Aplicando Privacidade Diferencial contra Ataques de Associação em Internet das Coisas

Davi Bezerra Yada da Silva[1], **Aldri Luiz dos Santos**[2], **Jeandro de M. Bezerra**[1,2]

[1]Campus de Quixadá - Universidade Federal do Ceará (UFC)

[2]Depto. de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)

davi09bezerra@alu.ufc.br,aldri@dcc.ufmg.br,jeandro@ufc.br

Abstract. *Deep learning models have been employed in intrusion detection systems to identify anomalies and classify attacks. These models are particularly useful in Internet of Things (IoT) environments, where devices face constant threats. However, deep learning models are vulnerable to attacks, such as membership inference, which can expose sensitive training data. In this context, a feedforward dense neural network was implemented for attack classification using the IoT-23 dataset. To mitigate data exposure risks, the DP-SGD (Differentially Private Stochastic Gradient Descent) algorithm was applied to the model. Additionally, the performance of a rule-based membership inference attack was used to evaluate the model's privacy level, with metrics such as accuracy, precision, and recall. Experiments with varying privacy levels revealed that attack effectiveness decreases proportionally to model efficiency. The use of DP-SGD reduced attack accuracy by 5%, without impacting its precision, and decreased recall by 11%.*

Resumo. *Os modelos de aprendizado profundo têm sido usados em sistemas de detecção de intrusão para detectar anomalias e classificar ataques para os ambientes de Internet das Coisas, onde os dispositivos são alvos de ameaças constantes. No entanto, os modelos de aprendizado profundo podem ser alvos de ataques como inferência de associação e expor dados sensíveis durante o treinamento. Nesse contexto, este trabalho implementou um modelo de redes neurais densas feedforward para classificação de ataques no conjunto de dados IoT-23. Para diminuir o risco de exposição de dados, aplicou-se o algoritmo DP-SGD (Differentially Private Stochastic Gradient Descent) no modelo. Além disso, utilizou-se o desempenho do ataque de inferência de associação baseado em regras para avaliar o nível de privacidade do modelo, com métricas de acurácia, precisão e recall. Também foram feitos experimentos com diferentes níveis de privacidade. Constatou-se que a eficiência do ataque diminui proporcionalmente com a eficiência do modelo. O uso do DP-SGD diminuiu a acurácia do ataque em 5%, sem impactar sua precisão, e o recall em 11%.*

1. Introdução

O crescimento da Internet das Coisas (IoT) impulsionou avanços em diversas áreas, como monitoramento ambiental, automação residencial e redes inteligentes. Dispositivos IoT geram grandes volumes de dados, utilizados para análises preditivas e personalização de

serviços [Vidal 2020]. Recentemente, modelos de aprendizado de máquina têm usado esses dados para identificar padrões complexos e construir sistemas inteligentes. Porém, sua utilização levanta preocupações sobre privacidade, devido ao risco de exposição de informações sensíveis. A privacidade é um conceito técnico que remete o controle de informações pessoais e a anonimização de dados para impedir que as informações de um indivíduo sejam vinculados a ele [Nissim and Wood 2018].

O aprendizado profundo tem aprimorado ferramentas de segurança para IoT [Javaid et al. 2016]. No entanto, esses modelos estão suscetíveis a ataques à privacidade, como a inferência de associação (do inglês, *membership inference attack* (MIA)), que podem revelar dados utilizados no treinamento, comprometendo informações como localização e hábitos dos usuários [Boulemtafes et al. 2020]. A extração de dados de usuários permite a inferência de hábitos e atividades como horário de sono [Pinheiro et al. 2020]; expondo os usuários a riscos de violação de privacidade.

A Privacidade Diferencial (PD) [Dwork 2006], uma modelagem matemática para manter utilidade e privacidade de dados, tem sido uma solução para mitigar esses riscos. Trabalhos anteriores empregaram o algoritmo Descida de Gradiente Estocástico Diferencialmente Privado (DP-SGD) para reduzir o risco de vazamento por meio de modelos. Essa abordagem é comumente utilizada para aprendizado federado [Pustozerova et al. 2023], mas pode ser melhor explorada para diferentes cenários e arquiteturas.

Neste trabalho, implementou-se o otimizador Estimativa de Momento Adaptativa Diferencialmente Privada (DP-ADAM) [Tang et al. 2024], para treinar um modelo *feedforward* na classificação de ataques em IoT. Avaliamos sua resistência ao ataque de inferência de associação. A abordagem proposta busca reduzir a complexidade do sistema, mantendo sua utilidade. Comparou-se um modelo *baseline* com diferentes versões sob privacidade diferencial, analisando a precisão e o impacto do ataque.

O restante do artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 discute os conceitos apresentados; a Seção 4 detalha a metodologia; a Seção 5 expõe os resultados e, por fim, a Seção 6 traz as conclusões.

2. Trabalhos Relacionados

Esta seção apresenta uma revisão dos trabalhos relacionados que abordam a proteção contra vazamentos de dados sensíveis em modelos de aprendizado profundo e apresenta uma comparação entre os principais trabalhos associados ao contexto desta pesquisa.

