Perturbação Controlada de Autovetores em PCA para Preservação de Privacidade em Dados Sensíveis

Ivo A. Pimenta[1], Kaynan S. Freitas[1], Evellin S. Moura [1], Erick S. Nascimento[1], Fabio A. Faria[2], Rafael L. Gomes[1]

[1]Universidade Estadual do Ceará (UECE)

{aguiar.pimenta, kaynan.freitas, evellin.moura,

erick.nascimento}@aluno.uece.br, rafa.lopes@uece.br

[2]Instituto Superior Técnico (IST/ULisboa)

fabio.faria@tecnico.ulisboa.pt

Abstract. *The growing adoption of digital solutions based on user data requires robust privacy mechanisms. Excessive perturbations can significantly degrade the performance of Machine Learning models. In this work, we propose PerturbPCA-α, an adjustable approach based on PCA. The technique introduces continuous perturbation to the principal eigenvectors of the dataset. This enables explicit control of the trade-off between privacy and utility. Consequently, the learning capability of ML models is preserved, as demonstrated by experiments on healthcare data that validate the effectiveness of the proposed method.*

Resumo. *A crescente adoção de soluções digitais baseadas em dados de usuários requer mecanismos robustos de privacidade. Perturbações excessivas podem degradar significativamente o desempenho de modelos de Aprendizado de Máquina. Neste trabalho, propomos o PerturbPCA-α, uma abordagem ajustável baseada em PCA. A técnica introduz perturbação contínua nos autovetores principais do conjunto de dados. Isso permite um controle explícito do trade-off entre privacidade e utilidade. Consequentemente, a capacidade de aprendizado dos modelos de ML é mantida, conforme demonstrado por experimentos com dados de saúde que validam a efetividade do método proposto.*

1. Introdução

Soluções digitais modernas se baseiam no paradigma da Internet das Coisas (*Internet of Things* - IoT) para coleta de dados do mundo real, onde uma ampla variedade de componentes físicos, incluindo unidades de sensoriamento, atuadores e dispositivos eletrônicos de uso cotidiano, é interconectada para possibilitar comunicação autônoma e compartilhamento de dados em sistemas do mundo real. Atualmente, aplicações de IoT são amplamente utilizadas em diversos setores, como cidades inteligentes, redes elétricas inteligentes, edifícios inteligentes, transporte inteligente, sistemas de monitoramento em tempo real, monitoramento ambiental, além de sistemas médicos e de saúde [Souza et al. 2024].

Um dos casos mais promissores para a sociedade é a aplicação de IoT no setor de saúde, o qual tem passado por uma transformação significativa impulsionada pelos

avanços em tecnologias da informação em saúde. A integração do paradigma IoT em ambientes médicos é chamada de Internet das Coisas Médicas (*Internet of Medical Things* - IoMT) [Kamalov et al. 2023], permitindo que dados clínicos sejam trocados de forma contínua entre uma ampla gama de dispositivos médicos interconectados. Esses dispositivos podem se comunicar pela mesma rede ou diretamente entre si, sem a necessidade de interação humano e máquina. Tal conectividade possibilita o monitoramento remoto e o diagnóstico de pacientes, ampliando a acessibilidade e a continuidade do cuidado em saúde [Boikanyo et al. 2023].

Este cenário oferece benefícios expressivos para pacientes, operadoras e profissionais de saúde ao viabilizar modelos de cuidado baseados em valor e proporcionar indicadores mais profundos sobre a condição clínica dos indivíduos. A pandemia da COVID-19 intensificou ainda mais a demanda por coleta remota de dados, impulsionando a adoção de soluções digitais nos ambientes médicos. Durante esse período, provedores de saúde passaram a depender fortemente da telemedicina para diagnósticos e consultas remotas, aumentando a demanda por tecnologias vestíveis, ferramentas de monitoramento remoto e plataformas de cuidado virtual. Esses sistemas permitem o monitoramento contínuo de sinais vitais, sintomas e outros indicadores de saúde, reduzindo visitas hospitalares e minimizando riscos de infecção [Ozcelik et al. 2025].

Os rápidos avanços tecnológicos na área de saúde também aceleraram a aplicação de técnicas de Inteligência Artificial (IA) neste setor, transformando os sistemas de saúde modernos em serviços essenciais para suporte aos profissionais e análise de dados [Razdan and Sharma 2022]. Esta transformação engloba ações como suporte à tomada de decisão, análise de exames, visualização de métricas sobre dados de saúde, dentre outras.

Entretanto, recentemente há um aumento na ocorrência de ataques e ações maliciosas sobre fontes de dados e serviços, resultando no vazamento de dados [Silveira et al. 2023]. Casos de vazamento de dados comprometem a confidencialidade organizacional e impactam diretamente o cumprimento das legislações de privacidade vigentes, tornando essencial a conformidade com regulações como o GDPR, na Europa, e a LGPD, no Brasil, a fim de evitar sanções legais e penalidades financeiras [Pimenta et al. 2024]. Considerando que os dados são um dos ativos mais valiosos para empresas e instituições públicas, sua proteção torna-se crítica [Pimenta et al. 2025]. Adicionalmente, marcos regulatórios como a HIPAA, nos Estados Unidos, e o próprio GDPR, na União Europeia, intensificaram a pressão por práticas seguras e confiáveis de tratamento de dados, especialmente em ambientes que lidam com informações pessoais sensíveis [Gong et al. 2022].

Essas preocupações são relevantes em ecossistemas que lidam com dados sensíveis, onde aspectos computacionais ainda precisam ser considerados, visto que há cenários caracterizados por capacidade de computação local limitada, restrições de comunicação e tarefas sensíveis à latência, o que motiva a necessidade de técnicas de preservação de privacidade que sejam leves, preservem a utilidade e sejam eficientes em termos de comunicação [Nobre et al. 2025]. Esse contexto acelera a implantação de tecnologias robustas que, por sua natureza, envolvem o processamento em larga escala de dados pessoais e clínicos, ampliando os riscos associados à privacidade.

