Predição de Atraso Alto em Infraestruturas de Rede através de Redes Neurais de Grafos

Janaina R. Santos[1], **Francisco V. J. Nobre**[1], **Ismael F. de Castro**[1], **Maria L. M. Linhares**[1], **Maria C. M. M. Ferreira**[1], **Rafael L. Gomes**[1]

[1]Universidade Estadual do Ceará (UECE)

{janaina.ribeiro, valderlan.nobre, clara.mesquita,ismael.fonteles

malu.linhares}@aluno.uece.br, rafa.lopes@uece.br

Abstract. *Managing large-scale network infrastructures is challenging, making the prediction of performance metrics, such as delay, crucial. However, existing predictive techniques struggle to understand how infrastructure characteristics and communication performance affect future behavior. Within this context, this work proposes an approach based on Graph Neural Networks (GNNs) for predicting high-delay situations. The infrastructure is modeled as a dynamic graph, where communication points are represented as nodes and their relationships are defined by historical delay correlations and topological similarity. The approach employs time windows and Bayesian hyperparameter optimization to enhance predictive capability. Experiments with real-world data from the Brazilian National Education and Research Network (RNP) demonstrate significant gains in the proactive detection of performance degradations.*

Resumo. *Gerenciar infraestruturas de rede de grande escala é desafiador, onde a predição de métricas de desempenho, tal como o atraso, é crucial. Contudo, as técnicas preditivas existentes possuem dificuldades em compreender como as características da infraestrutura e o desempenho das comunicações afetam o comportamento futuro. Dentro deste contexto, este trabalho propõe uma abordagem baseada em Redes Neurais de Grafos (GNNs) para predição de situações de atraso alto. A infraestrutura é modelada como um grafo dinâmico, no qual pontos de comunicação são representados como nós e suas relações são definidas por correlações históricas de atraso e similaridade topológica. A abordagem emprega aplicação de janelas temporais e otimização bayesiana de hiperparâmetros para aprimorar a capacidade preditiva. Experimentos com dados reais da Rede Nacional de Ensino e Pesquisa (RNP) demonstram ganhos significativos na detecção proativa de degradações de desempenho.*

1. Introdução

Serviços de monitoramento de rede são fundamentais para a compreensão do comportamento da infraestrutura e para o suporte ao planejamento estratégico em diversos ambientes operacionais, principalmente nos Provedores de Serviços de Internet (Internet Service Providers - ISPs) [Direito et al. 2023, Gomes et al. 2014]. Estes serviços fornecem medições regulares de desempenho, como vazão, atraso e perda de pacotes, que são essenciais para o planejamento de capacidade, gestão de Qualidade de Serviço (Quality os Service

- QoS) e conformidade com Acordos de Nível de Serviço (Service Level Agreement - SLA) [Ferreira et al. 2024, Souza et al. 2024].

O atraso ponta-a-ponta desempenha um papel crucial na avaliação do desempenho da rede, particularmente em redes de pesquisa e educação em larga escala [Brito et al. 2026]. Estudos empíricos utilizando dados operacionais reais da Rede Ipê da Rede Nacional de Pesquisa e Ensino (RNP)[1] do Brasil, uma infraestrutura basal que conecta mais de 500 instituições de pesquisa e educação, demonstraram que links com atraso persistentemente alto podem ser identificados através de análise baseada em correlação, revelando como a dinâmica do atraso em diferentes links impacta diretamente o desempenho geral da rede e a qualidade do serviço [Nobre et al. 2025].

Apesar da relevância das medições de atraso, o comportamento temporal do atraso de rede é inerentemente complexo e influenciado por múltiplos fatores, como mudanças de roteamento, padrões de congestionamento e instabilidade de caminho [Portela et al. 2024a, Portela et al. 2024b]. Estudos de medição em larga escala demonstraram que os atrasos de caminho na Internet exibem estruturas temporais heterogêneas, tornando sua caracterização e interpretação desafiadoras quando se baseia unicamente em análises estatísticas simples [Mouchet et al. 2020].

A crescente disponibilidade de medições de atraso contínuas e em larga escala amplia ainda mais esses desafios. Infraestruturas de medição como o RIPE Atlas geram volumes maciços de observações de atraso em pontos de vista geograficamente distribuídos, fornecendo visibilidade sem precedentes sobre o comportamento da rede ao longo do tempo. Contudo, a escala e a riqueza desses conjuntos de dados também expõem as limitações das abordagens analíticas tradicionais, que lutam para explorar totalmente as dependências temporais e as relações estruturais inerentes às redes de longa distância [Nosyk et al. 2025]. Neste contexto, a combinação de medições de atraso com informações topológicas extraídas de dados de *traceroute* torna-se particularmente relevante, pois permite a análise do comportamento do atraso não apenas ao longo do tempo, mas também através de caminhos de rede correlacionados. Uma visão integrada como essa é essencial para entender como as degradações de desempenho se propagam através de estruturas de roteamento compartilhadas, permitindo assim a antecipação de futuros eventos de atraso em redes de pesquisa operacionais.

Modelar o comportamento de atraso correlacionado representa desafios significativos para as abordagens tradicionais de aprendizado de máquina, que tipicamente assumem observações independentes ou dependem de representações de características fixas incapazes de capturar dependências estruturais. As infraestruturas de rede, entretanto, são inerentemente organizadas como grafos, onde nós e links estão interconectados por meio de relações complexas e dinâmicas que influenciam diretamente métricas de desempenho, como o atraso.

