Avaliação comparativa de redes neurais e de modelos baseados em árvores para a predição do grau de incapacidade física de pacientes com hanseníase

Pedro Henrique Correia Bezerra Silva[1], **Elisson da Silva Rocha**[1]
Patricia Takako Endo[1], **Eraylson Galdino da Silva**[1]

[1]Programa de Pós-Graduação em Engenharia da Computação (PPGEC),
Universidade de Pernambuco (UPE), Pernambuco, Brasil

{pedro.cbsilva, elisson.rocha, patricia.endo, eraylson.galdino}@upe.br

Abstract. *Leprosy is a significant public health concern due to its disabling potential and its substantial presence in Brazil. This study compared AI models applied to tabular data from the SINAN database to predict the final GIF in patients. Tree-based models (RF, LightGBM, and CatBoost) and neural networks (MLP, ResNet, and Transformer) were evaluated. LightGBM achieved superior performance and greater stability across classes, reaching an AUC OvO of 71.10%. The neural networks demonstrated a competitive performance, particularly the Transformer, which achieved an AUC OvO of 70.69%. To conclude, given the dataset used, tree-based models are more suitable for predicting GIF prognosis, while neural networks are alternatives for multimodal contexts.*

Resumo. *A hanseníase é um agravo relevante na saúde pública pelo seu potencial incapacitante e sua presença expressiva no Brasil. Este estudo comparou modelos de IA aplicados a dados tabulares do SINAN para prever o GIF final de pacientes. Foram avaliados modelos baseados em árvores (RF, LightGBM e CatBoost) e redes neurais (MLP, ResNet e Transformer). O LightGBM apresentou desempenho superior e maior estabilidade entre classes, com um AUC OvO de 71,10%. As redes neurais mostraram um desempenho competitivo, sobretudo o Transformer, com um AUC OvO de 70,69%. Conclui-se que, conforme a base utilizada, modelos baseados em árvores são mais adequados ao prognóstico do GIF, mas as redes neurais são alternativas para contextos multimodais.*

1. Introdução

As Doenças Tropicais Negligenciadas (DTNs) ainda são um desafio para sistemas de saúde, sobretudo por afetarem de forma desproporcional populações em contextos de vulnerabilidade socioeconômica nas regiões tropicais e subtropicais do mundo [World Health Organization 2020]. Entre essas enfermidades, a hanseníase permanece como um agravo de relevância clínica e epidemiológica, por seu potencial incapacitante e por demandar vigilância contínua e estratégias de detecção precoce e acompanhamento longitudinal dos pacientes [World Health Organization 2021].

No cenário global, em 2024, foram notificados 172.717 novos casos de hanseníase. Nas Américas, o Brasil concentra a maior parcela dos casos: 22.129 (12,8% de todos os casos), figurando também como o segundo país com a maior notificação de

casos novos no mundo, ficando somente atrás da Índia (com 100.957 de casos, representando 58,45%) [World Health Organization 2025].

O Grau de Incapacidade Física (GIF) é uma das principais métricas para monitorar o dano neural e as sequelas decorrentes da hanseníase. Conforme diretrizes clínicas [Organização Mundial da Saúde 2010], o GIF pode ser classificado em três níveis: grau 0, quando não há sinais de comprometimento nas mãos, pés e olhos; grau 1, quando se observa perda de sensibilidade e/ou redução de força muscular sem deformidades visíveis; e grau 2, quando há deformidades evidentes e/ou sequelas graves irreversíveis.

Na literatura, os trabalhos de Inteligência Artificial (IA) aplicada à hanseníase tem historicamente priorizado o suporte ao diagnóstico com base em imagens de lesões cutâneas, refletindo a forte expansão do aprendizado profundo (*deep learning*) em dados não-estruturados [Fernandes et al. 2024]. Há ainda um espaço para ampliar pesquisas voltadas para o uso de dados clínicos e sociodemográficos estruturados, com ênfase em aplicações prognósticas e de estratificação de risco [Andrade et al. 2025]. Essa lacuna torna-se particularmente interessante, quando se considera o Sistema de Informação de Agravos de Notificação (SINAN) como fonte de dados tabulares, com acesso aberto.

Este estudo avalia o desempenho de três arquiteturas de redes neurais (MLP, ResNet tabular e Transformer tabular) adaptadas para utilizar dados tabulares e de três algoritmos baseados em árvores (Random Forest, LightGBM e CatBoost) na predição multiclasse do GIF (0, 1 ou 2) ao final do tratamento de pacientes com hanseníase. Os modelos utilizam atributos disponíveis no momento do diagnóstico e durante o tratamento do paciente. Através desses modelos, pode ser possível apoiar decisão clínica e embasar estratégias de vigilância e planejamento em saúde, considerando a importância do prognóstico de incapacidade como componente da resposta ao agravo.