Diante desses desafios, este trabalho apresenta o *PerturbPCA*-α, um método de anonimização baseado em Análise de Componentes Principais (*Principal Component*

Analysis - PCA), projetado para aplicações baseadas em Aprendizado de Máquina (Machine Learning - ML) que lidam com dados sensíveis. Diferentemente de abordagens existentes (tais como as referências [Silveira et al. 2023, Aleroud et al. 2016]), que dependem da permutação completa dos autovetores do PCA, a técnica proposta introduz um parâmetro contínuo de perturbação $\alpha \in [0, 1]$, permitindo um controle refinado sobre o equilíbrio entre nível de privacidade e utilidade dos dados. Ao combinar gradualmente os espaços de autovetores original e perturbado de acordo com o valor de α, o *PerturbPCA*-α preserva estruturas estatísticas relevantes enquanto possibilita níveis ajustáveis de anonimização. Essa formulação adaptativa aborda diretamente a tensão entre privacidade e utilidade em sistemas inteligentes, assegurando que dados sensíveis possam ser protegidos sem comprometer o desempenho de modelos de aprendizado de máquina utilizados em tarefas de monitoramento, diagnóstico e suporte à decisão.

A fim de avaliar a proposta, foram realizados experimentos utilizando conjuntos de dados reais de dados biológicos de aplicações IoMT, onde foram avaliados aspectos de desempenho de inferência dos modelos de ML e de garantia de privacidade contra ataques. Os resultados dos experimentos demonstram a capacidade da proposta de manter os níveis de acurácia dos modelos, enquanto mitigam o impacto de possíveis ataques de privacidade sobre os dados.

As principais contribuições deste trabalho são: (i) a proposição de um mecanismo de anonimização ajustável baseado em PCA, denominado *PerturbPCA*-α, que permite controlar de forma contínua o compromisso entre privacidade e utilidade; (ii) a introdução de uma extensão sensível a *clusters* para lidar com dados heterogêneos; e (iii) uma avaliação experimental abrangente que considera simultaneamente métricas de utilidade preditiva e resistência a ataques de inferência de membros, demonstrando a eficácia da abordagem em cenários reais em dados sensíveis.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve o método de anonimização proposto incluindo uma descrição detalhada de sua implementação e da extensão baseada em clusters; a Seção 4 apresenta os experimentos, contemplando a caracterização dos conjuntos de dados utilizados e o protocolo experimental; a Seção 5 discute os resultados das avaliações de utilidade preditiva e de resistência a ataques de inferência de membros; por fim, a Seção 6 conclui o artigo e aponta direções para trabalhos futuros.

2. Trabalhos Relacionados

Esta seção apresenta e discute os principais trabalhos relacionados encontrados na literatura, abrangendo desde técnicas de anonimização para privacidade até métodos de geração de dados sintéticos, bem como proteção de bases de dados sensíveis, com foco especial em soluções que buscam equilibrar privacidade e utilidade.

Aleroud et al. [Aleroud et al. 2016] propõem uma técnica de anonimização que preserva prefixos combinada com um modelo baseado em condensação para gerar traços de rede sintéticos, porém estatisticamente coerentes. O método separa a anonimização de prefixos IP da estrutura em nível de *host* e aplica clusterização para preservar padrões comportamentais locais, ao mesmo tempo em que reduz riscos de reidentificação. Os resultados mostram melhor compromisso entre privacidade e utilidade em comparação com métodos tradicionais e demonstram que o processamento por clusters melhora o

desempenho de sistemas de detecção de intrusão, um princípio relevante para *datasets* biomédicos heterogêneos.

Thabit et al. [Thabit et al. 2021] apresentam um modelo de criptografia em duas camadas que integra operações lógicas e matemáticas com mecanismos inspirados na biologia molecular, produzindo textos cifrados com complexidade ajustável e resiliência em múltiplas camadas. Ao empregar transformações análogas aos processos de transcrição e tradução, o esquema aumenta a robustez contra ataques de força bruta e inferência estatística. Embora tenha sido projetado para confidencialidade criptográfica e não para anonimização, a ênfase em transformações parametrizadas e adaptativas informa diretamente estratégias de perturbação ajustável

Silveira et al. [Silveira et al. 2023] avançam a proteção de bases de dados ao combinar criptografia simétrica pesquisável com anonimização baseada em autovetores por meio do mecanismo PPM-Anon. O método permuta os autovetores da PCA para preservar propriedades estatísticas globais enquanto mitiga riscos de divulgação, e incorpora uma etapa de clusterização que permite a anonimização independente de subgrupos coerentes. Experimentos com cargas de trabalho reais demonstram que essa arquitetura híbrida mantém desempenho prático e reduz a perda de precisão.

Hyrup et al. [Hyrup et al. 2025] apresentam uma revisão sistemática sobre técnicas de preservação de privacidade para dados tabulares sintéticos na área da saúde. Eles analisam métodos que variam de modelos generativos a técnicas estatísticas de anonimização e destacam a persistente falta de padronização nas métricas de avaliação de privacidade e utilidade. Os autores relatam que, embora muitas técnicas de geração de dados sintéticos aleguem preservar a privacidade, há grande variação na forma como o risco de privacidade e a utilidade analítica são avaliados, dificultando comparações entre estudos. Essa incerteza reforça a necessidade de mecanismos de anonimização que permitam controle explícito da intensidade de distorção, ao mesmo tempo em que preservem a fidelidade dos dados em contextos sensíveis de saúde.

Considerando a literatura analisada, desde a anonimização de informações até a geração de dados sintéticos, torna-se evidente que a maioria das soluções existentes sacrifica a privacidade ou a fidelidade dos dados. Nossa proposta se diferencia ao combinar perturbação controlável no espaço latente via PCA com uma estratégia opcional de anonimização sensível a clusters. Esse desenho permite ajustar a intensidade da anonimização, preserva a estrutura estatística local e viabiliza análises downstream significativas, atendendo às lacunas de avaliação identificadas e oferecendo uma solução flexível de preservação de privacidade para dados sensíveis.

3. Proposta

O PerturbPCA-α, proposto neste trabalho, equilibra a proteção de dados e o desempenho de modelos de ML ao aplicar perturbações em autovetores via PCA. Ao ajustar o parâmetro α para transitar entre dados originais e aleatórios, a técnica retém padrões essenciais para o aprendizado. Experimentos com dados sensíveis comprovam a eficácia do método em sustentar a precisão dos modelos enquanto reduz significativamente riscos de ataques à privacidade, mostrando-se ideal para aplicações seguras.