O trabalho de [Oliveira 2024] implementa um modelo de aprendizado federado para IoT treinado sob DP-SGD. Os autores utilizam 3 tipos de ataque MIA: baseado em regras, caixa preta e inversão de modelo. Os autores avaliam o desempenho do modelo e do ataque sob diferentes níveis de privacidade.

No trabalho de [Pustozerova et al. 2023], avaliou-se o *trade-off* entre utilidade e privacidade em modelos de aprendizado federado. A comparação é feita com a privacidade do modelo sob treinamento com o otimizador DP-SGD e sob saída ofuscada. Os autores utilizam um ataque de inferência de associação de caixa-preta para avaliar a privacidade alcançada.

Em [Machooka et al. 2024], os autores comparam dois modelos de Aprendizado

Profundo para detecção de intrusão na Internet das Coisas: um modelo de Memória de Curto e Longo Prazo (LSTM) e um *Perceptron* Multicamadas (MLP), ambos treinados com o otimizador DP-SGD. O desempenho é avaliado usando as métricas dos modelos em quatro conjuntos de dados IoT: *MQTT-IOT-IDS2020 Intrusion Detection*, *RT Iot2022 dataset*, *CIC-Iot2023* e *Botnetiot Lo1 Label NoDuplicates*.

A pesquisa de [Cherubin et al. 2024] avalia a privacidade de modelos treinados com *DP-SGD* contra ataques de inferência de associação e de atributo. Os autores derivam limites fechados usando a métrica de Bayes β (um medidor de privacidade) para modelar o *DP-SGD* e garantir privacidade sem o parâmetro ϵ. Os resultados mostram que *DP-SGD* oferece proteção contra os ataques de inferência, principalmente para atributos. Os testes foram feitos com redes neurais totalmente conectadas nos conjuntos de dados *Adult* e *Purchase*.

Tabela 1. Comparação de trabalhos relacionados

Autor	**Modelos**	**Técnica de privacidade**	**avaliação**	**métricas**
[Oliveira 2024]	Aprendizado Federado	DP-SGD	Defesa contra MIA	acurácia multi-classe, acurácia binária, precisão e *recall*
[Pustozerova et al. 2023]	Aprendizado Federado	DP-SGD, saída ofuscada	Defesa contra MIA	Métricas de ataque, precisão, acurácia, *recall*, *f1-score*
[Machooka et al. 2024]	LSTM, MLP	DP-SGD	Comparação	acurácia, precisão, *recall*, *f1-score*
[Cherubin et al. 2024]	Redes neurais totalmente conectadas	*DP-SGD* modificado	Defesa contra MIA	β, precisão, taxa de positivos
Proposto. 2025	*feedforward*	DP-SGD	Defesa contra MIA	Métricas de ataque, precisão, acurácia, *recall*, *f1-score*

Os trabalhos relacionados demonstram o uso de *DP-SGD* para proteger modelos de aprendizado profundo em diferentes cenários, como aprendizado federado e detecção de intrusão, avaliando o impacto na privacidade e na utilidade dos modelos com ataques adversariais. Esta pesquisa propõe um modelo *feedforward* treinado com *DP-SGD* e testado contra um MIA baseado em regras. O uso de *DP-SGD* foi escolhido para tornar o modelo de classificação resistente a ataques adversariais, evitando o custo de adicionar ruído aos dados.

3. Fundamentação Teórica

Esta seção descreve de forma resumida os conceitos sobre redes neurais profundas *feedforward*, privacidade diferencial e a atuação do Ataque de Inferência de Associação.

3.1. Rede Neural Profunda feedforward

As redes neurais profundas são modelos computacionais compostos de múltiplas camadas de processamento totalmente interconectadas, capazes de representar e aprender abstrações de dados de forma progressiva [Reis 2021]. Cada camada recebe a saída da anterior, ajusta seus parâmetros para otimizar o aprendizado sobre os dados e, em seguida, passa uma nova abstração para a camada seguinte [Pouyanfar et al. 2018]. A arquitetura *feedforward* descreve o fluxo das conexões de aprendizado, indo da entrada até a saída, passando pelas camadas ocultas [Bezerra 2016]. Utilizou-se um modelo *feedforward* por

suas características de aprendizado. Esses modelos são capazes de aprender padrões complexos e representações em grandes volumes de dados para realizar classificações. Além disso, esse modelo é menos complexo computacionalmente em comparação a outros, reduzindo custos de memória e sobrecarga [Haddadi et al. 2010].

O treinamento de redes neurais requer o uso de um algoritmo de otimização que ajuste iterativamente os parâmetros do modelo para minimizar o erro de previsão. A Descida de Gradiente Estocástico (SGD) [Amari 1993], um dos algoritmos mais comuns nesse processo, realiza os seguintes passos: i) seleção de um *mini-batch* de dados aleatoriamente, ii) cálculo do gradiente de aprendizado para esse *mini-batch*, e iii) atualização dos parâmetros do modelo para minimizar o valor da função de perda.