2. Trabalhos Relacionados

Segundo [Fernandes et al. 2024], em sua revisão sistemática da literatura, a maior parte dos estudos investigados emprega imagens e algoritmos clássicos de aprendizado de máquina ou abordagens de *deep learning*, frequentemente sem ter a hanseníase como objetivo exclusivo do trabalho, mas como uma entre várias condições dermatológicas avaliadas.

Em outra revisão sistemática mais recente, dessa vez desenvolvida por [Andrade et al. 2025], sobre o uso de IA na hanseníase em diferentes eixos clínico-epidemiológicos, mostrou que os trabalhos ainda se concentram principalmente na identificação de sinais e sintomas da doença e no diagnóstico, com predomínio de abordagens baseadas em imagens. Os autores destacam que apenas um número reduzido de estudos explora de forma explícita dados tabulares (sociodemográficos, clínicos ou genéticos) para apoiar decisões no cuidado da hanseníase, evidenciando uma lacuna na diversidade de dados e na expansão de aplicações para além da imagem. Nesse conjunto, há exemplos isolados de uso de dados do SINAN para fins de predição epidemiológica e de diagnóstico [da Silva et al. 2018, De Souza et al. 2021, Dutra da Silva et al. 2018]. Contudo, tais iniciativas não se dedicam à predição clínica de desfechos funcionais no término do tratamento nem à comparação sistemática entre diferentes modelos de IA.

Nesse contexto, o presente estudo se diferencia por deslocar o foco da literatura do diagnóstico por imagem para uma tarefa prognóstica de relevância clínica e de vigilância:

a predição do GIF final de pacientes com base em registros tabulares do SINAN em escala nacional (2001–2024). Nesse sentido, o trabalho contribui simultaneamente para duas frentes: *(i)* o uso de base de dados pública e abrangente de saúde para inferência prognóstica de incapacidades decorrentes da hanseníase e *(ii)* o debate sobre o desempenho relativo de arquiteturas neurais tabulares e métodos tradicionais de árvores utilizando dados estruturados.

A comparação com estudos que também utilizam dados tabulares reforça a originalidade do recorte desta pesquisa. Em particular, [Freitas et al. 2025] avaliaram modelos de aprendizado de máquina para predizer diagnóstico tardio no Brasil a partir da presença de GIF 2 no momento do diagnóstico, utilizando dados de casos registrados no SINAN entre 2018 e 2022. Os modelos LightGBM e *ensembles* apresentam melhores desempenho; e os autores também apresentam uma análise de preditores relevantes como número de nervos afetados e forma clínica. Embora compartilhemos a mesma fonte de dados e também utilizemos modelos baseados em árvores, nosso estudo difere do trabalho de [Freitas et al. 2025] em três aspectos centrais: *(i)* alvo clínico (diagnóstico tardio versus incapacidade na alta); *(ii)* janela temporal (2018–2022 versus 2001–2024); e *(iii)* comparação com arquiteturas neurais tabulares como MLP, ResNet tabular e Transformers tabulares.

Ao expandir o horizonte temporal, explicitar o foco prognóstico da alta e realizar uma comparação sistemática entre diferentes famílias de modelos de IA, este trabalho complementa os estudos sobre a aplicação de IA para predizer o progresso de incapacidades físicas decorrentes da hanseníase.

3. Metodologia

Esta pesquisa adotou a metodologia de *Cross-Industry Standard Process for Data Mining* (CRIPS-DM) [Wirth and Hipp 2000] devido sua forte aceitação tanto em âmbitos acadêmicos quanto empresariais [Saltz 2021]. A pesquisa se desenvolveu seguindo cinco fases do ciclo de vida do CRISP-DM (entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem e avaliação), sem a fase de implantação, pois não se adequa aos objetivos deste estudo.

3.1. Entendimento do Negócio

O objetivo do estudo é avaliar e comparar redes neurais (MLP, ResNet tabular e Transformer tabular) e modelos de árvores (Random Forest, LightGBM e CatBoost) para a predição do GIF na alta de pacientes com hanseníase, a partir de dados tabulares do SINAN entre 2001-2024. Tal objetivo decorre da relevância clínica do GIF como indicador de dano neural e funcional e da necessidade de apoiar decisões de acompanhamento e prevenção de agravamento no momento de desfecho terapêutico [Organização Mundial da Saúde 2010, Secretaria de Vigilância em Saúde 2023]. A delimitação da tarefa de predição do GIF na alta, em oposição ao GIF no diagnóstico, busca explorar o potencial prognóstico do conjunto de variáveis registradas ao longo do tratamento, favorecendo o uso de modelos de IA como suporte à vigilância e à assistência.