Essa etapa inicial de processamento prepara os dados para a anonimização, abordando questões como diferenças de escala, ruído e inconsistências estruturais que são

comuns em conjuntos de dados sensíveis. Uma vez processados, os dados são encaminhados para a pipeline de anonimização, que opera antes de qualquer estágio analítico ou de aprendizado. Esse design garante que informações sensíveis sejam protegidas desde o início, reduzindo o risco de vazamento de dados durante o armazenamento, transmissão ou análise.

O mecanismo de anonimização pode ser aplicado globalmente a todo o conjunto de dados ou localmente por meio de uma estratégia de agrupamento (*clustering*). Essa flexibilidade permite que a estrutura se adapte a diferentes distribuições de dados, requisitos de privacidade e restrições computacionais comumente encontradas em cenários de aplicações que lidam com dados sensíveis.

3.1. Mecanismo de Anonimização PerturbPCA-α

A Figura 1 apresenta a visão geral do mecanismo de anonimização *PerturbPCA*-α, concebido para operar no espaço de representação latente dos dados em vez de modificar diretamente os atributos observados. Com a rápida expansão dos sistemas modernos, dados sensíveis são cada vez mais coletados, transmitidos e analisados em ambientes distribuídos, tornando a preservação da privacidade uma exigência regulatória e técnica. Entretanto, métodos tradicionais de anonimização frequentemente introduzem distorções tão severas que comprometem o desempenho de modelos de aprendizado de máquina subsequentes [Pimenta et al. 2024]. Observações semelhantes foram relatadas em estudos de anonimização de rastros, nos quais a remoção excessiva de informação estrutural inviabiliza análises comportamentais e de detecção [Aleroud et al. 2016].

Técnicas baseadas em PCA têm sido exploradas para preservar propriedades estatísticas globais dos dados, mas abordagens existentes normalmente dependem da permutação completa dos autovetores, sem oferecer um mecanismo gradual de controle da distorção [Silveira et al. 2023]. Em contraste, trabalhos recentes enfatizam a importância de estratégias parametrizadas que permitam calibrar continuamente o nível de perturbação [Thabit et al. 2021]. Nesse contexto, o *PerturbPCA*-α foi projetado para introduzir uma perturbação estrutural controlada na base de autovetores do PCA, possibilitando uma transição suave entre preservação de utilidade e proteção de privacidade por meio de um parâmetro $\alpha \in [0, 1]$.

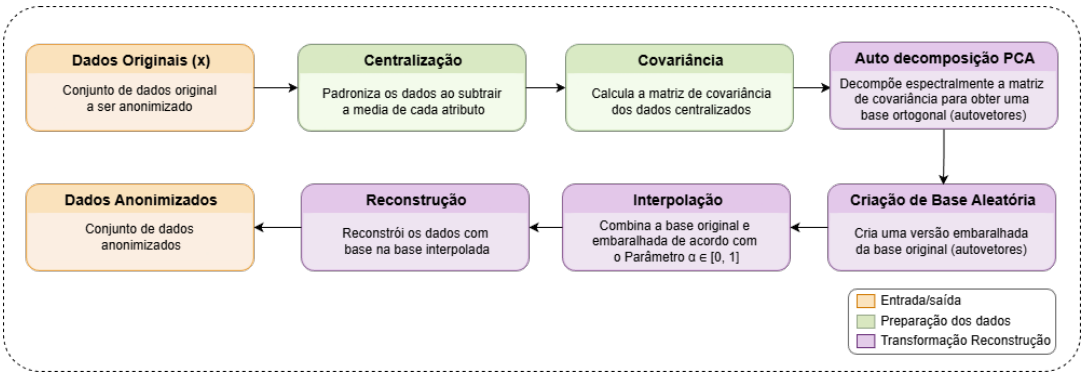

Figura 1. Visão geral do design de anonimização PerturbPCA-α.

O fluxo de processamento inicia-se com a centralização dos dados de entrada, na qual se subtrai a média de cada atributo. Em seguida, calcula-se a matriz de co-

variância dos dados centralizados e realiza-se a decomposição espectral para obter autovalores e autovetores ordenados em função da variância explicada. Esses autovetores definem uma base ortogonal que estrutura o espaço latente do PCA, permitindo manipular a representação dos dados de maneira organizada sem perturbar diretamente os valores brutos das características.

Conforme ilustrado na Figura 1, os dados centralizados são projetados nessa base PCA, gerando uma representação latente estatisticamente independente. Paralelamente, constrói-se uma versão perturbada da base de autovetores por meio do embaralhamento aleatório dos elementos de cada autovetor. Esse procedimento rompe correspondências diretas entre atributos e direções principais de variância, reduzindo o risco de reconstrução de informações sensíveis. Importante ressaltar que essa etapa não adiciona ruído aos dados, mas altera a própria geometria do espaço de representação.

Em vez de substituir completamente a base original pela versão embaralhada, o método realiza uma interpolação convexa entre ambas, controlada por α. Quando $\alpha = 0$, obtém-se a base original (sem anonimização); quando $\alpha = 1$, obtém-se a base totalmente perturbada; valores intermediários produzem uma transição contínua entre esses extremos, conforme sugerido pelo fluxo apresentado na figura.

Por fim, os dados projetados no espaço PCA são reconstruídos no espaço original utilizando a base interpolada $\mathbf{V}_\alpha$, com posterior reposição da média removida na etapa inicial. Como toda a perturbação ocorre no domínio dos autovetores, preservam-se padrões globais de variância relevantes para aprendizado de máquina, enquanto relações individuais entre atributos são parcialmente obscurecidas. Diferentemente de mecanismos baseados em ruído aditivo.

3.2. Extensão Baseada em Clusters

Conforme ilustrado na Figura 2, o *PerturbPCA-*α pode ser aplicado tanto de forma global quanto por meio de uma estratégia baseada em agrupamentos (clusters). Nesta subseção detalhamos essa segunda alternativa, que é particularmente adequada para conjuntos de dados heterogêneos.

