UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MATIAS ROMÁRIO PINHEIRO DOS SANTOS

CLASSIFICAÇÃO DE TRÁFEGO E DOS DISPOSITIVOS DE IOT ATRAVÉS DO FLUXO DE REDE E INSPEÇÃO DA CARGA ÚTIL DOS PACOTES

FORTALEZA

2018

MATIAS ROMÁRIO PINHEIRO DOS SANTOS

CLASSIFICAÇÃO DE TRÁFEGO E DOS DISPOSITIVOS DE IOT ATRAVÉS DO FLUXO DE REDE E INSPEÇÃO DA CARGA ÚTIL DOS PACOTES

Dissertação apresentada ao Curso de do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Redes de Computadores

Orientador: Prof. Dr. Arthur de Castro Callado

FORTALEZA

2018

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S236c Santos, Matias Romário Pinheiro dos.
Classificação de tráfego e dos dispositivos de IoT através do fluxo de rede e inspeção da carga útil dos pacotes / Matias Romário Pinheiro dos Santos. – 2018.
94 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação, Fortaleza, 2018.
Orientação: Prof. Dr. Arthur de Castro Callado.

1. Internet das Coisas. 2. Tráfego de rede. 3. Inspeção do conteúdo. 4. Classificação de tráfego. I. Título.

CDD 005

MATIAS ROMÁRIO PINHEIRO DOS SANTOS

CLASSIFICAÇÃO DE TRÁFEGO E DOS DISPOSITIVOS DE IOT ATRAVÉS DO FLUXO DE REDE E INSPEÇÃO DA CARGA ÚTIL DOS PACOTES

Dissertação apresentada ao Curso de do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Redes de Computadores

Aprovada em: 16 de Outubro de 2018

BANCA EXAMINADORA

Prof. Dr. Arthur de Castro Callado (Orientador)
Universidade Federal do Ceará (UFC)

Profa. Dra. Maria Rossana de Castro Andrade
Universidade Federal do Ceará (UFC)

Prof. Dr. Danielo Gonçalves Gomes
Universidade Federal do Ceará (UFC)

Profa. Dra. Judith Kelner
Universidade Federal de Pernambuco (UFPE)

À minha família, por sua total confiança em mim.
Mariana, minha esposa, amo-lhe mais que tudo
nessa vida.

AGRADECIMENTOS

Primeiramente agradeço minha esposa, Mariana, por sua paciência e companheirismo ao longo desses anos.

Aos meus pais, Pinheiro e Iranir, pela dedicação, carinho, amor e apoio incondicional.

Aos meus irmãos, Ronaldo, Tiago e Marcelo, pelo incentivo.

Aos meus queridos sogros.

Ao meu orientador, Arthur Callado, pelo incentivo, amizade, apoio, contribuição e acompanhamento.

Aos professores que me auxiliaram e puderam transmitir seus conhecimentos, com excelência, durante todo o mestrado.

À Universidade Federal do Ceará (UFC), pela oportunidade de fazer o mestrado e sua qualidade de ensino.

Aos amigos e colegas que fiz durante esse processo.

Aos amigos Anderson Almada, Tiago Martins, Priscylla e muitos outros que ajudaram direta e/ou inderetamente nessa conquista.

À todos que diretamente ou indiretamente fizeram parte desta conquista.

“Apesar dos nossos defeitos, precisamos enxergar que somos pérolas únicas no teatro da vida e entender que não existem pessoas de sucesso ou pessoas fracassadas. O que existe são pessoas que lutam pelos seus sonhos ou desistem deles.”

(Augusto Cury)

RESUMO

Internet das coisas surge como um paradigma computacional que promove a interconexão de objetos inteligentes à Internet e permite interação, eficiência operacional e comunicação. Com a crescente inclusão, na rede, de objetos inteligentes que possuem características como diversidade, heterogeneidade, mobilidade e baixo poder computacional, é fundamental desenvolver mecanismos que permitam gestão e controle. Além disso, é importante identificar se os ativos estão funcionando corretamente ou têm anomalias. As técnicas de classificação de tráfego são importantes para auxiliar a análise de rede e para lidar com muitos outros aspectos chave, como segurança, gerenciamento, controle de acesso e aprovisionamento de recursos. Mecanismos de classificação de tráfego ainda apresentam dificuldades quando aplicados em ambientes dinâmicos e sem conhecimento dos serviços, especialmente com criptografia. A fim de promover a classificação dos dispositivos de rede e o tráfego, especialmente de IoT, é apresentada uma técnica que utiliza a floresta aleatória (Random Forest), um algoritmo de aprendizado automático supervisionado, juntamente com a inspeção do conteúdo dos pacotes para este fim. Outrossim, é utilizado o mesmo algoritmo para executar a classificação do tráfego de rede através das características extraídas do fluxo de rede. Ao final desta dissertação, será apresentada a estratégia proposta em cenários de IoT e os resultados adquiridos.

Palavras-chave: Internet das Coisas. Tráfego de rede. Inspeção do conteúdo. Classificação de tráfego.

ABSTRACT

Internet of things arises as a computational paradigm that promotes the interconnection of intelligent objects to the Internet and allows interaction, operational efficiency, and communication. With the growing inclusion, in the network, of intelligent objects that have characteristics such as diversity, heterogeneity, mobility, and low computational power, it is essential to develop mechanisms that allow management and control. In addition, it is important to identify whether the assets are functioning properly or have anomalies. Traffic classification techniques are important to assist network analysis and to handle many other key aspects, such as security, management, access control, and resource supply. Traffic classification mechanisms still present difficulties when applied in dynamic environments and without knowledge of services, especially with cryptography. In order to promote the classification of the network devices and traffic, especially IoT, a technique is presented that uses the random forest (random forest), an automated learning algorithm supervised, together with the inspection of the contents of the packages To this end. Additionally, the same algorithm is used to perform the classification of network traffic through the characteristics extracted from the network flow. At the end of this dissertation, the proposed strategy will be presented in IoT scenarios and the results.

Keywords: Internet of Things. Network traffic. Packet Inspection. Traffic classification.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CART	*Classification and regression tree*
D2D	*Device to Device*
DDos	*Distributed Denial of Service)*
DNS	*Domain Name System*
DPI	*Deep Packet Inspection*
FN	*False Negative*
FP	*False Positive*
IANA	*Internet Assigned Numbers Authority*
IAT	*Inter-Arrival Time*
IDS	*Intrusion detection System*
IoT	*Internet of Things*
IPS	*Intrusion prevention system*
ISPs	*Internet service providers*
MAC	*Media Access Control*
ML	*Machine Learning*
P2P	*Peer-to-peer*
QoS	*Quality of Service*
RF	*Random Forest*
TN	*True Negative*
TP	*True Positive*
Voip	*Voice over Internet Protocol*

SUMÁRIO

1 INTRODUÇÃO

Esta dissertação apresenta uma estratégia para classificação de tráfego e dispositivos de Internet das coisas através de uma abordagem combinada entre *Machine Learning* (ML), por intermédio do algoritmo *Random Forest* (RF), e inspeção do conteúdo dos pacotes (*Deep Packet Inspection* (DPI)). Além disso, promove a classificação do tráfego de rede gerado ao fazer uso do mesmo algoritmo em características estatísticas extraídas do fluxo. As próximas seções deste capítulo estão estruturadas da seguinte forma: seção **1.1** apresenta a contextualização; em seguida, a seção **1.2** aborda a motivação, as questões que norteiam essa pesquisa e descreve a problemática abordada; a seção **1.3** estabelece os objetivos; a seção **1.4** descreve a metodologia empregada e os resultados esperados; e, por fim, a seção **1.5** apresenta a forma como estão organizados os demais capítulos da dissertação.

1.1 Contextualização

De acordo com Kurose e Ross (2012) e Solomon *et al.* (2014), a infraestrutura da rede tradicional possui em sua composição uma grande variedade de equipamentos, como *switches*, roteadores e dispositivos intermediários. Esses equipamentos promovem a construção de uma grande variedade de redes de dispositivos. De acordo com Atzori *et al.* (2010) e Lin *et al.* (2017), com o advento de Internet das Coisas (*Internet of Things* (IoT)), que conecta os dispositivos típicos do cotidiano à rede, características como a conectividade, a mobilidade e a heterogeneidade configuram-se como condições fundamentais à sua concepção em escala global. Ademais, as redes apresentam um crescimento de complexidade e proporcionam dificuldades a uma gestão eficaz à medida que novos usuários são adicionados; aumenta o consumo de dados; apresenta comportamento pervasivo; densidade de conexão e diversidade de perfis dos usuários. O monitoramento das redes, a detecção de anomalias, a avaliação e a análise de desempenho de rede resultam em mecanismos de suporte para gerentes e administradores (CALLADO *et al.*, 2009) (NGUYEN; ARMITAGE, 2008). Seu uso permite o entendimento do comportamento, da performance e da confiabilidade da rede. Dessa forma, esses recursos proporcionam suportes confiáveis e eficazes para resoluções de problemas complexos.

A Internet das coisas apresenta-se, de acordo com a literatura, como um novo paradigma computacional capaz de integrar uma grande variedade de sistemas heterogêneos, promovendo a conexão de dados, pessoas, objetos e aplicações através da Internet (ATZORI

et al., 2010). Segundo Lin *et al.* (2017) e Atzori *et al.* (2010), uma grande variedade de dispositivos móveis e objetos, tais como máquinas de lavar, geladeiras, *smartphones* e câmeras, está sendo incorporada à Internet em escala crescente. As capacidades de interação desses dispositivos com diversos tipos distintos de ambientes acabam por tornar estes ecossistemas em um paradigma de uso global e que aos poucos se apresenta incluída na rede mundial. Alguns desses dispositivos possuem capacidade de processamento e comunicação aliadas aos sensores, promovendo integração, interação e troca mútua de dados para realizar atividades em conjunto, o que pode ocorrer de forma autônoma, ubíqua e pervasiva (CHEN *et al.*, 2017). Além disso, a distribuição dos dispositivos de IoT ocorre nos mais variados ambientes, desde casas, escritórios, minas de ouro, petroleiros e em outros locais de difícil acesso (ATZORI *et al.*, 2010).

Decorrente da larga distribuição dos dispositivos de IoT, gerentes e administradores da infraestrutura podem não ter conhecimento de todos os ativos presente na rede, dificultando a sua gestão. Em um ambiente como *smart cities*, como afirma Zanella *et al.* (2014), por exemplo, é fundamental dispor de capacidade de gerenciamento da rede, visto que em tal paradigma tende a haver muitas aplicações e dados sensíveis, sendo fundamental criar mecanismos de suporte que garantam aos administradores da rede segurança, gerência, priorização do tráfego IoT (dependendo do cenário), status dos dispositivos, controle de acesso e detecção de anomalias. Os desafios mencionados podem ser atenuados através de uso de técnicas de classificação de tráfego de rede, que são utilizadas em diversos campos (*Machine Leaning*, por exemplo). De acordo com Wang (2013) e Callado *et al.* (2009), a classificação de tráfego permite uma visão refinada das aplicações que circulam na rede e o seu correto uso possibilita aos administradores suporte e segurança para garantir privacidade de acesso em aplicações ou até mesmo bloquear certos tipos de tráfego indesejado (*Peer-to-peer* (P2P), por exemplo). Ainda segundo Wang (2013), os mecanismos de segurança utilizam-se constantemente da classificação do tráfego como núcleo do sistema (*Firewalls* e Sistemas de detecção de intrusão, por exemplo), o intuito é impossibilitar atividades maliciosas como negação de serviço distribuída (*Distributed Denial of Service)* (DDos)), *Malwares* e acesso indevido. Recentes ataques foram realizados em ambientes de Internet das coisas, como ocorre em Tellez *et al.* (2016), o que permite afirmar que técnicas de intrusão, constantemente utilizadas em ataques na Internet, têm impacto direto em ecossistemas de IoT.

A crescente complexidade do tráfego de rede traz grandes desafios às técnicas de classificação, promovendo sempre novos estímulos decorrentes dos avanços na engenharia de

tráfego de redes. Com isso, à medida que a Internet evolui, novas propostas de técnicas de classificação de tráfego têm surgido para garantir acurácia em seus resultados, como ocorre desde o surgimento da classificação por portas até os modelos estatísticos largamente empregados atualmente (NGUYEN; ARMITAGE, 2008)(FINSTERBUSCH *et al.*, 2014).

A classificação por portas baseia-se em atribuir um tipo específico de tráfego a uma porta padrão (*Internet Assigned Numbers Authority* (IANA)). Essa metodologia possuía boa acurácia na classificação até o uso mais constante das portas aleatórias, o que provocou uma considerável queda na acurácia (ficando inferior a 70% nos melhores casos) (MOORE; PAPAGIANNAKI, 2005) e (MADHUKAR; WILLIAMSON, 2006). Em decorrência da grande queda na acurácia, surgiu a classificação do conteúdo dos pacotes, ou *payload*. Essa técnica, chamada DPI (*Deep Packet Inspection*), baseia-se na inspeção do conteúdo dos pacotes para classificá-los, contornando assim as deficiências da classificação por portas (quanto ao uso de portas aleatórias). Devido à necessidade de inspeção do conteúdo dos pacotes, essa metodologia acabou não despertando a atenção dos pesquisadores por infringir a privacidade dos usuários e por ser ilegal em diversos países (FINSTERBUSCH *et al.*, 2014). Além dos aspectos legais e da inconveniência de infringir a privacidade dos usuários, o uso mais frequente de protocolos de criptografia tornaram a sua acurácia menor. Essa nova problemática promoveu o surgimento de outros modelos de classificação, como o estatístico (NGUYEN; ARMITAGE, 2008) e (NAMDEV *et al.*, 2015), que baseia-se no uso de características estatísticas do fluxo do tráfego de rede (tamanho do pacote e tempo de chegada, por exemplo) para realizar a classificação.

Como já afirmado, a disseminação dos dispositivos IoT promove um significativo aumento no número de objetos, serviços e protocolos conectados. De acordo com Kawai *et al.* (2017), urge a necessidade de promover a identificação desses dispositivos, cujo objetivo é realizar uma série de ações, dentre as quais destacam-se as de segurança. Os autores afirmam que é possível identificá-los utilizando métodos estatísticos, aplicando algoritmos de aprendizado de máquina supervisionado em padrões de comunicação, uma vez que as características de comunicação de dispositivos de IoT diferem-se dos dispositivos tradicionais de rede (laptops e tablets, por exemplo). Ainda relacionado a ecossistemas de IoT, os autores Egea *et al.* (2017) descrevem a necessidade e a importância de classificar o tráfego em redes IoT com intuito de melhorar o desempenho, otimizar seus recursos e priorizá-lo. Os autores atestam que separar o tráfego dos sensores nos ambientes e realizar a priorização é fundamental para que em situações emergenciais faça-se o possível para evitar certas ocorrências, como danos materiais e humanos.

Após a leitura de (SIVANATHAN *et al.*, 2017)(MEIDAN *et al.*, 2017b), (MIETTINEN *et al.*, 2017) é possível afirmar que realizar a classificação dos dispositivos IoT permite uma série de vantagens e aplicações, dentre as quais pode-se citar: classificação do tráfego em IoT ou Não-IoT; identificação de padrões de comunicação; aprovisionamento de recursos; segurança. Várias propostas e técnicas, presentes na literatura, vêm sendo utilizadas com intuito de classificá-los, nesse contexto, o uso de técnicas baseadas em *fingerprinting* em comunicação *wireless* é uma das mais comuns. Porém, segundo Miettinen *et al.* (2017), o seu emprego apresenta desvantagens que limitam a usabilidade em diversas situações cruciais, principalmente devido ao uso excessivo de recursos e questões relacionadas à segurança, enquanto outras técnicas, focadas na identificação utilizando *hardware* ou drivers específicos (MAURICE *et al.*, 2013), possuem uso limitado para situações ou objetos específicos. Por exemplo, Kohno *et al.* (2005) propõem um método para *fingerprinting* de dispositivos físicos através da exploração da implementação do protocolo TCP. Os autores fazem uso da opção TCP *timestamp* na saída dos pacotes para explorar informações internas do *Clock*. Ao final, utilizam os desvios microscópicos no padrão do ciclo do *Clock* para identificá-los. Entretanto, uma limitação surge quando os dispositivos comunicam-se utilizando o UDP, como ocorre em redes de IoT que fazem uso de protocolos de camada de aplicação que a utiliza (CoAP, por exemplo) (BORMANN *et al.*, 2012).

Relacionado à classificação de dispositivos de IoT, a literatura apresenta abordagens e análises utilizadas com este intuito. Exemplo disso é Sivanathan *et al.* (2017), que analisaram a rede em um cenário emulado de *smart campus* e afirmam que dispositivos IoT tendem a usar um número limitado e específico de protocolos de camada de aplicação, além de uma quantidade de solicitações DNS singular e bem inferior aos dispositivos Não-IoT. Segundo Apthorpe *et al.* (2017), O uso do endereço MAC (*Media Access Control* (MAC)), na identificação de grupos de dispositivos, facilita a aplicação de análise de tráfego de rede e consultas de solicitações *Domain Name System* (DNS) no processo, uma vez que limita o número de possibilidades existentes.

A atual pesquisa incide sobre temas de segurança e controle de rede, principalmente relacionados aos dispositivos IoT, visto a necessidade de identificar os ativos presentes na rede corporativa ou residencial. As aplicações desenvolvidas através dos dados coletados dos dispositivos IoT são diversos, dentre os quais podemos citar o controle de temperatura do ambiente, monitoramento dos usuários e detecção de problemas cardíacos. Em um contexto no qual dispositivos estão sempre conectados, coletando e produzindo novos dados, detectar a indisponibilidade, ou mesmo a presença de dispositivos sem autorização de acesso, torna-se

um relevante problema de segurança. Exemplo disso é o trabalho de Miettinen *et al.* (2017), que motivados por questões de segurança em dispositivos IoT, realizaram a coleta de dados dos objetos em um cenário real, utilizando-os para identificar vulnerabilidades potenciais. Uma das formas de mitigar os riscos mencionados, que podem comprometer os dispositivos, é prover, intencionalmente, um maior controle da rede, restringindo o acesso a dispositivos não autorizados, e uma das formas é mediante a sua classificação.

1.2 Motivação

Hurlburt *et al.* (2012) mencionam que o início da Internet se deu para promover a comunicação entre grupos restritos de pessoas. As evoluções naturais da Internet a promoveram a um meio de comunicação entre pessoas e organizações, removendo os limites anteriores. Porém, surge um novo paradigma, chamado de IoT, capaz de adicionar objetos típicos do dia-a-dia aos meios de comunicação, promovendo novos mecanismos de interlocução e interação com o ambiente.

Segundo Lin *et al.* (2017) e Atzori *et al.* (2010), o desenvolvimento de ecossistemas de IoT gera uma expectativa de melhoria da qualidade de vida, aumento da segurança e expansão do desempenho em uma série de atividades. Entretanto, seu constante aprimoramento acaba promovendo sempre novos obstáculos a sua concepção, principalmente na utilização otimizada de seus recursos, uma vez que limitados, tornando-a objeto de estudo e desenvolvimento tanto na academia quanto na indústria. Segundo a Gartner (2017), a rápida proliferação dos dispositivos nos últimos anos resultou em uma elevada quantidade de objetos conectados e estima-se que até 2020 serão 20 bilhões de dispositivos conectados ao redor do mundo.

De acordo com Apthorpe *et al.* (2017), os dispositivos de IoT, conectados em rede, possuem sensores que estão sempre ativos e coletando dados dos usuários em seu cotidiano, transmitindo-os para meios externos, tipicamente a nuvem do fabricante (*Cloud Computing*). Exemplo disso são as atividades físicas do dia-a-dia, cotidiano das crianças ou até mesmo atividades conjugais. A coleta dos dados tende a se tornar um relevante problema de segurança, uma vez que os usuários dos dispositivos IoT possuem resistência quanto a certos comportamentos ou atividades das quais gostariam que não fosse feita a coleta ou acessadas, dentre elas temos a intimidade, uso de mídias, a aparência do usuário e a culinária (CHOE *et al.*, 2011). Nesse cenário, os usuários dos dispositivos esperam que os fabricantes os permitam maior controle. Porém, a rede à qual os dispositivos estão conectados é suscetível a diversas formas de invasões

de privacidade, dentre as quais podemos citar as inspeções do conteúdo dos pacotes pelas próprias ISPs e o WiFi *eavesdroppers* (APTHORPE *et al.*, 2017).

O risco de exposição dos usuários, mediante a análise do tráfego de rede, e uso de DPI pelos ISPs foi debatido e, em 2017, uma lei[1] nos Estados Unidos foi aprovada com intuito principal de coibir o acesso ao conteúdo dos usuário pelos *Internet service providers* (ISPs), bloqueando-os. Porém, limitá-las não é uma garantia total de segurança para os usuários desses dispositivos, visto que mesmo utilizando criptografia na camada de rede, o uso dos metadados e padrões de comunicação permitem realizar uma série de ações intrusivas, inclusive a identificação desse padrão, como afirma Apthorpe *et al.* (2017). Ainda assim, classificar os dispositivos e o tráfego permite uma série de vantagens essenciais, dentre as quais podemos citar: bloqueio de objetos indesejáveis; alocação e aprovisionamento de recursos; detecção de anomalias; identificação de perfis de usuários; identificação de padrões de comunicação. Ao fazer uso da classificação do tráfego de rede como uma ferramenta de suporte, tem-se a possibilidade de montar perfis de tráfego e entender o padrão de comunicação do entes da rede (ZHANG *et al.*, 2013). Assim, é possível desenvolver um melhor entendimento do comportamento da rede analisada (CALLADO, 2009).

Apesar de mecanismos de classificação de tráfego promoverem grandes capacidades em seu uso, existem desafios na sua implantação em cenários reais (ZHANG *et al.*, 2013b). É o que se verifica quando aplicado em ambientes dinâmicos, com grande quantidade de usuários e uma frequente inserção de novos dispositivos ou aplicações, pois devido à segurança surgem algumas restrições. Exemplo desses problemas são as soluções compostas por DPI, uma vez que necessitam acessar o conteúdo do pacote e acabam infringindo a segurança (privacidade) dos usuários e, em alguns países, desobedecendo leis de privacidade de dados. Ademais, algumas das técnicas exigem muitos recursos computacionais que frequentemente não estão disponíveis (processamento e armazenamento, por exemplo). De acordo com Zhang *et al.* (2009), Korczynski (2012), Zhang *et al.* (2013a), Finsterbusch *et al.* (2014) e Velan *et al.* (2015), alguns dos desafios da classificação da rede são relativos à composição do tráfego, à criptografia, ao uso de portas aleatórias e a novos protocolos e serviços adicionados dinamicamente. O fluxo da rede, de modo geral, tem aumentado muito, principalmente nesses últimos anos, e com isso surgem desafios para as metodologias de classificação, principalmente quanto aos valores de acurácia e precisão. Tais desafios são motivados principalmente pela grande aquisição de *tablets*, *Smartphones*, as

[1] https://www.congress.gov/115/plaws/publ22/PLAW-115publ22.pdf

redes multimídia e pela adição de dispositivos inteligentes à redes.

Ao longo dos últimos anos, grandes quantidades de propostas de classificação e identificação de tráfego surgiram e, com elas, surgem também muitos trabalhos para classificar tráfego de forma dinâmica (WANG, 2013), em tempo real (WICHERSKI *et al.*, 2013), com uso combinado ou independente de técnicas (CALLADO *et al.*, 2010) e para *Voice over Internet Protocol* (Voip) (YILDIRIM; RADCLIFFE, 2010). Os novos modelos emergentes garantem um leque de possibilidades para a construção de soluções em razão de domínios em progressão. Mesmo com os avanços já alcançados pela academia e pela iniciativa privada, não existe um modelo capaz de garantir a acurácia na classificação de tráfego em 100%, como afirmam Dainotti *et al.* (2011). Além disso, a presença de dispositivos IoT incita a construção de novos modelos que promovam uma acurácia mais elevada para esse tipo de tráfego. Segundo Ng *et al.* (2015), operadores de redes empresariais, interessados em *Quality of Service* (QoS), não conhecem todos os aplicativos que estão sendo executados em sua rede. O advento de IoT, que conecta objetos do cotidiano à rede, aumenta significativamente o problema mencionado. Ainda segundo os autores, existe a necessidade do desenvolvimento de soluções rápidas e automatizadas para monitorar, classificar e configurar de forma eficiente o tráfego de rede.

Nas pesquisas realizadas foram encontrados indícios de desafios e deficiências existentes no processo de classificação de dispositivos e de tráfego de redes IoT (NG *et al.*, 2015)(SIVANATHAN *et al.*, 2017)(MEIDAN *et al.*, 2017b)(APTHORPE *et al.*, 2017). Relacionado ao processo de classificação do tráfego de rede, identificar os novos serviços e dispositivos que são inclusos de forma dinâmica instiga a construção de soluções práticas. Dessa forma, necessita-se de novas abordagens ou técnicas para promover qualidade nos resultados. Além do mais, a literatura apresenta uma série de grandes desafios, inclusive para assegurar a privacidade e segurança dos dispositivos e da rede IoT (HAFEEZ *et al.*, 2017b)(HAFEEZ *et al.*, 2017a). Alguns trabalhos abordam a identificação dos dispositivos como uma premissa pertinente para classificar o tráfego em IoT ou não-IoT (SIVANATHAN *et al.*, 2017)(MEIDAN *et al.*, 2017b). Ainda relacionado aos dispositivos, identificá-los na rede é fundamental, inclusive para impossibilitar o acesso a objetos indesejáveis, além de dissociar o tráfego que é gerado por dispositivos de IoT, possibilitando a priorização.