Para enfrentar esses desafios, as Redes Neurais de Grafos (Graphical Neural Networks - GNNs) fornecem uma estrutura fundamentada para aprender a partir de dados estruturados em grafos, generalizando redes neurais para domínios não-Euclidianos e permitindo a propagação de informação entre entidades conectadas [Kipf and Welling 2017, Wu et al. 2021]. Ao modelar explicitamente as interações locais e globais por meio

[1] `https://www.rnp.br//`

de mecanismos de passagem de mensagens, as GNNs são adequadas para capturar como as variações de atraso em elementos específicos da rede se propagam através de estruturas de roteamento compartilhadas, tornando-as uma escolha natural para a modelagem e previsão de atraso em redes de comunicação em larga escala.

Embora estudos existentes tenham demonstrado com sucesso o valor das medições de atraso e da análise baseada em *traceroute* para identificar e caracterizar anomalias de desempenho, a maioria das abordagens permanece fundamentalmente reativa, baseando-se na análise retrospectiva de variações de atraso observadas. Além disso, trabalhos anteriores tipicamente tratam a dinâmica temporal e a topologia de rede como dimensões separadas, limitando a capacidade de capturar sua influência conjunta em eventos de degradação de atraso.

Dentro deste contexto, este artigo propõe uma solução baseada em GNN para prever degradações de latência em ISPs, superando as limitações de modelos preditivos tradicionais. Através da modelagem da rede como um grafo dinâmico e do uso de correlações históricas e similaridade topológica (*traceroute*), a arquitetura captura padrões espaço-temporais complexos. A metodologia inclui o tratamento de dados desbalanceados e otimização bayesiana. Quanto ao desenvolvimento, o framework GraphSAGE[2] foi utilizado com o objetivo de viabilizar infraestruturas de rede de grande porte, bem como grafos dinâmicos nos quais novos nós ou conexões mudam constantemente, em função dos caminhos utilizados na comunicação (informações de *traceroute*).

Validados com dados reais da Rede Nacional de Ensino e Pesquisa, os resultados comprovam que a união de dados topológicos com aprendizado de máquina relacional oferece uma ferramenta proativa superior para a manutenção do desempenho da rede e cumprimento de SLAs. Em horizontes de curto prazo (3 minutos), a proposta alcançou boas métricas e conforme o horizonte de predição foi expandido para 30 minutos, mantendo a robustez. Além da precisão técnica, a solução apresentou excelente calibração probabilística e viabilidade operacional com tempos de inferência de poucos segundos.

De maneira geral, as principais contribuições do artigo são: (1) Propõe uma arquitetura baseada em GNN que integra correlações de atraso e similaridade topográfica via *traceroute* para modelar a rede como um grafo dinâmico; (2) A solução mantém o desempenho em horizontes longos com degradação mínima, superando a instabilidade de modelos puramente temporais; (3) Otimização de Desempenho para alcançar alta precisão e calibração probabilística em predições de diversos horizontes; e (4) Eficiência operacional validada a viabilidade da solução em tempo real com dados reais, apresentando tempos de inferência compatíveis com os ciclos de coleta da rede.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve a arquitetura proposta; a Seção 4 detalha a metodologia experimental; a Seção 5 discute os resultados obtidos; e a Seção 6 apresenta as conclusões.

2. Trabalhos Relacionados

Esta seção apresenta uma análise crítica de estudos acadêmicos que exploraram técnicas e estratégias para modelar o desempenho de redes e prever atrasos usando abordagens

[2]https://snap.stanford.edu/graphsage/

baseadas em grafos e temporais. A revisão sistemática da literatura existente destaca as principais contribuições em GNN, modelagem espaço-temporal e análise de correlação, fornecendo a base para a abordagem proativa proposta neste trabalho. A Tabela 1 sintetiza os principais trabalhos analisados, destacando seus respectivos problemas e abordagens.

Uma taxonomia fundamental das GNNs é fornecida, categorizando-as em redes recorrentes, convolucionais, autoencoders e espaço-temporais. Sua pesquisa ressalta a capacidade das GNNs de lidar com dados não Euclidianos, onde relacionamentos complexos e interdependências entre objetos são prevalentes. Embora esta visão geral abrangente estabeleça o potencial teórico das GNNs, ela também aponta para desafios na escalabilidade e a necessidade de arquiteturas especializadas para capturar a natureza dinâmica das redes do mundo real [Wu et al. 2021].

As limitações dos simuladores de rede tradicionais e da análise matemática têm sido abordadas por meio da proposição de GNNs para a modelagem de redes [Farreras et al. 2023]. Esse estudo enfatiza a capacidade das GNNs de capturar relacionamentos complexos em dados estruturados como grafos, alcançando alta precisão com baixa sobrecarga computacional. No entanto, uma lacuna significativa identificada é a limitação na capacidade de generalização para topologias de rede maiores e não vistas, um desafio que permanece como ponto focal para pesquisadores que buscam implantar esses modelos em ambientes de produção.

No contexto de Redes Definidas por Software (SDN), são aplicadas Redes Convolucionais de Grafos Espaço-Temporais (STGCN) para prever o atraso de pacotes ponta a ponta [Ge et al. 2022]. Ao formular topologias de rede como grafos e utilizar o conjunto de dados da rede Abilene, o estudo demonstrou que as STGCNs superam métodos tradicionais de aprendizado de máquina, como Random Forest, ao capturar as dimensões espacial e temporal. De forma semelhante, foi proposta uma STGCN baseada em aprendizado em grafos para previsão de tráfego, introduzindo matrizes de adjacência dinâmicas com o objetivo de superar as limitações das estruturas de grafo fixas [Hu et al. 2022]. Embora eficazes para o fluxo de tráfego, esses modelos frequentemente assumem topologias estáticas ou predefinidas, o que pode não refletir plenamente as mudanças dinâmicas de roteamento em infraestruturas de longa distância, como a RNP.

