Caracterização Temporal de Vulnerabilidades de Dispositivos na Internet Brasileira usando Dados do Shodan

Isabelle Matos[1] **Giancarlo Teixeira**[1] **Lucas M. Ponce**[2] **Elverton Fazzion**[1,2]
Ítalo Cunha[2] **Cristine Hoepers**[3] **Klaus Steding-Jessen**[3]
Marcelo H. P. C. Chaves[3] **Dorgival Guedes**[2] **Wagner Meira Jr.**[2]

[1]DCOMP/UFSJ [2]DCC/UFMG [3]CERT.br/NIC.br

{isabelle.mattos25, giancarloprados}@aluno.ufsj.edu.br
{lucasmsp,cunha,dorgival,meira}@dcc.ufmg.br
{mhp,cristine,jessen}@cert.br fazzion@ufsj.edu.br

Abstract. *Understanding how vulnerabilities evolve in Internet-connected devices is important so that network operators can prioritize security efforts. In this paper, we perform a temporal characterization of vulnerabilities in devices exposed on the Brazilian Internet using Shodan data. The main challenge is to reliably identify the same physical device across multiple observations, since IP addresses may change over time. To address this, we combine the identification of static IPs, used as ground truth, with a device identifier derived from SSL/TLS attributes (the SHA-256 fingerprint and the common name field of X.509 certificates), validated with a specificity of 0.99 and an AUC of 0.72. Applying this approach to devices exposed between 2024 and mid-2025, we observe that critical vulnerabilities are the most persistent, remaining active for more than eight months, and that sectors outside the technology hub, such as Education and Public Administration, concentrate the highest proportions of critical risk.*

Resumo. *Compreender como vulnerabilidades evoluem em dispositivos conectados à Internet é importante para que operadores de rede possam priorizar esforços em segurança. Neste artigo, realizamos uma caracterização temporal de vulnerabilidades em dispositivos expostos na internet brasileira usando dados do Shodan. O principal desafio é identificar de forma confiável o mesmo dispositivo físico ao longo de múltiplas observações, já que endereços IP podem mudar com o tempo. Para isso, combinamos a identificação de IPs estáticos, usados como base verdade, com um identificador de dispositivo derivado de atributos SSL/TLS (impressão digital SHA-256 e campo common name de certificados X.509), validado com especificidade de 0,99 e AUC de 0,72. Aplicando a abordagem a dispositivos expostos entre 2024 e meados de 2025, observamos que as vulnerabilidades críticas são as mais persistentes, ativas por mais de oito meses em média, e que setores fora do polo de tecnologia, como Educação e Administração Pública, concentram as maiores proporções de risco crítico.*

1. Introdução

O número de dispositivos conectados à Internet tem crescido de forma contínua, abrangendo desde servidores e roteadores até câmeras de vigilância e equipamentos de automação [Antonakakis et al. 2017]. No Brasil, esse ecossistema é expressivo, e a

presença de dispositivos com vulnerabilidades não corrigidas representa um risco relevante para a segurança das redes nacionais [Lopes 2026]. Motores de busca especializados, como o Shodan[1] e o Censys[2], realizam varreduras contínuas nos espaços de endereçamento IPv4 e IPv6, coletando respostas de serviços que contêm metadados sobre *software*, versões, certificados criptográficos e, em alguns casos, identificadores de *hardware*. O conjunto de dados produzido por essas varreduras permite identificar dispositivos e suas vulnerabilidades sem a necessidade de sondagens ativas por parte do pesquisador.

Apesar do potencial dessas informações, a realização de análises temporais sobre a exposição a vulnerabilidades requer a identificação consistente de dispositivos ao longo do tempo. Endereços IP, frequentemente utilizados como identificadores de dispositivos, podem não ser estáveis para todos os tipos de conexão [Safi et al. 2022]. Por exemplo, provedores de acesso residencial comumente atribuem endereços dinâmicos, o que significa que um mesmo endereço IP pode corresponder a dispositivos distintos em momentos diferentes. Trabalhos anteriores tratam aspectos parciais desse problema: o DynMap [Cardoso et al. 2024] classifica blocos de endereços IP como estáticos ou dinâmicos, mas não identifica dispositivos individuais; já técnicas de sondagem ativa em múltiplas camadas [Liang et al. 2022] permitem identificar informações como o modelo e o *firmware*, mas sem considerar sua persistência ao longo de observações sucessivas.

Neste trabalho, apresentamos uma metodologia para a caracterização temporal de vulnerabilidades de dispositivos expostos na Internet brasileira. Para isso, propomos um mecanismo de detecção de mudança de estado do dispositivo baseado nos certificados SSL/TLS explorados pelo DynMap, incorporando sinais adicionais como a cadeia de certificados e o *fingerprint* JARM, o que permite rastrear equipamentos individuais mesmo sob endereçamento dinâmico. Em nossos experimentos, utilizamos dados de sondagens do Shodan coletados entre 2024 e 2025. Validamos o identificador proposto usando endereços IP estáticos como *ground truth*, alcançando especificidade de 0,99 e AUC de 0,72, o que demonstra que ele individualiza os dispositivos de forma mais precisa do que o rastreamento puramente por IP.

Com base nos registros de CVE do Shodan, classificamos o comportamento de cerca de 246 mil dispositivos segundo a estabilidade do seu perfil de vulnerabilidades ao longo de dois anos de observação. Os resultados mostram que o tempo médio de exposição varia conforme o tipo de equipamento, com os *firewalls* apresentando a maior proporção de vulnerabilidades críticas. Além disso, vulnerabilidades dessa severidade demonstram ser as mais persistentes, permanecendo ativas, em média, por mais de oito meses e envelhecendo sem mitigação junto aos dispositivos. Setores como Educação e Administração Pública concentram as maiores proporções de risco crítico.

As principais contribuições deste trabalho são: (i) uma estratégia de identificação que combina IPs estáticos e identificadores criptográficos SSL/TLS para rastrear dispositivos individuais ao longo do tempo; (ii) uma caracterização temporal, em larga escala, do comportamento de vulnerabilidades de dispositivos na Internet brasileira ao longo de dois anos de observação; e (iii) artefatos para mapear dispositivos nos dados do Shodan[3]. O restante do trabalho está organizado da seguinte forma: a seção 2 apresenta os trabalhos

[1]Shodan: https://shodan.io
[2]Censys: https://censys.io
[3]Disponível em: https://github.com/lucasmsp/tlhop-library

relacionados; a seção 3 descreve o funcionamento de um motor de busca de dispositivos conectados à Internet, como o Shodan; a seção 4 discute a metodologia proposta; a seção 5 apresenta os resultados; e, por fim, a seção 6 traz a conclusão e trabalhos futuros.