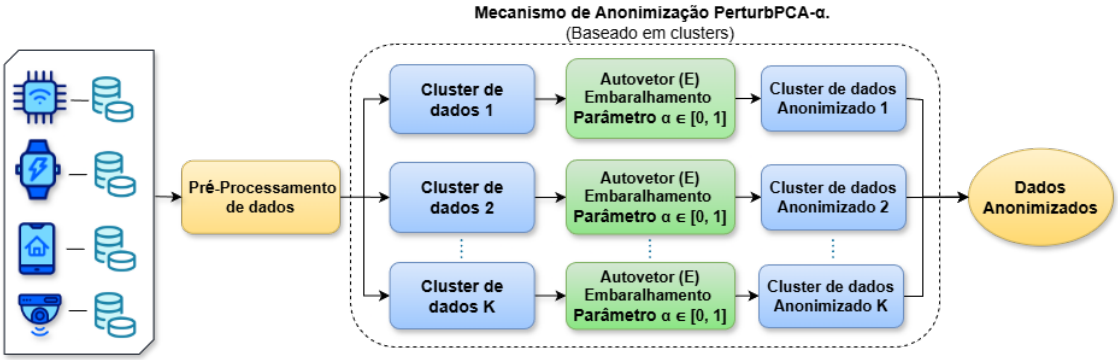

Figura 2. Visão geral do design de anonimização PerturbPCA-α.

Nos conjuntos de dados heterogêneos, as amostras frequentemente exibem padrões estatísticos ou contextos distintos que não podem ser adequadamente representados por uma única transformação global. Trabalhos anteriores combinados com agrupamento demonstram que agrupar registros estruturalmente semelhantes antes da

anonimização melhora a preservação das propriedades estatísticas e reduz a perda de precisão [Silveira et al. 2023]. A clusterização também se alinha ao princípio mais amplo de que a anonimização é mais eficaz quando as transformações respeitam as divisões naturais presentes nos dados, conforme observado em frameworks de anonimização existentes [Aleroud et al. 2016].

A extensão baseada em agrupamentos do *PerturbPCA*-α, representada esquematicamente na Figura 2, segue essa lógica, aplicando o mecanismo de anonimização de forma independente em cada grupo. O processo é composto por três estágios:

1. **Processamento dos Dados:** O conjunto de dados passa por uma fase inicial de pré-processamento para garantir a consistência entre as amostras. Essa etapa pode incluir normalização, tratamento de atributos ausentes e filtragem de ruído, possibilitando avaliações significativas de similaridade e evitando que os agrupamentos sejam moldados por artefatos em vez de estrutura intrínseca.
2. **Clusterização:** Os registros são agrupados de acordo com a similaridade dos atributos usando um método de clusterização baseado em distância, consistente com as estratégias utilizadas em sistemas anteriores de anonimização [Silveira et al. 2023]. Essa etapa garante que amostras com perfis estatísticos comparáveis sejam processadas juntas, permitindo que o mecanismo de anonimização se adapte à variabilidade local. A escolha do método de clusterização e da métrica de distância pode variar dependendo das características do conjunto de dados e do uso analítico pretendido.
3. **Anonimização por Cluster e Combinação:** Cada agrupamento é tratado como um subconjunto independente e passa pelo fluxo completo de anonimização. O PCA é calculado localmente para capturar os padrões dominantes de variância do grupo, após o qual a base de autovetores é perturbada de acordo com o nível de anonimização selecionado. O agrupamento é então reconstruído usando sua base perturbada, produzindo dados anonimizados que permanecem coerentes dentro do grupo, enquanto se beneficiam de transformações localmente otimizadas. Uma vez que todos os clusters tenham sido processados, suas saídas anonimizadas são recombinadas para formar o conjunto de dados final, com metadados ou rótulos preservados para manter a compatibilidade com tarefas subsequentes de aprendizado de máquina.

Essa extensão baseada em agrupamentos fortalece a privacidade ao gerar múltiplos espaços latentes perturbados independentemente, em vez de um único espaço global, reduzindo a suscetibilidade a ataques de inferência ou reconstrução. Ao mesmo tempo, melhora a utilidade ao garantir que o processo de anonimização seja sensível à estrutura local, evitando distorções que surgiriam ao aplicar transformações uniformes a dados biomédicos heterogêneos.

4. Experimentos

Durante os experimentos, avaliamos a eficácia do mecanismo de anonimização utilizando conjuntos de dados reais provenientes da literatura no contexto de dispositivos da Internet das Coisas Médicas (IoMT), considerando simultaneamente critérios de utilidade e privacidade. Diferentes formulações dos dados foram analisadas, incluindo cenários de

classificação binária e multiclasse, a fim de garantir uma avaliação abrangente sob distintas condições clínicas e estruturais.

A utilidade dos dados anonimizados foi avaliada por meio da comparação do desempenho de modelos de aprendizado de máquina treinados sobre os conjuntos de dados originais e sobre suas respectivas versões anonimizadas. Como métrica principal de desempenho, empregamos a acurácia balanceada, que oferece uma avaliação mais robusta em cenários com possível desequilíbrio entre classes, ao computar a média das taxas de acerto por classe e reduzir vieses associados à predominância de rótulos majoritários.

Adicionalmente, a preservação estatística dos dados após o processo de anonimização foi quantificada por meio da distância de Wasserstein. Essa métrica permite mensurar o deslocamento entre as distribuições dos dados originais e anonimizados, fornecendo uma interpretação geométrica da magnitude da perturbação introduzida pelo mecanismo proposto. Para uma análise mais robusta, a distância de Wasserstein foi computada individualmente por atributo e sumarizada tanto pela média quanto pela mediana, permitindo capturar, respectivamente, o efeito global da perturbação e seu impacto sobre as regiões centrais das distribuições. Ao analisar essas estatísticas em função do parâmetro de perturbação α, torna-se possível caracterizar de forma abrangente o compromisso entre privacidade e utilidade, evidenciando como incrementos controlados de anonimização afetam progressivamente a estrutura estatística dos dados.

Para avaliar a privacidade, realizamos um Ataque de Inferência de Membros (MIA), quantificando o grau em que a anonimização mitiga o risco de identificação adversária de amostras de treinamento. Esta avaliação combinada nos permite analisar a relação de compromisso entre privacidade e utilidade induzida por diferentes valores de α e determinar se é possível preservar a utilidade preditiva enquanto aumenta a resistência a ataques de inferência [Shokri et al. 2017].