A comparação entre árvores e redes neurais foi motivada por evidências de que métodos de *boosting* e *bagging* permanecem altamente competitivos em tarefas com dados tabulares, inclusive em cenários biomédicos, enquanto arquiteturas profundas recentes podem apresentar ganhos condicionados à configuração, ao pré-processamento e à

regularização [Gutta et al. 2024, Kadra et al. 2021]. Assim, avaliar esses grupos de modelos no domínio da hanseníase contribui para inferir se avanços recentes em *deep learning* tabular se traduzem em benefícios concretos em um problema de saúde pública com dados administrativos reais [Gorishniy et al. 2021, Shwartz-Ziv and Armon 2022, Grinsztajn et al. 2022].

3.2. Entendimento dos Dados

Os dados utilizados foram obtidos do SINAN, considerando registros de hanseníase entre 2001 e 2024. A base inicial era composta por 816.781 registros e 63 variáveis, das quais 51 se referem a informações gerais do paciente ou da notificação, enquanto 12 caracterizam aspectos epidemiológicos. Além disso, 13 colunas estavam no formato de data (dd/mm/aaaa), 12 eram variáveis numéricas contínuas e 38 categóricas, sendo 34 nominais e 4 ordinais. Após inspeção de consistência e completude, com foco na variável-alvo e em campos essenciais para caracterização clínica e operacional, foram removidos 371.042 registros, totalizando 445.739 observações válidas para análise. Essa depuração visou mitigar vieses decorrentes de ausência de informação crítica e assegurar maior qualidade dos dados a serem inseridos nos modelos.

Em conformidade com as evidências apresentadas por [da Costa et al. 2020, Véras et al. 2021, Véras et al. 2023, Bomtempo et al. 2023], o processo de exploração de dados concentrou-se nas variáveis com maior potencial de influenciar o avanço do GIF incluindo o GIF no momento da notificação e no momento de cura, a data de diagnóstico e de início de tratamento, além do perfil sociodemográfico do paciente, contemplando ano de nascimento, idade, sexo, estado gestacional, escolaridade e ocupação. Também foram consideradas variáveis epidemiológicas que refletem o comprometimento clínico, tais como o número de lesões, forma clínica, classificação operacional inicial, classificação atual, baciloscopia, esquema inicial e atual de tratamento, episódios reacionais e quantidade de nervos afetados.

Embora o processo de notificação do SINAN forneça padronização entre as diferentes instituições de saúde pública, ocorrem inconsistências decorrentes de ausência de preenchimento ou formatação inadequada conforme o dicionário de dados. Dentro da base utilizada, verificaram-se casos como datas fora do padrão, registros inválidos de escolaridade, idades incompatíveis com a regra de formatação, categoria de sexo preenchida numericamente em vez de alfabética e datas de início de tratamento que antecedem a data de diagnóstico.

3.3. Preparação dos Dados

Inicialmente, foram descartadas colunas com elevada proporção de valores ausentes ou sem pertinência clínica/operacional para a predição do GIF na alta. Essa decisão reduz ruído e evita que modelos aprendam padrões inadequados. As variáveis de data foram padronizadas e utilizadas para derivar a diferença temporal entre diagnóstico e início do tratamento, uma vez que atrasos assistenciais estão associados a maior risco de incapacidade e a piores desfechos funcionais em hanseníase [Santos et al. 2024]. Em registros com baciloscopia ausente, foi aplicado preenchimento dos dados conforme a orientação de sua relação com a classificação operacional [Ministério da Saúde 2022].

Variáveis categóricas passaram por *one-hot encoding*, enquanto as numéricas foram normalizadas entre 0 e 1 ou entre -1 e 1 no caso do número de lesões e de nervos

afetados a fim de preservar os dados das colunas, sem remover diversos registros da base, dado que possuíam uma quantidade considerável de dados faltantes (maior que 30%). Além disso, foram geradas duas versões do conjunto de dados: (i) uma base geral com variáveis associadas à fase inicial do tratamento e (ii) uma base com atributos relacionados a episódios reacionais e a informações clínicas mais avançadas associadas ao GIF na alta, como o esquema terapêutico atual e a classificação operacional atual. Essa estratégia permite avaliar se o acréscimo de variáveis tardias melhora desempenho preditivo e se há potencial de uso em diferentes momentos do cuidado.

3.4. Modelagem e Avaliação

No campo metodológico, o uso de IA em dados tabulares é marcado por um debate ainda ativo sobre o desempenho relativo entre modelos baseados em árvores e arquiteturas neurais. Evidências indicam que métodos de *gradient boosting* frequentemente alcançam resultados superiores em problemas tabulares tradicionais, mas também mostram que arquiteturas neurais modernas podem ser competitivas sob determinadas condições de modelagem e pré-processamento [Gorishniy et al. 2021, Shwartz-Ziv and Armon 2022]. Além disso, *benchmarks* recentes reforçam que não há um vencedor universal, o que torna pertinente avaliar diferentes famílias de modelos em domínios específicos de alta relevância social, como a saúde pública [Shmuel et al. 2025].