2. Trabalhos relacionados

As bases para a análise de segurança em larga escala na Internet foram estabelecidas por ferramentas de varredura como o ZMap [Durumeric et al. 2013] e por motores de busca como Censys e Shodan, que coletam e organizam serviços e protocolos em dispositivos conectados. A partir desses dados, diversos trabalhos vêm sendo desenvolvidos para explorar e avaliar aspectos da identificação de vulnerabilidades. O ShoVAT [Genge and Enăchescu 2016], por exemplo, combina resultados do Shodan com bases de vulnerabilidades conhecidas para enriquecer a busca por dispositivos vulneráveis. A ferramenta Scout [O'Hare et al. 2019] relaciona dados do Censys com o NVD para detecção passiva de vulnerabilidades (recurso não oferecido diretamente pelo Censys). Projetos como o TLHOP-Library [Ponce et al. 2024] propõem um arcabouço capaz de processar, em larga escala, dados de motores de busca de diversas fontes em uma interface única e amigável para operadores de rede e pesquisadores, reduzindo a complexidade de codificação. Já outros trabalhos [Bennett et al. 2021, Ponce et al. 2024] buscaram entender e comparar as limitações desses motores em diversos aspectos.

O problema da identificação de dispositivos em redes dinâmicas é abordado por Cardoso, Oliveira e Cunha [Cardoso et al. 2024] por meio do sistema DynMap. Os autores propõem o uso de dados públicos que podem ser obtidos por ferramentas de monitoramento, como o Shodan, para identificarem IP dinâmicos utilizando *fingerprints* de certificados X.509, obtidos através de protocolos HTTPS, e de chaves SSH. Esta técnica valida a premissa de que a estabilidade do identificador, neste caso, o IP estático, é o que permite a análise da persistência de vulnerabilidades ao longo de janelas temporais.

Para as técnicas de *fingerprint* focadas na extração de características em nível de dispositivo, o trabalho de Liayng *et. al* [Liang et al. 2022] aplica uma metodologia estruturada em três camadas distintas, analisando o equipamento sob as seguintes superfícies: camada de porta, camada de resposta de protocolo e a camada de recursos *web*. A primeira camada mapeia a superfície de exposição do dispositivo identificando portas abertas e protocolos de comunicação associados para inferir a categoria do dispositivo. A segunda camada avalia as regras de comunicação e a negociação desses protocolos, como as respostas de serviços Telnet ou SSH, para extrair informações mais específicas, como a marca e o modelo. Por fim, a terceira camada (recursos *web*) foca em aplicações HTTP/HTTPS, capturando metadados de páginas, URLs específicas e certificados SSL. Dessa forma, enquanto o DynMap soluciona o rastreamento em ambientes dinâmicos, a arquitetura em três camadas embasa a extração de metadados dos dispositivos.

Nosso trabalho se diferencia da literatura ao relacionar os identificadores de *hardware* com identificadores criptográficos para rastrear dispositivos individuais ao longo do tempo. Essa abordagem vai além da classificação por tipo de dispositivo e viabiliza a caracterização temporal de vulnerabilidades em equipamentos específicos.

3. Shodan

O Shodan é um dos motores de busca mais utilizados para o rastreamento de dispositivos e aplicações acessíveis pela Internet [Qu et al. 2024]. Motores como ele, baseia-se

na sondagem sistemática de endereços IP e portas abertas, com o objetivo de estabelecer conexões e coletar as respostas emitidas pelos serviços em execução nessas portas. Para ser capaz de se comunicar com os diversos protocolos e serviços disponíveis atualmente, o Shodan conta com diversos módulos de coleta, que implementam a lógica de comunicação e de extração de informação. Nesse contexto, cada sondagem bem-sucedida a um IP e a uma porta corresponde a um novo registro no Shodan, contendo o conteúdo da mensagem de resposta (*banner*), metadados de rede e vulnerabilidades inferidas sobre o serviço sondado e, em alguns casos, informações relacionadas, como sistema operacional, tipo de dispositivo, entre outros campos [Matherly 2015].

Capazes de sondar mais de 1500 destinos por segundo, motores como o Shodan geram grandes volumes de dados heterogêneos, com campos que variam entre aplicações e versões, o que dificulta análises longitudinais sem infraestrutura adequada. Para superar esse desafio, nossa pesquisa se baseia no arcabouço de processamento em larga escala proposto por [Ponce et al. 2024], que armazena e processa os dados em Apache Spark e permite enriquecê-los com diversas bases externas.

A escolha do Shodan, fundamenta-se em pesquisas que apontam uma maior taxa de sondagem em comparação a outros motores de busca [Bennett et al. 2021]. Os dados do Shodan utilizados neste trabalho foram disponibilizados pelo Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil (CERT.br) para o mapeamento de vulnerabilidades no espaço de endereçamento IPv4 brasileiro.

4. Metodologia

O processo metodológico deste trabalho parte de uma delimitação de três unidades centrais: dispositivo, serviço e vulnerabilidade. Um dispositivo é a unidade física ou virtual conectada à rede, identificada por um endereço IP; os serviços são aplicações que operam em portas lógicas. Quando um serviço usa versões de *software* com falhas conhecidas e não corrigidas, estabelece-se um estado de exposição prolongada, e a vulnerabilidade torna-se um risco persistente até que uma atualização seja aplicada. Assim, um único serviço pode expor simultaneamente um conjunto de n vulnerabilidades.

Para viabilizar o monitoramento desses dispositivos, nosso trabalho utiliza a metodologia do sistema DynMap para identificar a persistência de endereços IP na rede. A partir da extração da assinatura SHA-256 e do *common name* de certificados X.509 obtidos via serviços HTTPS (dupla que será identificada por **chave SSL** ao longo do artigo), é possível estabelecer um registro de identidade contínuo para o serviço, permitindo ampliar a análise para *banners* que não expõem identificadores físicos, como o MAC *address*.

