Transfer Learning para Doença Renal Crônica: Avaliando a Generalização de Modelos em Diferentes Cenários Clínicos

Rafael C. Pregardier[1], Luis A. L. Silva[1], Rafael A. da Costa[1], Gabriel V. S. Caetano[1], Joaquim V. C. Assunção[1], Carlos E. Poli-de-Figueiredo[2], Isabel C. Reinheimer[1]

[1]Curso de Ciência da Computação – Universidade Federal de Santa Maria (UFSM)
Av. Roraima nº 1000 – 97105-900 – Santa Maria -– RS – Brasil

[2]Escola de Medicina (ESMED) – Laboratório de Nefrologia
Pontifícia Universidade Católica do Rio Grande do Sul (PUC-RS)
90619-900 – Porto Alegre – RS – Brasil

{rcpregardier, luisalvaro, racosta, gvcaetano, joaquim}@inf.ufsm.br,

cepolif@pucrs.br, cristinareinheimer@gmail.com

Abstract. *Chronic Kidney Disease (CKD) requires robust predictive models in the face of population heterogeneity and data scarcity. This study evaluates Transfer Learning (TL) strategies on tabular data for CKD progression prediction under domain shift, comparing MLP with fine-tuning, TASC (proposed model), and TabPFN against models trained from scratch. Performance varied according to domain similarity: TASC stood out in clinical metrics under lower domain shift, while TabPFN showed high discrimination but inferior performance on balanced metrics. The results reinforce the importance of considering domain shift when choosing predictive models in nephrology.*

Resumo. *A Doença Renal Crônica (DRC) exige modelos preditivos robustos frente à heterogeneidade populacional e escassez de dados. Este estudo avalia estratégias de Transfer Learning (TL) em dados tabulares para predição de progressão da DRC sob domain shift, comparando MLP com fine-tuning, TASC (modelo proposto) e TabPFN a modelos treinados do zero. O desempenho variou conforme a similaridade entre domínios: o TASC destacou-se em métricas clínicas com menor domain shift, enquanto o TabPFN apresentou alta discriminação, mas desempenho inferior em métricas balanceadas. Os resultados reforçam a importância de considerar o domain shift na escolha de modelos preditivos em nefrologia.*

1. Introdução

A Doença Renal Crônica (DRC) constitui um dos principais desafios contemporâneos em saúde pública, associada à evolução silenciosa, complicações cardiovasculares e à necessidade de tratamento dialítico. A identificação precoce da progressão clínica é crucial para direcionar intervenções, otimizar recursos e reduzir desfechos adversos [Brasil. Ministério da Saúde 2014]. Embora técnicas de *Machine Learning* (ML) tenham sido amplamente aplicadas a esse problema (e.g., [Sanmarchi et al. 2023]), a transferência de conhecimento entre cenários clínicos reais segue limitada por fatores como variabilidade interinstitucional, desbalanceamento de classes e escassez de dados rotulados de alta qualidade.

Em pequenos *datasets* tabulares clínicos, modelos treinados do zero podem apresentar sobreajuste e desempenho instável, decorrentes de relações não lineares e dependências globais relevantes entre as variáveis [Hollmann et al. 2025]. Dentre as possíveis abordagens para mitigar as limitações de generalização em cenários com heterogeneidade e baixo volume amostral, o *Transfer Learning* (TL) busca reaproveitar as representações aprendidas em domínios relacionados. Oportunamente, modelos baseados em *Transformers* têm demonstrado desempenho competitivo em dados tabulares, através de mecanismos de autoatenção que permitem modelar interações globais entre atributos e capturar dependências não lineares entre variáveis [El-Melegy et al. 2024].

Diante disso, este trabalho investiga três estratégias de TL em subgrupos clínicos de DRC definidos por estágio da doença, etiologia e faixa etária, representando diferentes formas de *domain shift* observadas na prática. Avalia-se a transferibilidade de conhecimento entre modelos frente a distintas fontes de heterogeneidade populacional. Para isso, são explorados: i) *Multilayer Perceptron* (MLP) com *fine-tuning* (FT); ii) modelo baseado em *Transformers* com pré-treinamento auto-supervisionado (*Tabular Autoencoder Self-supervised Classifier* – TASC[1]); e iii) modelo do tipo *foundation model* para dados tabulares (*Tabular Prior- Data Fitted Network* – TabPFN). Os métodos são comparados a modelos treinados do zero nos diferentes cenários clínicos.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta uma revisão de trabalhos que utilizam *Transfer Learning* para predição na área clínica; a Seção 3 detalha o TASC e os outros modelos utilizados no trabalho; a Seção 4 apresenta experimentos e resultados; e, por fim, a Seção 5 discute conclusões, limitações e trabalhos futuros.

2. Trabalhos Relacionados

Modelos preditivos para DRC têm sido investigados na classificação de estágio, prognóstico de progressão e risco de desfechos adversos [Zhu et al. 2023, Okita et al. 2024]. Em dados clínicos tabulares, métodos tradicionais de ML costumam apresentar desempenho competitivo quando há um volume amostral adequado e uma distribuição estável das variáveis, mas degradam em cenários de escassez e heterogeneidade, especialmente em pacientes com estágios avançados e subtipos clínicos específicos ou perfis sub-representados[Borisov et al. 2024].