A avaliação experimental é realizada utilizando quatro conjuntos de dados distintos, denominados de D1 a D4. Esta seleção fornece uma variedade abrangente de sinais fisiológicos e cenários de diagnóstico para avaliar a robustez do método proposto. Os conjuntos de dados são descritos a seguir:

- **Dataset D1 (EEG Binário):** Conjunto de dados de EEG epiléptico derivado do repositório apresentado em [Andrzejak et al. 2001], no qual sinais cerebrais segmentados em vetores de 178 dimensões são formulados como um problema de classificação binária. A classe correspondente a eventos de crise é contrastada com as demais classes, representando atividade cerebral sem crise.
- **Dataset D2 (EEG Multiclasse):** Baseado na mesma fonte fisiológica do D1, este conjunto preserva a formulação multiclasse original, contemplando cinco condições clínicas distintas relacionadas à atividade cerebral, incluindo estados de crise, regiões tumorais e condições fisiológicas normais.
- **Dataset D3 (IoMT Sintético):** Conjunto de dados sintético projetado para simular cenários realistas da Internet das Coisas Médicas (IoMT) [Mandal 2025]. Ele integra atributos fisiológicos, como sinais vitais, e informações categóricas associadas a eventos e alertas de saúde, sendo adequado para a avaliação de modelos preditivos em ambientes controlados.
- **Dataset D4 (Registros Institucionais de Saúde):** Repositório de dados de saúde provenientes de múltiplas instituições médicas, inserido no contexto de aplicações

IoMT [Barman 2024]. O conjunto enfatiza parâmetros cardiovasculares e respiratórios, oferecendo um cenário representativo para o estudo de sistemas inteligentes aplicados à telemedicina e à resposta a emergências.

Por fim, é válido ressaltar que com o objetivo de garantir a reprodutibilidade científica, o código-fonte da solução proposta, bem como o conjunto de dados utilizado nos experimentos, estão publicamente disponíveis em um repositório online[1].

5. Resultados

A análise dos resultados é conduzida sob duas perspectivas fundamentais e interdependentes: a manutenção da utilidade preditiva para tarefas de análise de dados e a garantia de privacidade contra adversários. Primeiramente, na Seção 5.1, investigamos como a perturbação dos autovetores impacta o desempenho de classificadores de aprendizado de máquina. Em seguida, na Seção 5.3, quantificamos a eficácia da anonimização através de testes de resistência a ataques de inferência. O objetivo central é evidenciar como o ajuste do parâmetro α permite modular a relação de compromisso entre esses dois critérios.

5.1. Experimentos em Aprendizado de Máquina

Iniciamos a análise avaliando a robustez de cinco classificadores amplamente utilizados: Árvore de Decisão (DT), Gaussian Naïve Bayes (GNB), k-Vizinhos Mais Próximos (KNN), Regressão Logística (LR) e Perceptron Multicamadas (MLP), sob diferentes níveis de perturbação controlados pelo parâmetro α, considerando quatro conjuntos de dados com características distintas. Para assegurar uma comparação justa entre os modelos, foram adotadas divisões idênticas de treino e teste em todos os experimentos. O desempenho foi avaliado por meio da *acurácia balanceada*, métrica particularmente apropriada para cenários com possível desequilíbrio entre classes, pois considera a média das taxas de acerto por classe, mitigando vieses associados à predominância de rótulos majoritários. Os resultados dessa avaliação são apresentados na Figura 3.

Os resultados da 3 indicam que o aumento progressivo da perturbação introduzida pelo mecanismo de anonimização resulta em uma redução controlada da acurácia balanceada à medida que o parâmetro α cresce. Nos conjuntos de dados D1 e D2, a maioria dos classificadores mantém níveis elevados de desempenho mesmo sob valores mais altos de α, evidenciando que a perturbação afeta a representação dos dados de forma gradual, sem comprometer significativamente a separabilidade entre as classes. Esse comportamento confirma a capacidade do *PerturbPCA*-α de preservar a utilidade analítica em cenários de classificação binária e multiclasse estruturalmente bem definidos.

Para o conjunto de dados D3, a acurácia balanceada permanece elevada nos níveis iniciais de perturbação, apresentando uma degradação suave conforme α aumenta, com destaque para a maior estabilidade dos modelos MLP e GNB, que capturam padrões estatísticos globais ou representações distribuídas. Em contraste, o conjunto D4 apresenta maior sensibilidade ao aumento da perturbação, refletida por uma queda mais acentuada do desempenho, especialmente em modelos baseados em relações de proximidade. De forma consistente entre os cenários, DT e GNB exibem elevada robustez, o MLP alcança os maiores valores de acurácia balanceada, e o KNN apresenta a degradação mais pronunciada, comportamento esperado dada sua dependência direta de relações de distância.