3.2. Privacidade Diferencial

A Privacidade Diferencial é um modelo matemático robusto aplicado para garantir a privacidade de indivíduos em um conjunto de dados enquanto mantém sua utilidade. O equilíbrio entre a utilidade e a privacidade é alcançado por meio da adição de ruído aleatório, medido pelo parâmetro ϵ [Dwork 2006]. A definição de [Dwork 2006] para privacidade diferencial é a seguinte: Tendo os conjuntos de dados vizinhos D e D' diferindo-se em no máximo um elemento, o subconjunto S e um mecanismo A aplicado aos dados, diz-se que o mecanismo A é ϵ-diferencial se:

$$\Pr[A(D) \in S] \leq \epsilon \cdot \Pr[A(D') \in S]$$

Mostrando que a probabilidade de um resultado obtido pertencer a S a partir de D deve ser, no máximo, ϵ vezes a probabilidade de obter o mesmo resultado a partir de D'. Isso implica que um observador não pode inferir se o resultado de S veio de D ou D' e que a ausência ou presença de algum dado não afeta a saída.

O uso de privacidade diferencial, além de ofuscar dados, é aplicável a outros contextos que requerem a preservação de privacidade, como fluxo de dados de IoT para agregadores de serviço [Vidal 2020] e aprendizado profundo [Abadi et al. 2016]. Tornar o aprendizado de máquina diferencialmente privado envolve garantir que o modelo não exponha dados usados no treinamento. Para isso, pode-se adicionar ruído aos dados de treino, ofuscar as saídas do modelo ou usar um otimizador diferencialmente privado [Siachos et al. 2023], como ilustra a Figura 1.

O otimizador Descida de Gradiente Estocástico possui uma variante diferencialmente privada [Chua et al. 2024], projetada para evitar que o modelo memorize características dos dados e diminuir o risco de exposição. O DP-SGD implementa rodadas de corte e adição de ruído. Tendo o limite de corte c, o *micro-batche* de tamanho B, sendo uma divisão ainda menor de um *mini-batch*, e o multiplicador de ruído α, o otimizador segue os seguintes passos a cada iteração: i) Para cada *mini-batch*, seleciona um *micro-batch* B aleatoriamente; ii) Calcula o gradiente de aprendizado; iii) Adiciona ruído de acordo com α; iv) Reduz o gradiente ao limite c; e v) Atualiza os parâmetros.

3.3. Ataque de Inferência de Associação

O Ataque de Inferência de Associação se baseia em identificar se um registro de dado foi usado durante o treinamento do modelo. Ao ter acesso ao modelo e a uma amostra de dados, um atacante é capaz de utilizar suas predições para deduzir dados presentes no processo de treinamento [Hu et al. 2022]. Os atacantes definem um modelo-alvo,

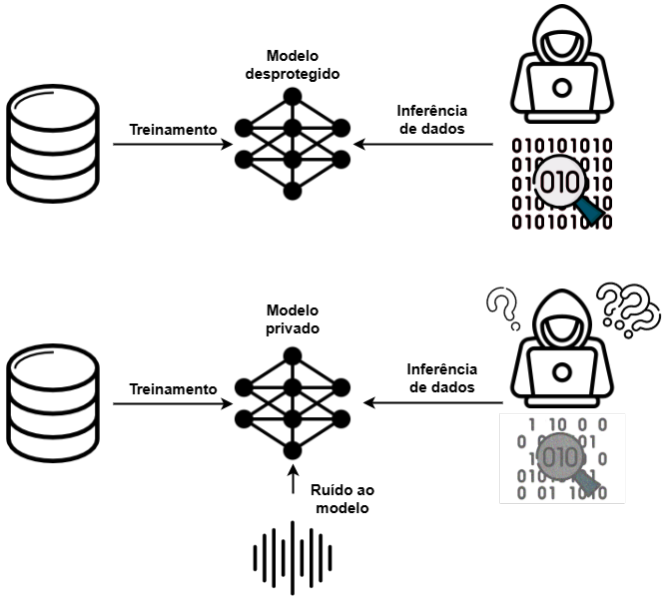

Figura 1. Representação de modelo diferencialmente privado

que normalmente é disponibilizado publicamente, e o usam para gerar previsões sobre uma amostra de dados. Conceitualmente, os dados presentes no conjunto de treinamento são chamados de membros. Já os dados não presentes são chamados de não-membros. Como um exemplo, suponha que um grupo de pesquisa utiliza um conjunto de dados com informações médicas sensíveis sobre pacientes para treinar um modelo de classificação, publicando em seguida o modelo usado e seus resultados em um repositório *online* para fins de pesquisa. Um atacante, então, pode utilizar esse modelo e realizar um ataque de inferência para identificar amostras presentes nos dados de treinamento, conseguindo expor os usuários e suas condições médicas [Zhang et al. 2022].