GNNs têm sido exploradas para a previsão do estágio de pré-roteamento de circuitos integrados [Lin et al. 2025]. Embora aplicadas a um domínio distinto, estratégias de aprendizado multitarefa para prever tanto o atraso da rede quanto o tempo de chegada (AT) oferecem indicadores relevantes sobre como lidar com modelos de atraso complexos com alta correlação. No entanto, a aplicação de tais técnicas a redes de telecomunicações em larga escala requer adaptações para lidar com o desequilíbrio de dados inerente e com a necessidade de integrar informações topológicas provenientes de medições ativas, como o *traceroute*.

Por fim, as abordagens baseadas em correlação ponderada pelo tempo também foram propostas para a identificação de links de alto atraso em Provedores de Serviços de Internet [Nobre et al. 2025]. Esses métodos isolam efetivamente os enlaces responsáveis pela degradação do desempenho ao correlacionar dados de *traceroute* com métricas de atraso. Apesar de fornecerem indicadores retrospectivos detalhados sobre o comportamento da rede, tais abordagens permanecem fundamentalmente reativas.

Tabela 1. Comparação entre trabalhos relacionados

Trabalho	**Abordagem**	**Principais Limitações**
Farreras et al. [Farreras et al. 2023]	GNN para modelagem de atraso	Baixa generalização para topologias não vistas
Ge et al. [Ge et al. 2022]	STGCN para atraso em SDN	Suposição de topologia estática
Hu et al. [Hu et al. 2022]	STGCN com adjacência dinâmica	Foco em tráfego, não em atraso
Lin et al. [Lin et al. 2025]	GNN multitarefa	Domínio distinto e sem tratamento de desbalanceamento
Nobre et al. [Nobre et al. 2025]	Correlação temporal com *traceroute*	Abordagem reativa
Este trabalho	GraphSAGE em grafo dinâmico	Predição proativa com adaptação topológica

A análise da literatura revela um conjunto de limitações recorrentes que dificultam a aplicação prática dos modelos atuais em ambientes operacionais de larga escala. Especificamente, as abordagens existentes frequentemente enfrentam desafios relacionados à escalabilidade e à adaptação a arquiteturas dinâmicas especializadas [Wu et al. 2021], carecem de capacidade de generalização para topologias maiores e não previamente observadas [Farreras et al. 2023] ou dependem de estruturas de grafo fixas e suposições estáticas que não refletem as mudanças de roteamento em cenários reais [Ge et al. 2022, Hu et al. 2022]. Além disso, técnicas oriundas de outros domínios falham em lidar adequadamente com o acentuado desequilíbrio de dados inerente às anomalias de rede [Lin et al. 2025], enquanto métodos baseados em correlação permanecem estritamente reativos, identificando problemas apenas após sua ocorrência [Nobre et al. 2025].

Nossa proposta aborda essas lacunas ao migrar de abordagens reativas de detecção e análise de degradações para uma estratégia de predição de degradações futuras, que modela pares origem–destino como nós em um grafo dinâmico. Diferente de trabalhos anteriores, nossa solução integra explicitamente correlações de atraso temporal com similaridade topológica derivada de medições ativas de *traceroute*, permitindo que o modelo se adapte a instabilidades de roteamento. Ao alavancar as arquiteturas de GNN, combinadas com otimização bayesiana de hiperparâmetros e estratégias robustas de balanceamento, supera-se as limitações de desequilíbrio de dados e generalização topológica, fornecendo uma ferramenta escalável e confiável para antecipar degradações em infraestruturas complexas como a RNP.

3. Proposta

O presente trabalho propõe uma solução para a previsão proativa de eventos de alta latência usando GNNs. As GNNs demonstraram forte eficácia na modelagem de sistemas complexos com dependências estruturais explícitas [Wu et al. 2021]. A metodologia integra informações de atraso temporal e características topológicas derivadas de dados de *traceroute*, permitindo a construção de um sistema capaz de antecipar, com boa antecedência, os pontos de comunicação na rede que são mais propensos a experimentar degradação futura de desempenho.

A modelagem proposta representa cada par de comunicação entre estados brasilei-

ros como um nó no grafo, que são derivados de arquivos contendo dados de medição para cada par, combinando latência e informações de *traceroute*. As arestas são estabelecidas a partir de uma combinação ponderada entre a correlação temporal do atraso e a similaridade topológica das rotas de endereço IP, controlada por um hiperparâmetro ajustável que permite balancear a influência relativa entre as dependências temporais e estruturais da rede. Essa estratégia permite capturar dependências espaciais entre pontos de comunicação que compartilham caminhos, roteadores intermediários ou tendências análogas de congestionamento. Essa formulação é inspirada em estudos que demonstram que os pontos de comunicação de rede podem exibir forte correlação temporal devido ao compartilhamento de infraestrutura e caminhos comuns, mesmo na ausência de conectividade direta explícita [Nobre et al. 2025]. O pipeline de execução geral é ilustrado na Figura 1.