Outra grande motivação em classificar redes com dispositivos IoT relaciona-se à iminente popularização dos ambientes *smart* (por exemplo, *smart cities*, *smart homes*, *smart campus*). Ao classificar o tráfego de rede com suporte à descoberta de tráfego IoT é possível,

inclusive, promover um entendimento maior sobre o impacto que esses dispositivos causam, além de descoberta de padrões na comunicação. No escopo desta dissertação está apresentada uma forma de classificar o tráfego de IoT através da identificação dos dispositivos, além de uma caracterização do tráfego de IoT.

1.2.1 Questões de pesquisa

Esta dissertação está voltada a promover a classificação de tráfego e dos dispositivos de IoT através do fluxo de rede e inspeção da carga útil dos pacotes, além de classificar o tráfego de rede gerado. Esta pesquisa está voltada à resolução de algumas questões:

1. É possível identificar dispositivos próprios de Internet das coisas ao utilizar técnicas de classificação, combinando a inspeção dos pacotes e *Machine Learning*?
2. Ao combinar duas técnicas é possível adquirir melhorias nos resultados?
3. Ao treinar uma rede utilizando ML (*Machine Learning*), é possível ter acurácia elevada na identificação de todos os dispositivos?
4. O *User-Agent*, presente no cabeçalho HTTP, promove suporte necessário para identificação de dispositivos específicos?
5. Quais características do tráfego IoT mais influenciam a classificação?
6. A literatura apresenta abordagens ou técnicas para identificar os dispositivos de Internet das coisas?

1.3 Objetivo

Este trabalho tem o objetivo de atenuar os problemas relacionados à classificação de tráfego de rede e dispositivos em ambientes IoT, focando principalmente em características típicas do ambiente, dentre as quais destacam-se a baixa taxa de dados, variabilidade de protocolos, tolerância ao atraso e armazenamento e processamento limitado. Este trabalho propõe uma estratégia que combina ML e inspeção de pacotes para classificar com acurácia e precisão elevadas dispositivos e tráfego nesses ambientes. Para alcançar o objetivo dessa pesquisa foram traçadas as seguintes metas:

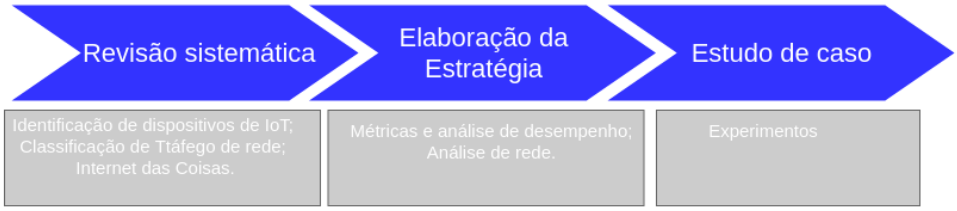
- Identificar na literatura trabalhos relacionados ao processo de classificação ou à identificação de dispositivos e tráfego de rede IoT;
- Identificar na literatura formas adequadas para avaliar a performance e suas métricas;

- Elaborar uma estratégia eficaz para proceder com as ações e demonstrar os resultados adquiridos;
- Avaliar a proposta mediante estudo de caso;
  - Identificar na literatura propostas atuais com avaliação em estudos de casos;
  - Analisar na literatura abordagens de classificação estatísticas de tráfego de rede utilizando ML;
  - Analisar os trabalhos relacionados, comparar seus resultados e avaliar o desempenho da estratégia proposta.
- Categorizar o tráfego gerado, assim como apresentar uma análise e o impacto dos dispositivos no processo de classificação em IoT e do tráfego de rede.

1.4 Metodologia

Os detalhes da metodologia científica e suas especificações estão descritas na Figura 1.

Figura 1 – Metodologia da pesquisa científica



Fonte: Autor

Neste trabalho é realizada uma revisão sistemática da literatura sobre os principais temas abordados na área de classificação de dispositivos e de tráfego de rede em ecossistemas de IoT. Os resultados adquiridos nessa revisão foram utilizados como base para formulação dos capítulos 1, 2 e 3, e os detalhes estão presentes no anexo A.

O mapeamento sistemático da literatura ocorreu através do uso de *strings* de busca, fontes de pesquisas e critérios de seleção pessoal (Qualis do veículo, ano de publicação e número de citações, por exemplo). A análise dos resultados encontrados é, especialmente, avaliada através da proximidade dos conteúdos: classificação de dispositivos com ML; classificação de tráfego de rede em ecossistemas de IoT; algoritmos de aprendizado automático supervisionados; uso de métricas de desempenho para avaliação da qualidade dos resultados.

A fase de elaboração da estratégia consiste no levantamento de abordagens para avaliar e validar corretamente a proposta, dentre as quais podemos citar as métricas presentes em (BOWES *et al.*, 2012), que faz uso da matriz de confusão para análise de *performance*. Como forma de avaliar a estratégia proposta para classificação dos dispositivos e do tráfego de rede, neste trabalho, um estudo de caso é conduzido como proposta de avaliação da eficácia através de avaliação por cálculos estatísticos. Os detalhes dos experimentos estão dispostos no Cap. 5 desta dissertação.

1.5 Organização da Dissertação

A dissertação apresenta-se assim distribuída: o capítulo 2 apresenta a fundamentação teórica, alicerçada nos temas: classificação de tráfego de rede, aprendizado automático e Internet das coisas; no capítulo 3 apresentamos os trabalhos relacionados, focando em temas de classificação com ML e identificação/classificação de dispositivos; o capítulo 4 apresenta a estratégia para classificar o tráfego de rede em ecossistemas de IoT; o capítulo 5 apresenta os resultdos adquiridos em testes de redes com presença de dispositivos de IoT; por fim, o capítulo 6 aborda a conclusão, os resultados científicos e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são abordados conceitos e definições fundamentais para o desenvolvimento desta pesquisa. A contextualização é fundamentada na revisão sistemática abordada na seção **1.4** e presente no anexo **A**. As próximas seções deste capítulo estão organizadas da seguinte forma: a seção **2.1** aborda a classificação de tráfego de rede; na sequência, a seção **2.2** aborda os conceitos de aprendizado de máquina; a seção **2.3** apresenta a fundamentação sobre Internet das coisas e suas tecnologias; por fim, a seção **2.4** apresenta as considerações finais.

2.1 Classificação do Tráfego de rede

Nesta seção estão apresentados os conceitos gerais e introdutórios relacionados à classificação de tráfego de rede, assim como as técnicas mais comumente utilizadas.

2.1.1 Visão geral

A classificação de tráfego é uma área na computação que tem despertado bastante interesse da comunidade acadêmica e da indústria em função das possibilidades de gerenciamento da rede como, por exemplo, *QoS* (*Quality of Service*), detecção de anomalias, gerenciamento de infraestruturas, aprovisionamento e alocação de recursos (CALLADO, 2009)(ZHANG *et al.*, 2013a)(VELAN *et al.*, 2015). Após leitura de Callado *et al.* (2009), Silvio *et al.* (2013) e Wang (2013), é possível afirmar que classificar o tráfego de rede não é uma simples coleta de pacotes ou fluxos, o processo todo está associado ao entendimento da dinâmica e do comportamento do tráfego das redes, promovendo a compreensão através da extração de características que possibilitam associar sua origem, sua formação, sua derivação, sua composição e seu impacto.

À medida que a Internet evolui e com ela os dispositivos e modelos computacionais, a complexidade de muitos de seus processos também crescem em ritmo acelerado. Atualmente já temos em grande escala a presença de streams de vídeos e áudios, jogos e compartilhamento de arquivos. Carela-Español (2014) afirma, em sua tese de doutorado, que operadores de redes, pesquisadores e até mesmo ISPs precisam conhecer as características de tráfego de suas redes para gerenciar os recursos ou mesmo cobrar os usuários com base no seu consumo. Desse tipo de necessidade surgiram os vários métodos de classificação de tráfego de rede. Os principais métodos encontradas na literatura estão apresentadas na Tabela 1, junto a algumas de suas principais características.

Tabela 1 – Técnicas Utilizadas na Classificação de Tráfego, baseada em (WANG, 2013).

Abordagem	Propriedades	Custo	Acurácia	Complexidade
Portas	Acesso às portas	Baixo	Baixa	Baixa
Estocástico	Assinaturas	Variável	Variável	Variável
DPI	*Payload*	Baixo	Variável	Alta
Estatística	Fluxos e Pacotes	Moderado	Alta	Alta

2.1.2 Classificação Baseada na Análise das Portas

Após a leitura de Callado (2009), Wang (2013), Korczynski (2012) e Carela-Español (2014), é possível afirmar que antigamente as empresas e ISPs (*Internet Service Providers*) conseguiam classificar o tráfego facilmente, principalmente devido à rede ser composta por poucos dispositivos e protocolos, possibilitando a classificação através do conhecimento do número das portas. Inicialmente, o seu uso era suficiente para classificar a rede, pois quase todas as aplicações utilizavam-se de portas fixas assinadas pela IANA. Por exemplo, aplicações Web através da porta 80 (HTTP), e-mails pela porta 25 (SMTP – *Simple Mail Transfer Protocol*) para enviar e pela porta 110 (POP3 – *Post Office Protocol*) para receber. A Tabela 2 apresenta alguns exemplos do que a literatura convencionou chamar de *portas bem conhecidas* Callado (2009).

Tabela 2 – Exemplos de portas bem conhecidas

Número da porta	Aplicação
20	FTP Dados
21	FTP Controle
22	SSH
23	Telnet
25	SMTP
53	DNS
80	HTTP
110	POP3
443	HTTPS
414	Syslog

Como mencionado na seção 1.1 e presente em Carela-Español (2014), este método de classificação se encontra ineficiente e ineficaz devido a sua imprecisão e à incompletude. Relacionado a precisão do modelo, é difícil estipular os valores, uma vez que as características da rede monitorada podem variar, mas existem estudos que apresentam sua acurácia entre 50%-70% nos melhores casos (MOORE; PAPAGIANNAKI, 2005) (MADHUKAR; WILLIAMSON,

2006).

2.1.3 Classificação Baseada em Payload

Segundo Li *et al.* (2016), o DPI originalmente foi projetado com intuito de melhorar a segurança da rede. Ele surgiu através da combinação das funcionalidades do sistema de detecção de intrusão (*Intrusion detection System* (IDS)) e dos sistemas de prevenção de intrusão (*Intrusion prevention system* (IPS)). Além disso, ele emergiu como alternativa ao problema da baixa acurácia na classificação por portas. Esse método faz a análise do conteúdo dos pacotes em busca de características ou assinaturas de aplicações. De acordo com Korczynski (2012) e Carela-Español (2014), a técnica visa identificar as aplicações que usam estratégias para se camuflar no tráfego, exemplo disso é o P2P (*Peer-To-Peer*).

Para Finsterbusch *et al.* (2014), há desafios chaves na utilização da classificação por DPI, dentre elas a necessidade de atualização contínua do conjunto das assinaturas de aplicativos para classificar novas aplicações e versões. Apesar disso, a técnica de classificação por DPI permanece como uma das técnicas mais utilizadas, como afirmam Finsterbusch *et al.* (2014) e Li *et al.* (2016). A Tabela 3 apresenta exemplos de padrões de assinaturas para algumas aplicações P2P utilizadas para classificá-los.

Tabela 3 – Exemplos de Strings de DPI, baseado em (KARAGIANNIS *et al.*, 2004)

Aplicação	**String Utilizada**	**Protocolo de transporte**
BitTorrent	"0x13Bit"	TCP
eDonkey	"0xe319010000"	TCP e UDP
Gnutella	"GNUTGIV"	TCP
Gnutella	"GND"	UDP

De acordo com Dainotti *et al.* (2012), devido à necessidade de acessar o conteúdo dos pacotes, a técnica deve lidar com grandes desafios de privacidade, exemplo disso é que alguns países restringem o acesso ao conteúdo da comunicação dos usuários através de regulamentações ou leis.

2.1.4 Classificação Baseada em Características do Fluxo

Segundo Dainotti *et al.* (2012), a classificação estatística do tráfego de rede surgiu como forma alternativa ao uso de DPI, decorrente da preocupação com as políticas de

privacidade e análise de carga útil dos pacotes. As características estatísticas são extraídas através do agrupamento dos pacotes em forma de fluxo. Ao final, a classificação consistirá na comparação estatística de tráfego desconhecido, ou gerado por uma fonte não analisada, com regras previamente estipuladas (NGUYEN; ARMITAGE, 2008). De acordo com Domingos (2012), os sistemas baseado em aprendizagem de máquina (*Machine Learning* – ML) aprendem através de dados empíricos e, dessa forma, associam automaticamente objetos com classes correspondentes. Ainda segundo o autor, os algoritmos podem ser divididos em supervisionado e não-supervisionado. De acordo com Dainotti *et al.* (2012), nos sistemas que utilizam os algoritmos de aprendizagem automática supervisionado as classes já estão previamente definidas pelo pesquisador, assim, os objetos de amostra são fornecidos ao sistema rotulados com suas respectivas classes; enquanto em algoritmos sem supervisão, o sistema identifica classes distintas e atribui objetos a elas por afinidade (por exemplo, utilizando técnicas de agrupamento).

O objetivo principal da classificação estatística do tráfego é categorizar o fluxo da rede de acordo com a aplicação geradora, sendo ela, por exemplo, baseada em análise de características da rede, como tamanho do pacote e tempo de intercalação (entre pacotes), como afirmam Nguyen e Armitage (2008), Zhang *et al.* (2009) e Korczynski (2012). Além disso, a abordagem estatística é caracterizada como um modelo de alta velocidade e acurácia, mas que apresenta uma complexidade elevada em seu desenvolvimento se comparada com as demais apresentadas (ZHANG *et al.*, 2013a).

2.1.5 Métodos híbridos para classificação de tráfego de rede

O uso de algoritmos de ML para classificação do tráfego baseada em características extraídas do fluxo recebe uma substancial atenção da academia. Da mesma forma, a identificação baseada em conteúdo, que faz uso de padrões de assinaturas, continuam sendo vastamente utilizadas, por exemplo, em IDS (Sistemas de Detecção de Intrusão). É perceptível, em estudos recentes, a construção de várias soluções híbridas para classificar o tráfego de rede baseadas em métodos de aprendizagem de máquinas junto às características extraídas do conteúdo, como os presentes em (CROTTI *et al.*, 2007), (SUN *et al.*, 2010) e (KORCZYńSKI; DUDA, 2012).

Há alguns anos, classificações baseadas apenas em análise simples de padrões de assinaturas e de comunicação obtinham resultados precisos (DAINOTTI *et al.*, 2012) (KORCZYńSKI; DUDA, 2012). No entanto, suas aplicações tornaram-se menos eficaz em casos no qual o tráfego de rede está criptografado. Dessa forma, outras propostas surgiram para mitigar esses problemas,

focados na substituição do sistema tradicional de verificação de padrões com métodos mais sofisticados de estatística ou mesmo baseados na combinação de técnicas. Como exemplo de propostas híbridas para classificação do tráfego de rede temos Lu *et al.* (2009). Os autores utilizaram, primeiramente, a classificação por assinatura (DPI) e para o tráfego classificado como *Unknown* foi aplicada a classificação por ML, utilizando árvores de decisão. A combinação de técnicas reforça a necessidade de desenvolvimento e aprimoramento das técnicas para classificação do tráfego de rede, visto que a Internet e sua complexidade tem aumentado significativamente.

2.2 Aprendizado de Máquina

Nesta seção serão abordados alguns conceitos relacionados ao Aprendizado de Máquina com foco em definições, modelos para avaliação de desempenho e seleção de características.

2.2.1 Visão geral

Segundo Mitchell (1997), aprendizado de máquina está associado às melhorias de desempenho de programas de computadores através da aquisição de conhecimento por intermédio das experiências em tarefas. Para Hastie *et al.* (2009), o processo de aprendizado estatístico desempenha um papel essencial em vários campos da ciência, desde tomadas de decisões a finanças, tendo o seu uso ocorrido em diversas situações emergenciais, como prevenção de ataques cardíacos e identificação de fatores de riscos. Um fator fundamental em aprendizado de máquina é o limiar de equilíbrio entre desempenho e qualidade. Dessa forma, Jain (1991) define avaliação de performance como o modo de utilização otimizada dos recursos computacionais através de medidas mensuráveis que permita identificar o seu dispêndio.

Domingos (2012) afirma que o processo de construção ou uso de sistemas baseado em aprendizado de máquina manifesta-se através da utilização de características (*features*), elas, segundo o autor, são processadas através de um modelo denominado engenharia de *features*. Ainda segundo o autor, utilizar apropriadamente as características é fundamental para a concepção de um projeto, apesar que para a sua correta construção é necessário um esforço muito grande, principalmente devido à complexidade do processo de montagem. Para Kulkarni (2017), os novos sistemas de ML focam na combinação otimizada dos recursos e sua maior capacidade surge pela correta relação entre a entrada de dados e a saída do algoritmo. Ainda segundo o autor,

ao fazer essa correta aplicação é possível verificar a redução da intervenção humana, melhoria do desempenho ao fazer uso de grande volume de dados e soluções mais completas para problemas complexos, como também afirma Kubat (2017).

A partir da leitura de Domingos (2012), pode-se concluir que o uso, na prática, de *Machine Learning* passa por uma metodologia estrita. Primeiramente, relacionada à escolha do algoritmo para aplicação em um determinado problema, é fundamental considerar a combinação de três componentes principais: 1 - Representação: o classificador deve ser representado em alguma linguagem que o computador entenda; 2 - Avaliação: uma função deve ser considerada para distinguir entre classificadores bons e ruins; 3 - Otimização: deve-se selecionar classificadores com maior desempenho e acurácia. A escolha da técnica de otimização é preponderante para uma maior eficiência do algoritmo.

2.2.1.1 Classificação Supervisionada

Para Sugiyama (2016), o método de classificação é tido como supervisionado quando extrai estruturas de aprendizado para classificar novas instâncias em classes predefinidas. Ainda segundo o autor, esse modelo consiste em realizar a classificação utilizando treinamento (base de dados já classificada) e posteriormente as saídas do algoritmo são dadas baseadas em sua correlação.

2.2.1.2 Classificação não-supervisionada

Segundo Sugiyama (2016), os métodos não supervisionados, diferente do supervisionado, não necessitam de um conjunto de dados rotulado completos para treinamento, pois o próprio método descobre a forma de associar os dados através de similaridades. Para Alpaydin (2014), esse método não contém um supervisor e, consequentemente, não ocorre um correto mapeamento entre as entradas e saídas desejadas. Porém, são identificados padrões de entradas mais frequentes para relacionar as similaridades e criar um agrupamento dos conjuntos de saída.

2.2.2 Random Forest

O Random Forest (RF), introduzido inicialmente por (BREIMAN, 2001), é um algoritmo de aprendizado automático supervisionado amplamente utilizado e que possui alto desempenho, como verificado em (ZIEGLER *et al.*, 2014). De acordo com o trabalho (WANG *et*

al., 2015), este algoritmo apresenta uma série de vantagens que o torna relevante em pesquisas, dentre elas temos: grande resistência a *overfiting*; necessita de uma porção pequena de parâmetros; possui baixa variação, o uso de múltiplas árvores reduz a chance de ocorrer falhas durante a classificação em decorrência da relação dos dados de treino e de teste.

O RF é utilizado em uma grande variedade de áreas de pesquisas, como em detecção de anomalias (PRASHANTH *et al.*, 2008), classificação de tráfego (WANG *et al.*, 2015) e, mais recentemente, na identificação de dispositivos IoT (SIVANATHAN *et al.*, 2017)(MEIDAN *et al.*, 2017a). Decorrente de seus ótimos resultados, de sua grande variedade de aplicação e de suas vantagens, este algoritmo é ideal para uso em classificação de tráfego de rede.

2.2.2.1 Formulação da árvore de decisão utilizada

O RF constrói várias árvores de decisão e os agrega para realizar a classificação. Segundo (HASTIE *et al.*, 2009), árvore de decisão é um modelo estatístico indutivo utilizado em aprendizado de máquina supervisionado. A classificação decorre, após a construção da árvore, do percorrimento do nó raiz ao nó folha. As árvores de decisão são baseadas no particionamento de recursos em determinados conjuntos e os ajustam a algum modelo simples, igual a uma constante. O modelo utilizado para a classificação em árvore foi o baseado em *Classification and regression tree* (CART), como o descrito em (BITTENCOURT; CLARKE, 2003), o qual é semelhante ao C4.5 (extensão do modelo ID3).

2.2.3 Métricas de Desempenho para Classificadores Estatísticos

De acordo com Bowes *et al.* (2012), para validar a qualidade dos resultados de uma classificação por ML é necessário utilizar a avaliação de desempenho por intermédio da matriz de confusão (Tabela 4). Os cálculos são realizados através do arranjamento dos valores na matriz e, então, calculado como mostrado na Tabela 5.

Tabela 4 – Matriz de Confusão

	Validação dos Testes		
Teste	Presente	Ausente	Total
Positivo	TP	FP	TP + FP
Negativo	FN	TN	FN + TN

De acordo com Alpaydin (2014), existem 4 formas possíveis de expressar as medidas,

são elas:

- ***True Positive* (TP):** – Verdadeiros positivos
- ***False Positive* (FP):** – Falsos positivos
- ***False Negative* (FN):** – Falsos negativos
- ***True Negative* (TN):** – Verdadeiros negativos

Através dos valores dispostos na matriz de confusão (Tabela 4), é possível produzir valores que representam as respostas para as avaliações pretendidas.

Tabela 5 – Métricas Utilizando da Matriz de Confusão

Nome	**Fórmula**
Acurácia	$\frac{TP+TN}{TP+TN+FP+FN}$
Precisão	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-Score	$\frac{2TP}{2TP+FP+FN}$
MCC	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
FPR	$\frac{FP}{TN+FP}$

O cálculo por matriz de confusão permite uma série de avaliações precisas quanto aos resultados obtidos. Entre os cálculos apresentados temos a acurácia, que permite identificar as proporções corretas de classificação independente de verdadeiro ou falso. Por outro lado, a precisão, que é análoga ao valor preditivo positivo (VPP), permite identificar a proporção de verdadeiros positivos em relação a todas as predições positivas. O *recall*, também conhecido como sensibilidade, corresponde à capacidade do sistema em predizer corretamente a condição para casos verdadeiros. Para compensar possíveis distorções na análise, utiliza-se o balanceamento *F1-Score*, que indica o ajuste de resultados em relação à acurácia e ao *recall*. Outra interessante métrica para análise é o coeficiente PHI, tambem conhecido como MCC (*Matthews correlation coefficient*), que realiza avaliação de qualidade em classificadores tendo o seu valor estipulado no intervalo entre -1 e 1. Quanto mais próximo de 1, maior a qualidade da predição e quanto mais próximo de -1 significa total desacordo entre a predição e a observação. Por fim, tem-se a razão de falso positivo (FPR) para avaliar a quantidade de falsos positivos em relação ao total que não deveria ter sido identificado.

2.2.4 Seleção de atributos estatísticos - Feature Selection

De acordo com Hastie *et al.* (2009), atributos estatísticos representam os dados com caráter qualitativo. Segundo Tang *et al.* (2014), a seleção de características, também conhecida

como seleção de variáveis ou de atributos estatísticos, é o processo de seleção do subconjunto de dados mais relevante para uso e construção do modelo que será utilizado em ML. Ainda segundo os autores, o seu uso é fundamental para simplificar a construção de modelos complexos, diminuir a dimensionalidade, reduzir o tempo de treino, reduzir a superposição (*overfitting*) e a variância.

Para Alpaydin (2014), a eficácia das estruturas de classificação depende fortemente da escolha correta dos atributos estatísticos para reduzir ruídos, aumentar o desempenho e remover possíveis redundâncias ou sobreposições. De acordo com Liu e Motoda (2007), os dados com dimensionalidade extremamente alta apresentam sérios desafios aos métodos de aprendizagem existentes. Ainda de acordo com os autores, a grande quantidade de atributos estatísticos tende a alterar o desempenho.

Devido ao problema da dimensionalidade, foram estudadas e desenvolvidas técnicas com o intuito de reduzi-la. O objetivo dessas técnicas é escolher um pequeno subconjunto dos atributos estatísticos com maior relevância, em conformidade com determinado critério de avaliação, promovendo um melhor desempenho, menor custo computacional e melhor formulação do modelo (TANG *et al.*, 2014).