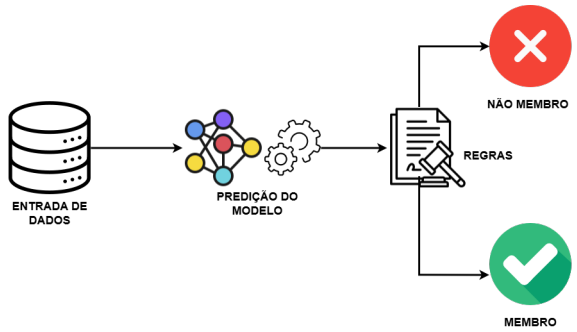

Figura 2. Diagrama de funcionamento de MIA

Os ataques de inferência de associação são classificados como Ataques de caixa preta, isto é quando o atacante tem acesso apenas ao modelo, e como Ataques de caixa branca, quando há acesso às saídas, parâmetros e configurações do modelo. Há uma variação chamada Ataque de caixa preta baseado em regras [AI 2024]. Essa variação se baseia em identificar amostras utilizando predições do modelo, definindo membros com base em uma regra definida. As regras são instruções usadas para analisar os dados e inferir se são membros ou não. A Figura 2 demonstra o seu funcionamento. Um exemplo

de regra consiste em definir uma amostra como membro se a predição do modelo for correta, com uma probabilidade acima de um limiar definido.

4. Implementação do Modelo Diferencialmente Privado

Esta seção apresenta as etapas e os procedimentos adotados para o desenvolvimento do modelo diferencialmente privado, descrevendo detalhadamente o ambiente experimental, as ferramentas e as estratégias empregadas para alcançar os objetivos propostos. A metodologia baseia-se na abordagem de [Oliveira 2024].

4.1. Conjunto de Dados

Definiu-se o conjunto de dados público IoT-23 [Sebastian Garcia 2020] para o treinamento e teste do modelo da rede neural. Esse conjunto de dados consiste em vinte e três capturas (cenários) de diferentes tipos de tráfego IoT, dos quais vinte cenários têm tráfego malicioso e três cenários têm tráfego benigno. A estratégia para o pré-processamento consistiu em gerar um conjunto de dados com amostras de cinquenta mil pacotes de cada cenário. A coleta das amostras mantém a sequência temporal baseada no atributo *timestamp* (ts). O propósito foi obter um conjunto de dados com maior representação de ataques para verificar o desempenho do modelo proposto.

Os atributos disponíveis no conjunto de dados são apresentados na Tabela 2. Utilizou-se os atributos *timestamp* para registro temporal, endereços IP de origem e destino e portas de origem e destino para identificar as comunicações, dispositivos e processos, número de *bytes* de pacotes IP para caracterização, protocolo usado e *label* para identificar o tráfego. Após a seleção dos atributos, o arquivo final ficou com a estrutura mostrada na Tabela 3.

Tabela 2. Atributos do conjunto de dados

ts	service	missed bytes
uid	duration	history
id.orig h	orig bytes	org pkts
id.orig p	resp bytes	orig ip bytes
id.resp h	conn state	resp pkts
id.resp p	local orig	resp ip bytes
proto	local resp	label

Tabela 3. Atributos selecionados para treinamento

ts	id.orig h	id.orig p	id.resp h	id.resp p	proto	orig ip bytes	label

A Tabela 4 resume os rótulos presentes no conjunto de dados inicialmente e suas quantidades. Esses rótulos são derivados de ataques como *command & control* (C&C) e *distributed denial of service* (DDoS) [Hoque et al. 2015]. Descartou-se do conjunto de treinamento as classes *C&C-HeartBeat*, *C&C-FileDownload*, *C&C-HeartBeat-FileDownload*, *attack*, *C&C-Torii*, *FileDownload* e *C&C-Mirai*, pois eram classes com

poucas amostras. Apesar da implementação de balanceamento de dados, descartamos essas classes por gerarem muitas réplicas pouco diversificadas, o que é responsável por gerar *overfitting*. Aplicou-se a técnica de *random oversampling* para replicar classes de dados minoritárias aleatórias e balancear proporcionalmente o conjunto de dados. A Tabela 5 indica as classes selecionadas e balanceadas para o treinamento do modelo. Aplicou-se a transformação de valores categóricos em numéricos. Implementamos a técnica de normalização *Label Encoder* no conjunto de dados e a técnica de normalização *MinMaxScaler* para normalizar os valores das amostras de dados. A divisão do conjunto de dados foi estabelecida em 50% para treinamento, 30% para teste e 20% para validação.

Tabela 4. Rótulos e suas quantidades

Rótulo	Quantidade	Rótulo	Quantidade
benigno	109177	PartOfAHorizontal PortScan	413024
C&C	15064	Attack	1387
C&C-HeartBeat	229	Okiru	131308
DDoS	76371	C&C-Torii	30
C&C-FileDownload	44	FileDownload	16
C&C-HeartBeat-FileDownload	11	C&C-Mirai	1

Tabela 5. Rótulos selecionados após balanceamento

Rótulo	Quantidade
benigno	104446
DDoS	104220
Okiru	104287
C&C	104298
PartOfAHorizontalPortScan	104209

4.2. Modelo de Classificação

O modelo diferencialmente privado foi desenvolvido baseado nas redes neurais profundas do tipo *feedforward* [Rimer and Martinez 2004]. Escolheu-se essa arquitetura por ser capaz de realizar tarefas de classificação com uma estrutura menos complexa e que gera menos consumo computacional. Mesmo sendo mais simples, o modelo é capaz de aprender as características dos dados, podendo generalizar e realizar classificações com alta taxa de acertos. Utilizou-se a função *Softmax* [Qi et al. 2017] para determinar a probabilidade de pertencimento do valor de entrada a cada uma das classes.