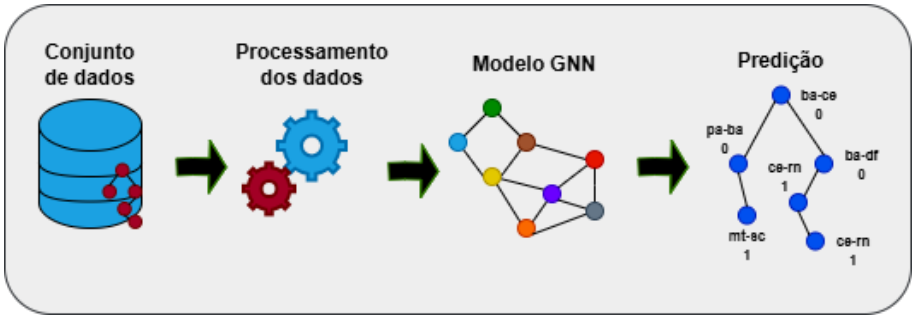

Figura 1. Visão geral da Solução.

3.1. Conjunto de Dados

O conjunto de dados utilizado neste trabalho foi construído a partir da integração de medições de atraso e informações de rota, coletadas ao longo de seis meses em 2023 por meio da API PerfSONAR[3] da RNP. O processo de preparação do conjunto de dados é composto por duas etapas principais: (i) a organização, associação e enriquecimento das medições de atraso e *traceroute* em um conjunto de dados unificado e (ii) a transformação desse conjunto de dados em uma sequência temporal de grafos, que serve como entrada para o modelo de GNN.

As medições de atraso foram coletadas em intervalos de um minuto, enquanto as medições de *traceroute* foram realizadas a cada dez minutos. Para cada par origem-destino monitorado, medições com carimbos de tempo próximos foram associadas, permitindo a correlação entre o comportamento temporal do atraso e as características topológicas do caminho de rede. Após essa etapa de associação e engenharia de atributos, foram gerados aproximadamente 650 arquivos por par de comunicação. Cada registro contém informações temporais, métricas de atraso fim a fim, número total de saltos, identificação dos primeiros e últimos nós do caminho, além do conjunto de endereços IPs e nomes de *host* intermediários observados ao longo da rota.

A partir desse conjunto de dados atraso–*traceroute* consolidado, procede-se à construção do conjunto de dados em formato de grafos temporais. Inicialmente, os arquivos correspondentes a cada par de comunicação são carregados e a frequência das medições é inferida automaticamente, utilizando-se a mediana dos intervalos temporais

[3]www.perfsonar.net

observados. Em seguida, aplica-se uma reamostragem uniforme para alinhar todas as séries temporais. Valores ausentes são tratados por interpolação linear, complementada por preenchimento para frente e para trás, garantindo continuidade temporal e consistência entre os diferentes pontos monitorados.

Para a definição dos rótulos, é calculado um limiar global de atraso τ, obtido a partir do percentil p da distribuição empírica de todos os valores de atraso observados no conjunto de dados. Formalmente, esse limiar é dado por

$$\tau = Q_p(\mathcal{D}), \tag{1}$$

em que $Q_p(\cdot)$ representa o p-ésimo operador de percentil e $\mathcal{D}$ o conjunto de todas as medições de atraso. Cada nó do grafo recebe um rótulo binário com base no valor futuro do atraso em um horizonte de predição: atrasos acima de τ são classificados como eventos de alto atraso, enquanto valores abaixo desse limiar são classificados como baixo atraso. O percentil p é tratado como um hiperparâmetro, sendo avaliados valores de 75, 80 e 85, com melhor desempenho observado para o percentil 85.

As séries temporais individuais são então integradas em um único quadro temporal global, no qual cada instante contém as medições de atraso, informações de rota e os respectivos rótulos de todos os pares de comunicação. A partir desse quadro, janelas temporais deslizantes são geradas para representar o histórico recente da rede, logo, cada janela temporal define um conjunto de estados do grafo, no qual os nós correspondem aos pares de comunicação e os atributos são extraídos da respectiva janela, sendo o rótulo associado a um horizonte de predição futuro. Essa estratégia é amplamente utilizada em tarefas de predição espaço-temporal. Para cada nó do grafo, são extraídas nove características a partir da janela temporal: cinco relacionadas ao comportamento do atraso (média, desvio padrão, valor máximo, valor mínimo e último valor observado) e quatro relacionadas às propriedades topológicas do caminho, derivadas do *traceroute* (média, mínimo, máximo e desvio padrão do número de saltos). Essa representação conjunta permite capturar simultaneamente a dinâmica temporal do atraso e aspectos estruturais e de estabilidade das rotas de rede.

A conectividade do grafo é definida de forma orientada a dados. As arestas são estabelecidas com base em uma métrica composta que combina a correlação temporal de Spearman entre as séries de atraso e a similaridade topológica entre os caminhos de rede, estimada a partir da sobreposição de rotas e da similaridade no número de saltos. Apenas pares de comunicação cuja similaridade excede um limiar mínimo são conectados, resultando em grafos esparsos que preservam relações relevantes. Os pesos das arestas refletem a intensidade dessas relações e influenciam o processo de agregação de informações durante o treinamento do modelo.

Cada amostra final é representada como um objeto *Data* da biblioteca PyTorch Geometric[4], contendo o tensor de atributos dos nós, a estrutura de arestas com seus respectivos pesos, os rótulos binários futuros e o carimbo de tempo associado. O conjunto de dados resultante consiste, portanto, em uma sequência temporal de grafos, em que cada grafo representa um estado instantâneo da rede, sendo utilizado como entrada para o treinamento e avaliação do modelo GNN.