De acordo com Liu e Yu (2005), métodos de seleção de características são divididos em quatro etapas principais: **1** - geração de subconjuntos, no qual um subconjunto candidato será escolhido com base em uma determinada estratégia de busca; **2** - avaliação de subconjuntos, onde ocorre a verificação de acordo com preceito de avaliação definido; **3** - critério de parada, nessa etapa o subconjunto que melhor se adequar ao critério de avaliação será escolhido entre todos os candidatos avaliados; **4** - validação de resultados, nela o subconjunto escolhido será validado usando um conjunto de validação.

2.3 Internet das Coisas

Nesta seção serão abordados conceitos, expectativas e tecnologias associadas aos ecossistemas de Internet das coisas.

2.3.1 Visão geral

A Internet das coisas apresenta-se, de acordo com a literatura, como um novo paradigma computacional capaz de integrar uma grande variedade de sistemas heterogêneos,

promovendo a conexão de dados, pessoas, objetos e aplicações através da Internet (ATZORI *et al.*, 2010). Por meio da coleta desses dados, podemos construir uma grande variedade de aplicações com possibilidade de incrementar funcionalidades através do uso de diversas técnicas como I.A. (Inteligência Artificial) e *Analytics*. Seu surgimento se deu em 1999 nos laboratórios do **MIT** (*Massachusetts Institute of Technology*), por meio de pesquisas com RFID patrocinados por Kelvin Ashton, co-fundador da Auto-ID (KRANENBURG; DODSON, 2008). Através de Farhan *et al.* (2017), é possível afirmar que IoT está promovendo uma grande revolução e vem avançando em diferentes áreas e domínios como sistemas embarcados e telecomunicações. Além disso, os principais fatores para a grande evolução de Internet das coisas são os objetos inteligentes, visto que possuem capacidade de processamento, conectividade e coleta de dados.

Desde o surgimento de IoT, diversos autores atribuiram-lhe muitas definições, entre elas a de Buyya e Dastjerdi (2016), na qual IoT concentra-se, principalmente, em conectividade e requisitos dos sensores de dispositivos conectados em ambientes típicos. Considerando que essas afirmações refletem os requisitos básicos de IoT, outras definições focam mais na necessidade de redes ubíquas e autônomas, em que a identificação e integração de serviços têm papel fundamental. Por exemplo, *Internet of Everything* (IOE) é um termo amplo utilizado pela Cisco para se referir a pessoas, coisas e lugares conectados à Internet global (LLC, 2013). Segundo López *et al.* (2012), IoT caracteriza-se como um paradigma que apresenta um alto grau de captura autônoma de dados, conectividade, interoperabilidade, mobilidade e transferências de eventos.

IoT possibilita aos usuário e empresas uma grande variedade de funcionalidades, promovendo um elevado aumento da capacidade de interação e comunicação com o ambiente, tudo isso através dos dispositivos inteligentes e a Internet. A IERC (2014) define IoT como uma infraestruturada dinâmica de rede global autônoma com suporte a interoperabilidade e uso de diversos protocolos padronizados, no qual objetos físicos e virtuais possuem identificadores, atributos, e comportamento próprio, através de interfaces otimizadas. De acordo com Lin *et al.* (2017), IoT surge como um modelo de coexistência de redes de dispositivos no qual todos são capazes de interagir entre si através de vários gateways e middlewares apoiado por um complexo plano de controle e gerenciamento. Assim, a infraestrutura de rede deve promover a integração dessas várias infraestruturas, dessa forma todos os sistemas ou aplicações, baseados em IoT, serão capazes de obter uma melhor performance no fornecimento de seus serviços através de um eficiente compartilhamento de informações e recursos.

IoT está relacionada à próxima geração de Internet. De acordo com Bahga e Madisetti (2014), a rede composta por dispositivos de IoT possuirá, em um futuro não muito distante, trilhões de nós, incluindo uma grande variedade de dispositivos ubíquos, dispostos em uma pluralidade de ambientes e dotados de sensores interconectados à rede. A partir da leitura de Zanella *et al.* (2014) e Buyya e Dastjerdi (2016), é possível afirmar que IoT está relacionada a uma série de tecnologias, dentre as quais podemos citar: Redes de sensores sem fio (RSSF); IPV6; computação em nuvem; e computação ubíqua. Decorrente da grande expectativa de investimentos, muitas projeções surgem sobre o seu estado atual, dentre elas temos a Gartner (2017), que apresentou uma estimativa de 8.6 bilhões de dispositivos até o final de 2017, crescimento de mais de 31% em relação a 2016 (6.3 bilhões), projeta para 2020 um total de 20 bilhões de dispositivos conectados (GARTNER, 2018). Ainda relacionado às projeções, a Forbers (COLUMBUS, 2017) afirma que o mercado de IoT crescerá anualmente cerca de 28.5%, isso corresponde a um crescimento bruto de 157 bilhões de dólares em 2016 para 428 bilhões em 2020.

De acordo com Farhan *et al.* (2017) e Bertino *et al.* (2016), alguns dos principais desafios no uso em larga escala de IoT e o seu emprego através da combinação de múltiplas técnicas está relacionado a:

- Coleta massiva de dados;
- Escalabilidade e diversidade;
- Exigências de segurança;
- Consumo de energia;
- BigData (*Data Collection and Analysis - DCA*);
- Tolerância a falhas;
- Analytics.

Esses desafios encontram-se em análise e desenvolvimento através de pesquisas pela academia e pela indústria com o intento de promover melhor integração das tecnologias. Outro fator importante nos ecossistemas IoT é que a sua composição depende essencialmente de três componentes principais (ATZORI *et al.*, 2010) (SANTOS *et al.*, 2016), que são:

- **Componentes Físicos** → Dispositivos eletrônicos, sensores que estão dispostos nos ambientes para realizar coletas de dados ou responder de acordo com a proposta, objetos inteligentes e atuadores.
- **Sistemas de Comunicação** → Tecnologias de transmissão de dados baseadas em redes

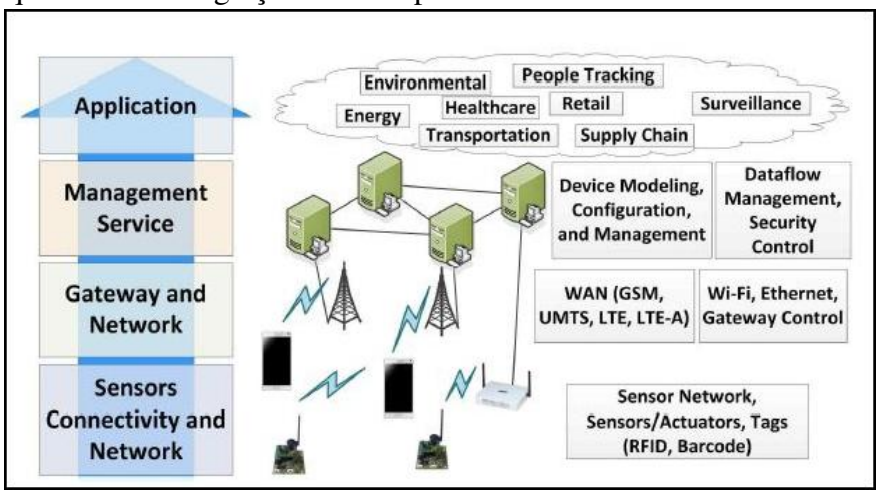
cabeadas ou sem-fio, estas móveis ou não.

- **Processamento da Informação** → Implementada através de programas, com o uso de I.A. ou não, para controlar e gerenciar os sistemas.

2.3.2 Arquiteturas IoT

Após leitura de Atzori *et al.* (2010), Buyya e Dastjerdi (2016) e Santos *et al.* (2016), pode-se afirmar que as arquiteturas de aplicações IoT são tipicamente subdivididas em quatro camadas principais (ver Figura 2). São elas:

Figura 2 – Arquitetura de integração dos componentes.



Fonte: (SUKANYA, 2015)

1. **Aplicação** → Web, Mobile
2. **Gerência de Serviços**→ Responsável pela segurança da informação, controle de segurança, gerenciamento de dispositivos, gerência e abstração dos dados.
3. **Gateway e Rede** → LAN, PAN, volume de dados massivo, *QoS*, escalabilidade.
4. **Conectividade e Sensores** → Baixo consumo, WSN (*Wireless Sensor Network*), baixa taxa de dados.

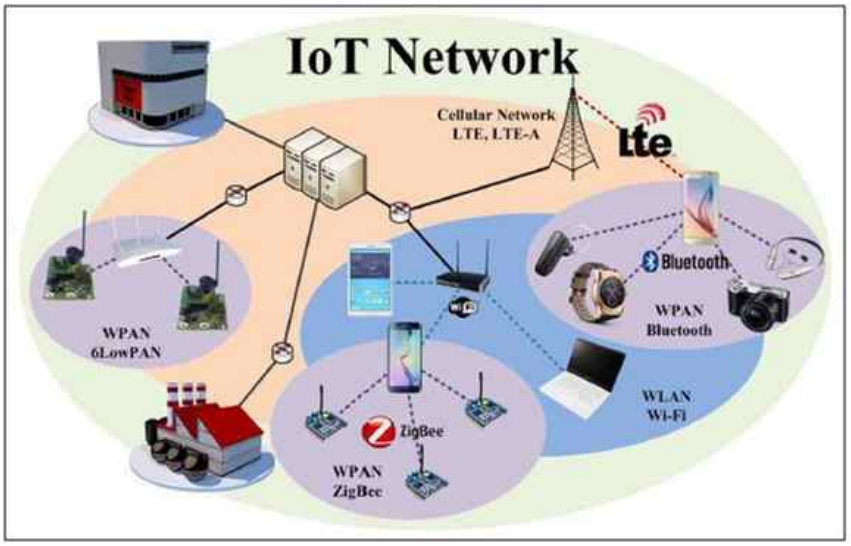
No começo da arquitetura estão presentes tecnologias como sensores, atuadores e *tags*. Essa camada é responsável pela coleta dos dados do ambiente. Na segunda camada existe o *Gateway* e a rede, responsáveis pelo roteamento dos dados coletados na camada inferior e o envio até a camada de gerência de serviços. Além do roteamento, ela é responsável por garantir a interoperabilidade dos sistemas, pois diversos dispositivos IoT comunicam-se de forma diferente através de protocolos diferentes. A terceira camada é chamada de gerência de serviços e é responsável pelas garantias de segurança, garantia de *QoS* e análise das informações. A última

camada faz uso dos dados coletados na forma de serviços para os usuários.

2.3.3 Tecnologias Associadas a IoT

IoT caracteriza-se por fazer uso de tecnologias heterogêneas (ver Figura 3) e à medida em que os dispositivos estão sendo inseridos, novos requisitos de escalabilidade, interoperabilidade e conectividade são adicionados. O IPv6 apresenta-se como um dos fatores preponderantes na adição dos dispositivos, além de protocolos com baixa taxa de dados como o padrão IEEE 802.15.4 (CHUNG *et al.*, 2013) (SERPANOS; WOLF, 2017).

Figura 3 – Heterogeneidade da rede IoT



Fonte: (SUKANYA, 2015)

2.3.4 Estrutura Básica Para Construção de Ecossistemas de IoT

IoT representa uma grande evolução que vem emergindo junto aos avanços em outros domínios da computação, o que promove e garante uma complementação tecnológica, além de desempenhar papel importante na integração e comunicação do físico com o virtual. Segundo Santos *et al.* (2016), algumas das principais estruturas para construção de ecossistemas de IoT são:

- **Identificação:** Necessita de mecanismos que garantam a identificação unívoca dos dispositivos na rede podendo ser utilizadas diversas tecnologias como o IP;
- **Comunicação:** Representa as tecnologias empregadas em ecossistemas IoT que promovem garantias de comunicação e interoperabilidade;

- **Serviços:** Existe uma grande variedade de serviços ofertados no contexto de IoT, essa estrutura representa essa diversidade;
- **Semântica:** IoT promove coleta de dados, esses dados necessitam de análise para que se possa extrair o máximo de conhecimento.

2.3.5 Uso de Dados em IoT

De acordo com o relatório da IHS (2016), a utilização otimizada de dados dos ecossistemas IoT é uma área de alta complexidade, dificultada principalmente pela necessidade de uma gestão e utilização otimizada. Ainda segundo a IHS (2016), existem sete situações chaves que são consideradas complexas na utilização desses dados, que são:

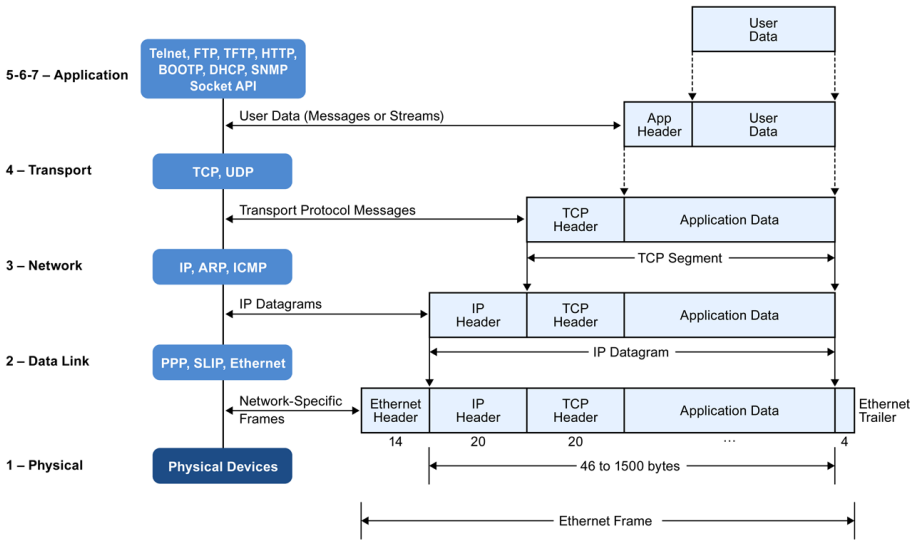
1. **Segurança:** Existe uma série de preocupações na utilização de dados em IoT, uma delas é a privacidade dos usuários. Um exemplo prático disso é apresentado por Hill (2012), que em uma reportagem da *Forbes* apresenta a identificação da gravidez de adolescentes apenas pelo padrão das compras através do uso de *Analytics*. A preocupações quanto à segurança em IoT é uma peça chave em sua aquisição, pois através de dados coletado dos usuários é possível cruzá-los, inferir e adquirir informações cruciais como viagens e preferências em negócios.
2. **Volume de Dados:** O desafio de gerenciamento e processamento de dados é fundamental no contexto IoT, o seu uso através de combinações de técnicas pode promover grandes mudanças nas perspectivas sociais.
3. **Diversidade dos Dados:** A complexidade está presente na grande variedade de fontes de dados, entre as quais temos os carros, geladeiras e câmeras, por exemplo. A grande heterogeneidade de dispositivos promove uma grande complexidade, principalmente na integração desses dados.
4. **Velocidade dos Dados:** Esse item promove a necessidade de construção de aplicações sensíveis ao tempo. Sistemas de tempo real terão que processar grandes volume de dados com uso de *Analytics* independentemente do contexto inserido, a complexidade tende a ser elevadíssima.
5. **Analytics:** Promove a conversão de dados brutos em informações de relevância aos usuários. Sua abordagem foca, inclusive, em melhorar os processos de trabalho e extrair valiosos *insights* acerca de comportamentos e tendências de mercado e dos consumidores, além de suas expectativas.

6. **Economia de Dados:** Os dados são utilizado para a construção e operação de aplicações, com a presença da diversidade, da velocidade de transmissão, de processamento e de *analytics*. A economia de dados vem da necessidade de conter as redundâncias e os desperdícios de recursos e sua complexidade vem da necessidade de profundo conhecimento sobre os dados para poder otimizar seu uso.
7. **Logística:** Representa o uso otimizado do ambiente para que proporcione as melhores experiências. Um exemplo de logística seria quando houvesse a integração do ecossistema IoT na nuvem e ocorresse processamento por *Fog Computing* ao invés do envio da totalidade dos dados, pois isso promoveria velocidade e menos congestionamento.

2.3.6 Protocolos IoT

A literatura apresenta constantes evoluções em protocolos que são projetados para minimizar ou resolver certos problemas pertinentes, entre eles a taxa de transmissão, salvaguarda dos dados e interoperabilidade (AL-FUQAHA *et al.*, 2015). O TCP/IP é considerada a suite fundamental de protocolos da Internet, onde o IP fornece conexão entre diversas redes, camada 3. O TCP e o UDP situam-se na camada 4 (transporte) (KUROSE; ROSS, 2012). Esses protocolos fundamentais possuem diversas formas de representação, entre elas temos a sua disposição na terceira e quarta camada do modelo em camadas, como ilustrado na Figura 4, utilizando os três últimos protocolos associados em sua representação.

Figura 4 – Modelo em camadas com o TCP e o IP



Fonte: (MICRIUM, 2017)

O que mais se percebe na construção e no aprimoramento de protocolos para IoT são

as grandes exigências que devem ser remediadas para atender a todos os requisitos de dispositivos, focando principalmente na alta latência e no uso cada vez menor de rede e de armazenamento. Muito das evoluções dos protocolos IoT estão presentes na camada de transferência de dados e na camada de aplicação. A seguir é apresentado um resumo de protocolos amplamente utilizados e pesquisados para IoT (AL-FUQAHA *et al.*, 2015).

- **CoAP:** *Constrained Application Protocol*, protocolo de camada de aplicação focado em dispositivos de Internet das coisas que possuem recursos limitado, possui como características o baixo consumo de recursos, tradução para HTTP, fácil implementação e suporte a *multicast*;
- **MQTT:** *Message Queuing Telemetry Transport*, protocolo bastante popular que baseia-se em protocolos *publish*/*subscribe*, sua principal aplicabilidade é para dispositivos de Internet das coisas em comunicação M2M (*Machine-to-Machine*) e dispositivos móveis;
- **HTTPS:** *Hyper Text Transfer Protocol Secure*, pode ser definida como uma implementação do HTTP com a adição de uma camada de segurança que faz uso de protocolos como SSL ou TLS, permitindo tráfego de dados com criptografia;
- **XMPP:** *Extensible Messaging and Presence Protocol*, protocolo de comunicação *open-source* e extensível focado em promover comunicação *people-to-people* interoperável. Sua principal característica é que sua comunicação é orietada a texto (XML);
- **Z-Wave:** É um protocolo de comunicação sem fios voltado ao envio de comandos de controle e de dados secundários (por exemplo, informações do tempo). Ele é concebido para uso em meio simples, confiável, de baixo consumo, de ondas rádio. Esse protocolo não apresenta suporte suficiente para envio de áudio nem de vídeo;
- **Outros:** Além dos protocolos abordados, para o universo de IoT temos uma grande quantidade deles desenvolvidos, desde infraestrutura (e.g., 6LowPAN, RPL), identificação de serviços e dispositivos (e.g., EPC, URIs), transporte e comunicação (e.g., Wi-fi, Bluetooth), descoberta (e.g., mDNS, DNS-SD) e semântica (e.g., Web Thing Model).

A literatura apresenta, constantemente, o surgimento ou a melhoria de protocolos para IoT. Dentre eles, podemos citar protocolos que apresentam certas singularidades em comunicação, como a taxa de dados e o tamanho de pacote:

1. **Enlace de Dados:**
    - IEEE 802.15.4
    - IEEE 802.11 AH

- LTE-A

2. **Protocolos de roteamento de camada de rede:**
    - RPL (Routing Protocol for Low-Power and Lossy Networks)
    - CORPL (cognitive RPL)
    - CARP (Channel-Aware Routing Protocol)
3. **Protocolos de encapsulamento de camada de rede:**
    - 6LoWPAN (IPv6 over Low power Wireless Personal Area Network)
    - IPv6 over Bluetooth Low Energy
4. **Protocolos da camada de sessão e aplicação:**
    - MQTT (Message Queue Telemetry Transport)
    - AMQP (Advanced Message Queuing Protocol)
    - CoAP (Constrained Application Protocol)

Apesar da diversidade de protocolos IoT que surgiu recentemente e estão surgindo, os focos principais em sua aplicação são os mesmos: desempenhar, aprimorar e desenvolver suporte suficiente para ecossistemas IoT voltados ao baixo poder computacional, tamanho reduzido (em sua maioria) dos pacotes, tolerância ao atraso, qualidade do enlace de comunicação, pilhas de protocolos, taxa geral de dados e interoperabilidade (NGUYEN *et al.*, 2015).

2.4 Considerações finais

Neste capítulo foram abordados conceitos e definições relacionados à classificação do tráfego de rede, aprendizado de máquina e Internet das coisas, foco desta dissertação.

Inicialmente, foram apresentadas, detalhadamente, as evoluções das técnicas de classificação de tráfego de rede e as formas como elas são utilizadas para classificar o tráfego. Além disso, foi apresentada a classificação por aprendizado de máquina e suas métricas para avaliação de desempenho. Ainda foram apresentadas as definições de aprendizado de máquina, a importância da engenharia de *features*, e as definições para avaliação das métricas por matriz de confusão.

O foco da contextualização empregada para IoT foi apresentar algumas das definições empregadas, os protocolos mais comumente utilizados, as expectativas quanto ao seu uso, a evolução e o grande potencial econômico dessa tecnologia.

Com base neste capítulo e no capítulo 3 (trabalhos relacionados), propomos uma estratégia para classificar o tráfego e os dispositivos em ecossistemas de IoT. Além disso, é

apresentada a caracterização da rede, detalhada nos capítulos 4 e 5.

3 TRABALHOS RELACIONADOS

Neste capítulo são apresentados os trabalhos relacionados. As seções estão estruturadas da seguinte forma: a seção **3.1** apresenta os trabalhos relacionados à classificação de tráfego de rede, destacando os que utilizam técnicas baseadas em ML; a seção **3.2** apresenta os trabalhos que abordam a classificação de tráfego de rede focados em ecossistemas de IoT e apresenta uma discussão sobre trabalhos focados em identificar ou classificar os dispositivos de IoT através do uso de técnicas variadas; a seção **3.3** apresenta um quadro comparativo entre os principais trabalhos destacados neste capítulo; e, por fim, a seção **3.4** apresenta as considerações finais.

3.1 Classificação de tráfego de rede

O uso da classificação do tráfego de rede surge como requisito fundamental para que gerentes e operadores de redes possam realizar uma série de ações, dentre elas a priorização do tráfego. Decorrentes dessa capacidade, muitas pesquisas foram realizadas, o que reflete o interesse da comunidade e da iniciativa privada pela gestão eficiente da rede. Durante a busca por trabalhos que lidam com a classificação do tráfego de rede, enfatizando a classificação estatística, foi identificada uma vasta literatura que aborda várias formas distintas de aplicações e o uso de diversos algoritmos. Relacionado ao uso de ML para a classificação, percebe-se que os autores aplicam-no para promover ações automatizadas e com alta precisão, exemplo disso são os trabalhos (ERMAN *et al.*, 2006), (NGUYEN; ARMITAGE, 2008), (CALLADO *et al.*, 2010), (ZHANG *et al.*, 2013b), (WANG, 2013), (NAMDEV *et al.*, 2015), (ZHANG *et al.*, 2015) e (MIDDLETON; MODAFFERI, 2016).

A classificação estatística do tráfego baseia-se na lógica de que, decorrente da diversidade da natureza das aplicações (*Video Streaming* vs *Chat*, por exemplo), é possível realizar sua classificação através da utilização de características generalistas com o intuito de identificar um comportamento determinístico (quantidade de Bytes, tamanho dos pacotes e taxa de transmissão típicos, por exemplo). Em (NGUYEN; ARMITAGE, 2008), os autores fornecem uma abrangente pesquisa, através de *Survey*, sobre técnicas e metodologias de classificação de tráfego de rede utilizando ML.

Ao longo dos anos, abordagens de classificação de tráfego de rede, utilizando algoritmos de ML, foram aplicadas e otimizadas para adaptar-se aos contextos de seu tempo. Foram selecionados alguns trabalhos em períodos diferentes para realizar uma pequena análise

dessa evolução. Em 2005, os autores (ZANDER *et al.*, 2005) abordaram a importância de desenvolver métodos alternativos para classificar e identificar o tráfego com eficiência, em detrimento da baixa acurácia apresentada em modelos que utilizam portas e DPI. Nesse caso, eles propuseram o uso de ML através de método não supervisionado, utilizando características extraídas do fluxo de diversas redes, foram elas: *Inter-Arrival Time* (IAT); tamanho médio e variância dos pacotes; volume do fluxo (bytes); duração. Ao final, a média de acurácia apresentada foi de 86,5%. Os autores utilizaram uma pequena quantidade de classes para avaliação, correspondendo a grande maioria das aplicações de rede da época (FTP, HTTP e Telnet, por exemplo).

Em 2010, os autores (SOYSAL; SCHMIDT, 2010), na busca por resultados com acurácia elevada para ambientes de alta velocidade e com criptografia compararam várias técnicas de ML empregadas em classificação do tráfego de rede. Foram utilizadas redes Bayesianas, árvores de decisão e *Multilayer Perceptrons*, todos através do software Weka[1]. A avaliação dos autores foi realizadas mediante a classificação das seguntes classes: P2P, *Cloud content delivery* (Akamai), Web (HTTP), Dados (FTP), serviços (DNS) e E-mail (SMTP, POP), correspondendo, segundo os autores, a mais de 90% das classes de aplicações da época. Ao final, o melhor resultado encontrado foi o de 99.2% na classificação utilizando árvores de decisão.