A implementação deste modelo ocorreu através da API de alto nível *Keras*, do *tensorflow*. Adotou-se uma estrutura de oito camadas totalmente conectadas (densas) com poucos nós de processamento. Aplicou-se a função *softmax* na camada de saída e a função de ativação Unidade Linear Retificada (ReLU) nas camadas de entrada e ocultas. A estrutura do modelo é apresentada na Figura 3 e a Tabela 6 resume os parâmetros de arquitetura e de treinamento.

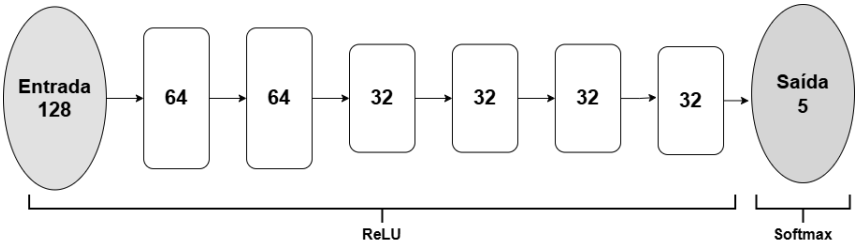

Figura 3. Estrutura do modelo de redes neurais profundas

Tabela 6. Parâmetros do modelo *baseline*

Tipo	Sequencial; Denso
Função de ativação	ReLU, Softmax
Otimizador	ADAM
Função de perda	*Sparse Categorical Crossentropy*
Épocas	20
batch size	128

Os rótulos no conjunto de treinamento guiam o modelo a aprender a classificar os dados. Isso deve-se ao modelo de redes neurais ter um aprendizado do tipo supervisionado. Os rótulos aprendidos pelo modelo são usados para classificar dados não rotulados posteriormente. O desempenho do modelo *baseline* com o conjunto de testes é resumido na Seção 5.

4.3. Privacidade Diferencial para Aprendizado Profundo

A solução é implementada baseada no algoritmo DP-ADAM, uma variação do DP-SGD que usa estimativas de momento, como média e variância, para atualizar os parâmetros de treinamento. O algoritmo DP-ADAM limita os gradientes de aprendizado e adiciona ruído aleatório para que o modelo não memorize características dos dados, evitando exposição na inferência. O DP-ADAM é um otimizador que torna o modelo diferencialmente privado por meio de ruído no processo de aprendizado do modelo, como citado na Seção 3.2.

A biblioteca *tensorflow privacy* foi utilizada para implementar o otimizador DP-ADAM integrado à API *Keras*. Esse otimizador participa ativamente do processo de treinamento do modelo e seus parâmetros definem o nível de privacidade alcançado. Seus parâmetros servem para, durante uma iteração do treinamento, separar os gradientes em lotes de tamanho fixo n, adicionar um nível de ruído e limitar o tamanho do gradiente de acordo com um limite definido [tensorflow privacy 2024].

5. Análise dos Resultados

Esta seção apresenta os resultados dos experimentos conduzidos com diferentes níveis de privacidade aplicados ao modelo de detecção de intrusão. Os experimentos avaliaram o impacto do uso de privacidade diferencial no desempenho do modelo e na eficácia do ataque de inferência de associação.

5.1. Cenário do Ataque de Inferência de Associação

O ataque de inferência de associação usado neste estudo [AI 2024] baseou-se em regras e teve acesso ao modelo treinado, além dos conjuntos de treinamento e teste. Para cada entrada no modelo, o ataque gerou um valor de predição e aplicou uma regra para determinar se a amostra pertencia ou não ao conjunto de treinamento. A Figura 4 ilustra o processo.

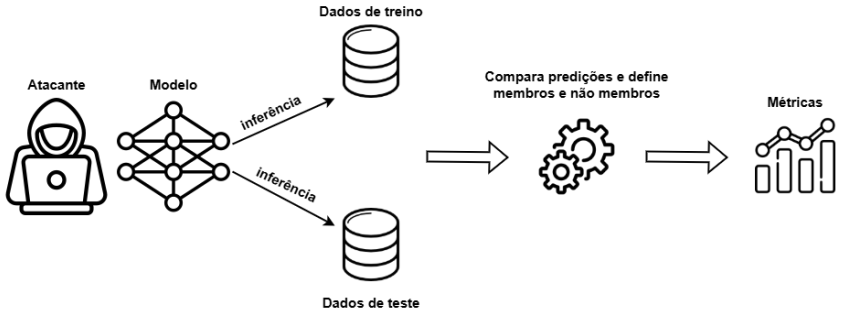

Figura 4. Funcionamento do ataque de inferência baseado em regras

Na avaliação do desempenho do ataque, foram usadas as métricas: acurácia de membros (*member acc*), acurácia de não membros (*non-member acc*), acurácia do ataque (*attack acc*), precisão e *recall*. Elas são definidas da seguinte forma:

$$1.\ member\ acc = \frac{membros\ corretamente\ inferidos}{total\ de\ membros}$$

$$2.\ non\text{-}member\ acc = 1 - \frac{não\ membros\ inferidos\ como\ membros}{total\ de\ não\ membros}$$

$$3.\ attack\ acc = \frac{member\ acc \times No.\ de\ membros + non\text{-}member\ acc \times No.\ de\ não\ membros}{total\ de\ amostras}$$

$$4.\ precisão = \frac{No.\ de\ positivos\ previstos\ corretamente}{total\ de\ positivos\ previstos}$$

$$5.\ recall = \frac{No.\ de\ positivos\ previstos\ corretamente}{total\ de\ positivos\ reais}$$