Nesse contexto, abordagens baseadas em TL emergem como uma alternativa promissora ao permitir o reaproveitamento de conhecimento entre domínios, melhorando a predição em contextos com poucos dados [Ebbehoj et al. 2022]. Suas estratégias incluem abordagens baseadas em modelo, instâncias, características e relações, destacando-se o *fine-tuning*[Tan et al. 2023]. Métodos como o proposto por [Segev et al. 2017] demonstram ganhos com a adaptação estrutural do Random Forest, indicando que o TL pode superar estratégias clássicas de transferência, como o Adaptive SVM.

Em cenários clínicos, o TL pode reduzir custos associados à coleta de dados e acelerar a modelagem, embora sua efetividade dependa da similaridade entre os domínios e do regime de adaptação adotado [Gao and Cui 2020]. Isso foi evidenciado por [Desautels et al. 2017] que aplicaram TL para predição de óbito em um hospital com escassez de dados, transferindo conhecimento entre instituições. O volume de dados para

[1] `https://github.com/Pre9o/TASC-Tabular-Autoencoder-Self-supervised-Classifier`

atingir AUC-ROC $>$ 0,80 foi reduzido de 4.000 pacientes para 220, assim como o tempo de coleta de dados clínicos.

No contexto da COVID-19, [Savalli et al. 2025] descrevem um estudo envolvendo 12 hospitais brasileiros, utilizando seu melhor modelo local como domínio-fonte e realizando adaptações incrementais com dados dos hospitais-alvo. A estratégia promoveu ganho de desempenho na maioria das instituições, embora seja sensível a discrepâncias na distribuição dos preditores. Já nos cenários de monitoramento clínico em tempo real, [Shih et al. 2020] desenvolveram um sistema de alerta precoce para complicações em hemodiálise utilizando dados provenientes de dispositivos vestíveis e equipamentos de diálise. O uso de TL elevou significativamente a precisão dos modelos desenvolvidos.

Dentre os modelos que podem ser utilizados para TL, abordagens baseadas em redes neurais profundas têm ampliado a capacidade de modelar relações não lineares e extrair representações latentes complexas nos dados clínicos. Neste contexto, *autoencoders* profundos podem viabilizar tanto o aumento do espaço amostral quanto o enriquecimento do espaço de características por meio da transferência de códigos latentes entre domínios, mitigando limitações de volume e desbalanceamento [Macias et al. 2022]. Especialmente, os *Transformers* que vêm sendo adaptados para dados tabulares por meio de *embeddings* de atributos e camadas de *self-attention*, obtendo resultados competitivos [Hollmann et al. 2023].

Embora o TL tenha apresentado resultados promissores em diferentes aplicações na saúde, sua utilização em dados clínicos tabulares ainda é relativamente limitada. Portanto, permanece a necessidade de investigar o comportamento de diferentes abordagens de TL em cenários de heterogeneidade clínica. Assim, este trabalho compara distintas estratégias de TL em dados tabulares de pacientes com DRC sob diferentes formas de *domain shift* clínico.

3. Metodologia

O *pipeline* do estudo foi estruturado em três etapas principais:

- **Exploração e Pré-processamento**: seleção e normalização de variáveis, definição dos subgrupos clínicos e construção dos domínios de transferência (fonte e alvo);
- **Arquiteturas e hiperparâmetros**: especificação dos modelos, definição do protocolo experimental e estratégia de otimização;
- **Treinamento dos modelos**: treinamento e avaliação de três estratégias – (i) MLP com FT, (ii) *Transformers* com pré-treinamento TASC e (iii) modelo do tipo *foundation model* para dados tabulares (TabPFN).

3.1. Exploração e Pré-processamento

Este trabalho utiliza um *dataset* com 51 variáveis formadas por dados clínicos estruturados de indivíduos com DRC (estágio 2 a 5), provenientes da coorte prospectiva CKD-ROUTE (dados disponíveis publicamente no repositório Dryad [Iimori et al. 2018]). O desfecho analisado é a progressão da DRC, definida como a perda de 50% da função renal basal ou a necessidade de diálise.

A seleção de variáveis utilizou o algoritmo Boruta [Kursa et al. 2010], aplicado exclusivamente aos dados de treinamento de cada subgrupo. O Boruta é um método de seleção de variáveis baseado em Random Forest (RF) que compara a importância de cada variável original com suas versões aleatorizadas (*shadow features*), confirmando como

relevantes aquelas que superam esse limiar. O classificador RF foi usado na execução do Boruta, com profundidade máxima de cinco níveis e número de estimadores "auto".

Além do Boruta, a seleção das variáveis considerou a avaliação de nefrologista, buscando um equilíbrio entre critérios computacionais e clínicos. Após a identificação de variáveis com maiores níveis de importância, os especialistas clínicos realizaram uma revisão adicional para eliminar redundância clínica entre atributos e prevenir vazamento de dados (ou seja, que nenhuma variável incorporasse informação direta ou indireta relacionada ao desfecho). A seguir, procedeu-se a normalização das variáveis contínuas.