Mais recentemente, em 2015, o trabalho de (NG *et al.*, 2015) aborda a classificação de tráfego de rede sobre a perspectiva de SDN (Redes Definidas por Software), haja vista que é um novo paradigma que possui grande impacto sobre futuras redes (e.g., IPs, 5G e sem fio), pois é uma abordagem inovadora para arquiteturas de redes que possibilita criar novas classes de funcionalidades de rede. O autor foca mais na construção de uma plataforma escalável, utilizando OpenFLow, com suporte, inclusive, a redes IoT.

Percebe-se, após essa análise, uma grande adaptação dos modelos e dos métodos para classificar o tráfego através de ML, assim como a adaptação e evolução das redes e da forma de avaliá-las.

3.2 Classificação de tráfego de rede focada em IoT

Com a evolução de IoT, uma grande variedade de dispositivos, tais como câmeras IP, impressoras, telefones IP e IPTVs, começaram a ser introduzidas às redes. Devido a essa disseminação, na literatura surgem trabalhos focados em classificar tráfego nesses ambientes,

[1] https://www.cs.waikato.ac.nz/ml/weka/

dedicados principalmente à resolução de problemas de segurança, *QoS* e priorização de tráfego de rede. Após a revisão de literatura, percebe-se um elevado número de trabalhos na busca pela introdução de mecanismos de segurança para detecção de ativos na rede, anomalias e dos tipos específicos de tecnologias utilizadas na comunicação dos dispositivos, identificando, inclusive, qual tráfego é IoT e o qual não é. Por exemplo, Meidan *et al.* (2017b) propõem um classificador utilizando múltiplos estágios. Os autores aplicam meta-classificação para distinguir tráfego gerado por dispositivos IoT e Não-IoT. Além disso, o autor faz uso das características estatísticas extraídas em forma de *4-tuple*[2] em uma classificação supervisionada.

Em relação às formas de identificar o tráfego gerado por dispositivos e diferenciá-los em IoT e Não-IoT, Apthorpe *et al.* (2017) explicam que o uso do endereço MAC facilita a identificação dos dispositivos e a classificação, em IoT e Não-IoT, através da aplicação de análise de tráfego de rede e consultas DNS, pois a análise do MAC limita o número de possibilidades existentes. Eles inclusive afirmam que os dispositivos IoT utilizam menos *DNS queries*.

Os autores Sharma *et al.* (2018), utilizando como base a classificação estatística do tráfego de rede, apresenta uma estratégia com o intuito de promover *QoS* voltada ao agendamento e à transmissão eficiente de pacotes por prioridade. A abordagem foca no controle e no gerenciamento da taxa de transmissão, do recebimento dos pacotes e do *Buffer*. Além disso, ela é baseada na análise da necessidade de banda larga dos componentes presentes na rede. O modelo de alocação da prioridade é feito utilizando cadeias de Markov, já o tipo de tráfego, IoT e Não-IoT, é classificado através da análise da taxa de transmissão dos dados e do tamanho médio dos pacotes.

No trabalho de (HAFEEZ *et al.*, 2017a), os autores propôs a implementação de uma plataforma para aprimorar a segurança na comunicação entre dispositivos D2D (*Device to Device* (D2D)). Foi utilizada lógica *Fuzzy*, no qual o modelo é treinado remotamente, junto a informações contextuais centradas nos dispositivos. A classificação do tráfego ocorreu para identificá-los entre maliciosos ou normal. O seu foco é promover segurança na comunicação entre dispositivos.

3.2.1 Classificação do tráfego IoT através da identificação dos dispositivos

Classificação de tráfego e identificação de dispositivos em ecossistemas de Internet das coisas, através de uso de ML, tem recebido bastante atenção da academia (SIVANATHAN

[2] *4-tuple*: IP origem, IP destino, porta origem e porta destino

et al., 2017)(SIBY *et al.*, 2017)(MEIDAN *et al.*, 2017b)(HAFEEZ *et al.*, 2017a). Além disso, há várias propostas de modelos desenvolvidos visando identificá-los para múltiplas finalidades, como é o caso de Wang *et al.* (2016), onde os autores propõem um método para identificar os dispositivos através do uso de identificação de camada física sem fio (*Wireless Physical Layer Identification* – WPLI), visando classificar os dispositivos autorizados com base em impressões digitais únicas de radiofrequência (RFFs – *radio frequency fingerprinting*). Aqui, o foco é garantir segurança na camada física de cada dispositivo individualmente.

Relacionado à análise e à classificação do tráfego de rede e à identificação de dispositivos de IoT, o trabalho de Ng *et al.* (2015) aborda alguns desafios que podem ocorrer no processo de classificação de tráfego de rede em ambientes com dispositivos inteligentes, principalmente pelas características dos ecossistemas IoT. Além disso, os autores afirmam que operadores de redes empresariais, interessados em QoS, não conhecem todos os aplicativos que estão sendo executados em sua rede. Inclusive, com o advento de IoT, que conecta objetos do cotidiano à rede, tem-se um aumento significativo do problema mencionado. Ainda segundo os autores, existe a necessidade do desenvolvimento de soluções rápidas e automatizadas para monitorar, classificar e configurar de forma eficiente o tráfego de rede.

Relativo à identificação de dispositivos, a literatura apresenta abordagens e análises específicas para esses ambientes. Exemplo disso é Sivanathan *et al.* (2017), onde os autores, ao analisarem a rede em um cenário emulado de *smart campus*, afirmam que dispositivos IoT tendem a usar um número limitado e específico de protocolos de camada de aplicação, além de consultas de DNS limitadas e inferiores, em quantidade, aos dispositivos Não-IoT. Relacionado a uma análise relativa à identificação de dispositivos de IoT, temos que, segundo Apthorpe *et al.* (2017), o uso do endereço MAC e de consultas DNS facilitam a aplicação de análise de tráfego de rede para realização da identificação dos dispositivos, uma vez que o seu uso limita o número de possibilidades existentes.

No artigo de Sivanathan *et al.* (2017), cujas capturas[3] (*.pcap.gz) utilizamos nesta dissertação, os autores realizaram, primeiramente, uma clusterização utilizando o *K-Means* (algoritmo de agrupamento interativo que objetiva particionar observações em grupos mais próximos da média) com o intuito de avaliar o comportamento de alguns recursos em relação aos dispositivos IoT e Não-IoT. Ademais, realizaram uma classificação supervisionada do tráfego de rede utilizando características extraídas do tráfego (número de servidores conectados

[3] http://149.171.189.1

e tempo de inatividade, por exemplo). Primeiramente, eles utilizaram o modelo de partição 10-*fold* na avaliação de um conjunto de dados. A média do resultado foi de 97% de acurácia e, posteriormente, foi utilizado um conjunto de dados não rotulados (correspondendo à semana seguinte), para teste e, ao final, o resultado foi de 95% de acurácia.

Outro trabalho (MEIDAN *et al.*, 2017b) apresenta um algoritmo desenvolvido para realizar a identificação de dispositivos de Internet das coisas através de fluxo da rede agrupados em 4-*tuple*[4], técnicas de meta classificação e aprendizado de máquina supervisionado. A abordagem baseia-se na rotulação (BenchMark) de parte do tráfego apenas com os dispositivos IoT para treinamento e aprimoramento do modelo. Após isso, aplicaram o modelo treinado na própria rede não classificada. Esse método para identificação de dispositivos apresentou ao final 99,281% de acurácia.

Em um outro artigo (MEIDAN *et al.*, 2017a) foi proposto um método utilizando ML para identificar dispositivos sem autorização de acesso em uma rede através do algoritmo *Random Forest*. O algoritmo selecionado foi aplicado a recursos extraídos do tráfego de rede, objetivando identificar com precisão tipos de dispositivos IoT permitidos e presentes em uma lista branca. Para treinar e avaliar o classificador, coletaram e rotularam manualmente os dados de tráfego de rede de 17 dispositivos distintos, representados por nove classes. Os testes foram realizados com a introdução de novos dispositivos, que não estavam presentes na lista de autorização, e, ao final, obteve, em média, 96% de acurácia na identificação dos dispositivos sem autorização e 99% dos dispositivos presentes na lista.

3.3 Comparativo entre trabalhos

O quadro comparativo na Tabela 6 apresenta uma visão geral das abordagens avaliadas neste capítulo. Os critérios utilizados para avaliá-los foram extraídos baseados em nossa proposta e nos principais pontos de cada um dos trabalhos.

1. Classifica utilizando ML;
2. Utiliza DPI;
3. Aborda classificação de dispositivos de IoT;
4. Classifica tráfego de rede em IoT e Não-IoT;
5. Utiliza seleção de atributos estatísticos;
6. Utiliza métricas de desempenho para avaliação dos resultados.

[4] IPs (origem e destino) e portas (origem e destino)

Tabela 6 – Tabela comparativa entre trabalhos relacionados

AUTORES	Ambiente	1	2	3	4	5	6
(ZHANG *et al.*, 2013a)	Não-IoT	✓	✓	x	x	✓	✓
(ZHANG *et al.*, 2015)	Não-IoT	✓	✓	x	x	✓	✓
(SIBY *et al.*, 2017)	IoT	x	x	✓	✓	x	✓
(APTHORPE *et al.*, 2017)	IoT	✓	x	✓	x	x	x
(SIVANATHAN *et al.*, 2017)	IoT	✓	x	✓	✓	x	✓
(MEIDAN *et al.*, 2017a)	IoT	✓	x	✓	x	✓	✓
(MEIDAN *et al.*, 2017b)	IoT	✓	✓	✓	✓	x	✓
(HAFEEZ *et al.*, 2017a)	IoT	✓	x	✓	x	✓	✓
(SHARMA *et al.*, 2018)	IoT	✓*	x	x	✓	x	✓

***Utiliza modelos probabilísticos**

Todos os trabalhos apresentados na Tabela 6 possuem relação estreita com esta proposta, seja pelo foco em IoT, seja pelo processo de identificação ou classificação através do uso de ML. Os trabalhos apresentados são atuais e lidam com a proposta de prover várias funcionalidades no uso da classificação do tráfego de rede, dentre as quais se destacam a segurança e *QoS*.

Comparadas as propostas, foi possível identificar que as soluções apresentadas não promovem, em conjunto, a classificação do tráfego em IoT, a identificação das classes de aplicação e caracterização do tráfego, a classificação dos dispositivos e a seleção de atributos estatísticos para aprimorar os resultados. Dessa forma, o próximo capítulo apresenta a solução proposta, sua arquitetura, as características utilizadas e o método de avaliação de desempenho empregado.

3.4 Considerações finais

Este capítulo apresentou os trabalhos relacionados focados na análise de diferentes métodos utilizados na literatura e voltados à classificação do tráfego de rede e dos dispositivos de IoT. Através da revisão da literatura foram encontrados vários trabalhos relacionados a esta proposta. Assim, foram elencadas as seguintes características para comparação entre as soluções: tipos de ambientes; classificação utilizando ML; utilização de DPI; conceitos de classificação de dispositivos de IoT; classificação de rede em IoT e Não-IoT; utilização de seleção de características; e utilização de métricas de desempenho.

4 ESTRATÉGIA PARA CLASSIFICAÇÃO DE TRÁFEGO DE REDE E DE DISPOSITIVOS EM AMBIENTES DE IOT

Ao longo deste capítulo é apresentada a formulação da estratégia, o algoritmo *Random Forest*, as características estatísticas extraídas dos pacotes organizados em *5-tuple*[1] e a metodologia proposta para classificar os dispositivos e o tráfego IoT. Além disso, é apresentada uma visão geral sobre o processo, explicando-o detalhadamente. Depois disso, são descritas as funcionalidades previstas e suas oportunidades de aplicação. Por fim, é realizada uma discussão acerca da estratégia junto às conclusões.

O restante deste capítulo está assim organizado: a seção **4.1** aborda os conceitos introdutórios, que destina-se a detalhar algumas questões acerca dos problemas abordados; em seguida, a seção **4.2** apresenta a arquitetura da aplicação desenvolvida para extração das características estatísticas que possibilita a classificação por análise passiva da rede; por fim, a seção **4.3** apresenta as considerações finais.

4.1 Conceitos introdutórios

Ao analisar a literatura, observa-se que ao utilizar a classificação de tráfego de rede é possível adquirir uma série de benefícios, pois a mesma possibilita agregar um número maior de informações sobre a dinâmica, as características e o comportamento da rede. Essas informações podem ser utilizadas para priorização de tráfego ou criação de políticas de segurança voltadas a uma gestão de rede mais eficaz. Decorrente das vantagens ao empregá-la, é proposta uma estratégia voltada, especialmente, para ambientes de IoT, ligada diretamente à identificação dos dispositivos na rede, fazendo uso combinado de técnicas de DPI e ML. Além disso, ao identificá-los, torna-se possível inferir uma série de ações com diversos propósitos, como identificar dispositivos sem autorização de acesso e distinguir entre tráfego estabelecido por todos os dispositivos, permitindo a análise de rede e dos padrões de comunicação, o que possibita, inclusive, a tomada de decisão em diversas situações.

Baseado em cenários atuais de IoT, há uma série de eventos urgentes que necessitam de atenção; como privacidade dos usuários, identificação de ativos na rede e o controle dos dados coletados pelos sensores. Decorrente do nível emergencial que essa situação impõe, torna-se indispensável a construção de soluções robustas para promover maior gerência e controle. É através desta meta que foi concebida a abordagem desta dissertação para classificar o tráfego de

[1] Além da *4-tuple*, usa-se a informação do protocolo da camada de transporte utilizado, TCP ou UDP.

rede em IoT ou não-IoT. Contudo, há algumas premissas para realizá-la, entre as quais tem-se a identificação dos dispositivos mediante varreduras de assinaturas nos pacotes. A subseção 4.2.4 apresenta alguns desses componentes inspecionados para uso na classificação. Informações relacionado ao uso de ML, o algoritmo utilizado e sua formulação estão presentes na subseção 4.2.5.

4.2 Arquitetura da aplicação

O uso de ferramentas para monitoramento de tráfego são comuns e amplamente utilizadas, visto que suas funcionalidades possibilitam a aquisição de informações para realizar o gerenciamento de redes. Suas principais aplicabilidades estão associadas à coleta dos dados para análise de informações voltadas, especialmente, para a resolução de problemas diversos.

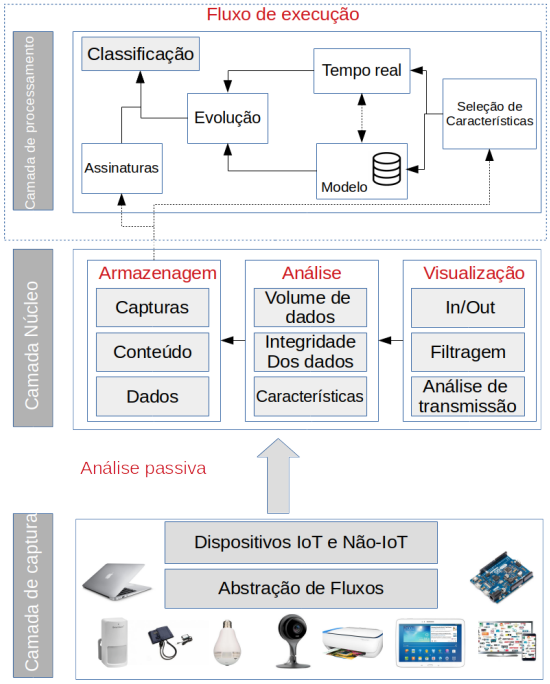
Análise de tráfego de rede surge como uma ferramenta de suporte para vários domínios, tais como controle de congestionamento, desenvolvimento de protocolos, entendimento do comportamento da rede e estudo de variabilidade. A forma de realizar a análise de rede pode ocorrer em várias perspectivas, como em nível de aplicação, da camada de transporte e da própria rede. A coleta dos dados, em muitos casos, são avaliadas através de mecanismos e técnicas desenvolvidas e refinadas continuamente, ocorrendo de forma passiva ou ativa.

A Figura 5 apresenta as composições do monitor de tráfego de rede (em camadas e seu fluxo de execução) desenvolvido para a aplicação junto à estratégia apresentada nesta dissertação. O seu desenvolvimento foi concebido para, inicialmente, extrair e analisar recursos primários, como a integridade dos dados capturados, tamanho total da coleta e suas características. Outrossim, foi projetado para a análise de rede mediante coleta promíscua de pacotes ou de carregamentos de arquivos de rastreamento (*.pcap) anteriormente coletados, o que tende a não promover impactos sérios no funcionamento normal, mesmo em uma análise em tempo real.

Em sua composição, há vários registros salvos com dados (User-Agent, Host e mDNS, por exemplo) de dispositivos, IoT e não IoT, para identificação e análise inicial por varredura de conteúdos específicos dos pacotes (assinaturas). Isto significa que os valores serão coletados e comparados com os presentes na base de dados, e uma vez igualados, serão salvos os valores $T = \{IP/MAC, Tipo, Dispositivo\}$ em forma de uma tupla. Esta tupla terá seu uso na classificação dos dispositivos e do tráfego de rede, em IoT e Não-IoT.

O núcleo da ferramenta (*Network Monitor Core*) apresenta uma série de funcionalidades iniciais voltada à visualização dos pacotes de entrada, à sua análise e à salva-guarda local.

Figura 5 – Arquitetura do monitor de tráfego de rede



Fonte: Baseado em (ANDREONI *et al.*, 2017)

Dentro de seu escopo, sobretudo para agrupamento dos pacotes em fluxo *5-tuple*[2] e a montagem da tupla T, tem-se, primeiramente, a ação de varredura nos pacotes de arquivos *.pcap, voltada à identificação, unívoca, do número máximo de objetos inteligentes, utilizando-os como premissa fundamental da classificação. Entretanto, a máxima eficiência da classificação do tráfego em IoT está na identificação de todos os dispositivos. Isto siginifica que a não identificação de os dispositivos nessa primeira etapa implicará na utilização da classificação estatística, fazendo uso dos fluxos extraídos de classes de dispositivos e, dessa forma, associando o tráfego a uma das classe de dispositivos, tudo isso de forma supervisionada.

A arquitetura apresentada é a conclusão da ferramenta proposta em (PINHEIRO; CASTRO., 2017), junto à estratégia apresentada em (SANTOS *et al.*, 2018). A sua concepção é voltada para análise de rede, classificação do tráfego e identificação dos dispositivos.

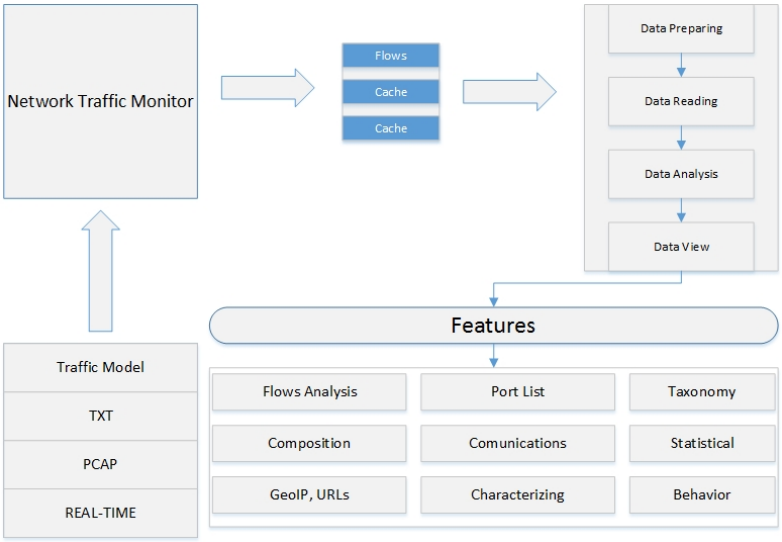
4.2.1 Funcionalidades

Dentre as muitas funcionalidades necessárias a um monitor de tráfego IoT, como realizar a extração de características de comunicação, identificar anomalias e verificar se os nós

[2] IP de Origem, Porta de Origem, IP de Destino, Porta de Destino e Protocolo (TCP ou UDP)

estão se comunicando corretamente, há o desafio de tornar simples a tarefa de gerenciamento da rede. Desse modo, é proposta uma forma simplificada de realizá-la, através do conhecimento dos dispositivos conectados em rede e se estes correspondem a IoT ou não. Na Figura 6 verificam-se informações extraídas para análise de redes, incluindo de IoT, dentre eles, a lista de portas, com intuito de identificar as mais utilizadas entre os dispositivos, a classificação taxonômica do tráfego de rede, voltada à identificação das classes de protocolos e, também, a análise de comunicação dos dispositivos, com o intuito de identificar todos os *endpoints*, além das cargas de tráfego. Esses recursos juntos permitem uma série de ações importantes, dentre as quais se destacam a segurança e o aprovisionamento de recursos.

Figura 6 – Funcionalidades associadas à ferramenta



Fonte: Autor

Além da coleta de recursos fundamentais ao gerenciamento de tráfego de rede IoT, a arquitetura proposta na Figura 5 permite, na funcionalidade *Network Traffic Monitor* (ver Figura 6), a coleta de tráfego em tempo real, convertendo-os em *.pcap, caso desejado, ou realizar a análise de pacotes de rastreamentos coletados anteriormente (*.pcap). Com esses dados são realizadas as extrações de informações como fluxo individual, User-Agent, Endpoint, DNS *queries*, filtragem de pacotes e taxa de explosão - adquiridas através do uso de *wrappers*[3] *Python* para o *tshark*[4]. A arquitetura é proposta, inclusive, para promover o auxílio e os recursos para classificação de tráfego e identificação dos dispositivos, tornando-a mais ágil, através de apropriação e reutilização de informações, podendo ser utilizadas posteriormente.

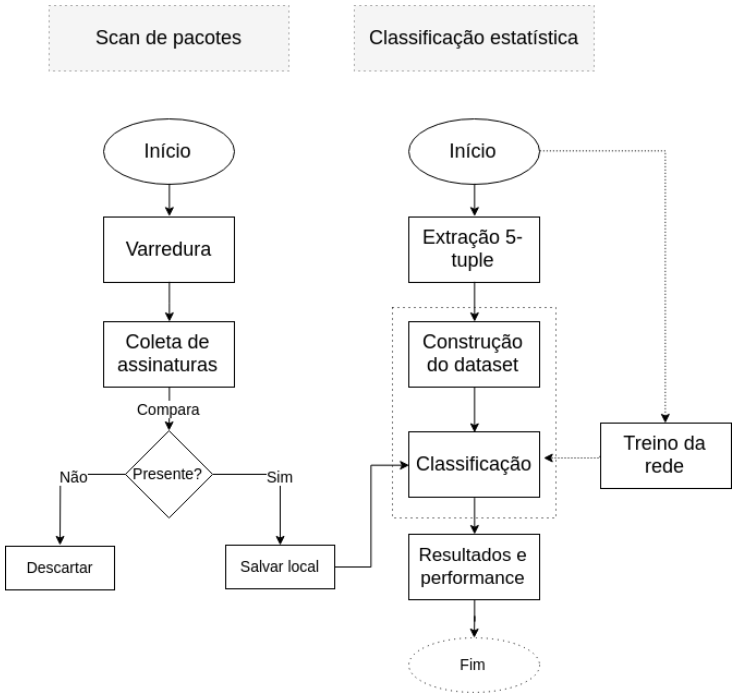
[3] Sub-rotina em uma biblioteca de software
[4] Protocolo para análise de rede

4.2.2 Procedimento

Em primeiro lugar, é feita uma varredura em todos os arquivos na busca por User-Agents, por DNS e mDNS *queries*, pelo MAC[5] e por IPs locais. Dessa forma, dispositivos automaticamente identificados por assinaturas, principalmente pelo User-Agent, comparando com uma base de dados Json, tem os correspondentes IP/MAC associados. É aplicada a classificação estatística para identificar os demais dispositivos conectados na rede local, utilizando os fluxos extraídos da fase de treino.

Realizar a identificação dos dispositivos permite corresponder o tráfego gerado, classificar o tráfego em IoT ou não e seu impacto na rede, principalmente. O processo ocorre como descrito na Figura 7. Será considerado tráfego de IoT, para propósito geral, aquele que foi originado ou destinado a dispositivos de IoT.

Figura 7 – Gráfico de fluxo



Fonte: Própria

4.2.3 Organização dos pacotes em forma de tupla

A CAIDA (2013) define fluxo como "um conjunto de pacotes que compartilham uma propriedade comum". Além disso, ele afirma que o tipo mais simples de fluxo é o denominado 5-

[5] A resolução do endereço MAC é feita através da conversão dos três primeiros bytes do endereço Ethernet para o nome da empresa de fabricação do equipamento.

tupla, consistindo em ip de origem e destino, porta de origem e de destino e o protocolo utilizado (UDP ou TCP). Inclusive, fluxos organizados dessa maneira são unidirecionais, onde seus pacotes viajam em uma mesma direção (o sentido inverso é definido como um fluxo diferente). Além disso, eles são referidos como *microfluxos* (TAYLOR, 2005).