5.2. Resultados

Os experimentos foram conduzidos com diferentes níveis de privacidade (ϵ) variando entre 0.08 e 1.5, considerando duas configurações de norma de corte: 1.3 e 1.5. O valor de norma de corte limita o tamanho do gradiente após a adição de ruído, controlando o aprendizado do modelo. A Tabela 7 apresenta as métricas do modelo *baseline*, ou seja, o modelo treinado sem privacidade diferencial. Ele atingiu uma acurácia de 99%, indicando um desempenho muito alto na classificação de ataques. Além disso, apresentou valores elevados de precisão (98%) e *recall* (99%), sem sinais evidentes de *overfitting*. A justificativa para isso está ilustrada na Figura 5. A curva de perda do modelo *baseline* indica que o modelo convergiu rapidamente, sem sinais de *overfitting*, corroborando os altos valores de acurácia observados.

Tabela 7. Métricas do modelo *baseline*

acurácia	precisão	recall	f1-Score
0.99	0.98	0.99	0.99

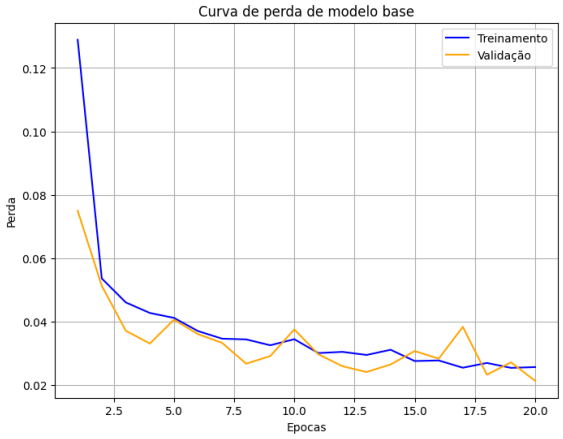

Figura 5. Curva de perda do modelo *baseline*

A Tabela 8 mostra os resultados do ataque de inferência de associação aplicado ao modelo *baseline*. O ataque alcançou 99% de acurácia na identificação de membros, com uma precisão de 70%. Isso demonstra que o modelo, sem proteção de privacidade diferencial, está altamente vulnerável a ataques de inferência.

Tabela 8. Métricas de ataque ao modelo *baseline*

member acc	non-member acc	attack acc	precisão	recall
0.99	0.01	0.7	0.7	0.99

A aplicação de privacidade diferencial reduziu a eficácia do ataque, mas também impactou a utilidade do modelo. A Tabela 9 mostra que para a norma de corte = 1.3, a acurácia caiu progressivamente de 91% para 53% conforme o nível de privacidade aumentou (ϵ = 1.5 a 0.08). O *recall* permanece alto inicialmente (93%-95%), mas também sofre impacto nos níveis mais altos de privacidade (ϵ = 0.08, *recall* de 65%). A precisão cai de 83% para 54%, indicando um aumento na taxa de falsas previsões.

A Tabela 10 mostra o impacto no ataque de inferência. Para $\epsilon = 1.5$ (menor privacidade), a acurácia do ataque é 66%, diminuindo para 51% quando $\epsilon = 0.08$. Isso indica que, embora o ataque ainda seja eficaz para valores altos de ϵ, níveis maiores de privacidade (valores baixos de ϵ) tornam a inferência de membros mais difícil.

Os experimentos também avaliaram o impacto da norma de corte no desempenho do modelo e do ataque. A Tabela 11 mostra os resultados para norma de corte = 1.5. O modelo mantém uma acurácia maior em níveis de privacidade elevados, variando entre 91% e 74%. De acordo com a Tabela 12, a precisão do ataque se mantém estável em 70%, independentemente do nível de privacidade. A acurácia do ataque diminui mais

Tabela 9. Métricas dos modelos privados com norma de corte = 1.3

ϵ	acurácia	precisão	recall	f1-score
1.5	0.91	0.83	0.93	0.85
1.0	0.87	0.78	0.93	0.83
0.6	0.90	0.81	0.95	0.86
0.4	0.86	0.81	0.93	0.85
0.29	0.78	0.70	0.89	0.74
0.12	0.62	0.59	0.81	0.6
0.08	0.53	0.54	0.65	0.54

lentamente em comparação com a norma de 1.3, indicando que essa configuração pode preservar melhor o equilíbrio entre utilidade e privacidade. A Tabela 12 reforça essa análise, mostrando que a acurácia do ataque permanece entre 66% e 59% para norma de corte 1.5, indicando que essa configuração pode ser mais vantajosa para preservar a utilidade do modelo enquanto reduz o risco de inferência.