Estudos contemporâneos já demonstram utilidade de modelos baseados em árvores ao explorar registros do SINAN para analisar incapacidade associada à hanseníase, especialmente quando o foco recai sobre marcadores de diagnóstico tardio e presença de grau 2 no momento da detecção [Freitas et al. 2025]. Contudo, ainda é limitada a produção científica que contraste, de modo sistemático e em escala nacional, o desempenho de redes neurais tabulares e de modelos baseados em árvores para prever o GIF especificamente no momento da alta.

O Random Forest foi adotado como *baseline* para dados tabulares, com bom desempenho em cenários multiclasses e heterogêneos [Breiman 2001]. LightGBM e CatBoost foram incluídos pela relevância contemporânea de métodos de *gradient boosting*, destacando-se o tratamento eficiente de não linearidades e interações complexas, com CatBoost oferecendo vantagens específicas em atributos categóricos [Ke et al. 2017, Dorogush et al. 2018]. Em relação às redes neurais, o MLP, o ResNet tabular e o Transformer tabular foram selecionados com base em evidências recentes sobre arquiteturas profundas adaptadas ao domínio tabular [Gorishniy et al. 2021].

Em relação a todos os modelos, a amostra foi particionada em 60% para treino, 20% para validação e 20% para teste. Essa estratégia assegura separação adequada para seleção de hiperparâmetros e avaliação final imparcial, minimizando risco de *overfitting* por ajuste excessivo no conjunto de teste. Considerando a distribuição desbalanceada do GIF na alta, foi aplicada uma estratégia de balanceamento híbrido no conjunto de treino, combinando subamostragem e técnicas de sobreamostragem. O uso de abordagens desse tipo é recomendado na literatura quando se busca reduzir viés de decisão em classes minoritárias, especialmente em problemas de saúde pública, onde o custo do falso negativo tende a ser alto [Salmi et al. 2024]. Inicialmente, a classe majoritária (0) contava com 203.639 amostras, enquanto a minoritária (2) possuía apenas 16.686. Por isso, realizou-se a sobreamostragem por meio da duplicação de registros e a subamostragem pela remoção

desses, ambas conduzidas de forma aleatória. Ao final, ambas passaram a ter a mesma quantidade da classe intermediária (1), totalizando 47.118 registros cada. Outra técnica de balanceamento efetuada foi a de aplicação de pesos nos modelos para que, durante o treinamento, eles atribuíssem uma importância variada aos possíveis valores da classe alvo (0, 1 e 2) conforme a sua presença na base de dados. Essa última abordagem influenciou na mitigação do viés de correção para com a classe majoritária e foi aplicada por meio do parâmetro *class_weights* ou *auto_class_weights* nos modelos.

Para os modelos de árvores, foi utilizado *GridSearchCV* com espaços de busca controlados, adequados à restrição de recursos computacionais e à necessidade de experimentação reprodutível. Para redes neurais, combinações de hiperparâmetros foram avaliadas no conjunto de validação, priorizando taxas de aprendizado, profundidade/capacidade do modelo e regularização.

Foram adotadas métricas que capturam diferentes aspectos de erro em cenários multiclasses e desbalanceados: precisão, *recall*, *F1-score* e *ROC-AUC* no esquema *one-vs-one* (AUC OvO). A escolha se justifica por permitir análise simultânea de desempenho global e sensibilidade por classe, sendo especialmente relevantes para monitorar o comportamento dos modelos em graus mais graves do GIF, uma vez que a base está desbalanceada em relação à classe alvo (0 = 76,04%; 1 = 17,69%; 2 = 6,27%).

4. Resultados e Discussão

Todos os resultados expostos a seguir foram obtidos conforme a saída dos modelos através da base de teste. Dessa forma, primeiramente, a Tabela 1 apresenta o desempenho dos modelos considerando exclusivamente variáveis disponíveis no início do tratamento e utilizando balanceamento por ponderação de classes. Dentro desse cenário, o LightGBM obteve o maior AUC OvO (70,16%), seguido pelo Transformer (69,77%) e pelo CatBoost (69,72%). O Random Forest, ao contrário, apresentou o menor *AUC OvO* (63,62%).

Tabela 1. Início do Tratamento - Balanceamento com Pesos

Modelo	Grau 0			Grau 1			Grau 2			AUC OvO
	P	R	F1	P	R	F1	P	R	F1	
Random Forest	0.79	0.89	0.83	0.27	0.16	0.20	0.13	0.08	0.10	0.6362
LightGBM	0.89	0.61	0.72	0.26	0.35	0.30	0.15	0.58	0.24	0.7016
CatBoost	0.89	0.59	0.71	0.24	0.30	0.27	0.14	0.63	0.24	0.6972
MLP	0.89	0.59	0.71	0.24	0.28	0.26	0.14	0.63	0.23	0.6891
ResNet	0.89	0.57	0.69	0.23	0.34	0.28	0.14	0.59	0.23	0.6871
Transformer	0.90	0.56	0.69	0.23	0.35	0.28	0.15	0.61	0.24	0.6977

Legenda: P = *Precision*, R = *Recall*, F1 = *F1-Score*.