A organização dos pacotes em forma de fluxo manifesta-se como uma necessidade de otimização na análise de pacotes. Para Nottingham e Irwin (2010), ela surge como requisito de desempenho para várias aplicações em redes IP, inclusive algoritmos de roteamento, além de fornecer maior agilidade em classificação de tráfego de rede. Essa melhoria no desempenho é adquirida pelo rearranjamento de milhões ou bilhões em um número reduzido de fluxos que compartilham uma propriedade comum. Para este trabalho, os fluxos serão considerados na forma bidirecional, onde é considerado IP de origem o endereço da máquina que enviou o primeiro pacote do fluxo.

4.2.4 Componentes para identificação de dispositivos

Realizar a análise do tráfego de rede para identificar componentes e dispositivos é uma técnica simples e executada em atividades forenses. Dessa forma, é feita a inspeção do conteúdo dos pacotes com o intuito de identificar, ao máximo, os dispositivos através da coleta de informações presentes nos pacotes. Os principais componentes utilizados para identificá-los foram:

- **User-Agent:** É uma String de requisição, presente no cabeçalho, que possibilita ao protocolo de rede (cliente) identificar o tipo de aplicação através de um "catálogo" com dados técnicos sobre o dispositivo, sistema operacional, aplicação, versão ou fornecedor do software;
- **Resolução do Endereço MAC:** É uma técnica de resolução que possibilita converter valores de formato numérico em algum modelo legível. Ela procede através da análise e conversão de códigos de manufatura atribuídas pela IEEE.
- **mDNS:** *multicast Domain Name System*, presente na RFC 6762, converte nomes de host para endereços IPs dentro de pequenas redes que não incluem um servidor de nomes local.
- **Resolução de nome TCP/IP:** Em redes TCP/IP, as máquinas possuem um arquivo (/etc/hosts) que contém informações de mapeamento de nome para Internet (*name-to-Internet-address*). O processo de obtenção de um endereço de Internet a partir de um nome de host, conhecido como resolução de nome, é feito pela sub-rotina *gethostbyname*,

presente em protocolos como o Tshark.

4.2.5 Random Forest

O RF é utilizado em uma grande variedade de áreas de pesquisas, como em detecção de anomalias (PRASHANTH *et al.*, 2008), classificação de tráfego (WANG *et al.*, 2015) e, mais recentemente, na identificação de dispositivos IoT (SIVANATHAN *et al.*, 2017)(MEIDAN *et al.*, 2017a). Decorrente de seus ótimos resultados, de sua grande variedade de aplicação e de suas vantagens, este algoritmo foi selecionado para uso em nossa estratégia.

Uma das principais características avaliadas para fazer uso deste algoritmo foi a sua simplicidade de implementação e o fato de que pode ser utilizado tanto para tarefas de classificação quanto para de regressão.

4.2.6 Aquisição das características estatísticas do fluxos

O fluxo é extraído e organizado em forma de *5-tuple*. As características estatísticas dos fluxos podem ser coletadas e transformadas em datasets (*.csv) através da ferramenta (PINHEIRO; CASTRO., 2017) e, inclusive, agrupadas de forma cronológica. Elas são extraídas e dispostas em vetores no *dataset*. Uma lista composta pelas principais características extraídas é apresentada na Tabela 7. Elas foram selecionadas para promover e identificar capturas associadas aos ecossistemas IoT, possibilitando os processos subsequentes. Para tanto, são consideradas, principalmente, características de tempo, tamanho (médio, mínimo e desvio padrão) e processamento dos pacotes (quantidade enviada, recebida, tempo ativo).

Tabela 7 – Características extraídas através dos fluxos 5-tupla TCP e UDP

Características	**Descrição**	**Característica**	**Descrição**
Ip_src	IP de origem	Ip_dst	IP de destino
src_port	Porta de origem	dst_port	Porta de destino
protocolo	TCP (6) ou UDP (17)	total_fPacote	total de pacotes na ida
total_bPacote	total de pacotes no sentido de volta	total_fBytes	total de bytes de ida
total_bBytes	total de bytes na volta	mim_fPacote	valor do menor pacote na ida
mim_bPacote	valor do menor pacote na volta	med_fPacote	tamanho médio dos pacotes na ida
med_bPacote	tamanho médio dos pacotes na volta	max_fPacote	tamanho máximo do pacote na ida
max_bPacote	tamanho máximo do pacote na volta	Dpd_fPacote	desvio padrão do tamanho na ida
Dpd_bPacote	desvio padrão dos pacotes na volta	IATmim_fPacote	menor tempo de envio na ida
IATmim_bPacote	menor tempo de envio na volta	IATmed_fPacote	tempo médio de envio na ida
IATmed_bPacote	tempo médio de envio na volta	IATmax_fPacote	tempo máximo de envio na ida
IATmax_bPacote	tempo médio de envio na volta	IATdpd_fPacote	desvio padrão de envio na ida
IATdpd_bPacote	desvio padrão na volta	duração	tempo total de comunicação

Ida: Relacionado à direção percorrida pelo pacote, relativo ao cliente (IP src_ip e porta src_port)
Volta: Relacionado ao percurso reverso dentro de um fluxo.(dst_ip e dst_port)

4.3 Considerações Finais

Neste capítulo foram apresentados a arquitetura do monitor de tráfego de rede desenvolvido junto a estratégia para classificação, os componentes dos pacotes utilizados para o processo de classificação, o detalhamento da estratégia desenvolvida para classificar os dispositivos e o tráfego em IoT e as características estatísticas extraídas do fluxo de rede.

Esta proposta visa resolver um problema pertinente relacionado à segurança e ao gerenciamento dos dispositivos de IoT em redes de forma prática. No capítulo seguinte (ver capítulo 5) será apresentado o uso prático da estratégia formulada, assim como os seus resultados quando avaliados em ambientes com dispositivos de IoT.

5 AVALIAÇÃO EXPERIMENTAL DA ESTRATÉGIA

Este capítulo apresenta a avaliação experimental da proposta. O intuito é demonstrar a efetividade da estratégia proposta na classificação de dispositivos de IoT e de tráfego de rede em alguns cenários. Os resultados dos experimentos foram gerados ao utilizar a estratégia proposta (ver capítulo 4) em arquivos de rastreamentos de pacotes (*.pcap) disponibilizados publicamente para pesquisas. Relativo aos testes, foram avaliados dois cenários distintos, o primeiro contendo apenas dispositivos de IoT, cujo intuito principal é avaliar a capacidade de identificação dos dispositivos através da classificação por assinaturas; e o segundo, correspondendo a uma rede emulada de *smart campus*, possui o intuito de avaliar a estratégia proposta em um cenário cotidiano típico.

O restante deste capítulo está assim organizado: seção **5.1** apresenta uma introdução aos experimentos; a seção **5.2** apresenta os atributos estatísticos utilizados; logo em seguida, a seção **5.3** apresenta a composição do primeiro cenário avaliado e seus resultados; a seção **5.4**, apresenta o segundo cenário e seus resultados; por fim, a seção **5.5** apresenta as considerações finais dos testes e da estratégia.

5.1 Introdução aos Experimentos

Esta seção apresenta a análise experimental sugerida neste trabalho para demonstrar como a estratégia proposta possibilita classificar dispositivos de IoT e, consequentemente, o tráfego gerado. Pretende-se, ainda, demonstrar como ela permite melhorar o desempenho da análise, se comparada com as propostas de classificação puramente estatísticas, como presente em Kawai *et al.* (2017).

Em vista disso, foram conduzidos experimentos utilizando análise passiva em uma rede totalmente IoT, utilizando várias tecnologias de comunicação, e em uma rede emulada com diversos objetos conectados em ambiente de *smart campus*, composta por uma diversidade de dispositivos, como câmeras e lâmpadas IP. A análise experimental foi realizada sobre uma perspectiva de monitoramento de tráfego de rede passiva, isto é, após o carregamento de algum arquivo (*.pcap) para realizar o processo. Nos experimentos, é avaliada a eficiência da estratégia através das métricas apresentadas na seção 2.2.3.

5.2 Seleção de características

Com o intuito de aumentar o desempenho e remover possíveis redundâncias ou sobreposições, principalmente em relação ao tempo de treino e aumento da acurácia, foi realizada a remoção de características através de algoritmos de seleção. Nesse caso, foi utilizado o algoritmo **CfsSubsetEval** (HALL, 1998), através do modelo *Best first Search*, que realiza a seleção do subconjunto de características através da capacidade preditiva de cada elemento, juntamente com seu grau de redundância.

Foi utilizado, em todos os testes, um conjunto formado pelas 4 características consideradas mais relevantes, haja vista que o uso de um número maior não acarreta em um aumento relevante nos valores de acurácia. Para a identificação dos dispositivos por características estatísticas do fluxo foram selecionados o tamanho máximo do pacote no sentido de ida, porta de origem, porta de destino e o tamanho médio dos pacotes. Por fim, para a classificação do tráfego de rede foram selecionados o total de bytes na ida, tamanho máximo do pacote na ida, total de bytes na volta e média dos pacotes na volta.

5.3 Cenário 1

Inicialmente, é aplicada a prosposta para classificação de dispositivos em um ambiente totalmente IoT, isto é, sem a presença ou inclusão de dispositivos Não-IoT. Nesses testes serão avaliadas várias tecnologias de comunicação e, ao final, serão comparados os resultados entre a classificação puramente com ML e a proposta nesta dissertação.

O primeiro cenário avaliado, desenvolvido e disponibilizado por (MIETTINEN *et al.*, 2017), contém 27 dispositivos de internet das coisas que utilizam várias tecnologias de comunicação, dentre elas detacam-se o Wifi, ZigBee e Z-wave (AL-SARAWI *et al.*, 2017).

A Tabela 8 apresenta a lista dos dispositivos presentes, assim como a especificação e as tecnologias de comunicação empregadas em cada um deles. Para a classificação por ML, extraímos os atributos estatísticos apresentados na subseção 4.2.6, do capítulo 4, para realizar a comparação dos resultados adquiridos na classificação por ML e na classificação apresentada nesta dissertação.

Tabela 8 – Lista de dispositivos de IoT e tecnologias avaliadas

Nome	Especificação	Wifi	ZigBee	Ethernet	Z-wave	Outro
Aria	Fitbit Aria WiFi-enabled scale	✓				
HomeMaticPlug	Homematic pluggable switch HMIP-PS					✓
Withings	Withings Wireless Scale WS-30	✓				
MAXGateway	MAX! Home automation sensors			✓		✓
HueBridge	Philips Hue Bridge model 3241312018		✓	✓		
HueSwitch	Philips Hue Light Switch PTM 215Z		✓			
EdnetGateway	Ednet.living Starter kit power Gateway	✓				✓
EdnetCam	Ednet Wireless indoor IP camera Cube	✓		✓		
EdimaxCam	Edimax IC-3115W Smart HD WiFi Network Camera	✓		✓		
Lightify	Osram Lightify Gateway	✓	✓			
WeMoInsightSwitch	WeMo Insight Switch model F7C029de	✓				
WeMoLink	WeMo Link Lighting Bridge model F7C031vf	✓	✓			
WeMoSwitch	WeMo Switch model F7C027de	✓				
D-LinkHomeHub	D-Link Connected Home Hub DCH-G020	✓		✓	✓	
D-LinkDoorSensor	D-Link Door & Window sensor				✓	
D-LinkDayCam	D-Link WiFi Day Camera DCS-930L	✓		✓		
D-LinkCam	D-Link HD IP Camera DCH-935L	✓				
D-LinkSwitch	D-Link Smart plug DSP-W215	✓				
D-LinkWaterSensor	D-Link Water sensor DCH-S160	✓				
D-LinkSiren	D-Link Siren DCH-S220	✓				
D-LinkSensor	D-Link WiFi Motion sensor DCH-S150	✓				
TP-LinkPlugHS110	TP-Link WiFi Smart plug HS110	✓				
TP-LinkPlugHS100	TP-Link WiFi Smart plug HS100	✓				
EdimaxPlug1101W	Edimax SP-1101W Smart Plug Switch	✓				
EdimaxPlug2101W	Edimax SP-2101W Smart Plug Switch	✓				
SmarterCoffee	Smarter SmarterCoffee coffee machine SMC10-EU	✓				
iKettle	Smarter iKettle 2.0 water kettle SMK20-EU	✓				

5.3.1 Identificação dos dispositivos por análise de rede

Através da inspeção da carga útil dos pacotes e análise de rede é possível realizar a identificação de dispositivos típicos de IoT. A Tabela 9 representa essa análise, correspondendo ao user-agent, DNS e mDNS *queries* de vários dispositivos.

Tabela 9 – Alguns resultados da varredura por assinaturas

Nome	MAC Name Resolution	User-Agent	DNS	mDNS
Aria	DeltaEle_ca:91:52	-	www.fitbit.com	-
D-LinkCam	D-LinkIn_25:5b:0e	ksoap2-android	signal.auto.mydlink.com	iotcam.local
D-LinkDayCam	D-LinkIn_1c:71:85	-	-	DCS-930LB_1C7185
D-LinkHomeHub	D-LinkIn_aa:fd:4e	ksoap2-android/2.6.0	r0802.dch.dlink.com	D-Link HNAP Service
D-LinkSensor	D-LinkIn_a8:e1:43	Wget	api.dch.dlink.com	D-Link HNAP Service
D-LinkSiren	D-LinkIn_dd:0d:60	Wget	tzinfo.dch.dlink.com	D-Link DCH-S220 Configuration
D-LinkSwitch	D-LinkIn_a9:3d:6f	Wget/1.13(linux-uclibc)	ntp1.dlink.com	D-Link HNAP Service
D-LinkWaterS	D-LinkIn_c5:17:5a	Wget	r0802.dch.dlink.com	D-Link HNAP Service
EdmaxCam	EdimaxTe_80:7a:08	-	www.myedimax.com	localcam.local
EdimaxPlug1101W	EdimaxTe_4a:76:49	-	www.myedimaxservice.com	-
EdnetGateway	Hi-Flyin_62:3c:6e	-	icomen.yunext.com	-
HueBridge	PhilipsL_24:76:ff	Dalvik/2.1.0(Linux;Android)	bridge.meethue.com	Philips hue-2476FF
Lightify	Osram_7b:5f:6b	-	lb01.arrayent.com	Lightify-017b5f6b
MAXGateway	Eq-3Entw_03:cb:be	-	homematic.com	-
TP-LinkPlugHS100	Tp-LinkT_00:fc:a3	-	devs.tplinkcloud.com	-
WeMoInsightSwitch	BelkinIn_41:c2:05	-	mtalk4.google.com.Belkin	-
WeMoLink	BelkinIn_cd:37:65	-	mobile-gtalk4.l.google.com	www.belkin.com
Withings	Withings_24:80:2a	Withings UserAgent	scalews.withings.net	-

Em suma, através da inspeção dos pacotes é possível identificar vários dispositivos conectados, e a classificação do tráfego em IoT pode ser realizada ao associá-los através do MAC ou do IP.

5.3.2 Resultado da classificação dos dispositivos por ML

A Tabela 10 apresenta os valores adquiridos após a realização dos testes na rede. Foi obtido um valor de acurácia de 85,12%, um pouco maior que os 81,5% obtidos pelo autor do trabalho que realizou a captura (MIETTINEN *et al.*, 2017). Ao realizar a classificação apresentada nesta dissertação, baseada no conjunto de componentes identificados na fase de varredura, para aprimorar os resultados, é obtido, ao final, os valores médios apresentados na Figura 9, ligeiramente melhores.

Tabela 10 – Avaliação do modelo 10-*fold*

Dispositivo	**TPR**	**FPR**	**Precisão**	**Recall**	**F1-score**	**Mcc**
Aria	1,000	0,001	0,969	1,000	0,984	0,984
TP-LinkPlugHSA	0,993	0,009	0,789	0,993	0,879	0,881
TP-LinkPlugHSB	0,368	0,003	0,560	0,368	0,444	0,450
D-LinkCam	0,807	0,003	0,893	0,807	0,848	0,845
D-Linkdaycam	0,985	0,000	0,970	0,985	0,977	0,977
D-LinkHomeHb	0,523	0,012	0,623	0,523	0,568	0,556
D-linkSensor	0,964	0,036	0,737	0,964	0,835	0,825
D-LinkSiren	0,848	0,020	0,713	0,848	0,775	0,764
D-LinkSwitch	0,429	0,010	0,664	0,429	0,521	0,518
D-LinkWather	0,257	0,010	0,461	0,257	0,330	0,329
DoorSensor	0,761	0,001	0,897	0,761	0,824	0,825
EdmaxCam	0,935	0,001	0,896	0,935	0,915	0,914
EdimaxPlug_A	0,549	0,010	0,532	0,549	0,541	0,531
EdmaxPlug_B	0,387	0,008	0,460	0,387	0,420	0,413
Ednetcam	0,971	0,001	0,930	0,971	0,950	0,949
HomeMatic	1,000	0,000	1,000	1,000	1,000	1,000
Huebrig	0,951	0,011	0,947	0,951	0,949	0,938
HueSwitch	0,979	0,003	0,987	0,979	0,983	0,978
ikettle	0,625	0,002	0,625	0,625	0,625	0,623
Light	0,833	0,001	0,893	0,833	0,862	0,862
MaxGatway	0,813	0,001	0,897	0,813	0,852	0,852
Smartcoffe	0,450	0,002	0,500	0,450	0,474	0,472
Wemoindhghtwitch	0,981	0,010	0,871	0,981	0,923	0,919
WemosSwitch	0,609	0,002	0,862	0,609	0,713	0,719
Withings	0,609	0,001	0,824	0,609	0,700	0,707
EdnetGatway	1,000	0,001	0,964	1,000	0,982	0,982
WeMoLink	0,790	0,003	0,904	0,790	0,843	0,840
Média ponderada	0,850	0,010	0,844	0,850	0,841	0,836

Ao fazer uso da classificação puramente estatística (ML), obteve-se para 10 dispositivos valor de precisão na identificação superior a 0,95 (95%), já os demais dispositivos apresentaram precisão em torno de 0,7 (70%). Esses testes mostram valores de erro médio absoluto em 0,0185 e teste com cobertura de 98,43%, com intervalos de confiança criados para um nível de confiança de 95%. O valor médio de acurácia foi de 85,12%, o que é um bom valor

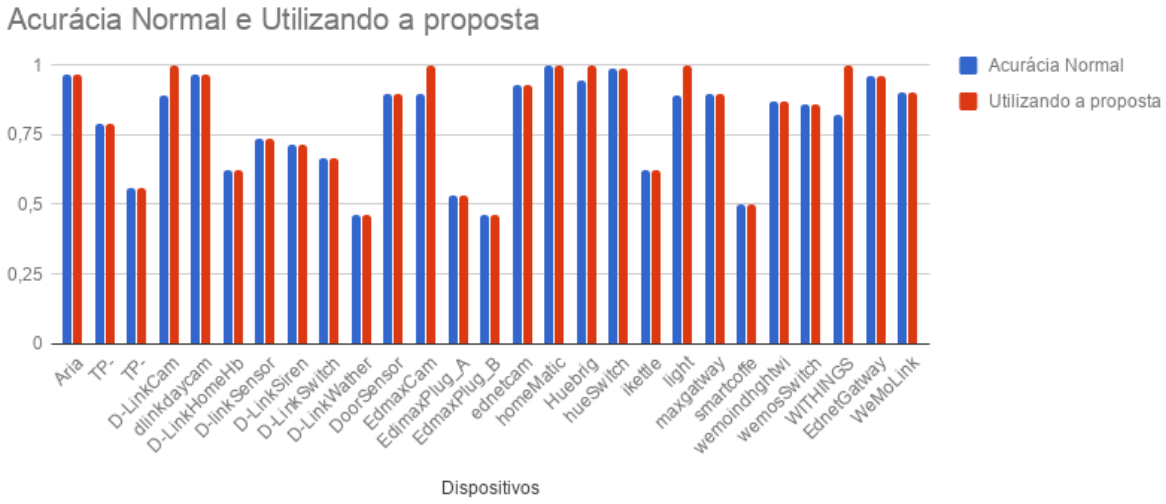
considerando a atribuição aleatória para os diferentes tipos existentes que seria em $\frac{1}{27} = 0,037\%$ de acurácia.

É possível observar que há um valor considerável de erros na identificação dos dispositivos. Uma das causas está na correspondência de objetos similares, entre os quais temos os 4 dispositivos D-Link, os 2 dispositivos Edimax e os 2 dispositivos TP-Link. Juntos, eles configuram-se como promovedores da diminuição acentuada da acurácia da classificação por ML.

5.3.3 Resultado da classificação através da estratégia proposta

A Figura 8 apresenta os valores individuais da classificação puramente estatística, utilizando o modelo *stratified 10-fold cross-validation*, e utilizando a estratégia proposta.

Figura 8 – Resultado individual das classificações dos dispositivos

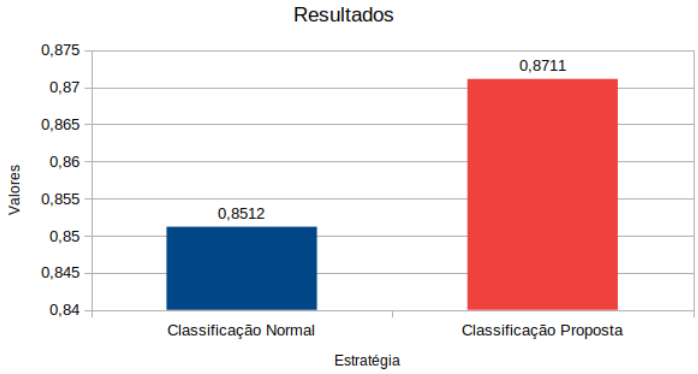


Fonte: Autor

Como resultado ao aplicar a classificação dos dispositivos, utilizando o modelo proposto, junto à classificação por ML, obteve-se, ao final, uma acurácia um pouco maior (87,11%). Essa leve melhoria foi possível por meio da identificação de vários dispositivos na fase de varredura, entre esses dispositivos destaca-se o *Withings*, que na classificação puramente com ML apresentou acurácia em aproximadamente 83% e com a estratégia proposta obteve 100% de acurácia.

O valor final adquirido no comparativo das acurácias demonstra uma siginificativa melhoria dos resultados ao aplicar a classificação proposta, com aumento em 1,99% no valor agregado (ver Figura 9).

Figura 9 – Comparativo da média de acurácia



Fonte: Própria

5.4 Cenário 2 e sua Arquitetura

O segundo cenário analisado possui sua visão geral representada na Figura 10. Ele corresponde a uma *smart campus* emulada, desenvolvida e disponibilizada por (SIVANATHAN *et al.*, 2017). Os autores realizaram a coleta da rede durante 3 semanas e todos os arquivos e algumas informações foram disponibilizados na Internet[1] publicamente. No total, os arquivos coletados possuem o tamanho de aproximadamente 12 Gb (*.pcap.gz).

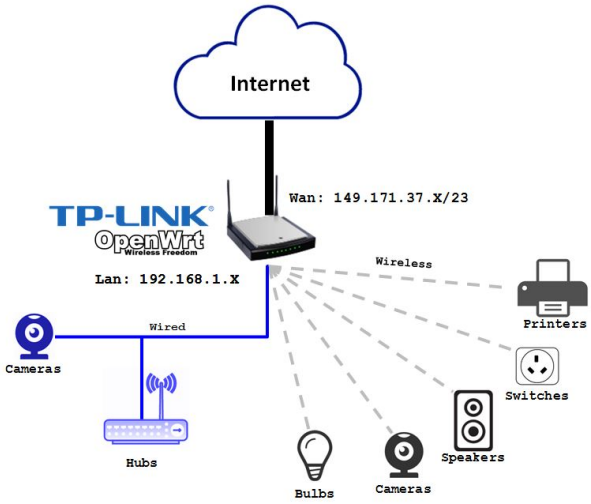
Os autores instrumentaram o ambiente com mais de 20 dispositivos IoT, coletaram os dados e os disponibilizaram publicamente. Segundo os autores, os dispositivos IoT foram configurados utilizandos os apps proprietários e instalados ao testbed. Uma variedade de dispositivos Não-IoT também foi agregada à rede, dentre os quais têm-se *Laptops*, *Smartphones* e *tablets*. O protocolo de camada de aplicação utilizado pelos dispositivos foi, principalmente, o HTTPS, seguido do HTTP e com uma pequena porção de MQTT. O uso do protocolo HTTP possibilita a extração de informações através da inspeção do conteúdo dos pacotes por não possuir criptografia, dentre eles temos o User-Agent, DNS *queries* e o Host. Porém, a inspeção do *payload* infringe a privacidade dos usuários, o que consolida crime em diversos países. Dessa forma, nossa abordagem busca, exclusivamente, a coleta de informações específicas de assinaturas, limitando o acesso e não produzindo prejuízos relevantes de privacidade.