[4]https://pytorch-geometric.readthedocs.io/en/latest/

3.2. Processamento dos Dados

O processamento de dados inclui as etapas de organização, particionamento temporal e balanceamento de amostras, a fim de garantir uma avaliação realista do modelo e mitigar os efeitos do desequilíbrio inerente aos dados da rede de comunicação. O particionamento do conjunto de dados segue uma estratégia estritamente temporal, preservando a ordem cronológica das amostras. Os dados mais antigos são usados para treinamento, enquanto as janelas temporais subsequentes são reservadas para validação e teste. Essa abordagem reflete um cenário de implantação realista, no qual o modelo é treinado com dados históricos e avaliado quanto à sua capacidade de generalizar para períodos futuros não vistos, evitando o vazamento de informações temporais que ocorreria em divisões aleatórias.

A primeira estratégia para balanceamento dos dados consiste em atribuir pesos diferenciados às classes na função de perda, calculados inversamente proporcionais à frequência das classes no conjunto de treinamento. Essa técnica aumenta a penalidade para erros associados à classe minoritária. Além disso, um amostrador ponderado é empregado para ajustar a probabilidade de seleção de cada grafo durante a formação dos lotes de treinamento, com base na proporção de nós rotulados como de alta latência em cada janela de tempo. Essa estratégia aumenta a exposição do modelo a exemplos mais informativos para prever eventos críticos. Além disso, a suavização de rótulos é aplicada para melhorar a calibração do modelo, visando probabilidades mais realistas e reduzindo o sobreajuste. Por fim, uma terceira abordagem é proposta, combinando simultaneamente o uso de pesos na função de perda e a amostragem ponderada, intensificando o foco do aprendizado em padrões associados à classe minoritária.

Os resultados experimentais apresentados na Seção 4 demonstram que as métricas de desempenho permanecem estáveis, indicando que a remoção dessas amostras não implica em perda de informação relevante, mas sim em uma redução do viés e um processo de aprendizado mais focado em padrões associados a atrasos elevados.

A escolha das estratégias de balanceamento é integrada ao processo de treinamento, sendo definida automaticamente durante a otimização de hiperparâmetros. O modelo é treinado diretamente no conjunto de dados de grafos gerado com amostragem ponderada, pesos na função de perda e suavização de rótulos, a fim de maximizar o desempenho e a generalização de acordo com as características do conjunto de dados.

3.3. Treinamento do Modelo

O treinamento do modelo é realizado utilizando o conjunto de dados em formato de grafos, gerado a partir das janelas temporais descritas na seção anterior, adotando-se uma estratégia de particionamento estritamente temporal. As amostras mais antigas compõem o conjunto de treinamento (70%), seguidas pelos conjuntos de validação (15%) e teste (15%). Essa estratégia assegura um cenário realista, no qual o modelo é treinado com dados históricos e avaliado quanto à sua capacidade de generalização para períodos futuros não observados.

A arquitetura do modelo é baseada em Redes Neurais de Grafos, sendo avaliadas diferentes variantes amplamente utilizadas na literatura, como GCN, GAT e GraphSAGE [Kipf and Welling 2017, Wu et al. 2021, Farreras et al. 2023]. Cada variante define um mecanismo distinto de agregação de informações entre nós vizinhos, permitindo analisar diferentes formas de propagação espacial do atraso na topologia da rede. O treinamento

ocorre ao longo de múltiplas épocas, com avaliação contínua no conjunto de validação e aplicação de *early stopping* com base no F1-Macro, evitando sobreajuste e reduzindo o custo computacional.

Os grafos são processados em *batches* utilizando o mecanismo nativo do PyTorch Geometric, que agrupa múltiplos grafos sem misturar suas conectividades. Cada grafo representa um instante temporal da rede, e cada nó corresponde a um par origem-destino, descrito por características estatísticas de atraso e informações de traceroute. Essa organização permite a predição paralela de múltiplos estados futuros da rede, sendo a predição realizada em nível de nó, isto é, individualmente para cada enlace lógico da rede.

A otimização do modelo ocorre pela minimização da função de perda durante o treinamento. O erro global é obtido pela média da perda no *batch*, e a retropropagação é utilizada para atualizar os parâmetros treináveis da rede. O otimizador *Adam* é empregado, com taxa de aprendizado e regularização definidas automaticamente durante o processo de otimização de hiperparâmetros. Além disso, um escalonador *ReduceLROnPlateau* ajusta dinamicamente a taxa de aprendizado quando não há melhoria no F1-Macro de validação, promovendo convergência mais estável.

3.4. Predição do Modelo

A predição é realizada aplicando o modelo treinado a janelas temporais recentes, operando em modo de avaliação para garantir comportamento determinístico. Os grafos de entrada mantêm a mesma estrutura e conjunto de características utilizados no treinamento. Para cada nó, o modelo produz probabilidades associadas às classes, sendo selecionada aquela de maior valor como predição final. As métricas de desempenho e os custos computacionais associados a esse processo são apresentados e discutidos na seção de avaliação experimental.

