85.99 [85.52–86.46]	93.05 [92.70–93.31]	96.52 [96.09–96.95]
	Prec. C1	68.59 [64.58–72.60]	38.33 [34.32–42.34]	73.30 [70.93–75.67]	87.71 [84.69–90.73]
	Rec. C1	46.66 [43.66–49.66]	55.37 [52.37–58.37]	42.99 [40.19–45.79]	75.06 [72.45–77.67]
	F1 C1	55.36 [52.49–58.23]	45.22 [42.35–48.09]	54.06 [51.69–56.43]	80.85 [78.42–83.28]

V. Conclusões e Trabalhos Futuros

Este trabalho vai além da comparação tradicional de métodos de imputação ao propor uma análise integrada entre estrutura local dos dados clínicos, custo computacional e impacto na predição de mortalidade em pacientes hospitalizados com COVID-19.

Inicialmente, avaliamos Média, KNN, GAIN, MICE e MissForest, mostrando que menor erro não implica melhor relação custo-benefício. A partir de uma análise estrutural baseada na similaridade entre pacientes e na entropia das variáveis, evidenciamos que o desempenho do KNN depende da organização local do espaço clínico. Com base nisso, propusemos o KNN++, que incorpora critérios estruturais na seleção de vizinhos, alcançando desempenho competitivo a um custo reduzido.

Em seguida, incorporamos o TabPFN, modelo de fundação para dados tabulares, que apresentou desempenho superior mesmo em regiões heterogêneas, com robustez e estabilidade computacional. A análise por quadrantes confirmou que a estrutura local influencia todos os métodos, mas o TabPFN é menos sensível a variações estruturais, confirmando sua robustez.

Por fim, demonstramos que a imputação não é um pré-processamento neutro: ela altera significativamente a predição de mortalidade. A imputação conjunta de treino e teste gerou os maiores ganhos, evidenciando que seu impacto depende da integração ao pipeline preditivo.

Como contribuição central, oferecemos uma visão estruturada e orientada à tarefa da imputação clínica, conectando estrutura dos dados, custo e desempenho preditivo. Como continuidade, propomos um sistema híbrido e adaptativo para ambientes hospitalares, capaz de selecionar dinamicamente a estratégia mais adequada ao contexto clínico.

References

[1] T. Shadbahr *et al.*, "The impact of imputation quality on machine learning classifiers for datasets with missing values," *Communications Medicine*, vol. 3, no. 1, p. 139, Oct. 2023.

[2] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second," arXiv:2207.01848, 2023.

[3] I. Ganem *et al.*, "Avaliando as Limitações e Potenciais do Algoritmo k-Vizinhos Mais Próximos (kNN) na Imputação de Dados Clínicos," in *Proc. Brazilian Symp. on Databases (SBBD)*, Fortaleza, Brazil, pp. 809–815, 2025.

[4] L. O. Joel, W. Doorsamy, and B. S. Paul, "A Comparative Study of Imputation Techniques for Missing Values in Healthcare Diagnostic Datasets," *Int. J. Data Science and Analytics*, vol. 20, no. 7, pp. 6357–6373, 2025.

[5] D. J. Stekhoven and P. Bühlmann, "MissForest—Non-parametric Missing Value Imputation for Mixed-Type Data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[6] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011.

[7] J. Yoon, J. Jordon, and M. van der Schaar, "GAIN: Missing Data Imputation using Generative Adversarial Nets," in *Proc. 35th Int. Conf. Machine Learning (ICML)*, pp. 5689–5698, 2018.

[8] N. Kouiroukidis and G. Evangelidis, "The effects of dimensionality curse in high dimensional KNN search," in *Proc. 15th Panhellenic Conf. Informatics*, pp. 41–45, 2011.

[9] J. Ma, A. Dankar, G. Stein, G. Yu, and A. Caterini, "TabPFGen – Tabular Data Generation with TabPFN," arXiv:2406.05216, 2024.

[10] M. S. Marcolino *et al.*, "Clinical characteristics and outcomes of patients hospitalized with COVID-19 in Brazil: Results from the Brazilian COVID-19 registry," *Int. J. Infectious Diseases*, vol. 107, pp. 300–310, June 2021.

[11] M. S. Marcolino *et al.*, "Explainable artificial intelligence for predicting cardiovascular events in hospitalised COVID-19 patients," *BMC Infectious Diseases*, vol. 25, no. 1, p. 1569, 2025.

[12] B. B. M. Paiva *et al.*, "Characterizing and Understanding Temporal Effects in COVID-19 Data," *Proc. Mach. Learn. Research*, vol. 184, pp. 33–40, 2022.

[13] F. Lana *et al.*, "Unraveling relevant cross-waves pattern drifts in patient-hospital risk factors among hospitalized COVID-19 patients using explainable machine learning methods," *BMC Infectious Diseases*, vol. 25, no. 1, p. 537, Apr. 2025.

[14] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. Wiley, 2019.

[15] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. van der Meulen, "Estimating differential entropy with kernel methods," *Int. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, 1997.