A divisão do *dataset* foi realizada em subgrupos clínicos, posteriormente organizados em domínios fonte e alvo, conforme descrito abaixo:

- **Estágio da DRC**: pacientes nos estágios 2, 3 e 4 compuseram o domínio fonte (n=929), enquanto pacientes em estágio 5 foram definidos como domínio alvo (n=209). Observou-se desbalanceamento entre os domínios: no conjunto fonte, a progressão da DRC ocorreu em 15,1% dos casos, enquanto no conjunto alvo essa proporção foi de 67,0%. As variáveis selecionadas incluíram Proteinuria, Etiologia da DRC, idade, hemoglobina (Hb), albumina (Alb), taxa de filtração glomerular estimada (eGFR), razão proteína/creatinina urinária (UPCR) e duração observacional.
- **Etiologia primária**: as causas de DRC foram agrupadas em dois conjuntos – etiologias metabólico-vasculares (diabetes e nefroesclerose), definidas como domínio fonte (n=738), e etiologias imunológicas e outras (glomerulopatias e demais causas), como domínio alvo (n=400). Observou-se diferença na distribuição do desfecho, com progressão em 28,3% no domínio fonte e 17,8% no domínio alvo. As variáveis selecionadas incluíram Proteinuria, idade, pressão arterial sistólica (SBP), índice de massa corporal (BMI), Hb, Alb, eGFR, UPCR e duração observacional.
- **Faixa etária**: os indivíduos foram divididos em idosos ($\geq$60 anos), definidos como domínio fonte (n=880), e adultos (18–59 anos), como domínio alvo (n=258). Observou-se progressão da DRC em 22,5% no domínio fonte e 31,8% no domínio alvo. As variáveis selecionadas incluíram Proteinuria, Etiologia da DRC, SBP, Hb, Alb, eGFR, UPCR e duração observacional.

A estratificação permitiu a construção de domínios com perfis epidemiológicos distintos, viabilizando a análise do impacto do *domain shift* e da aplicação de técnicas de TL entre subpopulações heterogêneas, o domínio fonte foi definido como o subgrupo de maior volume amostral e menor gravidade clínica, enquanto o alvo representa o subgrupo de interesse prático com dados mais escassos. Cada domínio (fonte e alvo) foi particionado em conjuntos de treino (70%), validação (15%) e teste (15%), com divisão estratificada em relação ao desfecho, preservando a proporção de casos com progressão da DRC. Valores faltantes foram imputados por modelos de RF, utilizando um classificador para variáveis categóricas e um regressor para variáveis contínuas.

3.2. Arquiteturas e hiperparâmetros

Foram avaliados modelos MLP (treinados do zero e com FT), TabPFN e TASC (modelo baseado em Transformer). Os hiperparâmetros foram otimizados via otimização bayesiana com Optuna [Akiba et al. 2019], tendo como função objetivo o F1-score macro no conjunto de validação. Os espaços de busca foram definidos para equilibrar capacidade de representação e controle de *overfitting*.

Para os modelos MLP (fonte e treinados do zero), o espaço de busca incluiu *batch size* $\{4, 8, 16, 32, 64\}$, até 1000 épocas, *learning rate* entre 10^{-7} e 10^{-2}, 2 a 4 camadas ocultas, 4 a 64 neurônios por camada, *dropout* de 0 a 0,5 e *weight decay* de 10^{-7} a 10^{-2}. Para o MLP alvo, considerou-se *learning rate* entre 10^{-9} e 10^{-2}, número de camadas congeladas de 0 a a L, além de *batch size* , épocas e *weight decay* nos mesmos intervalos. Todos os MLPs utilizaram função de ativação ReLU e ponderação de classes para mitigar desbalanceamento.

Para o TabPFN, foram avaliados de 2 a 32 estimadores e temperatura do *Softmax* dentre 0,1 e 5,0. Para o classificador proposto, no domínio fonte, o espaço de busca incluiu parâmetros estruturais e de regularização, como dimensão dos *embeddings*, tamanho da camada *feed-forward*, número de camadas e cabeças de atenção, além de *dropout*, proporção de máscara, *batch size*, épocas, taxa de aprendizado. No domínio alvo, a configuração foi ajustada para avaliar o classificador final, considerando *batch size*, épocas, taxa de aprendizado, dimensão da camada *feed-forward* e *weight decay*.

3.3. Treinamento dos modelos

Foi adotada a estratégia de *Early Stopping* com base no F1-score macro no conjunto de validação, com paciência de 50 épocas. As implementações foram realizadas em Python 3.12.3, utilizando TensorFlow (Docker 2.17.0).

MLPs treinados do zero: Os modelos foram treinados exclusivamente nos subconjuntos alvo de cada grupo clínico, Estágio 5, Etiologias Imunológicas e Outras e Faixa etária (adultos), resultando em três modelos independentes. Esses modelos foram utilizados como *baselines* para comparação com as demais abordagens.

- **Estágio 5**: *batch size*=32, 550 épocas, *learning rate*=$5{,}37 \times 10^{-4}$, arquitetura (16–16–8), *dropout*=(0,1; 0,3; 0) e *weight decay*=$2{,}28 \times 10^{-3}$.
- **Etiologias Imunológicas e Outras**: *batch size*=8, 950 épocas, *learning rate*=$4{,}23 \times 10^{-3}$, arquitetura (16–8–8), *dropout*=(0; 0,5; 0,5) e *weight decay*=$2{,}94 \times 10^{-3}$.
- **Faixa etária (adultos)**: *batch size*=64; 900 épocas, *learning rate*=$7{,}68 \times 10^{-3}$, arquitetura (32–16), *dropout*=(0,1; 0,3) e *weight decay*=$2{,}62 \times 10^{-3}$.