No que se refere ao *F1-score* por classe, o LightGBM manteve um bom desempenho no GIF 0 (72%) e apresentou melhor equilíbrio entre precisão e *recall* nos GIF 1 e 2 quando comparado aos demais modelos, ainda que em patamares modestos (F1 $\leq$ 30% para GIF 1 e $\leq$ 24% para GIF 2). De modo geral, todos os modelos demonstraram desempenho substancialmente superior para o GIF 0, enquanto os GIF 1 e 2 permaneceram com resultados inferiores.

A Tabela 2 apresenta os resultados para a mesma janela temporal, substituindo a ponderação por pesos pela estratégia de *hybrid sampling*. Novamente, o LightGBM manteve o melhor *AUC OvO* (70,12%), seguido pelo Transformer (69,74%). O Random Forest apresentou leve aumento de *AUC OvO* em relação à Tabela 1 (64,48%), enquanto o CatBoost apresentou comportamento instável, resultando em 0 para tanto o seu *recall* quanto para o *F1-Score* do GIF 2. Essa instabilidade potencialmente se instaurou devido à (i) possível distorção da distribuição original das variáveis categóricas ou ao (ii) sobreajuste da classe majoritária e da baixa prevalência do GIF 2 na base.

Tabela 2. Início do Tratamento - Balanceamento com *Hybrid Sampling*

Modelo	Grau 0			Grau 1			Grau 2			AUC OvO
	P	R	F1	P	R	F1	P	R	F1	
Random Forest	0.87	0.58	0.69	0.25	0.55	0.34	0.14	0.22	0.17	0.6448
LightGBM	0.89	0.60	0.72	0.26	0.36	0.30	0.15	0.57	0.24	0.7012
CatBoost	0.77	0.99	0.87	0.40	0.04	0.07	0.67	0.00	0.00	0.6836
MLP	0.89	0.60	0.72	0.25	0.31	0.28	0.14	0.59	0.23	0.6873
ResNet	0.89	0.55	0.68	0.25	0.27	0.26	0.12	0.65	0.20	0.6817
Transformer	0.89	0.60	0.72	0.25	0.25	0.25	0.13	0.67	0.22	0.6974

Legenda: P = *Precision*, R = *Recall*, F1 = *F1-Score*.

De modo geral, o *hybrid sampling* não promoveu melhora consistente no desempenho das classes minoritárias. O *F1-score* para os GIF 1 e 2 permaneceu $\leq 34\%$ para todos os modelos. Comparativamente à Tabela 1, apenas o Random Forest apresentou melhora relevante no GIF 1 (F1 = 34%), enquanto os demais mantiveram resultados semelhantes ou ligeiramente inferiores. Esses achados sugerem que, no cenário analisado, a reamostragem híbrida não foi suficiente para alterar substancialmente o padrão de desempenho observado.

A Tabela 3 incorpora variáveis coletadas durante o tratamento, mantendo o balanceamento por pesos. Por uma visão geral, percebe-se que houve um aumento discreto e consistente no *AUC OvO* para a maioria dos modelos. Por uma visão específica, o LightGBM apresentou o melhor *AUC OvO* (71,1%), seguido pelo Transformer (70,69%) e pelo CatBoost (70,59%).

Tabela 3. Durante o Tratamento - Balanceamento com Pesos

Modelo	Grau 0			Grau 1			Grau 2			AUC OvO
	P	R	F1	P	R	F1	P	R	F1	
Random Forest	0.79	0.90	0.84	0.30	0.17	0.21	0.14	0.07	0.09	0.6503
LightGBM	0.89	0.63	0.74	0.27	0.35	0.30	0.16	0.58	0.25	0.7110
CatBoost	0.90	0.62	0.73	0.25	0.30	0.27	0.15	0.63	0.24	0.7059
MLP	0.89	0.62	0.73	0.26	0.36	0.30	0.15	0.55	0.24	0.6992
ResNet	0.89	0.63	0.73	0.25	0.32	0.28	0.15	0.57	0.24	0.6965
Transformer	0.90	0.60	0.72	0.25	0.38	0.30	0.16	0.57	0.25	0.7069

Legenda: P = *Precision*, R = *Recall*, F1 = *F1-Score*.

Os *F1-scores* também apresentaram leve melhora nas classes minoritárias em comparação à janela inicial. Ainda assim, o padrão de concentração de desempenho no GIF 0 persiste. A melhora global é plausível do ponto de vista clínico-epidemiológico, pois variáveis coletadas mais tardiamente, como classificação operacional atual e episódios reacionais, refletem com maior fidelidade a evolução neurológica do paciente e sua resposta terapêutica. Tais elementos estão diretamente relacionados à classificação do GIF na alta, conforme a estratégia global de redução de incapacidades por hanseníase [World Health Organization 2021]. Apesar da melhora relativa, os valores permanecem aquém do ideal para suporte clínico individualizado, sobretudo nos GIF 1 e 2.