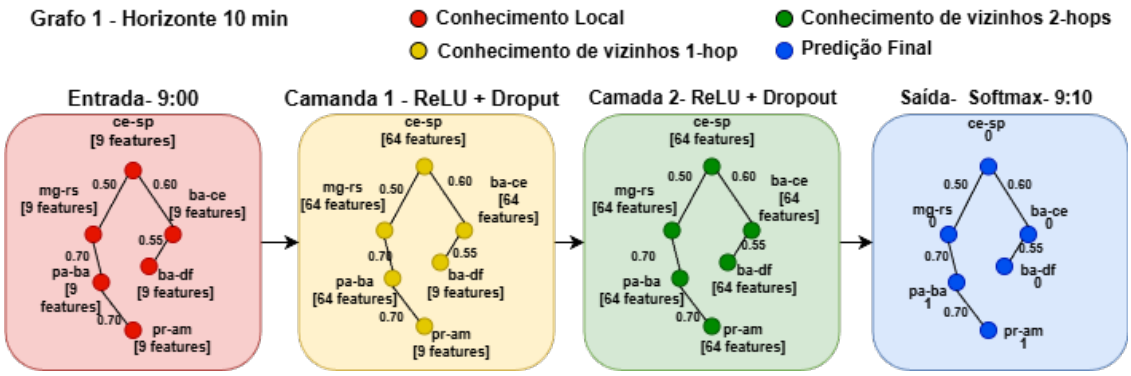

Figura 2. Predição do Modelo

A Figura 2 ilustra o pipeline de predição do modelo GNN para um horizonte de 10 minutos. Inicialmente, as características de cada nó representam apenas o conhecimento local do enlace. Na primeira camada convolucional, esse conhecimento é enriquecido com informações dos vizinhos de um salto (*1-hop*), enquanto a segunda camada expande o campo receptivo para vizinhos de até dois saltos (*2-hops*). Esse processo permite que cada nó incorpore informações topologicamente correlacionadas, capturando dependências espaciais relevantes para a dinâmica do atraso na rede. Após cada camada convolucional, são aplicadas funções de ativação *ReLU* e *dropout*, introduzindo não linearidade

e regularização. As representações finais dos nós são então projetadas por uma camada classificadora linear, gerando logits associados às classes de atraso baixo e alto. A aplicação da função *softmax* transforma esses valores em probabilidades, resultando na predição final do estado futuro de cada enlace no horizonte considerado.

4. Experimentos

Esta seção descreve o protocolo experimental adotado para a avaliação do modelo proposto, com foco na robustez estatística das métricas de classificação e na eficiência computacional. Visando garantir a reprodutibilidade científica, o código-fonte da solução está publicamente disponível em um repositório no GitHub[5].

4.1. Ambiente Computacional

Os experimentos foram conduzidos utilizando dados do serviço de monitoramento MonIPÊ[6] da RNP, que segue o padrão internacional perfSONAR. As medições de atraso são coletadas a cada minuto, enquanto os dados de *traceroute* são obtidos a cada dez minutos e as medições de taxa de transferência a cada quatro horas. A infraestrutura da RNP abrange os 26 estados brasileiros e o Distrito Federal, proporcionando diversidade geográfica e reforçando a validade externa do estudo. A Figura 3 apresenta uma visão geral da topologia da rede, incluindo nós, capacidade dos enlaces e distribuição geográfica.

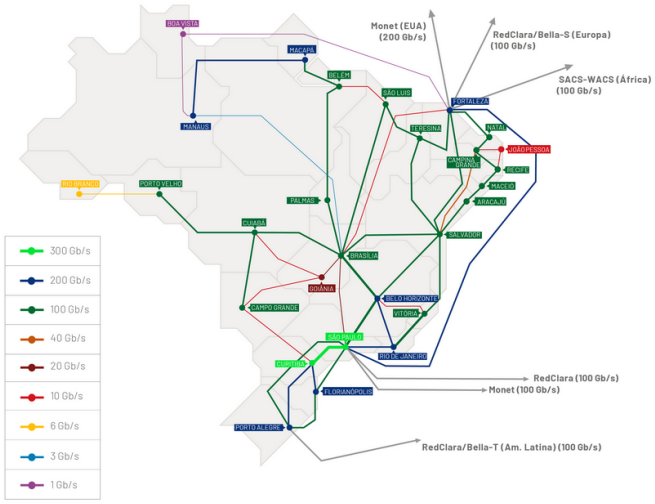

Figura 3. Visão geral da infraestrutura de rede RNP.

A avaliação experimental foi realizada em uma estação de trabalho equipada com processador Intel Core i9-14900, GPU NVIDIA RTX A4000 Ada com 20 GB de VRAM e 64 GB de RAM, utilizando os frameworks PyTorch e PyTorch Geometric. O conjunto de dados foi particionado de forma estritamente temporal, com 70% para treinamento, 15% para validação e 15% para teste, respeitando a ordem cronológica das amostras.

[5]`https://github.com/LarcesUece/SBRC-2026-GNN-Predict`
[6]https://monipe-central.rnp.br/

4.2. Aprendizado Profundo e Métricas de Avaliação

O modelo proposto baseia-se em GNNs, treinadas para a predição binária de eventos de alto atraso a partir de janelas temporais recentes de medições de atraso. Durante a fase de inferência, o modelo opera em modo de avaliação, assegurando comportamento determinístico e utilizando grafos com a mesma estrutura e conjunto de atributos empregados no treinamento. As probabilidades de saída são estimadas individualmente para cada nó, sendo a classe associada à maior probabilidade selecionada como predição final.