MLPs com FT: O pré-treinamento foi realizado nos subconjuntos de origem (estágios 2, 3 e 4; etiologias metabólico-vasculares; e faixa etária: idosos), permitindo que as MLPs aprendessem representações iniciais a partir de dados mais amplos e estruturalmente relacionados aos subconjuntos alvo. Essa etapa visa fornecer uma inicialização de pesos mais informativa do que a inicialização aleatória, explorando padrões compartilhados entre os domínios e favorecendo a transferência de conhecimento durante o *fine-tuning*. Os hiperparâmetros utilizados no treinamento em cada subconjunto de origem foram:

- **Estágios 2, 3 e 4:** *batch size*=8; 650 épocas; *learning rate*=10^{-3}; arquitetura (32–16–8); *dropout*=(0,1; 0,3; 0,3); *weight decay*=$4{,}20 \times 10^{-3}$.
- **Etiologias Metabólico-vasculares:** *batch size*=16; 500 épocas; *learning rate*=$7{,}95 \times 10^{-5}$; arquitetura (32–16–16); *dropout*=(0,1; 0,5; 0,1); *weight decay*=$5{,}34 \times 10^{-4}$.
- **Faixa etária (idosos):** *batch size*=4; 950 épocas; *learning rate*=10^{-4}; arquitetura (16–8–4); *dropout*=(0,3; 0,2; 0,3); *weight decay*=$2{,}58 \times 10^{-5}$.

Após o pré-treinamento, os modelos foram submetidos ao *fine-tuning*, no qual os pesos aprendidos nos subconjuntos de origem foram reutilizados e ajustados com dados

dos respectivos subconjuntos-alvo. Diferentemente do treinamento do zero, essa etapa permite refinar representações previamente aprendidas, adaptando-as às especificidades clínicas de cada domínio. Em alguns cenários, camadas iniciais foram mantidas congeladas, preservando padrões mais gerais, enquanto camadas finais foram atualizadas para especialização da tarefa. Essa abordagem permite avaliar o impacto do TL na estabilidade do treinamento e no desempenho preditivo em relação aos modelos treinados do zero. Os hiperparâmetros adotados no FT em cada subconjunto-alvo são apresentados a seguir:

- **Estágio 5:** *batch size*=64; 850 épocas; *learning rate*=10^{-3}; camadas congeladas=3; *weight decay*=$4{,}26 \times 10^{-3}$.
- **Etiologias Imunológicas e Outras:** *batch size*=16; 700 épocas; *learning rate*=$2{,}66 \times 10^{-4}$; camadas congeladas=1; *weight decay*=$2{,}96 \times 10^{-3}$.
- **Faixa etária (adultos):** *batch size*=8; 550 épocas; *learning rate*=$1{,}4 \times 10^{-3}$; camadas congeladas=0; *weight decay*=$3{,}36 \times 10^{-6}$.

TabPFN [Hollmann et al. 2023]: este modelo não requer treinamento nos dados da tarefa alvo, pois é previamente treinado em uma ampla coleção de tarefas sintéticas geradas a partir de um *prior* sobre dados tabulares. Durante esse processo, o modelo aprende a aproximar a inferência Bayesiana em problemas de classificação, a partir de conjuntos sintéticos com diferentes dimensões, tamanhos amostrais e relações estatísticas. Em cada tarefa, o modelo recebe exemplos rotulados e aprende a prever novas instâncias da mesma distribuição, internalizando estratégias de inferência reutilizáveis.

Na fase de inferência, o TabPFN opera em regime contextual, no qual os dados rotulados da tarefa alvo são utilizados apenas como contexto, sem atualização de parâmetros. O modelo recebe exemplos rotulados juntamente com as instâncias de teste e produz diretamente as probabilidades de classe, explorando o conhecimento adquirido no pré-treinamento.

Os parâmetros adotados em cada subconjunto-alvo foram:

- **Estágio 5:** estimadores=20; temperatura do *Softmax*=0,17.
- **Etiologias Imunológicas e Outras:** estimadores=2; temperatura do *Softmax*=0,30.
- **Faixa etária (adultos):** estimadores=15; temperatura do *Softmax*=3,0.

TASC: O classificador proposto baseia-se em um *autoencoder* com arquitetura *Transformer*, previamente treinado para aprender representações latentes de dados clínicos tabulares via aprendizado auto-supervisionado. Cada atributo é tratado como um *token*, permitindo modelar interações entre variáveis de forma análoga ao processamento de sequências. Variáveis categóricas são representadas por *embeddings*, enquanto variáveis contínuas são projetadas para o mesmo espaço por camadas lineares. As representações são concatenadas em uma sequência, à qual é adicionado um *token* de classificação (*[CLS]*) para capturar informações globais da amostra.

A sequência é processada por blocos *Transformer* compostos por *multi-head self-attention*, conexões residuais, *Layer Normalization* e camadas *feed-forward*. O mecanismo de *self-attention* permite capturar dependências entre atributos e modelar relações estruturais entre variáveis clínicas. O treinamento do *autoencoder* é realizado de forma auto-supervisionada por reconstrução de atributos mascarados, nos quais uma fração dos atributos é ocultada e o modelo aprende a inferir seus valores a partir das demais variáveis. Esse procedimento induz o aprendizado de representações latentes informativas, capturando dependências estruturais sem necessidade de rótulos.