A Tabela 4 apresenta os resultados da janela temporal de durante o tratamento sob *hybrid sampling*. Nesse contexto, o LightGBM novamente obteve o maior *AUC OvO* (71,04%), seguido pelo Transformer (70,63%). Além disso, o CatBoost voltou a apresentar *F1-score* igual a 0 para o GIF 2, reiterando a sua instabilidade sob esse tipo de balanceamento na base em questão.

Tabela 4. Durante o Tratamento - Balanceamento com *Hybrid Sampling*

Modelo	Grau 0			Grau 1			Grau 2			AUC OvO
	P	R	F1	P	R	F1	P	R	F1	
Random Forest	0.87	0.59	0.71	0.26	0.56	0.35	0.14	0.21	0.17	0.6576
LightGBM	0.89	0.63	0.74	0.27	0.36	0.31	0.16	0.58	0.25	0.7104
CatBoost	0.78	0.98	0.87	0.42	0.08	0.14	0.60	0.00	0.00	0.6924
MLP	0.89	0.62	0.73	0.26	0.33	0.29	0.15	0.57	0.23	0.6969
ResNet	0.89	0.62	0.73	0.26	0.24	0.25	0.13	0.64	0.22	0.6910
Transformer	0.90	0.62	0.74	0.26	0.24	0.25	0.14	0.68	0.23	0.7063

Legenda: P = *Precision*, R = *Recall*, F1 = *F1-Score*.

Comparando as Tabelas 3 e 4, observa-se que a alteração do método de balanceamento não produziu ganho sistemático. Alguns modelos apresentaram melhora pontual (como o Random Forest no GIF 1), enquanto outros tiveram piora (a exemplo do CatBoost nos GIF 1 e 2). De forma geral, a ponderação por pesos mostrou-se mais estável entre diferentes arquiteturas e janelas temporais.

Dessa forma, no conjunto de dados nacionais do SINAN analisado (2001-2024), modelos baseados em árvores, particularmente o LightGBM, apresentaram desempenho global superior em termos de *AUC OvO*. As redes neurais tabulares demonstraram desempenho próximo, compondo um segundo patamar competitivo, porém sem superar consistentemente os métodos de *boosting*.

Esses achados estão alinhados à literatura contemporânea sobre dados tabulares, na qual ensembles baseados em árvores permanecem altamente competitivos em cenários com variáveis heterogêneas e relações não lineares complexas. Arquiteturas neurais modernas, embora promissoras, tendem a demandar maior volume de dados balanceados ou desenho arquitetural específico para superar árvores de decisão com *gradient boosting* em contextos estruturados [Borisov et al. 2021, Grinsztajn et al. 2022, Gorishniy et al. 2021].

Contudo, independentemente da família de modelos, observou-se limitação consistente na predição dos GIF 1 e 2. Esse resultado sugere que as variáveis disponíveis no SINAN, embora abrangentes em termos administrativos e epidemiológicos, podem não capturar plenamente nuances fisiopatológicas associadas à progressão da incapacidade. Elementos clínicos mais detalhados, como a perda de sensibilidade térmica e tátil, força muscular segmentar, presença de nódulos e pápulas e parestesia nos pés, não estão integralmente representados na base analisada [Barbieri et al. 2022, Secretaria de Vigilância em Saúde 2023].

Além do mais, embora a ponderação por pesos tenha sido aplicada para mitigar o desbalanceamento, os resultados indicam que os modelos ainda concentram capacidade discriminativa na classe majoritária. Em termos clínicos, esse achado é relevante: a precisão mede a proporção de predições corretas entre os casos sinalizados como de determinado grau, influenciando diretamente a alocação de recursos assistenciais; já o *recall* indica a capacidade de identificar efetivamente os casos existentes. Isso é particularmente crítico em contextos clínicos, pois falsos positivos podem resultar em uso inapropriado de recursos de saúde, ansiedade desnecessária e potenciais intervenções inadequadas, enquanto falsos negativos podem levar à falta de tratamentos essenciais [DKE 2022]. Assim, apesar do desempenho global razoável em *AUC OvO*, os valores reduzidos de *F1* nas classes minoritárias indicam limitação para aplicação direta em contexto clínico.

Assim, os resultados indicam que, embora exista sinal preditivo relevante nos dados administrativos nacionais provenientes do SINAN, sua utilização isolada ainda não atinge robustez suficiente para aplicação clínica direta. O avanço nessa direção possivelmente dependerá da integração de dados clínicos mais granulares, de estratégias de modelagem sensíveis a desbalanceamento extremo e de abordagens multimodais que combinem informações tabulares e não estruturadas.