A metodologia completa é exemplificada na figura 1. A partir dos dados processados pelo TLHOP-Library, realizamos a identificação (seção 4.1) e a caracterização dos dispositivos (seção 4.2) por meio de análises textuais, de portas e de protocolos. Posteriormente, enriquecemos essas informações com fontes externas para a etapa de identificação de vulnerabilidades (seção 4.3). Cada uma dessas etapas será discutida a seguir.

4.1. Identificação do Dispositivo

Identificar dispositivos a partir de dados de rede é complexo: o endereço MAC, único por *hardware*, raramente é exposto, enquanto o IP sofre com ambiguidades

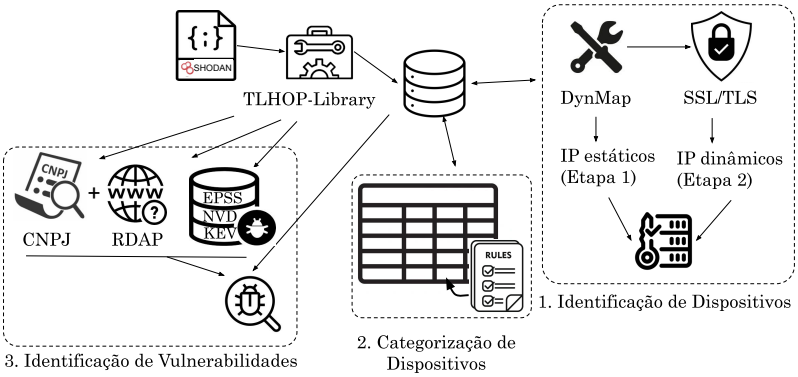

Figura 1. Esquema da arquitetura da metodologia proposta.

como endereçamento dinâmico e NAT. Para mitigar essa dinamicidade, estruturamos a identificação em duas etapas:

- **Etapa 1: Identidade baseada em IPs estáticos.** Aplicamos a metodologia do DynMap [Cardoso et al. 2024] para encontrar endereços IP estáticos no conjunto de dados. Com essa definição, foi possível delimitar uma amostra em que, com alta confiança, um IP está associado a apenas um dispositivo real.
- **Etapa 2: Identidade Criptográfica (SSL).** Visando expandir a análise para o cenário de IPs dinâmicos, foi definido um DEVICE_ID baseado no COMMON NAME (CN) de um certificado digital e outros dados do certificado SSL (seção 4.1.1). Esse identificador foi testado usando a identidade de IPs estáticos como *ground truth* e usado durante a análise de vulnerabilidades.

4.1.1. Usando informações SSL para identificar dispositivos

O identificador de dispositivo (DEVICE_ID) proposto consiste em um identificador temporal projetado para rastrear dispositivos individuais a partir de dados de varredura do Shodan. Em um cenário com IP dinâmicos, um mesmo endereço pode ser reatribuído a equipamentos diferentes ao longo do tempo, de forma que ele não pode ser usado como um identificador confiável de dispositivos. Baseado no método de identificação de serviços usado pelo DynMap [Cardoso et al. 2024], o DEVICE_ID oferece uma alternativa mais robusta, fundamentada no campo COMMON NAME do certificado SSL, enriquecido por múltiplos sinais de *fingerprinting* que permitem detectar substituições de dispositivos mesmo quando o CN permanece inalterado.

Para cada medição, são extraídos os seguintes atributos do certificado SSL: o COMMON NAME, o FINGERPRINT SHA256, a cadeia de certificados (CHAIN), o FINGERPRINT JARM e a data de expiração. Registros com valores nulos nos campos CN, *fingerprints* ou expiração são descartados. O identificador inicial é obtido por meio do *hash* MD5 do campo CN, seguindo a equação $\text{DEVICE\_ID}_0 = \text{MD5}(\text{CN})$.

Para distinguir dispositivos distintos que compartilham o mesmo CN ao longo do tempo, um mecanismo de detecção de mudança de estado é aplicado. Para cada varredura consecutiva de um mesmo dispositivo, ordenada temporalmente, um evento de mudança é sinalizado quando ao menos uma das seguintes condições é satisfeita entre o registro atual e o registro anterior: (i) a organização responsável pelo IP difere; (ii) o *fingerprint* JARM

difere, indicando alteração na configuração da pilha TLS; (iii) a cadeia de certificados difere; ou (iv) o *fingerprint* SHA256 do certificado difere e o certificado anterior tem um período de mais de 30 dias até a expiração. Esse período foi baseado na recomendação de renovação de certificado do *Let's Encrypt*[4] e é determinado porque, dado que a data de expiração está próxima, a mudança de certificado interpretado como uma renovação rotineira, preservando a identidade do dispositivo.

A cada evento de mudança, um contador c é incrementado, e o identificador final é computado como $\text{DEVICE\_ID} = \text{MD5}(\text{DEVICE\_ID}_0 \| c)$, onde $\|$ denota a concatenação. Dessa forma, mesmo que o CN permaneça inalterado, dispositivos distintos que operem sequencialmente sob um mesmo endereço IP recebem identificadores diferentes.

4.2. Análise Textual de Banners e Identificadores

Neste primeiro estágio, optou-se pela classificação de dispositivos IoT com base na metodologia de [Liang et al. 2022], centrada em *fine-grained identification*. O algoritmo utiliza portas abertas como indicadores da natureza do dispositivo, por exemplo, a porta 631 para identificar impressoras via IPP, e complementa a análise com a *Protocol-Response Layer*, extraindo informações de *banners*. Diferentemente da abordagem original, que busca detalhes como marca e modelo, nossa implementação usa os metadados textuais para validar a categoria funcional do dispositivo, garantindo uma rotulagem correta.

Para viabilizar a extração de informações a partir dos campos textuais, aplicou-se uma etapa de pré-processamento nos metadados extraídos pelo Shodan. O foco da busca por *banners* e assinaturas de fabricantes foi centralizado em quatro campos que fornecem informações complementares sobre os dispositivos, sendo essas:

- PRODUCT: Contém o nome do *software* ou *firmware* fornecedor do *banner*, sendo crucial para a identificação de marcas e modelos específicos.
- DEVICETYPE: Contém o nome do tipo de equipamento (*e.g.*, *router*, *printer*).
- DATA: Consiste no *banner* principal e bruto do serviço.
- CPE23: Refere-se ao *Common Platform Enumeration*, um texto estruturado que descreve sistemas operacionais e pacotes de *software* embutidos no dispositivo.