A função de perda no pré-treinamento é definida de forma híbrida para acomodar a heterogeneidade dos dados tabulares, combinando erro quadrático médio para atributos contínuos e entropia cruzada categórica esparsa para atributos categóricos. O treinamento utiliza o otimizador AdamW com regularização por *weight decay*. Os hiperparâmetros adotados são apresentados a seguir:

- **Estágios 2, 3 e 4:** *batch size*=16; *embedding*=48; 1 camada *Transformer* (4 cabeças); *feed-forward*=16; *dropout*=10^{-3}; *learning rate*=9.7×10^{-4}; 600 épocas; mascaramento=0,5%.
- **Etiologias Metabólico-vasculares:** *batch size*=32; *embedding*=48; 3 camadas *Transformer* (2 cabeças); *feed-forward*=48; *dropout*=0,22; *learning rate*=2.21×10^{-3}; 950 épocas; mascaramento=0,6%.
- **Faixa etária (idosos):** *batch size*=16; *embedding*=96; 3 camadas *Transformer* (6 cabeças); *feed-forward*=48; *dropout*=0,2; *learning rate*=5.09×10^{-3}; 1000 épocas; mascaramento=0,5%.

Após a etapa de pré-treinamento, apenas o componente *encoder* do *autoencoder* é reaproveitado para compor o classificador final. Nesse contexto, o *encoder* pré-treinado atua como um extrator de características, transformando o vetor de atributos de entrada em uma representação latente de maior nível semântico. Dessa forma, o conhecimento aprendido durante a tarefa de reconstrução é transferido para a tarefa supervisionada de classificação.

No classificador, o vetor de entrada é inicialmente processado pelo *encoder Transformer* pré-treinado. A representação latente correspondente ao *token* especial *[CLS]* é utilizada como resumo global da amostra, agregando informações relevantes provenientes de todos os atributos da entrada. Essa representação é então encaminhada para uma cabeça de classificação composta por uma camada de normalização , seguida por uma camada totalmente conectada com função de ativação ReLU e regularização por *dropout*. Por fim, uma camada densa com função de ativação sigmoide produz a probabilidade associada à classe positiva.

O treinamento do classificador é realizado utilizando a função de perda *binary cross-entropy* e o otimizador AdamW. Durante essa etapa, os parâmetros do *encoder* podem ser ajustados juntamente com a cabeça de classificação por meio de *fine-tuning*, permitindo adaptar as representações aprendidas no pré-treinamento à tarefa específica de predição. Os hiperparâmetros utilizados no treinamento dos classificadores para cada subconjunto de dados são apresentados a seguir.

- **Estágio 5:** *batch size*=8; 800 épocas; *learning rate*=3.2×10^{-3}; arquitetura (8); *dropout*=0,1; *weight decay*=9.2×10^{-5}.
- **Etiologias Imunológicas e Outras:** *batch size*=32; 900 épocas; *learning rate*=2.21×10^{-4}; arquitetura (32); *dropout*=0,4; *weight decay*=6.65×10^{-6}.
- **Faixa etária (adultos):** *batch size*=8; 700 épocas; *learning rate*=10^{-3}; arquitetura (32); *dropout*=0,2; *weight decay*=7.23×10^{-7}.

4. Experimentos e Resultados

Os experimentos avaliam o desempenho de modelos com TL em comparação ao treinamento do zero em três cenários clínicos de transferência de conhecimento: i) Estágio da DRC 5, ii) Etiologias Imunológicas e Outras e iii) Faixa etária (adultos).

As métricas principais são F1-score e Sensibilidade, dado o equilíbrio necessário entre a identificação de casos de progressão da DRC e a redução de falsos negativos. A AUC-ROC complementa a análise ao avaliar a capacidade discriminativa global, enquanto o Brier Score Loss mede a calibração das predições (quanto menor, melhor). Acurácia e Precisão são reportadas como métricas auxiliares, sendo interpretadas com cautela devido ao desbalanceamento de classes.

Tabela 1. Comparação da divergência de Jensen-Shannon (JSD) entre os *datasets* origem e alvo em cada grupo clínico.

Grupo	Prot.	Etiol. DRC	Idade	Hb	SBP	BMI	Alb	eGFR	UPCR	Duração obs.	Prog. DRC
Estágios	0.1301	**0.0401**	0.0657	**0.0266**	–	–	0.0560	0.1230	0.3275	0.5943	0.2133
Etiologias	**0.0011**	–	**0.0278**	0.0327	**0.0284**	**0.0167**	0.0934	**0.0335**	0.0226	0.1751	0.0114
Faixa etária	0.0113	0.0442	–	0.0385	0.0373	–	**0.0355**	0.0483	**0.0203**	**0.0820**	**0.0079**

A Tabela 1 apresenta a divergência de Jensen-Shannon (JSD) entre os domínios fonte e alvo em cada grupo clínico, utilizada como medida do grau de *domain shift*. Valores mais elevados indicam maior dissimilaridade entre as distribuições, potencialmente reduzindo a efetividade das estratégias de TL.

Tabela 2. Resultados no subconjunto Estágio 5.