A rede possui diversos dispositivos de IoT que podem ser agrupados da seguinte maneira:

- **Hubs:** Smart Things e Amazon Echo;
- **Câmeras:** Netatmo Welcome, TP-Link Day Night Cloud camera, Samsung SmartCam,

[1] http://149.171.189.1/

Figura 10 – Visão geral do Testbed mostrando os dispositivos e gateway de IoT



Fonte: (SIVANATHAN *et al.*, 2017)

Dropcam, 2 Insteon Cameras e Withings Smart Baby Monitor;

- **Switches e Triggers:** Belkin Wemo switch, TP-Link Smart plug, iHome e Belkin Wemo Motion Sensor;
- **Sensor de qualidade de ar:** NEST Protect Smoke Alarm e Netatmo Weather Station;
- **Dispositivos de saúde:** Withings Smart Scale, Blipcare Blood Pressure Monitor e Withings Aura Smart Sleep Sensor;
- **Lâmpadas:** Light Bulbs LiFX Smart Bulb;
- **Electronicos:** Triby Speaker, PIX-STAR Photo-frame e HP Printer.

5.4.1 Análise da rede através do monitor

A avaliação da proposta é feita através da divisão da coleta em duas partes: treino e teste. Foram utilizados 13 dias para treino e 7 para os testes. Ao realizar a análise da rede e extração das características, correspondente a 13 dias dos arquivos, obteve-se, ao final, um total de 11337216 pacotes para um montante de 5253079111 bytes ditribuidos em 187138 fluxos TCP (4791973948 bytes ou 91,22% do tráfego) e 110571 fluxos UDP (461105163 bytes ou 8,88% do tráfego), o que corresponde a 297709 fluxos TCP e UDP no total.

Os arquivos correspondentes aos outros 7 dias têm em sua composição um total de 8666343 pacotes para um montante de 5141507199 bytes distribuídos em 65208 fluxos TCP (4469739728 bytes ou 86,9% do tráfego) e 72576 fluxos UDP (671767471 bytes ou 13,1% do tráfego).

5.4.2 Resultados da identificação por análise do conteúdo dos pacotes

Os resultados apresentados na Tabela 11 demonstram como é possível identificar os dispositivos em rede local através da análise de pacotes. Foi possível identificar praticamente todos os dispositivos Não-IoT através do User-Agent. Em contrapartida, poucos dispositivos IoT foram identificados (e.g., impressora HP e Câmera Netatmo).

Tabela 11 – Demonstrativo da identificação através da inspeção dos pacotes

Dispositivo	Tipo	Mac com resolução	Endereço Mac	User-Agent
Insteon Camera	IoT	00:62:6e:51:27:2e	00:62:6e:51:27:2e	I/r/n
Insteon Camera	IoT	Shenzhen_19:de:4f	e8:ab:fa:19:de:4f	I/r/n
Amazon Echo	IoT	AmazonTe_56:cc:d3	44:65:0d:56:cc:d3	-
PIX-STAR Photo-f.	IoT	AmpakTec_33:bb:85	e0:76:d0:33:bb:85	Mozilla/5.0
Iphone	N-IoT	Apple_df:a1:e1	d0:a6:37:df:a1:e1	CaptiveNetworkSupport-346 wispr
MacBook	N-IoT	Arunans-MacBook-Pro.local	ac:bc:32:d4:6f:2f	Mac OS X/10.11.5 (15F34)
iHome	IoT	hap-29D71D.local	74:c6:3b:29:d7:1d	WMSDK/r/n
Belkin Wemo switch	IoT	BelkinIn_79:f4:89	ec:1a:59:79:f4:89	Portable SDK for UPnP devices
Belkin motion sensor	IoT	BelkinIn_83:28:11	ec:1a:59:83:28:11	Linux/2.6.21, UPnP/1.0, Portable SDK
Dropcam	IoT	Dropcam_2f:e4:b2	30:8c:fb:2f:e4:b2	-
Dropcam	IoT	Dropcam_b6:ea:45	30:8c:fb:b6:ea:45	-
MacBook/Iphone	N-IoT	Hassans-MacBook-Pro.local	f4:5c:89:93:cc:85	Mac OS X/10.11.6 (15G1004)
HP Printer	IoT	HP705A0FE49BC0.local	70:5a:0f:e4:9b:c0	HP-Inkjet-WebUpdate-1.1-nb
Android Phone	N-IoT	Htc_a7:a3:c2	b4:ce:f6:a7:a3:c2	Dalvik/2.1.0 (HTC_PN071 Build/LRX22G)
Triby Speaker	IoT	Invoxia_02:20:44	18:b7:9e:02:20:44	Lavf/57.2.100
LiFX Smart Bulb	IoT	LifiLabs_01:83:08	d0:73:d5:01:83:08	-
NEST smoke alarm	IoT	NestLabs_25:be:e4	18:b4:30:25:be:e4	-
Netatmo weather station	IoT	Netatmo_03:b8:ac	70:ee:50:03:b8:ac	Dalvik/1.6.0 (AOSP Netatmo-Camera Build)
Netatmo Welcome	IoT	Netatmo_18:34:43	70:ee:50:18:34:43	Dalvik/1.6.0 (AOSP Netatmo-Camera Build)
Smart Things	IoT	Physical_00:67:5e	d0:52:a8:00:67:5e	-
Blipcare Blood Pressure	IoT	Rezolt_00:2e:25	74:6a:89:00:2e:25	-
Samsung Galaxy Tab	N-IoT	android-537f8-lan.local3	08:21:ef:3b:fc:e3	Dalvik/2.1.0 (Android 6.0.1; SM-T810)
Samsung SmartCam	IoT	SamsungE_ab:6b:88	00:16:6c:ab:6b:88	-
TP-Link Smart plug	IoT	Tp-LinkT_00:56:39	50:c7:bf:00:56:39	-
TP-Link Router	N-IoT	Tp-LinkT_51:33:ea	14:cc:20:51:33:ea	-
TP-Link Cloud	IoT	Tp-LinkT_93:51:f1	f4:f2:6d:93:51:f1	hehe
Withings Monitor	IoT	Withings_11:18:a8	00:24:e4:11:18:a8	Withings UserAgent
Withings Smart scale	IoT	Withings_1b:6f:96	00:24:e4:1b:6f:96	Withings UserAgent
Withings Aura smart	IoT	Withings_20:28:c6	00:24:e4:20:28:c6	Withings UserAgent
Laptop	N-IoT	Dan-PC.local	74:2f:68:81:69:42	Microsoft-Windows/6.1 UPnP/1.0
Android Phone	N-IoT	MurataMa_ff:1e:da	40:f3:08:ff:1e:da	Dalvik/2.1.0 (Android 5.0.1; GT-I9505)

Quanto à resolução do endereço MAC e IP, temos um suporte visual que pode ser aplicado e possibilita identificar dispositivos ou a empresa de manufatura de dispositivos que não apresentaram User-Agent, como é o caso da DropCam[2]. Apesar do suporte, o uso da resolução do endereço MAC permite, genericamente, identificar a empresa responsável por sua fabricação, o que permite limitar o número de possibilidades em uma eventual análise de rede.

5.4.3 Identificação dos dispositivos utilizando a análise de DNS e mDNS queries

Alternativas para identificação dos dispositivos foram introduzidas para monitoramento de dispositivos de IoT, entre os quais temos a identificação do host, mDNS[3] (*Multicast Domain Name Service*) e DNS *queries*. Ambos os componentes deixam rastros que permitem identificá-los através de análise minuciosa.

[2] Empresa americana que desenvolve câmeras para transmissão de vídeos através de Wi-Fi.
[3] Permite resolver nomes para endereços IP em redes que não possuem um servidor de nome local

Dispositivo	DNS queries	MDNS queries
Amazon Echo	pindorama.amazon.com	-
	device-metrics-us.amazon.com	-
Belking Wemo Motion Sense	www.belkin.com	-
Ihome	api.evrything	hap-29D71D.local,163.1.168.192.in-addr.arpa
	mqtt.evrythng.com	iHome SmartPlug-29D71D._hap._tcp.local
DropCam Camera	nexus.dropcam.com	-
	oculus689-vir.dropcam.com	-
Insteon Camera	connect.insteon.com	-
HP printer	www.hp.com	HP
	xmpp006.glb1.hp.com	HP705A0FE49BC0.local
	chat.hpeprint.com	-
Netatmo Welcome	netcom.netatmo.net	-
NEST Protect smoke alarm	frontdoor.nest.com	-
	log-rts08-iad01.devices.nest.com	-
Netatmo weather station	netcom.netatmo.net	-
Blipcare Blood Pressure meter	tech.carematix.com	-
Pix-Star Photo Frame	iptime.pix-star.com	-
	pix-star.com	-
LiFX Smart Bulb	lb.lifx.co	-
	v2.broker.lifx.co	-
Smart Thing	DC.connect.smartthings.com	-
Samsung SmartCam	www.samsungsmartcam.com	SAMSUNG-SNH-P6410BN _http._tcp.local
Triby Speaker	whatismyip.akamai.com.edgesuite.net	triby-2044.local
	blue.invoxia.io	_spotify-connect._tcp.local
TP-Link Smart plug	devs.tplinkcloud.com	-
TP-Link Day Night Cloud camera	aps1-relay.tplinkcloud.com	Little Cam-9351f1._tcp.local
Withings Smart scale	scalews.withings.net, withings.net	-
Withings Aura smart sleep sensor	scalews.withings.net	_withings-aura-bridge._tcp.local
	xmpp.withings.net	WSD-28C6._withings-aura._tcp.local

Os resultados presentes na Tabela 12 apenas reforçam a capacidade de identificação dos dispositivos através da análise dos pacotes. A classificação do tráfego de rede em IoT, em nossa proposta, possui como premissa a identificação dos dispositivos. O processo de identificação de dispositivos ocorre através de seu mapeamento mediante análise dos domínio DNS. Porém, é possível afirmar que o uso da nuvem da Amazon (AWS), pelo gerente dos dispositivos, impôs uma dificuldade acentuada ao camuflá-los, mas, ainda assim, muitos dos dispositivos emitiram consultas (DNS *queries*) para mais de um domínio, permitindo obter os resultados presentes na Tabela 12.

5.4.4 Análise e caracterização dos dispositivos em rede

Foi realizada a análise e caracterização do comportamento dos dispositivos da rede. Neste caso, foi utilizada uma comparação baseada na quantidade de *Endpoints*, na média do tamanho dos pacotes, no número total de pacotes DNS e na quantidade de requisições únicas DNS (*DNS Unique Request*).

Essa análise permite modelar um perfil visual do comportamento dos dispositivos em um rede com tráfego agregado. Além disso, permite fazer uma imagem apropriada das características utilizadas, ou que o podem ser, para realizar a classificação da rede em IoT.

Tabela 13 – Análise de características específicas entre dispositivos de IoT e Não-IoT

Análise dos dispositivos					
Classe	Nome	EndPoints	Av.pkt Mean	DNS Packets	Uniq. DNS Req.
Cameras	TP-Link Cloud camera	57	358,13	2344	>0 e $\leq$20
	Samsung Camera	119	290,8	52029	
	Dropcam*	16	120	30	
	Netatmo Welcome	76	292,47	2883	
	Baby Monitor	11	75,03	11970	
	Insteon Camera	48	119,99	55509	
Air Sensor	NEST smoke alarm	20	228,3	146	>0 e $\leq$3
	Netatmo weather station	18	154,81	11593	
Hubs	Amazon	73	258,74	85509	>0 e $\leq$3
	Smart Things	489	69,94	6672	
Switchers	Belkin Wemo switch	22	242,37	4259	>0 e $\leq$3
	TP-Link Smart plug	102	126,1	1500	
	iHome	31	96,48	281	
	Belkin wemo motion sensor	24	243,19	4230	
Health Care	Withings Smart scale	5	214,51	78	>0 e $\leq$3
	Aura smart sleep sensor	15	203,82	8176	
	Blipcare Blood Pressure	5	123,55	9	
Light	LiFX Smart Bulb	573	94,45	19838	>0 e $\leq$3
Eletronics	Triby Speaker	67	112,41	8554	>0 e $\leq$3
	PIX-STAR Photo-frame	10	197,36	11216	
	HP Printer	43	242,15	242	
Não-IoT	Sansung Galaxy Tab	920	520,99	29956	>50 e $\leq$195
	Android Phone	347	691,48	2528	

***Apenas uma amostra foi utilizada**

Duas características fundamentais apresentaram um padrão de comportamento deter-

minístico que pode promover a dissociação entre os tipos de dispositivos. Entre eles, a quantidade de endpoints (servidores) e de nomes de *queries* únicas entre dispositivos, IoT e Não-IoT, que apresentou maior variação entre os tipos de dispositivos, como é possível observar na Tabela 13. A variação do tamanho dos pacotes entre eles e a quantidade de endpoints também podem ser utilizadas para realização da classificação do tráfego de rede, em IoT e Não-IoT, visto que a média dos pacotes IoT é inferior à média de Não-IoT e a quantidade de endpoints, na grande maioria dos IoT, também é inferior.

5.4.5 Resultados da classificação do fluxo de rede

O uso do modelo *stratified 10-fold cross-validation*, que é uma melhora da validação cruzada, garante que cada dobra (fold) possua estruturas de entrada igualmente representada para classificação. O modelo 10-fold (ou *K-fold*) divide o subconjunto de dados em 10 partes mutualmente exclusivas (subconjuntos disjuntos) e, dessa forma, utiliza 9 parte para estimação dos parâmetros (ou K-1) e 1 para o teste do modelo (ocorrendo K vezes) (ZHANG *et al.*, 2015). O resultado dessa validação está presente na Tabela 14.

Tabela 14 – Avaliação do modelo 10-*fold*

Dispositivo	TPR	FPR	Precisão	Recall	F1-score	Mcc
Belkin Wemo switch	0,999	0,000	0,999	0,999	0,999	0,998
Samsung SmartCam	0,997	0,000	0,998	0,997	0,997	0,997
Withings Smart Baby Monitor	0,987	0,001	0,986	0,987	0,986	0,985
TP-Link Day Night Cloud camera	0,951	0,000	0,972	0,951	0,961	0,961
PIX-STAR	0,983	0,001	0,975	0,983	0,979	0,979
Amazon Echo	0,969	0,002	0,975	0,969	0,972	0,969
Netatmo weather station	1,000	0,000	0,999	1,000	0,999	0,999
Netatmo welcome	0,919	0,002	0,900	0,919	0,909	0,907
Withings Smart scale	0,846	0,000	1,000	0,846	0,917	0,920
NEST Protect smoke alarm	0,964	0,000	1,000	0,964	0,982	0,982
Belkin wemo motion sensor	0,972	0,000	0,963	0,972	0,967	0,967
HP Printer	0,814	0,000	0,973	0,814	0,886	0,889
Blipcare Blood Pressure meter	0,750	0,000	1,000	0,750	0,857	0,866
Triby Speaker	0,813	0,000	0,861	0,813	0,837	0,837
DropCam	0,261	0,000	0,286	0,261	0,273	0,273
TP-Link Smart plug	0,915	0,001	0,890	0,915	0,903	0,902
Smart Things	0,999	0,000	1,000	0,999	1,000	1,000
Withings Aura smart sleep	0,962	0,001	0,971	0,962	0,966	0,965
IHome	0,937	0,000	1,000	0,937	0,967	0,968
LiFX Smart Bulb	0,995	0,000	0,993	0,995	0,994	0,994
Insteon Camera	0,997	0,000	0,999	0,997	0,998	0,998
Não-IoT	0,982	0,005	0,975	0,982	0,978	0,974
Média	0,986	0,001	0,987	0,986	0,986	0,985

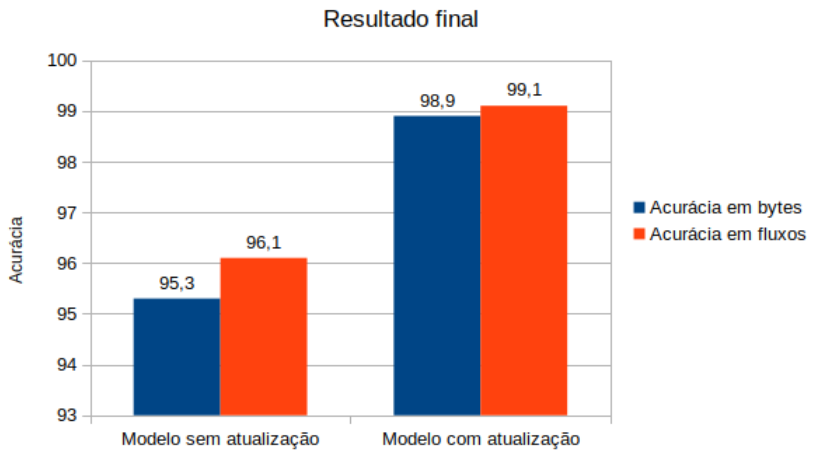
O pior resultado encontrado foi o relacionado à Dropcam. Isso ocorreu por que outro dispositivo similar (Nest Dropcam) possuía características similares (O autor da captura (SIVANATHAN *et al.*, 2017) não considerou o Nest Dropcam como dispositivo de IoT). O teste

apresentou, ao final, a acurácia de 98,64% com intervalo de confiança em 95%, desvio padrão de 1,03 e erro padrão de 0,00257. Esses resultados permitem afirmar, quanto à capacidade do modelo, que existe um suporte confiável para a predição de dispositivos através das características utilizadas.

Os resultados apresentados na Figura 11 referem-se ao uso do *dataset* de teste. Sua comparação está relacionada ao modelo sem atualização, que representa, exclusivamente, o uso das características estatísticas extraídas do fluxos da rede presente na Tabela 7, e o modelo atualizado, que corresponde ao uso em conjunto das informações extraídas através da inspeção dos pacotes (IP, MAC e User-Agent) na primeira etapa. O processo ocorre como descrito na Figura 7 do capítulo 4.

Utilizando o conjunto de dados anterior como treino e aplicando-o ao restante dos dias para teste, obteve-se um valor de acurácia muito próximo ao presente na Tabela 14. A avaliação do modelo ocorreu através da análise de acurácia quanto aos fluxos de dispositivos identificados corretamente e o total de bytes correspondente. Porém, vale ressaltar que os endereços IP extraídos e presentes nos *datasets* não são utilizados pelo algoritmo na classificação estatística, para evitar *Overfitting* (superestimação dos parâmetros).

Figura 11 – Resultados da identificação dos dispositivos para fluxos e bytes



Fonte: Autor

Ao aplicar os componentes extraídos na identificação dos dispositivos na fase de busca por assinaturas, obteve-se, ao final, acurácia de 99,1% com relação a fluxos e de 98,9% em relação a bytes, o que reduziu significativamente o número de FPR nos dispositivos identificados (Principalmente Não-IoT, 0,002). A Figura 11 apresenta o resultado e a comparação da acurácia após a utilização dos resultados de inspeção de pacotes. Esse resultado tem variação alta

decorrente da identificação de quase 100% dos dispositivos Não-IoT (PCs, Smarthones e tablet foi de 100%). Os dispositivos Não-IoT foram os que apresentaram menor acurácia no teste, principalmente com relação a bytes (quando não foram aplicados os recursos extraídos da inspeção de pacotes).

5.4.6 Resultos da classificação do tráfego

Classificar o tráfego gerado pelos diferentes dispositivos permite montar perfis de tráfego ou mesmo a construção de um padrão de comunicação. Além disso, a caracterização do tráfego de rede é importante não apenas para a construção de relatórios de caracterização de tráfego, mas também para sua aplicação em segurança da informação. Pensando nisso, aplicamos a classificação do tráfego de rede, através de classificação estatística, para avaliar classes de protocolos que se tornaram mais presentes e a taxa de tráfego referente.

A classificação de tráfego de rede foi realizada através do uso de classificação supervisionada. Foram utilizados tanto a validação cruzada quanto o método *holdout*. O método *holdout* consiste em dividir o conjunto total de dados em dois subconjuntos mutuamente exclusivos - um para treinamento e outro para teste (DOMINGOS, 2012). A validação cruzada ocorreu como nos testes anteriores, a título de validação e generalização dos dados utilizados. Porém, para a classificação do tráfego, composto pelo *dataset* de teste, foi utilizado um conjunto de dados correspondente ao *dataset* de treino com presença de pequenas porções do tráfego para treino, e ao final possibilitar a predição das classes de aplicações para o conjunto completo.

Tabela 15 – Resultados da classificação do tráfego de rede

Método	Acurácia	MCC	TPR	FPR	F1-Score
10-Fold	0,9904	0,987	0,990	0,003	0,990
Holdout	0,984	0,975	0,980	0,004	0,96

Ao final dos testes, o modelo stratified 10-fold cross-validation apresentou, para um nível de confiança de 95%, erro padrão de 0,0028 e desvio padrão de 0,92, média de 99,04% de acurácia na classificação dos fluxos. Além disso, ao utilizar o treino para construção do modelo e aplicado os testes, ao final, teve-se 98,4% de acurácia. Os resultados mostram que o tráfego de rede pode ser classificado com alta acurácia utilizando o conjunto de dados extraído pela ferramenta (PINHEIRO; CASTRO., 2017), treinar a rede com pequenas porções do tráfego e utilizando algoritmos de seleção de características (seção 5.2). A composição de tráfego

encontrada através da classificação para a rede está apresentada na Tabela 16 e tem seus valores correspondentes organizados em porcentagem de fluxos e de bytes. É importante ressaltar que dentro das classes temos uma variedade de classes de tráfego que foram agrupadas. Por exemplo, para a classe Web, temos ssl.youtube, ssl.google, http.apple e outros.

Tabela 16 – Caracterização do tráfego de rede

Classes	% Fluxos	% Bytes	Protocolos principais
Serviços	56,51%	1,14%	DNS e NTP
Web	32,35%	64,46%	HTTP, HTTPS e MQTT
Descoberta	4,3%	4,16%	SSDP e UPnP
E-mail	2,73%	0,16%	IPOP (3), SMTP e IMAP (3)
Configuração	1,72%	0,31%	DHCP (v6) e NetBIOS
VPN	1,71%	0,01%	VPN e OpenVPN
Interativo	0,01%	29,66%	SSH e Telnet
Banco de Dados	0,03%	0,003%	PostgreSQL e MySQL
P2P	0,63%	0,08%	BitTorrent

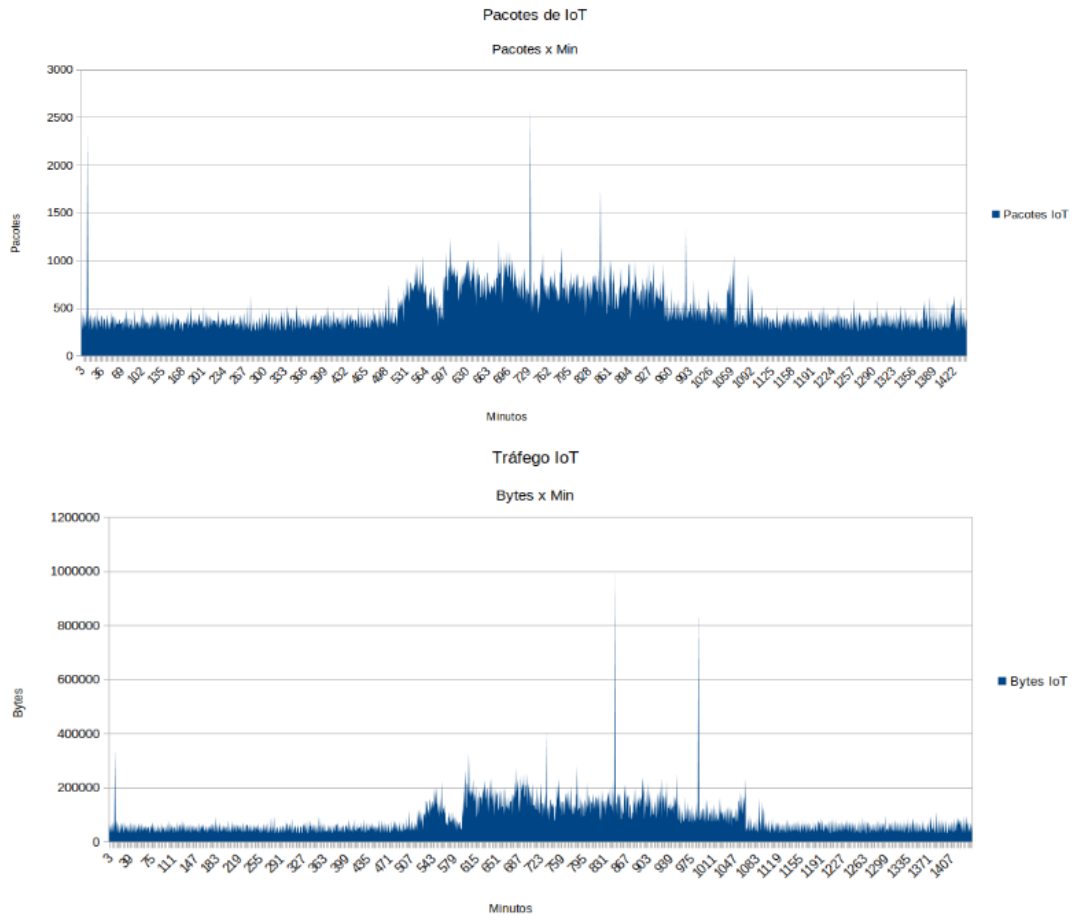
5.4.7 Resultados da classificação em IoT

Na Figura 12 é apresentada uma demonstração de como o tráfego de rede IoT se comporta em relação à captura completa. Nela é mostrado um comparativo de tráfego de rede total e o tráfego gerado apenas pelos dispositivos de IoT, utilizando como exemplo o dia 16 de outubro de 2016. A figura demonstra que o comportamento típico de IoT (tolerância ao atraso, processamento e armazenamento limitados, por exemplo) promovem uma grande disparidade no tamanho e na quantidade dos pacotes/bytes.