Em seguida, aplicou-se análise de densidade de termos nesses campos para medir a frequência e relevância das informações, identificando quais apresentavam conteúdo suficiente para uma classificação confiável. Permitindo assim, aplicar um dicionário com os termos *router, webcam, firewall, printer* e *nas* para verificar se o conteúdo textual confirma a categoria sugerida pelas portas e protocolos. Esse direcionamento garante que a classificação final seja baseada em evidências textuais relevantes e discriminativas.

Conforme exposto na tabela 1, a identificação dos dispositivos não pode depender exclusivamente dos campos textuais devido à disparidade na qualidade dos dados. Para essa análise, o registro representa cada entrada individualmente preenchido em um campo, descartando-se valores nulos ou vazios. Além disso, a quarta coluna da tabela expressa o percentual de registros úteis (que contêm as palavras-chave no dicionário) em relação ao total de registros preenchidos especificamente naquelo campo.

Os resultados revelam um contraste importante: o campo DEVICETYPE é o que contém mais informações diretas sobre os dispositivos, com cerca de 74,83% de seus

[4]https://letsencrypt.org

CAMPO	# REGISTROS	# REG. COM PALAVRAS-CHAVE	%
PRODUCT	550.710.500	2.367.404	0,43
DEVICETYPE	1.037.656	776.503	74,83
DATA	665.114.536	5.907.691	0,89
CPE23	493.595.698	971.750	0,20

Tabela 1. Ocorrência de palavras-chave de categorias IoT nos campos textuais com percentual em relação ao total de registros preenchidos no respectivo campo.

dados indicando uma categoria IoT. No entanto, ele é infrequente. Já o campo DATA possui um volume massivo de 665 milhões de registros, mas a grande maioria de seu conteúdo não contém informações relacionadas aos dispositivos IoT, visto que os termos relacionados a essas categorias aparecem em menos de 1% do total preenchido.

Dessa forma, essa limitação dos dados textuais justifica a importância de integrar múltiplas camadas de verificação, tal como proposto na metodologia de [Liang et al. 2022]. Como os campos de texto isolados não oferecem volume de informações suficientemente abrangente para todo o conjunto de dados, a incorporação da análise de portas e protocolos torna-se essencial para preencher essas lacunas. Contudo, para adequar essa lógica à natureza do nosso contexto, introduzimos adaptações estruturais no algoritmo original.

4.2.1. Algoritmo de Classificação para Dispositivos IoT

Adaptamos o algoritmo de [Liang et al. 2022] para a classificação de dispositivos IoT como um sistema de decisão. O objetivo é analisar o histórico de um dispositivo e tentar rotulá-lo em cada instante de observação. Ao final, o algoritmo aplica um consenso de classificação, em que o dispositivo recebe uma categoria final se houver unanimidade entre todos os seus registros de coleta ao longo do tempo. Essa etapa é importante para garantir a persistência da identidade do dispositivo, assegurando que apenas aqueles com comportamento homogêneo sejam classificados. O funcionamento do algoritmo é dividido em duas fases principais:

1. **Classificação por Hierarquia de Evidências:** Para cada instante de coleta no histórico do dispositivo (TIMESTAMP), o algoritmo realiza a extração e a concatenação dos metadados de portas, protocolos e campos textuais. A busca por assinaturas textuais é realizada exclusivamente sobre o conjunto de campos selecionados previamente representadas na tabela 1. A classificação em cada instância segue a ordem:
   - **Identificação por Coerência de Porta e Assinatura:** Quando uma porta padrão de algum serviço é detectada (ex.: 631 para impressoras), o algoritmo realiza uma varredura completa em todos os campos textuais em busca de assinaturas de categorias IoT. A classificação só é realizada caso haja uma correspondência única e exata: o texto deve indicar apenas uma categoria e esta deve ser idêntica à prevista para aquela porta. Caso existam termos contraditórios no texto ou múltiplas categorias detectadas, a classificação é marcada como UNKNOWN.
   - **Identificação Baseada em Fabricante e Perfil de Portas:** Na ausência de uma porta característica isolada, o algoritmo busca por assinaturas baseadas em marcas específicas. Por exemplo, a detecção das portas 81 e 82 de forma combinada constitui um indicador técnico de dispositivos Hikivision [Liang et al. 2022]. Portanto, o

algoritmo utiliza esses padrões de portas como um mecanismo para cruzar o fabricante identificado com o dicionário de assinaturas textuais. Esse processo garante que a classificação atribuída seja consistente com o *hardware*.

- **Análise de Protocolos:** Nos casos em que a configuração de portas não permite uma classificação definitiva, o algoritmo avalia os protocolos de aplicação utilizados pelo dispositivo. Certos protocolos possuem uma finalidade característica do ecossistema IoT, por exemplo, o uso do protocolo RTSP é um indicativo de dispositivos de vídeo, como *webcams*. Assim, o sistema realiza uma varredura nos metadados em busca de termos que confirmem a natureza do dispositivo com base na funcionalidade do protocolo detectado. Ao final, a classificação só é consolidada se as informações extraídas dos campos textuais validam a finalidade técnica do protocolo.
- **Identificação por Extração Textual:** Para os casos em que a camada de rede não é suficiente para classificar os dispositivos, o algoritmo realiza uma busca textual no conjunto de campos selecionados. Para evitar falsos positivos, a classificação só é validada se houver uma correspondência única e exclusiva para uma das categorias IoT do dicionário. Caso o texto apresente ambiguidade ou múltiplas referências conflitantes, o dispositivo é classificado como UNKNOWN.

2. **Consolidação por Consenso Temporal:** O dispositivo só recebe uma categoria final se todos os rótulos válidos gerados em seu histórico forem idênticos. Dessa forma, se um endereço IP apresentar comportamentos divergentes em momentos distintos, por exemplo, comportando-se como uma impressora em uma varredura e como um roteador em outra, o algoritmo invalida a classificação, rotulando o dispositivo como UNKNOWN. Isso assegura que a base de dados resultante seja composta apenas por dispositivos cuja função na rede é estável e verificável.