[1] `https://github.com/IvoAP/pca_perturb_sbrc2026`

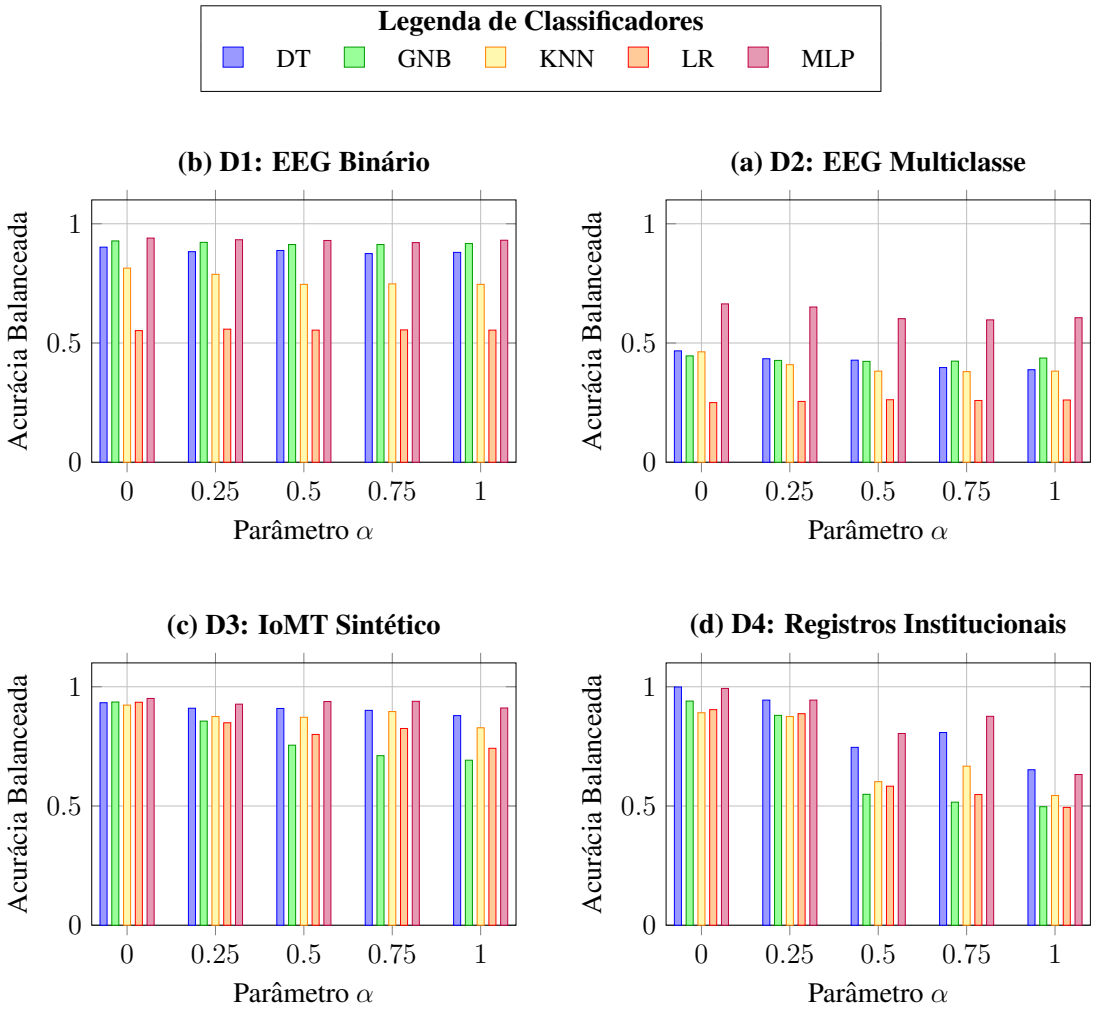

Figura 3. Impacto da anonimização na acurácia balanceada.

Em conjunto, esses resultados demonstram que o *PerturbPCA*-α promove um compromisso suave e ajustável entre utilidade e anonimização, sendo adequado para aplicações que demandam controle fino do impacto da perturbação sobre o desempenho preditivo.

5.2. Experimentos de Distância de Wasserstein

A Figura 4 apresenta a evolução da média da distância de Wasserstein entre os dados originais e anonimizados em função do parâmetro de perturbação α. Essa métrica fornece uma interpretação geométrica do deslocamento entre distribuições, sendo particularmente adequada para quantificar o impacto da perturbação no espaço contínuo das características. Observa-se que, para todos os conjuntos de dados, a distância de Wasserstein cresce monotonicamente à medida que α aumenta, indicando que o mecanismo de anonimização introduz perturbações progressivas e controladas, sem alterações abruptas no comportamento estatístico dos dados.

Ao analisar a *média* da distância de Wasserstein, é possível observar que os conjuntos de dados D1 e D2 apresentam valores absolutos reduzidos ao longo de todo o intervalo do parâmetro α, indicando que as perturbações introduzidas pelo *PerturbPCA*-α não alteram de maneira significativa a estrutura estatística global dessas distribuições.

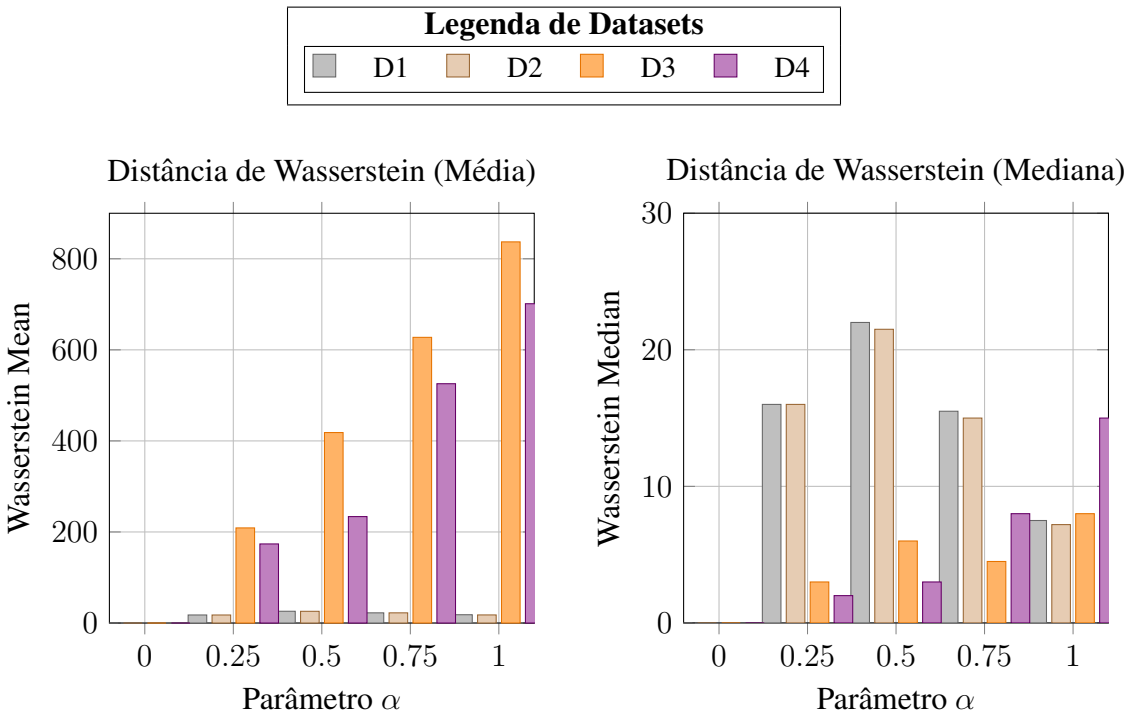

Figura 4. Média e mediana da distância de Wasserstein

Esse comportamento é esperado, visto que os dados de D1 e D2 apresentam correlação temporal e espacial, com variância concentrada em poucos componentes principais, o que limita o impacto das perturbações, mesmo quando α aumenta. Em contrapartida, os conjuntos D3 e D4 mostram um crescimento mais pronunciado e aproximadamente linear da distância de Wasserstein à medida que α aumenta, o que reflete uma maior variabilidade e amplitude de seus atributos. O comportamento monotônico observado na média indica que o *PerturbPCA*-α promove um aumento controlado da divergência, permitindo um ajuste progressivo do nível de anonimização.