A figura serve para ilustrar o comportamento da rede. Sua formação se deu após a identificação dos dispositivos através da estratégia proposta. O tráfego de IoT, ao final, corresponde a todos os pacotes e bytes que possuem como origem ou destino um dos dispositivos IoT. A representação é feita através dos dados coletados e agrupados, cronologicamente, em minutos, onde a relação utilizada para demonstração é feita como pacotes x minuto ou bytes x minutos.

A figura apresenta uma grande diferença nas características de tráfego. Por exemplo, o tráfego formado por todos os dispositivos da rede (Figura 12) apresenta um pico de pacote em cerca de 20000 pacotes e um volume de aproximadamente 110 Mb. Enquanto isso, na (Figura 13) verifica-se cerca de 2600 pacotes para um volume de aproximadamente 1 Mb, relacionados exclusivamente ao tráfego de Internet das coisas.

Figura 12 – Tráfego da rede analisada



Fonte: Autor

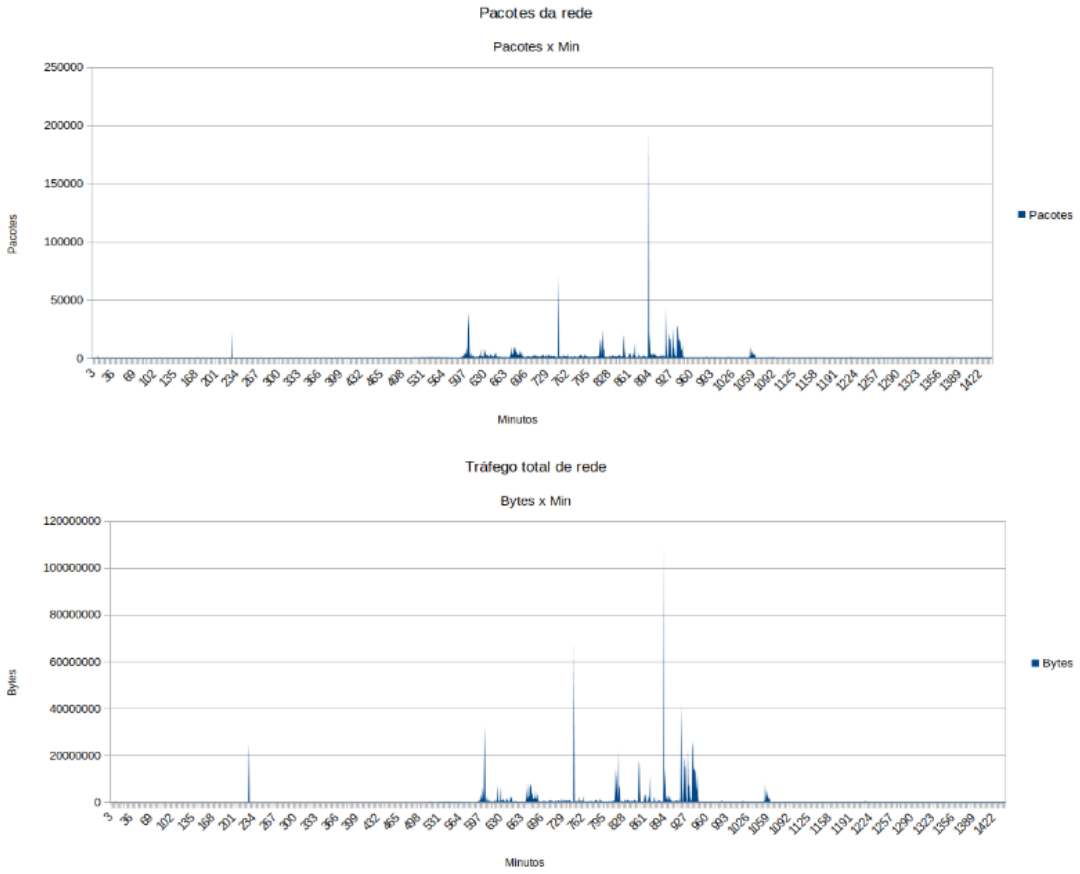
5.4.8 Caracterização da rede

As observações apresentadas nessa seção demonstram a análise dos dados através de observações do nível de pacotes no tráfego dos mais de 20 dispositivos presentes nessa rede e observações apresentadas em (SIVANATHAN *et al.*, 2017). Nesta fase da análise foram avaliadas de forma ampla algumas características de tráfego IoT, incluindo a composição do fluxo, distribuição dos pacotes e protocolos mais amplamente utilizados.

A composição do tráfego IoT identificado foi de mais de 13 milhões de pacotes, dos quais 88,934% correspondendo a IPV4 (ARP apresentou 8,4% dos pacotes). Com relação ao tráfego IoT, 83,5% dos fluxos eram TCP e 16,5% correspondem a UDP. Após o processo de *Benchmark*, aplicando inferências aos protocolos na camada de aplicação, é possível afirmar que as portas de destino mais amplamente utilizadas pelos dispositivos de IoT foram:

1. **TCP 443:** HTTPS - Implementação do protocolo HTTP sobre camadas adicionais de segurança (SSL/TLS);

Figura 13 – Tráfego geral da rede analisada



Fonte: Autor

2. **TCP 80:** HTTP - Protocolo de transferência de hipertexto;
3. **UDP 1900:** UPnP/SSDP - Protocolo de descoberta de serviço simples, focado na publicação e descoberta de serviços na rede. Ela utiliza mecanismos de configuração simples para realizar essa atividade como o DHCP ou o DNS.

O tráfego gerado pelos dispositivos de IoT na rede analisada está criptografado em 55% de sua totalidade (HTTPS) e outros 11% correspondem a tráfego HTTP. Dentro da rede é possível identificar o protocolo SSDP (TCP 1900). Sua presença possibilita afirmar que os dispositivos de internet das coisas estão anunciando sua presença para fazer a descoberta de outros dispositivos. A composição do tráfego IoT por SSDP é de 8% do volume total e sua comunicação se faz através de Multicast (endereço 239.255.255.250). A porta TCP de número 1935, associada a vídeo, corresponde a 7% de todo o tráfego IoT e associa-se, como já foi informado, ao dispositivos **baby monitor camera**. Outros dois protocolos que se apresentam constantes são o NTP (UDP 123) que é responsável pela sincronização dos relógios e representa 2% do tráfego total de internet das coisas na rede analisada.

5.5 Considerações Finais

Este capítulo mostra como é possível identificar dispositivos de Internet das coisas e o fluxo gerado por eles através de análise de tráfego utilizando duas técnicas associadas. O método apresentado possiblitou a identificação do tráfego de rede gerado por dispositivos em aproximadamente 99% e classificou tráfego de rede com acurácia. A abodagem pode ser aplicada em ambientes corporativos. Além disso, foi demonstrado que características típicas para classificação estatística da rede podem ser utilizadas para a identificação de padrões e classes de tráfego.

6 CONCLUSÃO

Este capítulo está estruturado da seguinte forma: a seção **6.1** responde as questões de pesquisas anteriormente apresentadas; a seção **6.2** lista os resultados alcançados através da abordagem apresentada; na sequência, a seção **6.3** discute os artigos produzidos durante a realização da pesquisa; a seção **6.4** discute as limitações deste trabalho; e por fim, a seção **6.5** apresenta as oportunidades e possibilidades de aplicação para trabalhos futuros.

6.1 Respostas às questões de pesquisas

Inicialmente, foram apresentadas algumas questões de pesquisas que nortearam esta dissertação, após os testes realizados e a estratégia desenvolvida temos, agora, a sua resolução.

1. É possível identificar dispositivos próprios de Internet das coisas ao utilizar técnicas de classificação, combinando a inspeção dos pacotes e Machine Learning?
    - Sim, a classificação por ML já permite um resultado considerável na identificação, desde que a rede seja previamente treinada com classes ou tipos específicos apropriados, como foi apresentado no capítulo 5. Contudo, é possível afirmar que o uso combinado permite aumentar, significativamente, a acurácia e precisão no processo.
2. É possível adquirir melhores resultados ao combinar duas técnicas para classificação de dispositivos e tráfego de rede?
    - Sim, como demonstrado no capítulo 5 e presente na Figura 11, por exemplo, os resultados de acurácia foram melhores que a classificação por ML.
3. Ao treinar uma rede utilizando ML (*Machine Learning*), é possível ter acurácia elevada na identificação desses dispositivos?
    - Sim, os testes, apresentados no capítulo 5, apresentaram resultados significativos na identificação dos dispositivos utilizando unicamente ML.
4. O User-Agent, presente no cabeçalho HTTP, promove suporte necessário para identificação de dispositivos específicos?
    - Sim, alguns dispositivos apresentam User-Agent que promovem a identificação unívoca deles, como apresentado nas Tabela 9 e 11 do capítulo 5.
5. Quais características do tráfego IoT mais influenciam a classificação?
    - As características que mais influenciaram a classificação estão detalhadas na seção 5.2, e foram: o total de bytes na ida, o tamanho máximo do pacote na ida, o total de

bytes na volta e a média dos pacotes na volta.

6. A literatura apresenta abordagens ou técnicas para identificar os dispositivos de Internet das coisas?
    - A literatura aborda técnicas e métodos para classificação de tráfego em IoT, além dos dispositivos; isso é perceptível em trabalhos como os de (MIETTINEN *et al.*, 2017), (SIVANATHAN *et al.*, 2017) e (MEIDAN *et al.*, 2017b).

6.2 Resultados Alcançados

A grande quantidade de dispositivos que são inseridos na rede promove grande complexidade à sua gestão eficaz, à segurança e à privacidade. Baseados nesses desafios, propomos uma estratégia combinada para identificar os dispositivos e classificar a rede.

Neste trabalho, assim como em (SIVANATHAN *et al.*, 2017) e (MEIDAN *et al.*, 2017b), foi atacado o problema da identificação dos dispositivos através do uso de ML e a classificação do tráfego em IoT, promovendo identificação automática e com bons resultados. A eficácia do modelo é comprovada baseado no uso de avaliação de métricas de desempenho, como as presentes em (JAIN, 1991). Adicionalmente, esse trabalho aborda o desafio de identificar o tráfego gerado por diferentes dispositivos, IoT e não-IoT, incluindo a análise do padrão de comunicação e caracterização do tráfego. A abordagem proposta inclui o uso de ML com a inspeção do conteúdo dos pacotes para realizar a classificação.

Ao final, esta dissertação demonstra como é possível identificar dispositivos de Internet das coisas e o fluxo gerado por eles através de análise de tráfego utilizando duas técnicas associadas. O método apresentado possibilitou a identificação dos dispositivos IoT em aproximadamente 99% e classificou tráfego de rede com relevante acurácia. A abordagem pode de ser aplicada em ambientes coorporativos no processo de identifcação dos dispositos conectados. Além disso, foi demonstrado que características típicas para classificação estatística da rede podem ser utilizadas para a identificação de padrões e classes de tráfego.

O quadro comparativo na Tabela 17, também presente no capítulo 3 na Tabela 6, apresenta uma visão geral das principais abordagens avaliadas nesta dissertação. Os critérios utilizados para avaliá-los foram extraídos baseados em sua proximidade com a dissertação. Os principais pontos para comparação de cada um dos trabalhos foi elencado da seguinte maneira:

1. Classifica utilizando ML;
2. Utiliza DPI;

3. Aborda classificação de dispositivos de IoT;
4. Classifica tráfego de rede em IoT e Não-IoT;
5. Utiliza seleção de atributos estatísticos;
6. Utiliza métricas de desempenho para avaliação dos resultados.

Tabela 17 – Tabela comparativa entre trabalhos relacionados

AUTORES	Ambiente	1	2	3	4	5	6
(ZHANG *et al.*, 2013a)	Não-IoT	✓	✓	x	x	✓	✓
(ZHANG *et al.*, 2015)	Não-IoT	✓	✓	x	x	✓	✓
(SIBY *et al.*, 2017)	IoT	x	x	✓	✓	x	✓
(APTHORPE *et al.*, 2017)	IoT	✓	x	✓	x	x	x
(SIVANATHAN *et al.*, 2017)	IoT	✓	x	✓	✓	x	✓
(MEIDAN *et al.*, 2017a)	IoT	✓	x	✓	x	✓	✓
(MEIDAN *et al.*, 2017b)	IoT	✓	✓	✓	✓	x	✓
(HAFEEZ *et al.*, 2017a)	IoT	✓	x	✓	x	✓	✓
(SHARMA *et al.*, 2018)	IoT	✓*	x	x	✓	x	✓
Dissertação	IoT	✓	✓	✓	✓	✓	✓

***Utiliza modelos probabilísticos**

Na Tabela 17 é perceptível que a dissertação compreende todas os pontos elecandos, o que demonstra uma melhoria em relação aos demais trabalhos avaliados.

6.3 Produção Bibliográfica

Durante o período de mestrado, março de 2016 a outubro de 2018, foram produzidos 2 trabalhos, ambos em conferências e relacionados a esta dissertação. A seguir, as publicações estão detalhadas:

1. **Matias Romário Pinheiro dos Santos**, Arthur Callado. *An Architecture Proposal for Network Traffic Monitoring with IoT Traffic Classification Support* In. **First IEEE Summer School on Smart Cities – S3C**. *Qualis: Sem Qualis*

   O artigo propõe o desenvolvimento de uma ferramenta de monitoramento de tráfego de rede que possibilite extrair atributos estatísticos de fluxo de rede e classificar o tráfego em IoT e não-IoT. Os testes apresentados nesse artigo foram realizados através de análise de arquivos *.pcap da indústria disponibilizados publicamente (PINHEIRO; CASTRO., 2017).
2. **Matias R. P. Santos**, Danielo G. Gomes, Rossana M. C. Andrade e Arthur C. Callado. *An efficient approach for device identification and traffic classification in IoT ecosystems*. In. **IEEE Symposium on Computers and Communications – ISCC**. *Qualis: A2*

O artigo apresenta uma estratégia para identificação automática dos dispositivos de IoT através do uso de classificação estatística com ML e inspeção do conteúdo dos pacotes, além de classificar o tráfego, taxonomicamente, demonstrando a sua composição e o impacto dos dispositivos no processo. Os testes foram realizados através de uma rede *smart campus* emulada (SANTOS *et al.*, 2018).

6.4 Limitações

A estratégia proposta, apesar de apresentar uma melhoria significativa na acurácia na classificação de tráfego de rede e de dispositivos, necessita realizar varreduras em busca por componentes específicos (assinaturas) dos pacotes. Dessa forma, os dispositivos precisam se comunicar utilizando meios não criptografados como o HTTP, um meio de transferência de dados não seguro.

É possível concluir que os experimentos foram limitados em relação aos protocolos analisados, haja vista que há uma grande quantidade de protocolos estudados e desenvolvidos para a construção de ecossistemas de IoT, como é o caso do CoAP.

Associado à estratégia está o processo de varredura para coleta de assinaturas, que configura-se como uma limitadora de aplicação em vários países, mesmo que essas buscas por conteúdos sejam específicas, tornando-se uma desvantagem.

Finalmente, outra limitação é a necessidade de treinar previamente a rede para identificar os demais dispositivos conectados, para, posteriomente, classificar o tráfego IoT. Isso decorre do uso da classificação por ML supervisionada.

6.5 Trabalhos Futuros

Durante o processo desta pesquisa, foi possível perceber a necessidade do desenvolvimento de estratégias alternativas e, ao mesmo tempo, generalistas para a classificação do tráfego em ambientes de IoT. A eminente popularização dos ambientes smart apresenta um grande desafio à gestão eficaz da rede, além de apresentar-se como um empecilho à classificação com resultados de acurácia elevada. Muitos trabalhos já empregam, no processo de classificação, a estratégia inicial da identificação dos ativos, haja vista que promove maior controle e acurácia na distinção e classificação do tráfego gerado.

Como proposta para trabalhos futuros, é fundamental promover o aprimoramento

desta pesquisa para a generalização da classificação do tráfego de IoT, contornando as limitações apresentadas. Além disso, decorrente da necessidade de generalização mais acentuada da proposta, será necessário utilizar uma série de protocolos amplamente presentes na literatura para realizar uma validação mais genérica e com mais ampla cobertura. Ao final, disponibilizar a ferramenta desenvolvida para a academia, *open source*, para que seja possível à comunidade propor melhorias ou até mesmo o refinamento da técnica.

REFERÊNCIAS

AL-FUQAHA, A.; GUIZANI, M.; MOHAMMADI, M.; ALEDHARI, M.; AYYASH, M. Internet of things: A survey on enabling technologies, protocols, and applications. **IEEE Communications Surveys & Tutorials**, v. 17, n. 4, p. 2347–2376, Fourthquarter 2015. ISSN 1553-877X.

AL-SARAWI, S.; ANBAR, M.; ALIEYAN, K.; ALZUBAIDI, M. Internet of things (iot) communication protocols: Review. In: **2017 8th International Conference on Information Technology (ICIT)**. [S.l.: s.n.], 2017. p. 685–690.

ALPAYDIN, E. **Introduction to Machine Learning**. 3. ed. [S.l.]: The MIT Press, 2014. ISBN 0262028182, 9780262028189.

ANDREONI, M.; SANZ, I. J.; MENEZES, D.; PUJOLLE, G.; DUARTE, O. C. M. B. Catraca: uma ferramenta para classificação e análise de tráfego escalável baseada em processamento por fluxo. In: . [S.l.: s.n.], 2017.

APTHORPE, N.; REISMAN, D.; SUNDARESAN, S.; NARAYANAN, A.; FEAMSTER, N. Spying on the smart home: Privacy attacks and defenses on encrypted iot traffic. **CoRR**, abs/1708.05044, 2017. Disponível em: <http://arxiv.org/abs/1708.05044>. Acesso em: 28 jun. 2018.

ATZORI, L.; IERA, A.; MORABITO, G. The internet of things: A survey. **Computer Networks**, v. 54, n. 15, p. 2787 – 2805, 2010. ISSN 1389-1286.

BAHGA, A.; MADISETTI, V. **Internet of Things: A Hands-On Approach:**. Arsheep Bahga & Vijay Madisetti, 2014. ISBN 9780996025515. Disponível em: <https://books.google.com.br/books?id=JPKGBAAAQBAJ>. Acesso em: 14 mar. 2018.

BERTINO, E.; CHOO, K.-K. R.; GEORGAKOPOLOUS, D.; NEPAL, S. Internet of things (iot): Smart and secure service delivery. **ACM Trans. Internet Technol.**, ACM, New York, NY, USA, v. 16, n. 4, p. 22:1–22:7, december 2016. ISSN 1533-5399. Disponível em: <http://doi.acm.org/10.1145/3013520>. Acesso em: 18 ago. 2018.

BITTENCOURT, H. R.; CLARKE, R. T. Use of classification and regression trees (cart) to classify remotely-sensed digital images. In: **IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)**. [S.l.: s.n.], 2003. v. 6, p. 3751–3753 vol.6.

BORMANN, C.; CASTELLANI, A. P.; SHELBY, Z. Coap: An application protocol for billions of tiny internet nodes. **IEEE Internet Computing**, v. 16, n. 2, p. 62–67, March 2012. ISSN 1089-7801.

BOWES, D.; HALL, T.; GRAY, D. Comparing the performance of fault prediction models which report multiple performance measures: Recomputing the confusion matrix. In: **Proceedings of the 8th International Conference on Predictive Models in Software Engineering**. New York, NY, USA: ACM, 2012. (PROMISE '12), p. 109–118. ISBN 978-1-4503-1241-7. Disponível em: <http://doi.acm.org/10.1145/2365324.2365338>. Acesso em: 18 mar. 2018.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, Oct 2001. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1010933404324>. Acesso em: 21 jun. 2018.

BUYYA, R.; DASTJERDI, A. V. **Internet of Things: Principles and Paradigms**. 1st. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2016. ISBN 012805395X, 9780128053959.

CAIDA. **Tipos de fluxos**. 2013. Disponível em: <https://www.caida.org/research/traffic-analysis/flowtypes/>. Acesso em: 12 fev. 2018.

CALLADO, A.; KAMIENSKI, C.; SZABO, G.; GERO, B. P.; KELNER, J.; FERNANDES, S.; SADOK, D. A survey on internet traffic identification. **IEEE Communications Surveys & Tutorials**, v. 11, n. 3, p. 37–52, rd 2009. ISSN 1553-877X.

CALLADO, A.; KELNER, J.; SADOK, D.; KAMIENSKI, C. A.; FERNANDES, S. Better network traffic identification through the independent combination of techniques. **Journal of Network and Computer Applications**, v. 33, n. 4, p. 433 – 446, 2010. ISSN 1084-8045.

CALLADO, A. de C. **Traffic identification in IP networks**. Tese (Doutorado) — Universidade Federal de Pernambuco, Pernabuco, Brazil, 2009.

CARELA-ESPAñOL, V. **Network Traffic Classification: From Theory to Practice**. 1-197 p. Tese (Doutorado) — Universitat Politècnica de Catalunya Barcelona, 2014.

CHEN, L.; THOMBRE, S.; JäRVINEN, K.; LOHAN, E. S.; ALéN-SAVIKKO, A.; LEPPäKOSKI, H.; BHUIYAN, M. Z. H.; BU-PASHA, S.; FERRARA, G. N.; HONKALA, S.; LINDQVIST, J.; RUOTSALAINEN, L.; KORPISAARI, P.; KUUSNIEMI, H. Robustness, security and privacy in location-based services for future iot: a survey. **IEEE Access**, v. 5, p. 8956–8977, 2017. ISSN 2169-3536.

CHOE, E. K.; CONSOLVO, S.; JUNG, J.; HARRISON, B.; KIENTZ, J. A. Living in a glass house: A survey of private moments in the home. In: **Proceedings of the 13th International Conference on Ubiquitous Computing**. New York, NY, USA: ACM, 2011. (UbiComp '11), p. 41–44. ISBN 978-1-4503-0630-0. Disponível em: <http://doi.acm.org/10.1145/2030112.2030118>. Acesso em: 20 jul. 2018.

CHUNG, J.-M.; LEE, D.; SONG, W. J.; CHOI, S.; LIM, C.; YEOUM, T. Enhancements to fpmipv6 for improved seamless vertical handover between lte and heterogeneous access networks. **IEEE Wireless Commun.**, v. 20, n. 3, p. 1–0, 2013. Disponível em: <https://doi.org/10.1109/MWC.2013.6549290>. Acesso em: 04 abr. 2018.

COLUMBUS, L. **2017 Roundup Of Internet Of Things Forecasts**. 2017. Disponível em: <https://www.forbes.com/sites/louiscolumbus/2017/12/10/2017-roundup-of-internet-of-things-forecasts/>. Acesso em: 07 mai. 2018.

CROTTI, M.; DUSI, M.; GRINGOLI, F.; SALGARELLI, L. Traffic classification through simple statistical fingerprinting. **SIGCOMM Comput. Commun. Rev.**, ACM, New York, NY, USA, v. 37, n. 1, p. 5–16, jan. 2007. ISSN 0146-4833. Disponível em: <http://doi.acm.org/10.1145/1198255.1198257>. Acesso em: 19 jun. 2018.

DAINOTTI, A.; PESCAPE, A.; CLAFFY, K. C. Issues and future directions in traffic classification. **IEEE Network**, v. 26, n. 1, p. 35–40, January 2012. ISSN 0890-8044.

DAINOTTI, A.; PESCAPÉ, A.; SANSONE, C. Early classification of network traffic through multi-classification. In: ____. **Traffic Monitoring and Analysis: Third International**

Workshop, TMA 2011, Vienna, Austria, April 27, 2011. Proceedings. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 122–135. ISBN "978-3-642-20305-3".

DOMINGOS, P. A few useful things to know about machine learning. **Commun. ACM**, ACM, New York, NY, USA, v. 55, n. 10, p. 78–87, out. 2012. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/2347736.2347755>. Acesso em: 06 jan. 2018.

EGEA, S.; REGO, A.; CARRO, B.; SANCHEZ-ESGUEVILLAS, A.; LLORET, J. Intelligent iot traffic classification using novel search strategy for fast based-correlation feature selection in industrial environments. **IEEE Internet of Things Journal**, PP, n. 99, p. 1–1, 2017.

ERMAN, J.; MAHANTI, A.; ARLITT, M. Qrp05-4: Internet traffic identification using machine learning. In: **IEEE Globecom 2006**. [S.l.: s.n.], 2006. p. 1–6. ISSN 1930-529X.