4.3. Identificação de Vulnerabilidades

Após a delimitação temporal e a definição da janela mínima de observação, cruzamos os dados com repositórios de vulnerabilidades e dados de identificação organizacional. Essa filtragem tem como objetivo isolar os IPs que apresentavam vulnerabilidades catalogadas e que pudessem ser atribuídos a organizações formalmente estabelecidas. O fluxo de dados foi estruturado nas seguintes etapas:

- **Etapa 1: Atribuição de Propriedade:** Os IPs foram validados para remover anomalias de formatação, descartando registros sem endereço IP válido. Em seguida, realizamos o mapeamento da titularidade da infraestrutura cruzando os identificadores de metadados com os registros do protocolo RDAP (Registration Data Access Protocol). Além disso, os identificadores de propriedade, fornecidos pelo RDAP, são cruzados com a base de dados da Receita Federal. Dessa forma, apenas os endereços IP mapeados para um Cadastro Nacional da Pessoa Jurídica (CNPJ) ativo e portadores do código de Classificação Nacional de Atividades Econômicas (CNAE) foram mantidos. Essa filtragem foi adotada para garantir que o conjunto de dados contivesse apenas informações consistentes e passíveis de validação oficial. Ao restringir a análise a registros vinculados em um CNPJ ativo, o estudo assegura a rastreabilidade e a autenticidade dos dados analisados.
- **Etapa 2: Análise sobre tuplas:** Como um único serviço pode possuir múltiplas falhas de segurança, a análise temporal das vulnerabilidades foi estruturada pela tupla (IP, PORTA, CVE), para garantir que cada vulnerabilidade em um serviço específico pudesse ser rastreada de forma independente.

- **Etapa 3: Enriquecimento:** Para avaliar o risco associado a cada vulnerabilidade, realizamos o cruzamento dos dados com as bases NVD e FIRST EPSS, obtendo métricas como o Common Vulnerability Scoring System (CVSS) e o Exploit Prediction Scoring System (EPSS). Adicionalmente, integramos as informações com o catálogo KEV (Known Exploited Vulnerabilities) da agência governamental norte-americana CISA, com o objetivo de identificar vulnerabilidades que possuem exploração ativa registrada.

A aplicação desse conjunto de regras reduziu o número de dispositivos selecionados para 246.227. Dessa forma, o conjunto de dados resultante combina dados de propriedade consistente com métricas atualizadas de severidade e evidências de vulnerabilidades já catalogadas como ameaças ativas.

5. Resultados

Nesta seção, avaliamos os resultados da nossa proposta de identificação de dispositivos e vulnerabilidades. Inicialmente, na seção 5.1, apresentamos uma caracterização do nosso conjunto de dados. Na seção 5.2, validamos nossa abordagem de identificação de dispositivos pelo identificador SSL. Para a seção 5.3, avaliamos os resultados por meio de uma caracterização conjunta dos dispositivos identificados que engloba tanto os equipamentos com IPs estáticos quanto dinâmicos. Por fim, a seção 5.4 detalha a análise temporal das vulnerabilidades identificadas na base de dados cujo objetivo é rastrear o ciclo de vida das falhas de segurança antes de uma eventual correção.

5.1. Caracterização dos Dados

Definimos a análise temporal de vulnerabilidades como a capacidade de rastrear, de forma contínua, o estado de uma falha em um serviço específico ao longo do tempo. Para isso, é preciso compreender o comportamento das varreduras: a tabela 2 resume métricas como número de sondagens, dias sem sondagem e total de portas identificadas.

MÉTRICA	MÉDIA	MÍNIMO	MÁXIMO
Sondagens (Qtd. de vezes)	6,54	1	185.691
Dias sem Sondagem (Dias avulsos)	100,55	0	577
Maior Janela entre Sondagens Consecutivas (Dias)	58,28	0	577
Total de Portas Distintas (Qtd.)	1,86	1	3.181
Total de Módulos Distintos (Qtd.)	1,79	1	132

Tabela 2. Estatísticas descritivas dos dispositivos (N = 100.567.049).

No contexto da tabela 2, uma sondagem equivale à coleta de um *banner*, ou seja, o registro de uma resposta de um serviço associado àquele endereço IP em um dado instante. A presença de *outliers* na base, como um IP recebendo mais de 185 mil sondagens e expondo 3 mil portas, representa comportamentos atípicos que refletem em infraestruturas com propósitos singulares ou configurações de rede complexas, que acabam destoando do comportamento majoritário dos dispositivos expostos na internet.

Dessa forma, para que o cálculo de tempo de exposição a uma vulnerabilidade seja sustentável, é necessário que o dispositivo possua um histórico mínimo de observações, uma vez que dispositivos efêmeros podem introduzir ruídos na análise do ciclo de vida das

vulnerabilidades. Portanto, considerando que o Shodan realiza, em média, seis sondagens por dispositivo, conforme demonstrado na tabela 2, incorporamos a análise da Função de Distribuição Acumulada (CDF), na figura 2, para obter uma delimitação mais precisa do número de sondagens dos IPs. A partir dessa distribuição, estabelecemos como critério um limite mínimo de 10 sondagens para cada dispositivo. Esse intervalo proporciona a captura de ativos com um número de varreduras superior à média global, permitindo um rastreamento temporal consistente.

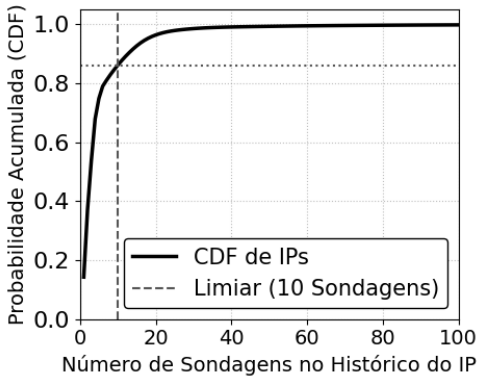

Figura 2. Distribuição Acumulada do número de sondagens por endereço IP.

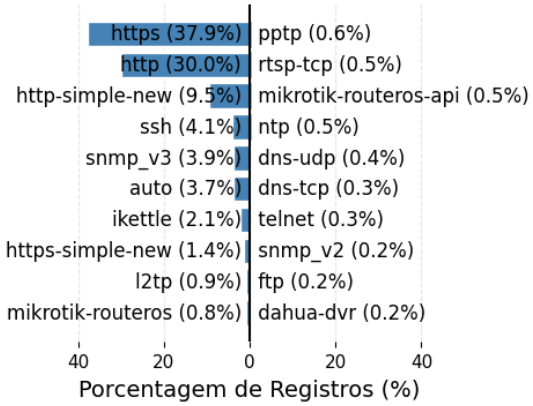

Figura 3. Módulos frequentes.