5. Conclusão

Este estudo realizou uma avaliação comparativa entre modelos baseados em redes neurais em contexto de dados tabulares (MLP, ResNet e Transformer) e modelos baseados em árvores (Random Forest, LightGBM e CatBoost) para a predição do GIF final de pacientes com hanseníase. Para esse propósito, foram estruturadas duas bases representando janelas temporais distintas do cuidado: uma contendo exclusivamente variáveis disponíveis no início do tratamento e outra incorporando informações coletadas durante o acompanhamento terapêutico. Adicionalmente, investigaram-se duas estratégias de mitigação do desbalanceamento da variável-alvo: ponderação por pesos e *hybrid sampling*.

Os resultados evidenciaram superioridade consistente dos modelos baseados em *gradient boosting*, especialmente do LightGBM, que apresentou os maiores valores de *AUC OvO* em ambas as janelas temporais e sob diferentes estratégias de balanceamento. O Transformer tabular apresentou desempenho competitivo e estável, compondo, juntamente com o CatBoost e as demais redes neurais, um segundo patamar de resultados, porém sem superar sistematicamente o LightGBM. O Random Forest, embora tradicionalmente robusto em dados tabulares, apresentou desempenho global inferior na maioria dos cenários avaliados.

Ainda mais, observou-se que o balanceamento por ponderação de pesos se mostrou mais estável do que o *hybrid sampling*, produzindo resultados mais consistentes en-

tre modelos e janelas temporais. O *hybrid sampling* não promoveu ganho sistemático nas classes minoritárias e, em alguns casos, notadamente no CatBoost, esteve associada a instabilidades relevantes na edição do GIF 2. Esses achados reforçam que a escolha da estratégia de balanceamento deve ser orientada por validação empírica específica ao modelo e ao contexto dos dados.

A inclusão de variáveis coletadas durante o tratamento resultou em melhora discreta e consistente no desempenho global, sugerindo que informações mais próximas do desfecho capturam com maior fidelidade a evolução clínica e neurológica do paciente. Ainda assim, independentemente da arquitetura empregada, verificou-se limitação persistente na predição dos GIF 1 e 2.

Esses resultados sugerem que as variáveis disponíveis atualmente na base do SINAN podem não capturar integralmente aspectos fisiopatológicos determinantes da progressão da incapacidade, como a perda segmentar de sensibilidade e comprometimento motor específico. Além disso, o forte desbalanceamento estrutural da variável alvo impõe desafios adicionais à modelagem. Dessa forma, os achados sustentam que modelos baseados em árvores, particularmente o LightGBM, são mais adequados para a predição final do GIF levando em consideração a base utilizada neste estudo. Contudo, o desempenho ainda não atinge robustez suficiente para aplicação clínica direta, sobretudo em relação às classes mais graves.

Como perspectivas futuras, planejamos *(i)* incorporar variáveis clínicas mais granulares provenientes de formulários de avaliação neurológica; *(ii)* investigar arquiteturas multimodais que integrem dados tabulares e informações não estruturadas; e *(iii)* aplicar técnicas de explicabilidade (XAI) para analisar os atributos mais relevantes no aprendizado dos modelos. Tais direções podem contribuir para o desenvolvimento de ferramentas preditivas mais seguras, equitativas e clinicamente aplicáveis no contexto da vigilância e do cuidado em hanseníase.

Referências

Andrade, H. G. V. d., Rocha, E. d. S., Monteiro, K. H. d. C., de Morais, C. M., dos Santos, D. C. M., Nascimento, D. C., Dourado, R. A., Lynn, T., and Endo, P. T. (2025). On the usage of artificial intelligence in leprosy care: A systematic literature review. *PLOS Computational Biology*, 21(6):e1012550.

Barbieri, R. R., Xu, Y., Setian, L., Souza-Santos, P. T., Trivedi, A., Cristofono, J., Bhering, R., White, K., Sales, A. M., Miller, G., et al. (2022). Reimagining leprosy elimination with ai analysis of a combination of skin lesion images with demographic and clinical data. *The Lancet Regional Health–Americas*, 9.

Bomtempo, C. F., Ferrari, S. M. F., de Faria Grossi, M. A., and Lyon, S. (2023). Evolução do grau de incapacidade física e do escore olhos, mãos e pés em casos novos de hanseníase: do diagnóstico à alta medicamentosa. *Hansenologia Internationalis: hanseníase e outras doenças infecciosas*, 48:e37331–e37331.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2021). Deep neural networks and tabular data: A survey. *arXiv preprint arXiv:2110.01889*.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

da Costa, N. M. G. B., Barbosa, T. d. C. S., Queiroz, D. T., Oliveira, A. K. A., Montemezzo, L. C. D., and do Couto Andrade, U. (2020). Perfil sociodemográfico e grau de incapacidade do portador de hanseníase em um centro de referência no estado do ceará. *Brazilian Journal of Development*, 6(6):41439–41449.

da Silva, Y. E. D., Salgado, C. G., Conde, V. M. G., and Conde, G. A. B. (2018). Application of clustering technique with kohonen self-organizing maps for the epidemiological analysis of leprosy. In *Advances in Intelligent Systems and Computing*, pages 295–309. Springer.