FARHAN, L.; SHUKUR, S. T.; ALISSA, A. E.; ALRWEG, M.; RAZA, U.; KHAREL, R. A survey on the challenges and opportunities of the internet of things (iot). In: **2017 Eleventh International Conference on Sensing Technology (ICST)**. [S.l.: s.n.], 2017. p. 1–5.

FASTFORMAT. **Mapeamento Sistemático da Literatura: Como fazer?** 2015. <https://blog.fastformat.co/revisao-da-literatura-2/>. Acessado em: 22-03-2018.

FINSTERBUSCH, M.; RICHTER, C.; ROCHA, E.; MULLER, J. A.; HANSSGEN, K. A survey of payload-based traffic classification approaches. **IEEE Communications Surveys & Tutorials**, v. 16, n. 2, p. 1135–1156, Second 2014. ISSN 1553-877X.

GARTNER. **Gartner Says 8.4 Billion Connected "Things"Will Be in Use in 2017, Up 31 Percent From 2016**. 2017. Disponível em: <https://www.gartner.com/newsroom/id/3598917>. Acesso em: 15 jan. 2018.

GARTNER. Leading the iot. In: ____. **Leading the IoT**. Gartner, 2018. Disponível em: <https://www.gartner.com/imagesrv/books/iot/iotEbook_digital.pdf>. Acesso em: 6 fev. 2018.

HAFEEZ, I.; DING, A. Y.; ANTIKAINEN, M.; TARKOMA, S. Toward secure edge networks: Taming device-to-device (d2d) communication in iot. **CoRR**, abs/1712.05958, 2017. Disponível em: <http://arxiv.org/abs/1712.05958>. Acesso em: 03 jul. 2018.

HAFEEZ, I.; DING, A. Y.; TARKOMA, S. Ioturva: Securing device-to-device (d2d) communication in iot networks. In: **Proceedings of the 12th Workshop on Challenged Networks**. New York, NY, USA: ACM, 2017. (CHANTS '17), p. 1–6. ISBN 978-1-4503-5144-7. Disponível em: <http://doi.acm.org/10.1145/3124087.3124093>. Acesso em: 03 set. 2018.

HALL, M. A. **Correlation-based Feature Subset Selection for Machine Learning**. Tese (Doutorado) — University of Waikato, Hamilton, New Zealand, 1998.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. 2. ed. [S.l.]: Springer-Verlag New York, 2009. ISSN 0172-7397. ISBN 978-0-387-84857-0.

HILL, K. **How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did**. 2012. Disponível em: <https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#2838b83d6668>. Acesso em: 18 mar. 2018.

HURLBURT, G. F.; VOAS, J.; MILLER, K. W. The internet of things: A reality check. **IT Professional**, v. 14, n. 3, p. 56–59, May 2012. ISSN 1520-9202.

IERC. **Internet of Things**. 2014. Disponível em: <http://www.internet-of-things-research.eu/about_iot.htm>.

IHS. **IoT platforms: enabling the Internet of Things**. 2016. Disponível em: <https://cdn.ihs.com/www/pdf/enabling-IOT.pdf>. Acesso em: 14 abr. 2018.

JAIN, R. **The Art of Computer Systems Performance Analysis**. 1st. ed. [S.l.]: John Wiley E Sons Inc., New York, 1991. 685 p. ISSN 1050-916X. ISBN 0471-50336-3.

KARAGIANNIS, T.; BROIDO, A.; FALOUTSOS, M.; CLAFFY, K. Transport layer identification of p2p traffic. In: **Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement**. New York, NY, USA: ACM, 2004. (IMC '04), p. 121–134. ISBN 1-58113-821-0. Disponível em: <http://doi.acm.org/10.1145/1028788.1028804>. Acesso em: 10 jan. 2018.

KAWAI, H.; ATA, S.; NAKAMURA, N.; OKA, I. Identification of communication devices from analysis of traffic patterns. In: **2017 13th International Conference on Network and Service Management (CNSM)**. [S.l.: s.n.], 2017. p. 1–5.

KITCHENHAM, B. A.; BRERETON, P.; TURNER, M.; NIAZI, M.; LINKMAN, S. G.; PRETORIUS, R.; BUDGEN, D. Refining the systematic literature review process - two participant-observer case studies. **Empirical Software Engineering**, v. 15, n. 6, p. 618–653, 2010. Disponível em: <https://doi.org/10.1007/s10664-010-9134-8>. Acesso em: 28 set. 2018.

KOHNO, T.; BROIDO, A.; CLAFFY, K. C. Remote physical device fingerprinting. **IEEE Transactions on Dependable and Secure Computing**, v. 2, n. 2, p. 93–108, April 2005. ISSN 1545-5971.

KORCZYNSKI, M. **Classifying Application Flows and Intrusion Detection in Internet Traffic**. 1-138 p. Tese (Doutorado) — DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE, 2012.

KORCZYńSKI, M.; DUDA, A. Classifying service flows in the encrypted skype traffic. In: **2012 IEEE International Conference on Communications (ICC)**. [S.l.: s.n.], 2012. p. 1064–1068. ISSN 1550-3607.

KRANENBURG, R. van; DODSON, S. **The Internet of Things: A Critique of Ambient Technology and the All-seeing Network of RFID**. Institute of Network Cultures, 2008. (Network notebooks). ISBN 9789078146063. Disponível em: <https://books.google.com.br/books?id=PilgkgEACAAJ>. Acesso em: 03 jul. 2018.

KUBAT, M. **An Introduction to Machine Learning**. 2. ed. [S.l.]: Springer International Publishing, 2017. ISBN 978-3-319-63912-3.

KULKARNI, M. **Reverse Hypothesis Machine Learning: A Practitioner's Perspective**. 1. ed. [S.l.]: Springer International Publishing, 2017. 138 p. ISSN 1868-4394. ISBN 978-3-319-55312-2.

KUROSE, J. F.; ROSS, K. W. **Computer Networking: A Top-Down Approach (6th Edition)**. 6th. ed. [S.l.]: Pearson, 2012. ISBN 0132856204, 9780132856201.

LI, G.; DONG, M.; OTA, K.; WU, J.; LI, J.; YE, T. Deep packet inspection based application-aware traffic control for software defined networks. In: **2016 IEEE Global Communications Conference (GLOBECOM)**. [S.l.: s.n.], 2016. p. 1–6.

LIN, J.; YU, W.; ZHANG, N.; YANG, X.; ZHANG, H.; ZHAO, W. A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications. **IEEE Internet of Things Journal**, v. 4, n. 5, p. 1125–1142, Oct 2017.

LIU, H.; MOTODA, H. **Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)**. [S.l.]: Chapman / Hall/CRC, 2007. ISBN 1584888784.

LIU, H.; YU, L. Toward integrating feature selection algorithms for classification and clustering. **IEEE Transactions on Knowledge and Data Engineering**, v. 17, n. 4, p. 491–502, April 2005. ISSN 1041-4347.

LLC, L. R. **"An Introduction to the Internet of Things (IoT)**. 2013. Disponível em: <https://www.cisco.com/c/dam/en_us/solutions/trends/iot/introduction_to_IoT_november.pdf>. Acesso em: 01 fev. 2018.

LóPEZ, T. S.; RANASINGHE, D. C.; HARRISON, M.; MCFARLANE, D. Adding sense to the internet of things. **Personal Ubiquitous Comput.**, Springer-Verlag, London, UK, UK, v. 16, n. 3, p. 291–308, mar. 2012. ISSN 1617-4909. Disponível em: <http://dx.doi.org/10.1007/s00779-011-0399-8>. Acesso em: 02 jun. 2018.

LU, W.; TAVALLAEE, M.; GHORBANI, A. A. Hybrid traffic classification approach based on decision tree. In: **GLOBECOM 2009 - 2009 IEEE Global Telecommunications Conference**. [S.l.: s.n.], 2009. p. 1–6. ISSN 1930-529X.

MADHUKAR, A.; WILLIAMSON, C. A longitudinal study of p2p traffic classification. In: **14th IEEE International Symposium on Modeling, Analysis, and Simulation**. [S.l.: s.n.], 2006. p. 179–188. ISSN 1526-7539.

MAURICE, C.; ONNO, S.; NEUMANN, C.; HEEN, O.; FRANCILLON, A. Improving 802.11 fingerprinting of similar devices by cooperative fingerprinting. In: **2013 International Conference on Security and Cryptography (SECRYPT)**. [S.l.: s.n.], 2013. p. 1–8.

MEIDAN, Y.; BOHADANA, M.; SHABTAI, A.; OCHOA, M.; TIPPENHAUER, N. O.; GUARNIZO, J. D.; ELOVICI, Y. Detection of unauthorized iot devices using machine learning techniques. **CoRR**, abs/1709.04647, 2017. ISSN 1709.04647. Disponível em: <http://arxiv.org/abs/1709.04647>. Acesso em: 14 mai. 2018.

MEIDAN, Y.; BOHADANA, M.; SHABTAI, A.; GUARNIZO, J. D.; OCHOA, M.; TIPPENHAUER, N. O.; ELOVICI, Y. Profiliot: A machine learning approach for iot device identification based on network traffic analysis. In: **Proceedings of the Symposium on Applied Computing**. New York, NY, USA: ACM, 2017. (SAC '17), p. 506–509. ISBN 978-1-4503-4486-9. Disponível em: <http://doi.acm.org/10.1145/3019612.3019878>. Acesso em: 13 mai. 2018.

MICRIUM. **Designing the Internet of Things**. 2017. Disponível em: <https://www.micrium.com/iot/internet-protocols/>. Acesso em: 29 mar. 2018.

MIDDLETON, S. E.; MODAFFERI, S. Scalable classification of qos for real-time interactive applications from ip traffic measurements. **Computer Networks**, v. 107, n. Part 1, p. 121 – 132, 2016. ISSN 1389-1286.

MIETTINEN, M.; MARCHAL, S.; HAFEEZ, I.; ASOKAN, N.; SADEGHI, A. R.; TARKOMA, S. Iot sentinel: Automated device-type identification for security enforcement in iot. In: **2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)**. [S.l.: s.n.], 2017. p. 2177–2184. ISSN 1063-6927.

MITCHELL, T. M. **Machine Learning**. [S.l.]: WCB McGraw-Hill, 1997.

MOORE, A. W.; PAPAGIANNAKI, K. Toward the accurate identification of network applications. In: ____. **Passive and Active Network Measurement: 6th International Workshop, PAM 2005, Boston, MA, USA, March 31 - April 1, 2005. Proceedings**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. p. 41–54. ISBN 978-3-540-31966-5.

NAMDEV, N.; AGRAWAL, S.; SILKARI, S. Recent advancement in machine learning based internet traffic classification. **Procedia Computer Science**, v. 60, n. Supplement C, p. 784 – 791, 2015. ISSN 1877-0509. Knowledge-Based and Intelligent Information & Engineering Systems 19th Annual Conference, KES-2015, Singapore, September 2015 Proceedings.

NG, B.; HAYES, M.; SEAH, W. K. G. Developing a traffic classification platform for enterprise networks with sdn: Experiences amp; lessons learned. In: **2015 IFIP Networking Conference (IFIP Networking)**. [S.l.: s.n.], 2015. p. 1–9.

NGUYEN, K. T.; LAURENT, M.; OUALHA, N. Survey on secure communication protocols for the internet of things. **Ad Hoc Netw.**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 32, n. C, p. 17–31, set. 2015. ISSN 1570-8705. Disponível em: <http://dx.doi.org/10.1016/j.adhoc.2015.01.006>. Acesso em: 14 jan. 2018.

NGUYEN, T. T. T.; ARMITAGE, G. A survey of techniques for internet traffic classification using machine learning. **IEEE Communications Surveys & Tutorials**, v. 10, n. 4, p. 56–76, Fourth 2008. ISSN 1553-877X.

NOTTINGHAM, A.; IRWIN, B. Parallel packet classification using gpu co-processors. In: **Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists**. New York, NY, USA: ACM, 2010. (SAICSIT '10), p. 231–241. ISBN 978-1-60558-950-3. Disponível em: <http://doi.acm.org/10.1145/1899503.1899529>. Acesso em: 12 jan. 2018.

PINHEIRO, S. M. R. .; CASTRO., C. A. de. An architecture proposal for network traffic monitoring with iot traffic classification support. In: **2017 IEEE First Summer School on Smart Cities (S3C)**. [S.l.: s.n.], 2017. p. 55–60.

PRASHANTH, G.; PRASHANTH, V.; JAYASHREE, P.; SRINIVASAN, N. Using random forests for network-based anomaly detection at active routers. In: **2008 International Conference on Signal Processing, Communications and Networking**. [S.l.: s.n.], 2008. p. 93–96.

SANTOS, B. P.; SILVA, L. A. M.; CELES, C. S. F. S.; NETO, J. B. B.; PERES, B. S.; VIEIRA, M. A. M.; VIEIRA, L. F. M.; GOUSSEVSKAIA, O. N.; LOUREIRO, A. A. F. Internet das coisas: da teoria à prática. In: **Anais SBRC 2016**. [S.l.: s.n.], 2016.

SANTOS, M. R. P.; ANDRADE, R. M. C.; GOMES, D. G.; CALLADO, A. C. An efficient approach for device identification and traffic classification in iot ecosystems. In: **2018 IEEE Symposium on Computers and Communications (ISCC)**. [S.l.: s.n.], 2018. p. 00304–00309. ISSN 1530-1346.

SERPANOS, D.; WOLF, M. **Internet-of-Things (IoT) Systems Dimitrios Serpanos Marilyn Wolf Architectures, Algorithms, Methodologies**. [S.l.]: Springer International Publishing, 2017. 95 p. ISBN 978-3-319-69714-7.

SHARMA, R.; KUMAR, N.; SRINIVAS, T. Markov chain based priority queueing model for packet scheduling and bandwidth allocation. In: KUMAR, N.; THAKRE, A. (Ed.). **Ubiquitous Communications and Network Computing**. Cham: Springer International Publishing, 2018. p. 91–103. ISBN 978-3-319-73423-1.

SIBY, S.; MAITI, R. R.; TIPPENHAUER, N. O. Iotscanner: Detecting privacy threats in iot neighborhoods. In: **Proceedings of the 3rd ACM International Workshop on IoT Privacy, Trust, and Security**. New York, NY, USA: ACM, 2017. (IoTPTS '17), p. 23–30. ISBN 978-1-4503-4969-7. Disponível em: <http://doi.acm.org/10.1145/3055245.3055253>. Acesso em: 11 set. 2018.

SILVIO, V.; D., R.; A., D.; A., P.; A., F.; M., M. Reviewing traffic classification. v. 7754, p. 123–147, 2013.

SIVANATHAN, A.; SHERRATT, D.; GHARAKHEILI, H. H.; RADFORD, A.; VISHWANATH, C. W. A.; SIVARAMAN, V. Characterizing and classifying iot traffic in smart cities and campuses. p. 1–6, Mai 2017.

SOLOMON, M. G.; KIM, D.; CARRELL, J. L. **Fundamentals Of Communications And Networking**. 2nd. ed. USA: Jones and Bartlett Publishers, Inc., 2014. ISBN 1284060144, 9781284060140.

SOYSAL, M.; SCHMIDT, E. G. Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison. **Performance Evaluation**, v. 67, n. 6, p. 451 – 467, 2010. ISSN 0166-5316. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0166531610000027>. Acesso em: 14 mar. 2018.

Introduction to statistical machine learning. In: SUGIYAMA, M. (Ed.). 1. ed. Boston: Morgan Kaufmann, 2016. ISBN 978-0-12-802121-7. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128021217000017>. Acesso em: 19 jul. 2018.

SUKANYA. **opentechdiary**. 2015. Disponível em: <https://opentechdiary.wordpress.com/2015/07/18/part-4-a-walk-through-internet-of-things-iot-basics/>. Acesso em: 03 jul. 2018.

SUN, G. L.; XUE, Y.; DONG, Y.; WANG, D.; LI, C. An novel hybrid method for effectively classifying encrypted traffic. In: **2010 IEEE Global Telecommunications Conference GLOBECOM 2010**. [S.l.: s.n.], 2010. p. 1–5. ISSN 1930-529X.

TANG, J.; ALELYANI, S.; LIU, H. **Feature Selection for Classification: A Review**. 2014.

TAYLOR, D. E. Survey and taxonomy of packet classification techniques. **ACM Comput. Surv.**, ACM, New York, NY, USA, v. 37, n. 3, p. 238–275, set. 2005. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/1108956.1108958>. Acesso em: 11 mai. 2018.

TELLEZ, M.; EL-TAWAB, S.; HEYDARI, M. H. Iot security attacks using reverse engineering methods on wsn applications. In: **2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)**. [S.l.: s.n.], 2016. p. 182–187.

VELAN, P.; VCERMAK, M.; CELEDA, P.; DRA, M. A survey of methods for encrypted traffic classification and analysis. **Netw.**, Wiley-Interscience, New York, NY, USA, v. 25, n. 5, p. 355–374, sep 2015. ISSN 0028-3045. Disponível em: <https://doi.org/10.1002/nem.1901>. Acesso em: 16 abr. 2018.

WANG, C.; XU, T.; QIN, X. Network traffic classification with improved random forest. p. 78–81, Dec 2015.

WANG, W.; SUN, Z.; REN, K.; ZHU, B. Increasing user capacity of wireless physical-layer identification in internet of things. In: **2016 IEEE Global Communications Conference (GLOBECOM)**. [S.l.: s.n.], 2016. p. 1–6.

WANG, Y. **Automatic network traffic classification**. 1-187 p. Tese (Doutorado) — Deakin University, 2013.

WICHERSKI, G.; WEINGARTEN, F.; MEYER, U. Ip agnostic real-time traffic filtering and host identification using tcp timestamps. In: **38th Annual IEEE Conference on Local Computer Networks**. [S.l.: s.n.], 2013. p. 647–654. ISSN 0742-1303.

YILDIRIM, T.; RADCLIFFE, P. Voip traffic classification in ipsec tunnels. In: **2010 International Conference on Electronics and Information Engineering**. [S.l.: s.n.], 2010. v. 1, p. V1–151–V1–157.

ZANDER, S.; NGUYEN, T.; ARMITAGE, G. Automated traffic classification and application identification using machine learning. In: **The IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05)l**. [S.l.: s.n.], 2005. p. 250–257. ISSN 0742-1303.

ZANELLA, A.; BUI, N.; CASTELLANI, A.; VANGELISTA, L.; ZORZI, M. Internet of things for smart cities. **IEEE Internet of Things Journal**, v. 1, n. 1, p. 22–32, Feb 2014. ISSN 2327-4662.

ZHANG, J.; CHEN, C.; XIANG, Y.; ZHOU, W.; VASILAKOS, A. V. An effective network traffic classification method with unknown flow detection. **IEEE Transactions on Network and Service Management**, v. 10, n. 2, p. 133–147, June 2013. ISSN 1932-4537.

ZHANG, J.; CHEN, X.; XIANG, Y.; ZHOU, W.; WU, J. Robust network traffic classification. **IEEE/ACM Transactions on Networking**, v. 23, n. 4, p. 1257–1270, Aug 2015. ISSN 1063-6692.

ZHANG, J.; XIANG, Y.; WANG, Y.; ZHOU, W.; XIANG, Y.; GUAN, Y. Network traffic classification using correlation information. **IEEE Transactions on Parallel and Distributed Systems**, v. 24, n. 1, p. 104–117, Jan 2013. ISSN 1045-9219.

ZHANG, J.; XIANG, Y.; ZHOU, W.; WANG, Y. Unsupervised traffic classification using flow statistical properties and ip packet payload. **Journal of Computer and System Sciences**, v. 79, n. 5, p. 573 – 585, 2013. ISSN 0022-0000. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0022000012001729>. Acesso em: 10 mai. 2018.

ZHANG, M.; JOHN, W.; CLAFFY, K.; BROWNLEE, N. State of the art in traffic classification: A research review. In: **PAM Student Workshop**. [S.l.: s.n.], 2009.

ZIEGLER; ANDREAS; KöNIG, I. R. Mining data with random forests: current options for real-world applications. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Periodicals, Inc, v. 4, n. 1, p. 55–63, 2014. ISSN 1942-4795. Disponível em: <http://dx.doi.org/10.1002/widm.1114>. Acesso em: 07 jan. 2018.

ANEXO A – METODOLOGIA DA REVISÃO DE LITERATURA

O mapeamento sistemático de literatura é um método para construir estruturas de classificação e estruturar a área de pesquisa selecionada. Em um mapeamento sistemático, a análise tem por objetivo demonstrar a quantidade de publicações categorizadas em esquemas. O processo ocorreu como descrito em (FASTFORMAT, 2015) e em (KITCHENHAM *et al.*, 2010)

Foram realizadas pesquisas nas bases de dados *Science Direct*, *IEEE*, *Scopus*, *Web of Science* e *ACM Digital Library* na busca de literatura que abordasse temas relacionados a esta dissertação, mediante Revisão Sistemática de Literatura (RSL). A pesquisa foi realizada nos dias 3 e 4 de fevereiro de 2018.

Foram selecionados apenas trabalhos que possuem menção clara ao problema abordado nesta dissertação. Ao final, foram escolhidas as referências publicadas em inglês e português utilizando como critérios: proximidade do conteúdo; trabalhos indexados com estrato indicativos de qualidade (Qualis Capes); número de citação; ano de publicação. Contudo, não foi utilizada literatura sem a análise do resumo/*abstract*, das avaliações e do estudo de caso.

As palavras-chaves utilizadas e o número de estudos identificados em cada base estão apresentados nas tabelas a baixo.

Base de dados:	**Scopus** - 123 documentos
String de busca:	TITLE-ABS-KEY (((iot OR "Internet of things") AND (device* OR objetc) AND (identification OR classification) AND ("machine learning"OR statistical OR dpi OR "packet inspection")) OR ((network) AND (iot OR "Internet of things") AND (identification OR classification) AND ("Statistical*")))

Base de dados:	**IEEE** - 21 documentos
String de busca:	(((iot OR "Internet of things") AND (device* OR object) AND (identification OR classification) AND ("machine learning"OR statistical OR dpi OR "packet inspection")) OR ((network) AND (iot OR "Internet of things") AND (identification OR classification) AND ("Statistical*")))

Foram identificados nas bases de dados 291 documentos que somam-se a mais 9 trabalhos identificados em busca manual (Google acadêmico), totalizando 300. Entretanto, ao cruzar os resultados para remover duplicações totalizou 221 documentos. Dentre esses

Base de dados:	**ScienceDirect** - 2 documentos
String de busca:	Title, abstract, keywords: TS = (TS = (IoT OR "Internet of things") AND TS= (identification OR classification) AND TS = (network* OR traffic*) AND TS= (Statistical*)) OR (TS=(IoT OR "Internet of things") AND TS = (device* OR object) AND TS = (identification OR classification) AND TS= ("machine learning"OR statistical OR dpi OR "packet inspection"))

Base de dados:	**ACM DL** - 145 documentos
String de busca:	Searched for (+(network +OR +traffic) +AND +("IoT"+OR +"Internet of Things") +AND+ ("identification"+OR +"classification") +AND+ (machine learning +OR+ statistical)) +OR+ ((+device* +OR+ object) +AND +("IoT"+OR +"Internet of Things") +AND+ ("identification"+OR +"classification") +AND+ ("machine learning"+OR+ statistical +OR+ dpi +OR+ "packet inspection"))

documentos, apenas 22 foram selecionados baseado nos critérios apresentados aqui e no capítulo 1.

Os artigos identificados manualmente, através de buscas no Google acadêmico, foram:

- *Detection of Unauthorized IoT Devices Using Machine Learning Technique.* (MEIDAN *et al.*, 2017a). Este trabalho foi **utilizado** por abordar a identificação dos dispositivos utilizando o algoritmo RF.
- *IoTurva: Securing device-to-device(d2d) communication in iot networks* (HAFEEZ *et al.*, 2017b). **Utilizado**.
- *Toward secure edge networks: Taming device-to-device(d2d) communication in iot*. Arxiv. (HAFEEZ *et al.*, 2017a) **Utilizado**.
- *A Model for Anomalies Detection in Internet of Things (IoT) Using Inverse Weight Clustering and Decision Tree.* Dissertação de mestrado, Dublin Institute of Technology, 2017. doi:10.21427/D7WK7S. Esta dissertação **Não foi utilizado**.
- *Appliance Recognition on Internet-of-Things Devices.* Gérôme, 2014. **Não foi utilizado.**
- *Object Classification based Context Management for Identity Management in Internet of Things.* Parikshit 2013. **Não foi utilizado**.
- *Appliance Recognition on Internet-of-Things Devices.* Gérôme Bovet, Antonio Ridi, Jean Hennebert. 2014. **Não foi utilizado**.
- *Identity Management Framework for Internet of Things.* Tese de doutorado. Mahalle, Parikshit N.Aalborg Universitet. 2014. **Não foi utilizado**.

- *IoT Security Techniques Based on Machine Learning.* Arxiv. Liang et. al, 2018. **Não foi utilizado**.

Todos os trabalhos identificados foram analisados exaustivamente durante um período de 5 meses. Os trabalhos não utilizados foram excluidos após análise quanto a adequação à dissertação, quanto aos resumos, quanto a quantidade de citações e/ou por não possuir caso de uso que se adequasse ao escopo.