Complementando a validação temporal, é importante compreender também quais serviços operam majoritariamente nesses IPs. A figura 3 detalha a prevalência dos módulos de varredura do Shodan identificados, evidenciando que a maior parte da infraestrutura responde a requisições *web* (HTTP e suas variações).

5.2. Validação do Identificador SSL

Para avaliar o acerto do DEVICE_ID proposto, um subconjunto de endereços IP estáticos, identificados por meio do DynMap e contendo valores SSL, é utilizado como *ground truth*. A premissa é que, em IPs estáticos, a associação entre a identidade real do dispositivo e seu endereço IP é mais confiável, permitindo a construção de pares para validação.

O método de avaliação consistiu em gerar, a partir dos endereços IP estáticos, uma base com pares de medições de dispositivos e comparar a correspondência entre a identificação por IP e o DEVICE_ID. Para gerar os pares, foi usado um produto cartesiano em uma amostra de medições: foram amostrados 4.376 endereços IP, a partir dos quais foram coletadas 9.393 medições, que resultaram em aproximadamente 88 milhões de pares. Verificamos que a métrica de *especificidade*, que mede o sucesso em separar dispositivos diferentes, resultou em $0,99$ e a *sensibilidade*, capacidade de agrupar medições de um mesmo dispositivo, resultou em $0,46$. O escore AUC (Area Under the ROC Curve), que consolida a *especificidade* e *sensibilidade* em um único valor, resultou em $0,72$.

A baixa *sensibilidade* indica que o identificador está separando muitos dispositivos (alta taxa de falsos negativos). Para a análise temporal, isso pode gerar uma quebra de continuidade em alguns perfis, já que o mesmo dispositivo será separado em diferentes

CATEGORIA	# IPS ESTÁTICOS	# IPS DINÂMICOS
Firewall	22.327	151.374
Printer	156.019	39.486
Webcam	29.137	16.837
Router	21.906	15.759
NAS	698	2.863
Other IoT	13.526	15.088
Unknown	100.323.436	7.122.520
TOTAL GERAL ANALISADO	100.567.049	7.363.927

Tabela 3. Distribuição dos dispositivos identificados de IP estáticos e dinâmicos.

DEVICE_ID ao longo do tempo. Entretanto, devido à alta especificidade, os recortes tendem a ser coesos, o que aumenta a confiança na estabilidade do dispositivo ao analisar seu perfil de vulnerabilidades.

No entanto, é preciso considerar que, embora abordagens baseadas em varreduras públicas, como o DynMap, utilizem a assinatura SHA256 de certificados HTTPS como um identificador estável para rastrear serviços, observamos que essa premissa pode falhar em infraestruturas de nuvem altamente dinâmicas, como as associadas ao domínio `googleusercontent.com` (AS15169). Nossos resultados demonstram casos em que o FINGERPRINT SHA256 do certificado permanece inalterado para um mesmo par de IP e porta durante longos períodos, o que levaria a uma inferência de estaticidade pela lógica do DynMap. No entanto, a análise do FINGERPRINT JARM nesses mesmos registros revela variações temporais frequentes. Como o JARM captura nuances da implementação do SSL/TLS e do *hardware* subjacente, sua mutabilidade na presença de um certificado persistente indica a rotação de instâncias de *hardware* ou mudanças na configuração do servidor por trás de balanceadores de carga e *proxies* de borda [Bennett et al. 2021].

Ao identificar essa instabilidade, o DEVICE_ID apresenta uma granularidade mais fina, sendo capaz de contornar a falsa percepção de um *host* único e revelar a natureza compartilhada de dispositivos em nuvem. Associando isso ao fato de que, como discutido em [Bai and Fossaceca 2025], um AUC Score acima de 0,7 é aceitável, dado a complexidade desse tipo de tarefa.

5.2.1. Relação do identificador com a classificação IoT

Nesta etapa, analisou-se em que medida o DEVICE_TYPE identificado complementa o identificador SSL. Para isso, foi criado um rótulo composto pela concatenação do identificador com a categoria IoT do dispositivo. As quantidades de cada tipo de dispositivo com alocação dinâmica são apresentadas na tabela 3.

A análise comparativa revelou que os índices de falsos positivos e falsos negativos permaneceram inalterados após a introdução da classificação IoT. Isso indica que o DEVICE_ID baseado em SSL possui, por si só, alta capacidade de discriminação entre categorias IoT, reforçando a detecção de *hardware* discutida anteriormente. Além disso, o DEVICE_ID adiciona granularidade à análise ao identificar dispositivos específicos, e não apenas classes gerais de funcionamento.

DATA		PRODUCT		CPE23		DEVICETYPE	
TERMO	%	TERMO	%	TERMO	%	TERMO	%
gmt	68,58	httpd	62,37	cpe	69,51	router	0,10
cache	64,77	cloudfront	61,01	cloudfront	61,01	wap	0,01
error	61,85	mikrotik	4,19	amazon	61,01	switch	0,01
net	61,67	sshd	3,63	linux	3,94	misc	0,00
via	61,46	dropbear	3,61	kernel	3,85	security	0,00
cf	61,25	test	2,08	jquery	1,47	pbx	0,00
pop	61,20	bandwidth	2,08	nginx	0,99	firewall	0,00
amz	61,18	nginx	0,94	f5	0,98	webcam	0,00
cloudfront	61,16	akamaighost	0,94	apache	0,88	device	0,00
915	61,00	apache	0,91	bootstrap	0,53	media	0,00

Tabela 4. Top 10 termos frequentes por atributo textual para os dispositivos da classe UNKNOWN com percentual calculado sobre o total de registros da base de dispositivos classificados como desconhecidos.

5.3. Classificação Final dos Dispositivos

Após adaptar a metodologia de [Liang et al. 2022], o modelo foi aplicado aos IPs estáticos e dinâmicos da base (tabela 3). Observa-se que mais de 107 milhões de registros foram classificados como UNKNOWN (99,76% dos IPs estáticos), resultado do critério conservador da seção 4.2.1, que exige unanimidade no histórico do dispositivo para priorizar a confiabilidade em detrimento da cobertura.