Métricas	MLP do zero	Fine-tuning	TabPFN	TASC (Proposto)
Acurácia	0.8452 ± 0.0316	$\mathbf{0.8516 \pm 0.0158}$	0.7290 ± 0.0158	$\mathbf{0.8516 \pm 0.0258}$
Precisão	0.8259 ± 0.0311	$\mathbf{0.8371 \pm 0.0288}$	0.7660 ± 0.0112	0.8296 ± 0.0273
Sensibilidade	0.8491 ± 0.0431	0.8276 ± 0.0012	0.7947 ± 0.0121	$\mathbf{0.8538 \pm 0.0270}$
F1-Score	0.8300 ± 0.0348	0.8294 ± 0.0115	0.7262 ± 0.0145	$\mathbf{0.8373 \pm 0.0275}$
AUC-ROC	0.8800 ± 0.0286	0.8781 ± 0.0103	$\mathbf{0.9276 \pm 0.0114}$	0.8781 ± 0.0197
Brier Score Loss	0.1934 ± 0.0215	0.1984 ± 0.0220	$\mathbf{0.1410 \pm 0.0034}$	0.1587 ± 0.0212

Com base na Tabela 2, o modelo TASC apresentou o melhor desempenho nas métricas de maior relevância clínica, com F1-score de 0,8373 e Sensibilidade de 0,8538. Em relação ao método de *fine-tuning*, observou-se ganho de aproximadamente 3,2% em Sensibilidade, o que significa, concretamente, que a cada 100 pacientes com progressão real da DRC, o TASC identificaria cerca de 3 casos adicionais — volume que, em nefrologia, pode representar intervenções precoces clinicamente relevantes, como ajuste terapêutico ou encaminhamento para diálise. Embora o FT tenha alcançado Acurácia e Precisão ligeiramente superiores (0,8516 e 0,8371, respectivamente), esse resultado é menos desejável em cenários clínicos nos quais o custo de um falso negativo — um paciente com progressão não identificada — supera o custo de um falso positivo.

O MLP treinado do zero apresentou F1-score de 0,8300 e Sensibilidade de 0,8538, sugerindo ganho limitado das estratégias de TL neste cenário. Esse resultado é consistente com o elevado *domain shift* observado na Tabela 1 para o grupo de ”Estágios”, com altos valores de JSD em variáveis-chave como UPCR (0,3275), duração observacional (0,5943) e proteinúria (0,1301), além de 0,2133 no desfecho. Esse padrão reflete a heterogeneidade entre pacientes em estágio terminal e aqueles em estágios iniciais, indicando aproveitamento limitado de representações aprendidas no pré-treinamento.

O TabPFN destacou-se em AUC-ROC (0,9276) e Brier Score Loss (0,1410), com ganhos de aproximadamente 5,4% e 27% em relação ao MLP treinado do zero, respectivamente, indicando melhor discriminação global e calibração probabilística. Contudo, sua AUC elevada reflete boa capacidade de ordenação de risco, mas não implica desempenho adequado mesmo com o limiar de decisão ajustado: o menor F1-score (0,7262) entre os modelos indica que, mesmo com calibração adicional do limiar, o TabPFN tenderia a gerar mais falsos negativos em uso clínico real, comprometendo sua aplicabilidade prática neste cenário.

Tabela 3. Resultados no subconjunto Etiologias Imunológicas e Outras.

Métricas	MLP do zero	Fine-tuning	TabPFN	TASC (Proposto)
Acurácia	$\mathbf{0.8333 \pm 0.0149}$	0.7733 ± 0.0170	0.7833 ± 0.0000	0.8167 ± 0.0149
Precisão	$\mathbf{0.7526 \pm 0.0095}$	0.6920 ± 0.0130	0.7141 ± 0.0000	0.7313 ± 0.0135
Sensibilidade	$\mathbf{0.8627 \pm 0.0171}$	0.7837 ± 0.0358	0.8321 ± 0.0000	0.8314 ± 0.0195
F1-Score	$\mathbf{0.7778 \pm 0.0119}$	0.7044 ± 0.0132	0.7283 ± 0.0000	0.7545 ± 0.0160
AUC-ROC	0.9306 ± 0.0152	0.8687 ± 0.0182	$\mathbf{0.9373 \pm 0.0045}$	0.9054 ± 0.0131
Brier Score Loss	0.1365 ± 0.0125	0.1897 ± 0.0077	$\mathbf{0.0807 \pm 0.0051}$	0.1785 ± 0.0249

Para o cenário clínico de Etiologias Imunológicas e Outras, os resultados (Tabela 3) indicam que o MLP treinado do zero apresentou o melhor desempenho nas métricas balanceadas, com F1-score de 0,7778 e Sensibilidade de 0,8627. O TASC obteve F1-Score de 0,7545 e Sensibilidade de 0,8314, valores inferiores ao MLP do zero, porém superiores ao *fine-tuning* tradicional em aproximadamente 7,1% no F1-Score e 6,1% na Sensibilidade.

O *fine-tuning* apresentou o pior desempenho neste cenário, com F1-score de 0,7044 e Sensibilidade de 0,7837 — queda de aproximadamente 8,2% em relação ao MLP do zero, o que equivale a cerca de 8 casos adicionais não identificados a cada 100 pacientes com progressão real. Esse resultado é consistente com o *negative transfer* decorrente da adaptação a partir de um domínio fonte de etiologias metabólico-vasculares, fenômeno discutido por [Gao and Cui 2020] e [Ebbehoj et al. 2022], que apontam a similaridade entre domínios como determinante para o sucesso da transferência. De fato, embora a dissimilaridade global seja relativamente baixa (Tabela 1), variáveis como albumina (JSD = 0,0934) e duração observacional (JSD = 0,1751) exibem maior divergência, sugerindo que o conhecimento extraído do domínio fonte atuou como ruído na adaptação, em vez de fornecer representações transferíveis.