A definição da arquitetura e dos parâmetros de treinamento foi orientada por um processo de otimização bayesiana de hiperparâmetros, conduzido por meio do *framework* Optuna. O processo explorou diferentes configurações arquiteturais e de treinamento, tendo como critério de seleção a maximização da métrica F1-Macro no conjunto de validação, de modo a garantir desempenho equilibrado entre as classes. Adicionalmente, foram avaliadas estratégias específicas para lidar com o desbalanceamento inerente aos dados de atraso em redes, incluindo ponderação de classes na função de perda, amostragem balanceada em lote e suavização de rótulos, visando aumentar a sensibilidade do modelo a eventos raros de alto atraso sem comprometer a estabilidade do treinamento.

A avaliação do desempenho do modelo prioriza métricas robustas ao desbalanceamento de classes, destacando-se a F1-Macro e a Acurácia Balanceada, que atribuem igual importância a eventos normais e anômalos. A métrica F1-Ponderado é utilizada como indicador complementar do desempenho global, refletindo a contribuição da classe majoritária. Além disso, o *Brier Score* é empregado para avaliar a calibração probabilística das predições, permitindo verificar a confiabilidade das probabilidades estimadas pelo modelo, aspecto fundamental em cenários de monitoramento proativo e tomada de decisão baseada em limiares de risco.

5. Resultados

A Tabela 2 apresenta o desempenho preditivo do modelo proposto para diferentes horizontes de predição, variando de 3 a 30 minutos. Observa-se que o modelo mantém valores estáveis de F1-Macro e F1-Ponderado, com degradação limitada à medida que o horizonte de predição aumenta, evidenciando robustez frente à redução da correlação temporal.

Tabela 2. Métricas de previsão do modelo por horizonte

Horizonte	F1-Ponderado	F1-Macro	Acurácia_Balanceada	Brier Score
3 min	93.57	88.80	92.37	0.0688
10 min	93.47	88.56	91.65	0.0679
30 min	92.93	86.47	83.03	0.0076

Em previsões de curto prazo (3 minutos), o modelo atinge elevado desempenho, refletindo sua capacidade de discriminar eventos normais e de alto atraso. À medida que o horizonte se estende para 10 e 30 minutos, nota-se uma redução gradual nas métricas, comportamento esperado em cenários de maior incerteza temporal. Ainda assim, os valores de *Brier Score* indicam boa calibração probabilística, especialmente no horizonte de 30 minutos.

A comparação com um modelo LSTM, apresentada na Tabela 3, evidencia diferenças claras entre abordagens puramente temporais e modelos sensíveis à topologia. En-

quanto o LSTM apresenta desempenho superior em horizontes muito curtos, seu desempenho degrada de forma mais acentuada à medida que o horizonte de predição aumenta. Em contraste, o modelo baseado em GNN demonstra maior estabilidade, mantendo desempenho competitivo em horizontes médios e longos.

Tabela 3. Comparação das métricas de desempenho entre LSTM e GNN em diferentes horizontes de previsão.

Horizonte	Modelo	F1-Pondeado	F1-Macrro	Acurácia_Balanceada	Brier Score
3 min	LSTM	99.97	99.62	99.87	0.0013
	GNN	93.57	88.80	92.37	0.0571
10 min	LSTM	92.73	87.61	92.86	0.0550
	GNN	93.47	88.56	91.65	0.0571
30 min	LSTM	90.94	84.80	90.83	0.0684
	GNN	92.93	86.47	83.03	0.0076

Tabela 4. Porcentagem de degradação da F1-Macro

Modelo	3 → 10 min	3 → 30 min	Estabilidade
LSTM	-12.01 p.p.	-14.82 p.p.	Alto
GNN	-0.24 p.p.	-2.33 p.p.	**Baixo**

Essa tendência é quantificada na Tabela 4, que apresenta a degradação percentual da métrica F1-Macro em relação ao horizonte de 3 minutos. Os resultados evidenciam que o modelo baseado em grafos apresenta uma redução substancialmente inferior de desempenho quando comparado ao LSTM, o que sugere que a modelagem explícita de relações espaciais atua como um mecanismo de regularização implícita diante do aumento da incerteza temporal. Nesse contexto, a noção de estabilidade está associada à capacidade do modelo de sustentar seu desempenho ao longo de diferentes horizontes temporais, sendo operacionalizada pela magnitude da degradação observada, ou seja quanto menor a variação negativa da F1-Macro, maior a estabilidade do modelo. Assim, os resultados indicam que o modelo em grafos apresenta maior robustez temporal em relação ao LSTM.

Os resultados confirmam que a integração entre informações temporais de atraso e características topológicas extraídas de dados de *traceroute* é determinante para o desempenho do modelo. Configurações que incorporam apenas correlação estatística superam o modelo puramente temporal, enquanto a inclusão explícita da topologia física da rede resulta nos melhores desempenhos observados.

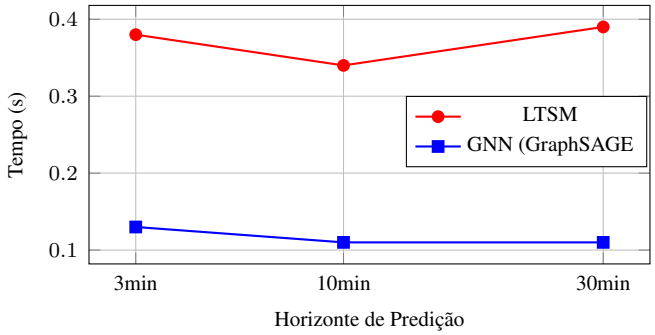

Figura 4. Tempo de predição por ponto de comunicação.