A fim de compreender a natureza desses dispositivos e identificar possíveis padrões não mapeados, utilizamos uma função de extração de termos frequentes, mapeando as palavras mais comuns presentes nos campos mencionados em 4.2. O resultado, apresentado na tabela 4, revela uma dominância de vocabulários relacionados a infraestruturas em nuvem. Vocabulários atrelados a provedores de hospedagem como "cloudfront" e "amazon" ao longo dos campos DATA, PRODUCT e CPE23, aliados à alta frequência de serviços de rede genéricos, como "httpd", indicam que uma parcela dos dados pertence a servidores respondendo por portas padrão. Esses dispositivos, mesmo classificados como UNKNOWN, permanecem integrados ao ecossistema de dados do estudo. Contudo, devido à ambiguidade dos metadados textuais, o algoritmo 4.2.1 opta por não lhes atribuir uma categoria funcional IoT específica.

5.4. Análise Temporal da Exposição dos Dispositivos

A seguir apresentamos um estudo de caso das vulnerabilidades expostas nos dispositivos identificados, relacionando fatores como tipo de dispositivo e setores economicos das organizações responsáveis. Nesse contexto, a tabela 5 detalha as características de exposição por tipo de dispositivo e a proporção daqueles que apresentam um alto CVSS.

Em relação ao tempo de exposição, o DEVICE_ID demonstrou uma capacidade superior na individualização dos dispositivos durante as sondagens. Esse impacto é visto na exposição média dos dispositivos: enquanto o rastreamento por IP estático resulta em uma média de 193 dias, a identificação por meio do DEVICE_ID apresenta uma média de 156 dias. Analisando os resultados na tabela, a categoria UNKNOWN representa, majoritariamente, infraestruturas em nuvem e servidores *web*, contudo, a taxa de vulnerabilidade de alta severidade é consideravelmente baixa. Esse comportamento é coerente com o rigor do monitoramento de segurança inerente a esse setor. Em contraste, equipamentos

TIPO	EXPO. MÉDIA	EXPO. MÁXIMA	# DISPOSITIVOS	% CVSS ≥ 9
Router	370	564	16	$\leq$ 0,00
Webcam	291	579	182	17,03
NAS	289	567	20	85,00
Printer	211	570	65	43,08
Firewall	185	543	16	6,25
Other IoT	391	579	695	90,65
Unknown	219	579	245.233	11,53

Tabela 5. Exposição de dispositivos (por dia) e registro CVSS por dispositivo.

SEVERIDADE	EXPOSIÇÃO MÉDIA	EXPOSIÇÃO MÁXIMA	# CVE
Crítica	258,27	579	32
Alta	215,87	579	12
Média	224,41	579	119
Baixa	191,54	579	76

Tabela 6. Métricas de exposição (em dias) por nível de severidade.

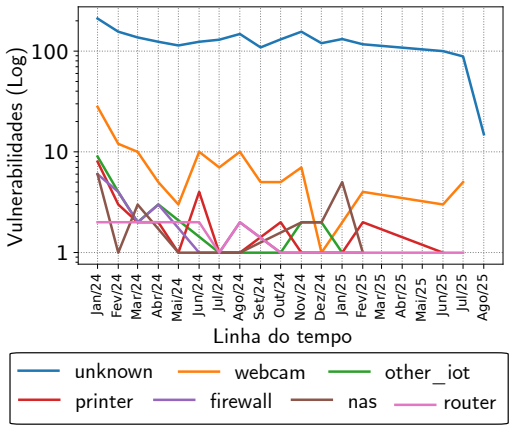

Figura 4. Vulnerabilidades por tipo de dispositivo.

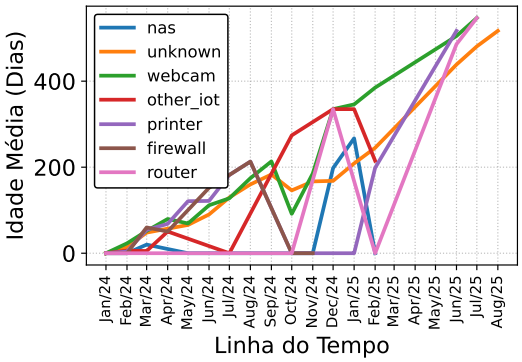

Figura 5. Progressão da idade das vulnerabilidades por tipo de dispositivo.

periféricos como impressoras, por não serem tradicionalmente vistos como alvos prioritários, tendem a permanecer expostos por períodos mais longos.

Para relacionar o tempo de exposição à severidade das vulnerabilidades, a pontuação CVSS foi dividida em quatro classes: *crítica* (maior que 9), *alta* (maior que 8), *média* (maior que 6) e *baixa* caso contrário. A tabela 6 apresenta as informações de exposição para cada uma dessas categorias. Observa-se que as vulnerabilidades de severidade crítica possuem um tempo médio de exposição superior, apresentando uma diferença de aproximadamente dois meses quando comparadas àquelas de baixo risco.

Além disso, a análise temporal das vulnerabilidades por tipo de dispositivo pode ser vista de forma conjunta nas figuras 4 e 5. A análise dos gráficos revela que, embora o número de vulnerabilidades se mantenha estável ao longo do tempo, essas falhas tendem a envelhecer junto com os dispositivos, indicando uma ausência de mitigação nos dispositivos afetados. Esse cenário torna-se particularmente crítico ao considerar que a severidade média das vulnerabilidades mapeadas é alta.

Analisando as métricas por setor, a tabela 7 revela que a seção de Informação e Comunicação concentra a maioria dos equipamentos, o que é justificado por sua

SETOR (CNAE)	# DISPOSITIVOS	% CVSS ≥ 9
J - Informação e Comunicação	220.653	10,75%
P - Educação	6.252	27,98%
G - Comércio e Reparação	6.054	17,53%
S - Outras Atividades de Serviços	3.397	24,93%
O - Administração Pública	2.961	20,91%
N - Atividades Administrativas	2.426	13,40%
M - Atividades Profissionais	1.368	12,28%
C - Indústrias de Transformação	912	18,64%
K - Atividades Financeiras	570	14,21%
F - Construção	560	11,96%

Tabela 7. Distribuição do volume de dispositivos e porcentagem de vulnerabilidades críticas por setor CNAE (Top 10).

composição, que engloba divisões de telecomunicação e tecnologia da informação. Apesar do alto volume de dispositivos, a criticidade nesse setor mantém-se baixa (10,75%), indicando uma gestão de segurança mais proativa. Em contrapartida, a concentração de mais de 20% de falhas críticas nos setores de Educação e Administração Pública é consistente com desafios do setor público, como restrições orçamentárias e burocracia nos processos de aquisição de infraestrutura de TI [Georg et al. 2023].