O TabPFN novamente apresentou melhor desempenho em AUC-ROC (0,9373) e Brier Score Loss (0,0807), este último cerca de 40,8% inferior ao MLP treinado do zero (0,1365), indicando superior calibração probabilística. Entretanto, seu F1-score (0,7283) foi aproximadamente 6,4% inferior ao MLP do zero, reforçando o padrão observado no cenário de estágios, no qual o TabPFN apresenta boa discriminação global, porém desempenho inferior nas métricas balanceadas mais relevantes para o contexto clínico.

A Tabela 4 apresenta os resultados para o cenário de pacientes adultos. O TASC obteve o melhor desempenho em todas as métricas, com F1-score de 0,9020, Sensibilidade de 0,9038 e AUC-ROC de 0,9639 — ganhos de aproximadamente 4,5% no F1-score e 2,2% na Sensibilidade em relação ao MLP do zero, equivalendo à identificação de cerca

Tabela 4. Resultados no subconjunto Faixa etária: Adultos.

Métricas	MLP do zero	Fine-tuning	TabPFN	TASC (Proposto)
Acurácia	0.8718 ± 0.0000	0.8872 ± 0.0125	0.8359 ± 0.0126	$\mathbf{0.9128 \pm 0.0126}$
Precisão	0.8533 ± 0.0000	0.8699 ± 0.0137	0.8172 ± 0.0139	$\mathbf{0.9037 \pm 0.0139}$
Sensibilidade	0.8846 ± 0.0000	0.8923 ± 0.0094	0.8462 ± 0.0172	$\mathbf{0.9038 \pm 0.0211}$
F1-Score	0.8628 ± 0.0000	0.8775 ± 0.0121	0.8247 ± 0.0139	$\mathbf{0.9020 \pm 0.0149}$
AUC-ROC	0.9485 ± 0.0110	0.9479 ± 0.0180	0.9609 ± 0.0029	$\mathbf{0.9639 \pm 0.0032}$
Brier Score Loss	0.0848 ± 0.0102	0.0930 ± 0.0182	0.1639 ± 0.0020	$\mathbf{0.0756 \pm 0.0122}$

de 2 pacientes adicionais com progressão real a cada 100 casos. Em contexto ambulatorial, esse ganho pode antecipar decisões como intensificação do controle pressórico ou encaminhamento para preparo à terapia renal substitutiva. O resultado é consistente com o baixo *domain shift* observado na Tabela 1, com valores reduzidos de JSD em variáveis-chave como UPCR (0,0203) e proteinúria (0,0113), indicando maior similaridade distribucional entre os domínios fonte (idosos) e alvo (adultos).

O método tradicional de *fine-tuning* também apresentou desempenho competitivo, com F1-Score de 0,8775 e Sensibilidade de 0,8923, superando o MLP do zero em aproximadamente 1,7% e 0,9%, respectivamente. Esse resultado reforça que, em cenários com maior similaridade entre domínios, estratégias de TL tendem a produzir ganhos mais consistentes — contrastando diretamente com o *negative transfer* observado no cenário de etiologias, onde o mesmo método apresentou queda expressiva de desempenho. Tomados em conjunto, esses dois cenários sugerem que valores de JSD superiores a aproximadamente 0,15 em variáveis clinicamente relevantes podem funcionar como sinal de alerta para o risco de *negative transfer*, orientando a escolha entre TL e treinamento do zero na prática.

Neste ensaio, o TabPFN não apresentou o melhor desempenho, diferentemente do padrão observado nos cenários clínicos anteriores (estágios e etiologias). Embora tenha alcançado AUC-ROC de 0,9609, valor próximo ao do TASC (0,9639), indicando elevada capacidade discriminativa global, como apresentado em [Hollmann et al. 2025], seu F1-score (0,8247) foi o mais baixo entre os modelos avaliados, aproximadamente 4,4% inferior ao MLP treinado do zero. Além disso, o Brier Score Loss do TabPFN (0,1639) foi o mais elevado entre os modelos, indicando pior calibração probabilística, já que valores menores correspondem a melhor desempenho. Assim, apesar da boa capacidade de ordenação global, o modelo apresentou limitações em termos de desempenho clínico e confiabilidade probabilística.

5. Conclusão

A DRC exige o desenvolvimento de modelos preditivos capazes de apoiar decisões clínicas em cenários marcados por heterogeneidade populacional, limitação de dados e mudanças na distribuição entre subpopulações. Nesse contexto, este estudo investigou a efetividade de três estratégias de *Transfer Learning* em dados clínicos tabulares, analisando o comportamento dos modelos frente a diferentes graus de *domain shift*, além de propor o modelo TASC, que combina pré-treinamento auto-supervisionado baseado em *Transformers* com *fine-tuning* supervisionado.

Os resultados indicam que o desempenho das abordagens é diretamente associado

à similaridade distribucional entre os domínios. Em cenários de menor *domain shift*, como na estratificação por faixa etária, o TASC apresentou ganhos consistentes em métricas clinicamente relevantes; em contrapartida, contextos de maior heterogeneidade, como nos estágios da doença, limitaram os benefícios do TL. O TabPFN demonstrou elevada capacidade discriminativa em todos os cenários e boa calibração nos estágios e etiologias; porém, teve desempenho inferior em métricas balanceadas, o que pode restringir sua aplicabilidade em contextos que exigem maior sensibilidade. Em geral, os achados reforçam que a escolha da estratégia deve considerar o grau de *domain shift* e as métricas mais relevantes à decisão clínica, contribuindo para sistemas de apoio mais robustos na nefrologia.