De Souza, M. L. M., Lopes, G. A., Branco, A. C., Fairley, J. K., and Fraga, L. A. D. O. (2021). Leprosy screening based on artificial intelligence: Development of a cross-platform app. *JMIR MHealth and UHealth*, 9(4):e23718.

DKE (2022). *German Standardization Roadmap on Artificial Intelligence*. DIN e. V., Berlin, 2 edition.

Dorogush, A. V., Ershov, V., and Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Machine Learning (ICML) – Workshop on Challenges for Categorical Data*. PMLR.

Dutra da Silva, Y. E., Salgado, C. G., Gomes Conde, V. M., and Barros Conde, G. A. (2018). Data mining using clustering techniques as leprosy epidemiology analyzing model. In *Lecture Notes in Computer Science*, pages 284–293. Springer.

Fernandes, J. R. N., Teles, A. S., Fernandes, T. R. S., Lima, L. D. B., Balhara, S., Gupta, N., and Teixeira, S. (2024). Artificial intelligence on diagnostic aid of leprosy: A systematic literature review. *Journal of Clinical Medicine*, 13(180):1–22.

Freitas, L. R. S., Freitas, J. A. O., Penna, G. O., and Duarte, E. C. (2025). Evaluating machine learning models for predicting late leprosy diagnosis by physical disability grade in brazil (2018–2022). *Tropical Medicine and Infectious Disease*, 10(5):131.

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. (2021). Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems (NeurIPS)*.

Gutta, V., Ganakammal, S. R., Jones, S., Beyers, M., and Chandrasekaran, S. (2024). Unnt: A novel utility for comparing neural net and tree-based models. *PLOS Computational Biology*, 20(4):1–11.

Kadra, A., Lindauer, M., Hutter, F., and Grabocka, J. (2021). Well-tuned simple nets excel on tabular datasets. NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA.

Ministério da Saúde (2022). Protocolo clínico e diretrizes terapêuticas da hanseniase. Publicação oficial; Accessed 2025-12-09.

Organização Mundial da Saúde (2010). *Estratégia Global Aprimorada para Redução Adicional da Carga da Hanseníase : 2011–2015 — Diretrizes Operacionais (Atualizadas)*. Organização Pan-Americana da Saúde, Brasília.

Salmi, M., Atif, D., Oliva, D., Abraham, A., and Ventura, S. (2024). Handling imbalanced medical datasets: review of a decade of research. *Artificial Intelligence Review*, 57(10):273.

Saltz, J. S. (2021). Crisp-dm for data science: strengths, weaknesses and potential next steps. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 2337–2344. IEEE.

Santos, G. M. C. d., Byrne, R. L., Cubas-Atienzar, A. I., and Santos, V. S. (2024). Factors associated with delayed diagnosis of leprosy in an endemic area in northeastern brazil: a cross-sectional study. *Cadernos de Saúde Pública*, 40(1):e00113123.

Secretaria de Vigilância em Saúde (2023). Formulário para avaliação neurológica simplificada e classificação do grau de incapacidade física em hanseníase.

Shmuel, A., Glickman, O., and Lazebnik, T. (2025). A comprehensive benchmark of machine and deep learning models on structured data for regression and classification. *Neurocomputing*, 655:131337. Available online 3 September 2025.

Shwartz-Ziv, R. and Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*.

Véras, G. C. B., da Silva, L. H., Sarmento, W. M., de Moraes, R. M., dos Santos Oliveira, S. H., and Soares, M. J. G. O. (2023). Características sociodemográficas e epidemiológicas relacionadas ao grau de incapacidade física em hanseníase no estado da paraíba, brasil. *Hansenologia Internationalis: hanseníase e outras doenças infecciosas*, 48:e38999–e38999.

Véras, G. C. B., Lima Júnior, J. F., Cândido, E. L., and Maia, E. R. (2021). Risk factors for physical disability due to leprosy: a case-control study. *Cadernos Saúde Coletiva*, 29:411–423.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester.

World Health Organization (2020). *Ending the Neglect to Attain the Sustainable Development Goals: A Road Map for Neglected Tropical Diseases 2021–2030*. World Health Organization, Geneva. Electronic version. Licensed under CC BY-NC-SA 3.0 IGO.

World Health Organization (2021). Global leprosy (hansen's disease) strategy 2021–2030: Towards zero leprosy. Technical report, World Health Organization, Geneva. Global strategy document.

World Health Organization (2025). Global leprosy (hansen disease) update, 2024: Beyond zero cases – what elimination of leprosy really means. *Weekly Epidemiological Record*, 100(37):365–384. Published 12 September 2025.