Nas Tabelas 10 e 12, a precisão se mantém em 70% independente do nível de ruído. Apesar de o número de identificações do ataque ser afetado pelo DP-SGD, o critério usado para identificação de membros não é afetado pela privacidade, o que mantém a precisão estável.

Tabela 10. Métricas de ataque aos modelos privados com norma de corte = 1.3

ϵ	member acc	non-member acc	attack acc	precision	recall
1.5	0.91	0.08	0.66	0.7	0.91
1.0	0.87	0.12	0.64	0.7	0.87
0.6	0.89	0.09	0.65	0.7	0.9
0.4	0.85	0.14	0.64	0.7	0.85
0.29	0.77	0.22	0.61	0.7	0.77
0.12	0.61	0.38	0.54	0.7	0.61
0.08	0.53	0.47	0.51	0.7	0.53

Tabela 11. Métricas dos modelos privados com norma de corte = 1.5

ϵ	acurácia	precisão	recall	f1-score
1.5	0.91	0.84	0.94	0.88
1.0	0.9	0.84	0.96	0.88
0.6	0.88	0.79	0.95	0.84
0.4	0.88	0.79	0.94	0.84
0.29	0.79	0.72	0.88	0.77
0.12	0.74	0.68	0.86	0.73
0.08	0.74	0.65	0.82	0.69

A Figura 6 compara a acurácia do ataque e do modelo para diferentes valores de ϵ. À medida que ϵ diminui, a acurácia do modelo cai (indicando perda de desempenho), enquanto a acurácia do ataque também reduz (indicando maior proteção da privacidade). A norma de corte 1.5 oferece um melhor equilíbrio entre privacidade e desempenho, já que o modelo mantém uma acurácia mais alta para níveis de privacidade elevados.

Tabela 12. Métricas de ataque aos modelos privados com norma de corte = 1.5

ϵ	member acc	non-member acc	attack acc	precision	recall
1.5	0.91	0.08	0.66	0.7	0.91
1.0	0.9	0.09	0.66	0.7	0.9
0.6	0.88	0.12	0.65	0.7	0.87
0.4	0.88	0.11	0.65	0.7	0.88
0.29	0.79	0.2	0.61	0.7	0.79
0.12	0.73	0.26	0.59	0.7	0.73
0.08	0.73	0.26	0.59	0.7	0.73

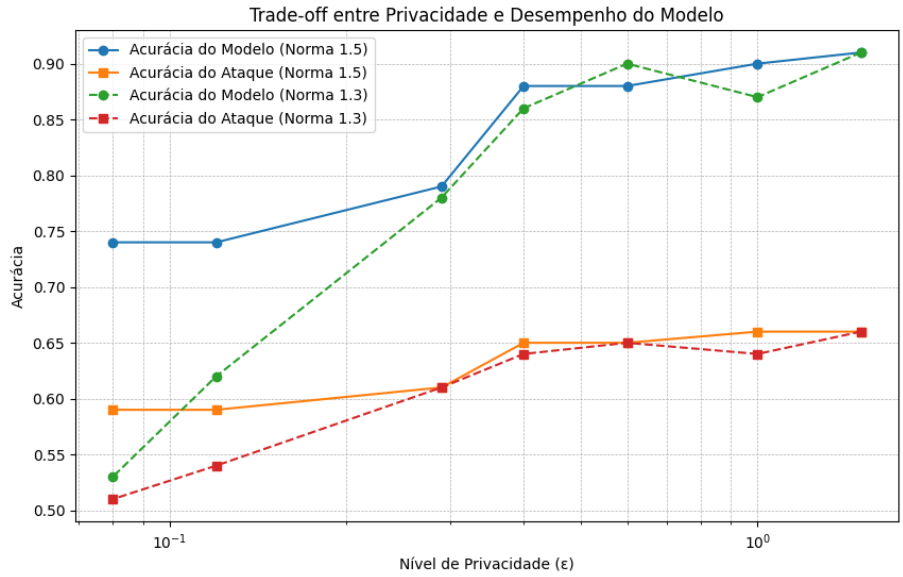

Figura 6. Comparação de acurácia de ataques e modelos

Os resultados confirmam o *trade-off* entre privacidade e utilidade. Enquanto a privacidade diferencial reduz a eficácia dos ataques de inferência, ela também impacta o desempenho do modelo de detecção de intrusões. A norma de corte 1.5 se mostrou a melhor opção para manter o desempenho do modelo enquanto reduz o risco de inferência.