A análise com base na mediana da distância de Wasserstein revela um padrão consistente com os resultados médios, mas com magnitudes inferiores, sugerindo que o impacto da perturbação sobre as regiões centrais das distribuições é mais uniforme. Para os conjuntos D1 e D2, a mediana permanece relativamente estável em todos os níveis de α, o que reforça que as transformações aplicadas não afetam significativamente a estrutura central dessas distribuições. Por outro lado, para D3 e D4, observa-se um crescimento gradual na mediana, embora este seja menos influenciado por variações extremas. Esses resultados confirmam que o *PerturbPCA*-α introduz perturbações de maneira homogênea e controlada, preservando a organização estatística central dos dados, especialmente em conjuntos altamente estruturados como D1 e D2.

5.3. Experimentos de Inferência de Membros

Esta seção avalia a robustez do mecanismo proposto frente a ataques de inferência de membros (MIA), conforme descrito na Seção 4. Nesse cenário, modelos de aprendizado de máquina treinados atuam como vítimas, enquanto um atacante de caixa preta explora informações de confiança extraídas das saídas desses modelos para inferir a participação

de amostras no conjunto de treinamento. Os experimentos são conduzidos sobre os conjuntos de dados D1–D4 sob diferentes níveis de anonimização controlados pelo parâmetro α, sendo a eficácia do ataque mensurada pela Área Sob a Curva ROC (AUC). Valores de AUC próximos a 0,50 indicam resistência efetiva ao ataque, ao passo que valores mais elevados evidenciam maior vazamento de privacidade. Os resultados consolidados na Figura 5 demonstram que o nível de proteção fornecido pelo mecanismo varia de forma sistemática com o grau de perturbação e com a complexidade dos dados, confirmando a capacidade do método de mitigar ataques MIA de maneira progressiva.

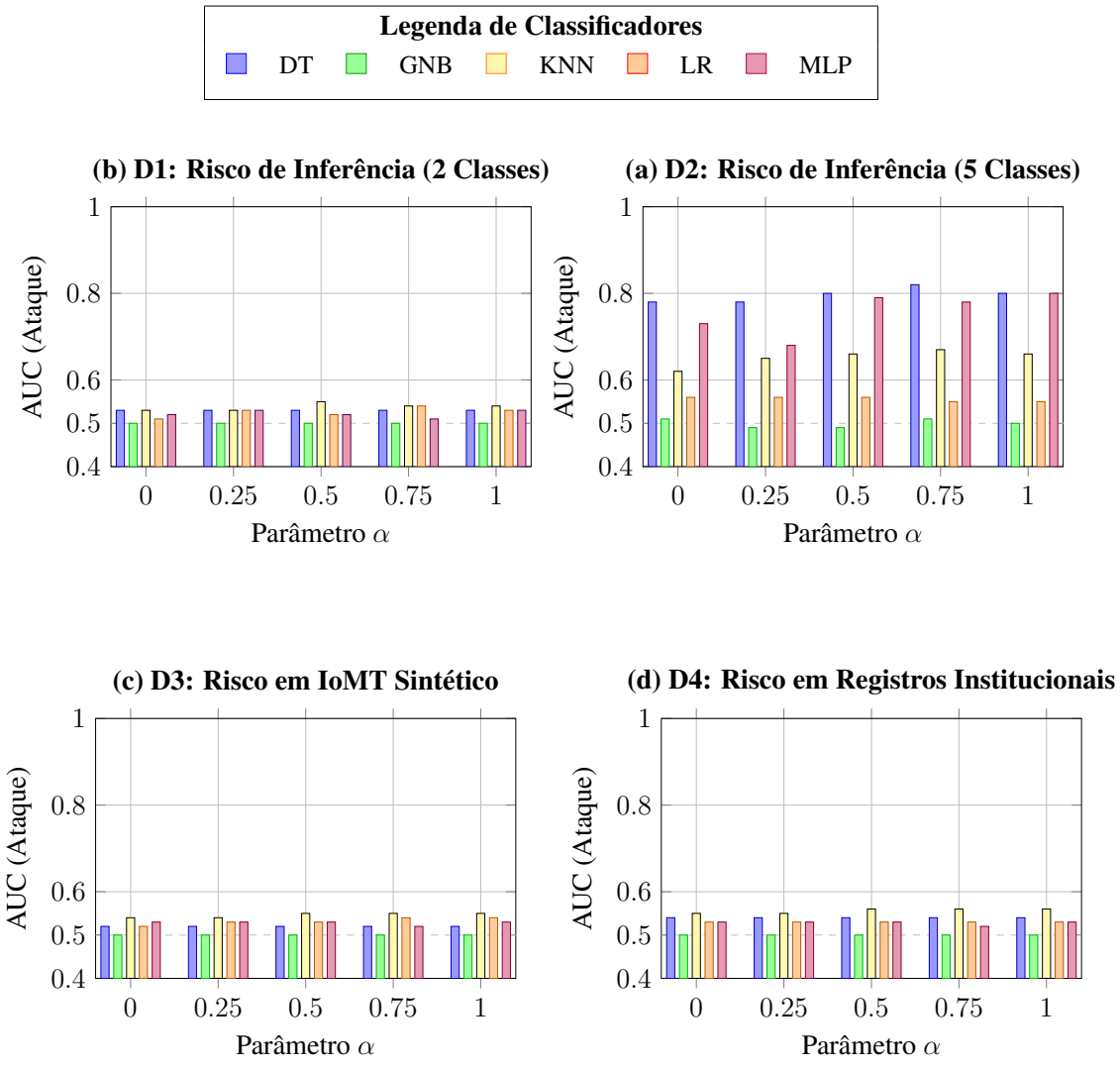

Figura 5. Desempenho do Ataque de Inferência de Membros (MIA)