Além disso, os experimentos foram conduzidos exclusivamente na coorte prospectiva CKD-ROUTE, de origem japonesa, o que pode restringir a generalização dos achados para populações com perfis epidemiológicos distintos. Ademais, a ausência de validação prospectiva ou multicêntrica impede afirmações mais amplas sobre o comportamento dos modelos em ambientes clínicos reais. Por fim, as formas de *domain shift* investigadas limitam-se a três estratificações clínicas predefinidas: estágio da doença, etiologia e faixa etária, sem contemplar variações de origem institucional ou temporal, que representam fontes adicionais relevantes de heterogeneidade.

Como trabalhos futuros, propõe-se: a validação dos modelos com coortes brasileiras ou de outros países, avaliando a transferibilidade intercultural das representações aprendidas; além da aplicação de técnicas de explicabilidade, como SHAP e LIME, ao TASC, visando à interpretação clínica das predições e ao aumento da confiança dos profissionais de saúde no modelo; e a extensão do *pipeline* para dados longitudinais ou multimodais em nefrologia, explorando a capacidade do TASC de integrar séries temporais de exames laboratoriais e informações de múltiplas fontes clínicas.

Este trabalho foi apoiado pela CAPES, Edital nº 14/2023. Poli-de-Figueiredo é Bolsista PQ-CNPq (Processo nº 308250/2022-2).

Referências

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proc. of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2024). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519.

Brasil. Ministério da Saúde (2014). Diretrizes clínicas para o cuidado ao paciente com doença renal crônica – drc no sistema Único de saúde. Secretaria de Atenção à Saúde, Departamento de Atenção Especializada e Temática.

Desautels, T., Calvert, J., Hoffman, J., Mao, Q., Jay, M., Fletcher, G., Barton, C., Chettipally, U., Kerem, Y., and Das, R. (2017). Using transfer learning for improved mortality prediction in a data-scarce hospital setting. *Biomedical informatics insights*, 9:1178222617712994.

Ebbehoj, A., Thunbo, M. Ø., Andersen, O. E., Glindtvad, M. V., and Hulman, A. (2022). Transfer learning for non-image data in clinical research: A scoping review. *PLOS Digital Health*, 1(2):e0000014.

El-Melegy, M., Mamdouh, A., Ali, S., Badawy, M., El-Ghar, M. A., Alghamdi, N. S., and El-Baz, A. (2024). Prostate cancer diagnosis via visual representation of tabular data and deep transfer learning. *Bioengineering*, 11(7):635.

Gao, Y. and Cui, Y. (2020). Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nature communications*, 11(1):5131.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. (2023). Tabpfn: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations 2023*.

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326.

Iimori, S., Naito, S., Noda, Y., Sato, H., Nomura, N., Sohara, E., Okado, T., Sasaki, S., Uchida, S., and Rai, T. (2018). Prognosis of chronic kidney disease with normal-range proteinuria: the ckd-route study. *PLoS One*, 13(1):e0190493.

Kursa, M. B., Jankowski, A., and Rudnicki, W. R. (2010). Boruta–a system for feature selection. *Fundamenta informaticae*, 101(4):271–285.

Macias, E., Lopez Vicario, J., Serrano, J., Ibeas, J., and Morell, A. (2022). Transfer learning improving predictive mortality models for patients in end-stage renal disease. *Electronics*, 11(9):1447.

Okita, J., Nakata, T., Uchida, H., Kudo, A., Fukuda, A., Ueno, T., Tanigawa, M., Sato, N., and Shibata, H. (2024). Development and validation of a machine learning model to predict time to renal replacement therapy in patients with chronic kidney disease. *BMC nephrology*, 25(1):101.

Sanmarchi, F., Fanconi, C., Golinelli, D., Gori, D., Hernandez-Boussard, T., and Capodici, A. (2023). Predict, diagnose, and treat chronic kidney disease with machine learning: a systematic literature review. *Journal of nephrology*, 36(4):1101–1117.

Savalli, C., Carneiro, A. H. A., Barcellos Filho, F., Bigoto, M. A. R., Wichmann, R. M., Chiavegatto Filho, A. D. P., et al. (2025). Transfer learning for covid-19 predictive modeling: A multicenter study of 12 hospitals. *Annals of Epidemiology*, 108:1–7.

Segev, N., Harel, M., Mannor, S., Crammer, K., and El-Yaniv, R. (2017). Learn on source, refine on target: A model transfer learning framework with random forests. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1811–1824.

Shih, C., Youchen, L., Chen, C.-h., and Chu, W. C.-C. (2020). An early warning system for hemodialysis complications utilizing transfer learning from hd iot dataset. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 759–767.

Tan, Z., Luo, L., and Zhong, J. (2023). Knowledge transfer in evolutionary multi-task optimization: A survey. *Applied Soft Computing*, 138:110182.

Zhu, Y., Bi, D., Saunders, M., and Ji, Y. (2023). Prediction of chronic kidney disease progression using recurrent neural network and electronic health records. *Scientific reports*, 13(1):22091.