6. Conclusão e Trabalhos futuros

Neste trabalho, apresentamos uma metodologia para identificar, classificar e monitorar, em larga escala, dispositivos expostos na Internet brasileira. Para contornar a ambiguidade dos endereços de rede, integramos identidades baseadas em IPs estáticos (via DynMap) a identidades criptográficas derivadas de certificados SSL/TLS, viabilizando o rastreamento de equipamentos individuais mesmo sob endereçamento dinâmico. Sobre essa base, adaptamos um algoritmo de classificação em múltiplas camadas que cruza portas, protocolos e metadados textuais, o que permitiu categorizar centenas de milhares de equipamentos, com destaque para a forte presença de *firewalls* e impressoras.

A caracterização revelou que a maioria dos dispositivos não classificados (UNKNOWN) corresponde, provavelmente, a infraestruturas de nuvem e servidores *web*, como indica a predominância de termos como “cloudfront” e “amazon” e do serviço “httpd”. Quanto ao panorama de vulnerabilidades, observamos um contraste claro entre setores: ambientes de nuvem e o setor de tecnologia exibem baixas taxas de falhas de alta severidade, enquanto dispositivos periféricos e equipamentos de setores não tecnológicos, como Educação, concentram as maiores proporções de risco crítico. A análise temporal agrava esse quadro ao mostrar que as falhas críticas não só permanecem expostas por mais tempo, como também envelhecem continuamente, sem a devida correção.

Para trabalhos futuros, propomos estender a classificação para caracterizar infraestruturas de nuvem e servidores, avaliando o uso de aprendizado de máquina em etapas desse processo. Como essas classes concentram a maioria dos dispositivos do conjunto de dados, o objetivo é aprimorar o cruzamento de perfis de portas e protocolos com um novo dicionário de assinaturas voltado a esses tipos de infraestrutura.

Declaração sobre uso de Inteligência Artificial

Ferramentas de Inteligência Artificial foram utilizadas exclusivamente para apoio editorial. Todas as decisões de pesquisa, métodos, análises e interpretações foram realizadas integralmente pelos autores, sem que essas ferramentas influenciassem o conteúdo científico ou as conclusões do trabalho.

Agradecimentos

Este trabalho foi parcialmente financiado pelo NIC.br, RNP, FAPEMIG, CNPq, CAPES, MASWEB, INCT-Cyber e IAIA (INCT para IA).

Referências

Antonakakis, M., April, T., Bailey, M., Bernhard, M., Bursztein, E., Cochran, J., Durumeric, Z., Halderman, J. A., Invernizzi, L., Kallitsis, M., Kumar, D., Lever, C., Ma, Z., Mason, J., Menscher, D., Seaman, C., Sullivan, N., Thomas, K., and Zhou, Y. (2017). Understanding the Mirai Botnet. In *26th USENIX Security Symposium (USENIX Security 17)*.

Bai, K. Z. and Fossaceca, J. M. (2025). EM-AUC: A Novel Algorithm for Evaluating Anomaly Based Network Intrusion Detection Systems. *Sensors*, 25(1):21.

Bennett, C., Abdou, A., and van Oorschot, P. C. (2021). Empirical scanning analysis of Censys and Shodan. In *Proceedings of the 2021 Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb)*.

Cardoso, G., Oliveira, L., and Ítalo Cunha (2024). Identificação de Endereços IP Dinâmicos com Dados Públicos. In *Anais do XXIV Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*.

Durumeric, Z., Wustrow, E., and Halderman, J. A. (2013). ZMap: Fast internet-wide scanning and its security applications. In *Proceedings of the 22nd USENIX Security Symposium*.

Genge, B. and Enăchescu, C. (2016). ShoVAT: Shodan-based vulnerability assessment tool for internet-facing services. *Security and Communication Networks*, 9(15):2696–2714.

Georg, M., Rodrigues, W., Alves, C., Silveira Junior, A., and Nunes, R. (2023). Os desafios da Segurança Cibernética no setor público federal do Brasil: estudo sob a ótica de gestores de tecnologia da informação. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, E54:602–616.

Liang, C., Yu, B., Xie, W., Wang, B., and Peng, W. (2022). Fine-grained identification for large-scale iot devices: A smart probe-scheduling approach based on information feedback. *Applied Sciences*, 12(16).

Lopes, B. S. (2026). Segurança da Informação em Ambientes de Internet das Coisas (IoT): Desafios, Vulnerabilidades e Estratégias de Proteção. *Revista Ibero-Americana de Humanidades, Ciências e Educação*, 12(3):1–22.

Matherly, J. (2015). *Complete Guide to Shodan: Collect. Analyze. Visualize. Make Internet Intelligence Work for You*. Leanpub.

O'Hare, J., Macfarlane, R., and Lo, O. (2019). Identifying Vulnerabilities Using Internet-Wide Scanning Data. In *Proceedings of the IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGS3)*.

Ponce, L., Cunha, I., Matos, I., Ítalo Cunha, Fazzion, E., Hoepers, C., Steding-Jessen, K., Chaves, M., Guedes, D., and Meira Jr., W. (2024). Arcabouço Multi-motor para Detecção de Vulnerabilidades na Internet Brasileira. In *Anais do XLII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. SBC.

Qu, J., Ma, X., Liu, W., Sang, H., Li, J., Xue, L., Luo, X., Li, Z., Feng, L., and Guan, X. (2024). On Smartly Scanning of the Internet of Things. *IEEE/ACM Transactions on Networking*, 32(2):1019–1034.

Safi, M., Dadkhah, S., Shoeleh, F., Mahdikhani, H., Molyneaux, H., and Ghorbani, A. (2022). A Survey on IoT Profiling, Fingerprinting, and Identification. *ACM Transactions on Internet of Things*, 3(4):1–22.