De acordo com os dados apresentados na Figura 5, para o conjunto D1, correspondente à tarefa de classificação binária de EEG, todos os modelos avaliados apresentam valores de AUC consistentemente próximos a $0{,}50$ para todo o intervalo do parâmetro α. Esse comportamento indica que os classificadores treinados não expõem sinais significativos de pertinência, mesmo na ausência de anonimização, e que o aumento progressivo da intensidade da perturbação não degrada esse perfil de privacidade. Tendência semelhante é observada nos conjuntos D3 (monitoramento de saúde) e D4 (alertas de pacientes), nos quais, independentemente do modelo ou nível de anonimização, os valores de AUC

permanecem próximos à linha de base de palpite aleatório, sugerindo que a estrutura estatística desses dados, em conjunto com a perturbação aplicada, não permite a distinção confiável entre amostras de treinamento e de teste.

Em contraste, o conjunto D2, que representa um cenário multiclasse mais complexo, apresenta maior suscetibilidade a ataques de inferência de membros, com um comportamento não monotônico em relação ao aumento do parâmetro α. Nesse caso, a perturbação no espaço latente pode modificar as fronteiras de decisão de determinados classificadores, ora revelando, ora obscurecendo padrões associados à pertinência das amostras. Ainda assim, o Gaussian Naive Bayes demonstra elevada robustez ao longo de todos os níveis de perturbação, mantendo valores de AUC próximos a $0{,}50$, possivelmente devido à sua natureza probabilística e menor propensão ao *overfitting*. De forma geral, o *PerturbPCA*-α não amplifica os riscos de inferência de membros, preservando níveis próximos ao ideal de privacidade nos conjuntos D1, D3 e D4 e evitando a exacerbação de vulnerabilidades inerentes mesmo no cenário mais desafiador representado por D2.

6. Conclusão

Este trabalho apresentou o *PerturbPCA*-α, um mecanismo de anonimização ajustável baseado em PCA, projetado para enfrentar os desafios de privacidade inerentes aos sistemas inteligentes que lidam com dados sensíveis. Ao aplicar uma perturbação contínua à base de autovetores do PCA, em vez de depender da permutação completa de características, a abordagem proposta oferece um compromisso suave e controlável entre privacidade e utilidade. Isso possibilita a proteção de dados sensíveis, preservando a estrutura estatística essencial necessária para análises posteriores, tornando o método particularmente adequado para ambientes modernos e com restrição de recursos.

Trabalhos futuros estenderão este mecanismo para cenários dinâmicos e distribuídos, incluindo aprendizado federado, processamento de dados em *streaming* e IoT em dispositivo.

Agradecimentos

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

Referências

Aleroud, A., Chen, Z., and Karabatis, G. (2016). Network trace anonymization using a prefix-preserving condensation-based technique (short paper). In *OTM Confederated International Conferences"On the Move to Meaningful Internet Systems* , pages 934–942. Springer.

Andrzejak, R. G., Lehnertz, K., Mormann, F., Rieke, C., David, P., and Elger, C. E. (2001). Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907.

Barman, P. (2024). IoMT Dataset for ML-Based Health Monitoring. `https://www.kaggle.com/dsv/7736523`. [Online]. Available: Kaggle Dataset.

Boikanyo, K., Zungeru, A. M., Sigweni, B., Yahya, A., and Lebekwe, C. (2023). Remote patient monitoring systems: Applications, architecture, and challenges. *Scientific African*, 20:e01638.

Gong, X., Chen, Y., Wang, Q., Wang, M., and Li, S. (2022). Private data inference attacks against cloud: Model, technologies, and research directions. *IEEE Communications Magazine*, 60(9):46–52.

Hyrup, T., Lautrup, A. D., Zimek, A., and Schneider-Kamp, P. (2025). A systematic review of privacy-preserving techniques for synthetic tabular health data. *Discover Data*, 3(1):1–32.

Kamalov, F., Pourghebleh, B., Gheisari, M., Liu, Y., and Moussa, S. (2023). Internet of medical things privacy and security: Challenges, solutions, and future trends from a new perspective. *Sustainability*, 15(4).

Mandal, G. (2025). Patient Data for Healthcare Monitoring System. `https://www.kaggle.com/dsv/12719441`. [Online]. Available: Kaggle Dataset.

Nobre, F. V. J., Silva, D. d. S., Ferreira, M. C. M. M., Brito, M. L. M. L., de 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.

Ozcelik, M. M., Kok, I., and Ozdemir, S. (2025). A survey on internet of medical things (iomt): Enabling technologies, security and explainability issues, challenges, and future directions. *Expert Systems*, 42(5):e70010.

Pimenta, I., Silva, D., Moura, E., Silveira, M., and Gomes, R. L. (2024). Impact of data anonymization in machine learning models. In *Proceedings of the 13th Latin-American Symposium on Dependable and Secure Computing*, pages 188–191.

Pimenta, I. A., Lee, M. H., Bittencourt, L. F., and Gomes, R. L. (2025). Adaptive privacy based on mutual information for machine learning in edge-cloud environments. *IEEE Networking Letters*.

Razdan, S. and Sharma, S. (2022). Internet of medical things (iomt): Overview, emerging technologies, and case studies. *IETE Technical Review*, 39(4):775–788.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Silveira, M. M., Portela, A. L., Menezes, R. A., Souza, M. S., Silva, D. S., Mesquita, M. C., and Gomes, R. L. (2023). Data protection based on searchable encryption and anonymization techniques. In *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, pages 1–5. IEEE.

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.

Thabit, F., Alhomdy, S., and Jagtap, S. (2021). A new data security algorithm for the cloud computing based on genetics techniques and logical-mathematical functions. *International Journal of Intelligent Networks*, 2:18–33.