6. Conclusão

Este trabalho propôs uma aplicação de privacidade diferencial em um modelo *feedforward* para classificação de ataques, utilizando DP-SGD e avaliando o impacto de um ataque de inferência de associação baseado em regras. O conjunto de dados IoT-23 foi utilizado para treinamento e teste do modelo, explorando dois valores de norma de corte (1.3 e 1.5) e sete níveis de privacidade (ϵ variando de 0.08 a 1.5). Os resultados indicam que o desempenho do ataque está diretamente relacionado à acurácia do modelo. O uso de DP-SGD reduz a eficiência do ataque, mas compromete a utilidade do modelo. Um valor de norma de corte maior (1.5) demonstrou um melhor *trade-off*. Para trabalhos futuros, recomenda-se validar o comportamento da aplicação em cenários reais de redes IoT baseadas em IP e outros conjuntos de dados representativos. Apesar de mitigar vazamentos de dados em modelos, o DP-SGD não ofusca os dados, o que pode ser um risco caso um atacante tenha

acesso indevido. Além disso, melhores configurações são necessárias para que haja um *trade-off* menos restritivo entre privacidade e desempenho.

Agradecimentos

Este trabalho possui financiamento parcial de: PIBIC-UFC (2024-2025) e FAPESP (2018/23098-0). Partes deste artigo foram auxiliadas por ferramentas de inteligência artificial generativa, como o ChatGPT, para reformulação de trechos textuais e revisão linguística. Todo o conteúdo técnico foi elaborado e revisado pelos autores.

Referências

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

AI, T. (2024). Adversarial robustness toolbox repository. https://github.com/Trusted-AI/adversarial-robustness-toolbox.

Amari, S.-i. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196.

Bezerra, E. (2016). Introdução à aprendizagem profunda. *Artigo–31º Simpósio Brasileiro de Banco de Dados–SBBD2016–Salvador*.

Boulemtafes, A., Derhab, A., and Challal, Y. (2020). A review of privacy-preserving techniques for deep learning. *Neurocomputing*, 384:21–45.

Cherubin, G., Köpf, B., Paverd, A., Tople, S., Wutschitz, L., and Zanella-Béguelin, S. (2024). {Closed-Form} bounds for {DP-SGD} against record-level inference attacks. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 4819–4836.

Chua, L., Ghazi, B., Kamath, P., Kumar, R., Manurangsi, P., Sinha, A., and Zhang, C. (2024). How private is dp-sgd? *arXiv preprint arXiv:2403.17673*.

Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Haddadi, F., Khanchi, S., Shetabi, M., and Derhami, V. (2010). Intrusion detection and attack classification using feed-forward neural network. In *2010 Second international conference on computer and network technology*, pages 262–266. IEEE.

Hoque, N., Bhattacharyya, D. K., and Kalita, J. K. (2015). Botnet in ddos attacks: trends and challenges. *IEEE Communications Surveys & Tutorials*, 17(4):2242–2270.

Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., and Zhang, X. (2022). Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37.

Javaid, A., Niyaz, Q., Sun, W., and Alam, M. (2016). A deep learning approach for network intrusion detection system. In *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS)*, pages 21–26.

Machooka, D., Yuan, X., Roy, K., and Chen, G. (2024). Comparison of lstm and mlp trained under differential privacy for intrusion detection. In *2024 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, pages 1–10. IEEE.

Nissim, K. and Wood, A. (2018). Is privacy privacy? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170358.

Oliveira, J. A. d. (2024). F-nids: sistema de detecção de intrusão baseado em aprendizado federado.

Pinheiro, A. J., de Araujo-Filho, P. F., Bezerra, J. d. M., and Campelo, D. R. (2020). Adaptive packet padding approach for smart home networks: A tradeoff between privacy and performance. *IEEE Internet of Things Journal*, 8(5):3930–3938.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Chen, S.-C., and Iyengar, S. S. (2018). A survey on deep learning: Algorithms, techniques, and applications. *ACM computing surveys (CSUR)*, 51(5):1–36.

Pustozerova, A., Baumbach, J., and Mayer, R. (2023). Differentially private federated learning: Privacy and utility analysis of output perturbation and dp-sgd. In *2023 IEEE International Conference on Big Data (BigData)*, pages 5549–5558. IEEE.

Qi, X., Wang, T., and Liu, J. (2017). Comparison of support vector machine and softmax classifiers in computer vision. In *2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, pages 151–155. IEEE.

Reis, C. H. (2021). Otimização de hiperparâmetros em redes neurais profundas. *Minas Gerais*.

Rimer, M. and Martinez, T. (2004). Softprop: softmax neural network backpropagation learning. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, volume 2, pages 979–983. IEEE.

Sebastian Garcia, Agustin Parmisano, M. J. E. (2020). Iot-23: A labeled dataset with malicious and benign iot network traffic (version 1.0.0). http://doi.org/10.5281/zenodo.4743746.

Siachos, I., Kaltakis, K., Papachristopoulou, K., Giannoulakis, I., and Kafetzakis, E. (2023). Comparison of machine learning algorithms trained under differential privacy for intrusion detection systems. In *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 654–658. IEEE.

Tang, Q., Shpilevskiy, F., and Lécuyer, M. (2024). Dp-adambc: Your dp-adam is actually dp-sgd (unless you apply bias correction). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15276–15283.

tensorflow privacy (2024). tensorflow privacy - dpkerasadamoptimizer.

Vidal, I. d. C. (2020). Protecting: garantindo a privacidade de dados gerados em casas inteligentes localmente na borda da rede.

Zhang, Z., Yan, C., and Malin, B. A. (2022). Membership inference attacks against synthetic health data. *Journal of biomedical informatics*, 125:103977.