Por fim, a eficiência computacional dos modelos é analisada na Figura 4, que apresenta o tempo médio de predição para diferentes horizontes de predição. Observa-se que o modelo baseado em GNN apresenta tempos de inferência consistentemente inferiores aos da abordagem baseada em LSTM, mantendo-se na ordem de centésimos de segundo para todas as janelas avaliadas.

Além disso, o tempo de predição da GNN mostra-se estável mesmo com o aumento do horizonte temporal, enquanto a LSTM apresenta maior variação e custo computacional. Esses resultados evidenciam que a formulação baseada em grafos alia eficiência computacional à robustez preditiva, sendo plenamente compatível com os requisitos de monitoramento contínuo e predição proativa de degradação de desempenho em redes de comunicação reais.

6. Conclusão

Este trabalho apresentou uma abordagem para a predição proativa de eventos de alta latência na RNP utilizando Redes Neurais Gráficas (GNNs). A principal contribuição reside na integração de métricas temporais de atraso com informações topológicas extraídas de traceroute, permitindo capturar dependências espaciais que métodos tradicionais de séries temporais não modelam adequadamente. A representação de pares de comunicação como nós e a definição de arestas com base em correlação temporal e similaridade estrutural mostraram-se eficazes na identificação de gargalos em larga escala, enquanto estratégias de balanceamento foram essenciais para lidar com o desequilíbrio de dados.

Os resultados demonstram que a arquitetura GNN, com agregação de até dois saltos, produz predições probabilísticas confiáveis e maior robustez em horizontes de predição crescentes quando comparada a uma linha de base LSTM temporal, evidenciando a importância da incorporação da topologia da rede para monitoramento proativo. Como trabalhos futuros, destacam-se o uso de GNNs dinâmicas, a extensão para predição multiclasse ou regressão de atrasos e a investigação de estratégias de escalabilidade para redes de maior porte.

Agradecimentos

Pesquisa parcialmente financiada pelo CNPq (Processos 305946/2025-0 e 405940/2022-0) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 88887.954253/2024-00 e 88887.972043/2024-00.

Referências

Brito, M. L. L., Ferreira, M. C. M., Portela, A. L. C., and Gomes, R. L. (2026). Ai-based estimation of bandwidth availability for data offloading in edge-cloud computing. *IEEE Networking Letters*, 8:69–73.

Direito, R., Matos, D., Gomes, D., Gomes, D., and Aguiar, R. (2023). A monitoring system to measure the impact of a network application in a 5g network. In *Proceedings of the IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, pages 72–78.

Farreras, M., Soto, P., Camelo, M., Fàbrega, L., and Vilà, P. (2023). Improving network delay predictions using gnns. *Journal of Network and Systems Management*, 31(65):1–36.

Ferreira, M. C., Ribeiro, S. E., Nobre, F. V., Linhares, M. L., Araújo, T. P., and Gomes, R. L. (2024). Mitigating measurement failures in throughput performance forecasting. In *Proceedings of the 20th International Conference on Network and Service Management (CNSM)*, pages 1–7.

Ge, Z., Hou, J., and Nayak, A. (2022). Gnn-based end-to-end delay prediction in software defined networking. In *International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 372–378.

Gomes, R. L., Bittencourt, L. F., and Madeira, E. R. M. (2014). A similarity model for virtual networks negotiation. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC '14, page 489–494, New York, NY, USA. Association for Computing Machinery.

Hu, N., Zhang, D., Xie, K., Liang, W., and Hsieh, M.-Y. (2022). Graph learning-based spatial–temporal graph convolutional neural networks for traffic forecasting. *Connection Science*, 34(1):429–448.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Lin, Z., Zhang, H., Gao, P., Yu, F., Wu, T., Xiong, X., and Cai, S. (2025). Gnn-based timing prediction in prerouting stage with multitask learning strategy. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 44(8):3154–3164.

Mouchet, M., Vaton, S., Chonavel, T., Aben, E., and den Hertog, J. (2020). Large-scale characterization and segmentation of internet path delays with infinite hidden markov models. *IEEE Access*, 8:175000–175015.

Nobre, F. V. J., Silva, D. d. S., Ferreira, M. C. M. M., Brito, M. L. M. L., Araújo, T. P., and Gomes, R. L. (2025). Time-weighted correlation approach to identify high delay links in internet service providers. *Journal of Internet Services and Applications*, 16(1):419–430.

Nosyk, F. et al. (2025). A day in the life of ripe atlas: Operational insights and applications in network measurements. *arXiv preprint arXiv:2511.22474*.

Portela, A., Linhares, M. M., Nobre, F. V. J., Menezes, R., Mesquita, M., and Gomes, R. L. (2024a). The role of tcp congestion control in the throughput forecasting. In *Proceedings of the 13th Latin-American Symposium on Dependable and Secure Computing*, LADC '24, page 196–199, New York, NY, USA. Association for Computing Machinery.

Portela, A. L. C., Ribeiro, S. E. S. B., Menezes, R. A., Araújo, T. P. d., and Gomes, R. L. (2024b). T-for: An adaptable forecasting model for throughput performance. *IEEE Transactions on Network and Service Management*, 21(3):2791–2801.

Souza, M. S., Ribeiro, S. E. S. B., Lima, V. C., Cardoso, F. J., and Gomes, R. L. (2024). Combining regular expressions and machine learning for sql injection detection in urban computing. *Journal of Internet Services and Applications*, 15(1):103–111.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24.