Segurança de Dados em Nuvem através de Aprendizado de Máquina: uma Revisão Sistemática da Literatura

Matheus Soares de Lacerda[1], **Robson Gonçalves Fechine Feitosa**[1] **(Orientador)**

[1]Sistemas de Informação – Instituto Federal do Ceará - Campus Crato (IFCE)
Rodovia CE 292, KM 15, Gisélia Pinheiro – Crato – CE – Brazil

the.matheus6@gmail.com, robsonfeitosa@ifce.edu.br

Abstract. *This work aims to aggregate, compare and synthesize (through a systematic review) the works present in the literature, which use machine learning to deal with security threats, in the context of cloud computing. As main results, a knowledge base was synthesized and demands were observed that indicate possible relevant research questions, and, consequently, more in-depth studies, such as: the investigation of Security as a Service (SecaaS).*

Resumo. *O presente trabalho tem como objetivo agregar, comparar e sintetizar (por meio de uma revisão sistemática) os trabalhos presentes na literatura, que utilizam aprendizado de máquina para lidar com ameaças de segurança, no contexto da computação em nuvem. Como principais resultados, sintetizou-se uma base de conhecimento e foram observadas demandas que indicam possíveis questões relevantes de pesquisa, e, consequentemente, estudos mais aprofundados, como: a investigação da Security as a Service (SecaaS).*

1. Introdução

Computação em nuvem (CN), do inglês *Cloud Computing*, é um modelo computacional que surge com a evolução de elementos técnicos, lógicos, físicos e arquiteturais, onde a capacidade de *hardware*, a virtualização, a Internet e a computação pervasiva são fornecidas de forma simples, com pouca configuração e baixo custo [Mell et al. 2011]. [Lin and Chen 2012], [Pereira et al. 2016] e [Modi et al. 2013] evidenciam que através da combinação de tais tecnologias foi possível o desenvolvimento e evolução de arquiteturas que se diferem do modelo padrão, ou seja *in-house*.

Conforme detalhado em [Subramanian and Jeyaraj 2018], a CN pode ser classificada em alguns níveis de abstração, sendo eles: IaaS (*Infrastructure as a Service*), PaaS (*Platform as a Service*) e SaaS (*Software as a Service*). O nível IaaS fornece vários conjuntos de recursos para o desenvolvimento de novos negócios, como o fornecimento de uma estrutura completamente virtualizada através da Internet. Ele permite que organizações de Tecnologia da Informação (TI) e desenvolvedores de *software* aumentem ou diminuam o número de máquinas virtuais em funcionamento dependendo da carga de trabalho, promovendo eficiência no uso de recursos de TI (e, consequentemente, redução de custos). São exemplos: o *Elastic Compute* da *Amazon Cloud* (EC2) e o *Mosso Hosting Cloud*.

A contratação de serviços em uma plataforma, como um modelo de serviço PaaS, é realizada por um provedor de serviços em nuvem com ferramentas, infraestrutura e sistemas operacionais, que possibilitem o desenvolvimento focado apenas nos produtos finais.

PaaS simplifica significativamente o *design* de aplicativos da *Web*; toda a manutenção dos serviços de *back-end* é realizado pelo provedor PaaS [Lin and Chen 2012]. O nível de serviço SaaS pode ser entendido como uma aplicação desenvolvida sobre um PaaS, onde os clientes podem pagar, alugar ou assinar aplicativos ou serviços de provedores de nuvem acessados pela Internet [Lin and Chen 2012]. Ou seja, [Mell et al. 2011] definem SaaS como "a capacidade fornecida ao consumidor de usar os aplicativos do provedor em execução em uma infraestrutura de nuvem".

Segundo [Tan et al. 2014], as tecnologias presentes na CN, além de fornecer vantagens econômicas e práticas, herdam vulnerabilidades que tornam os sistemas em nuvem mais suscetíveis a ataques. Diversos autores [Ashktorab et al. 2012, Modi et al. 2013, Pereira et al. 2016, Patil 2018] apresentam diferentes problemas e formas de mitigá-los e, em sua maioria, afirmam que problemas relacionados à segurança são as principais preocupações durante a gestão de plataformas na nuvem.

Visando a correção destes problemas, é necessário ampliar as soluções de segurança tradicionais, como: *firewall*; e, sistemas de detecção/prevenção de intrusões, para que possam lidar com o tráfego de rede de alta velocidade, bem como, com a configuração dinâmica da rede na nuvem. Dessa forma, soluções baseadas em *Machine Learning* (ML) também contribuem nas questões mencionadas. Ademais, a preservação da integridade, confidencialidade e não repúdio de dados utilizando ML em CN é uma área emergente, pois reúne algumas das principais técnicas de computação atuais.

Logo, o presente trabalho tem como objetivo analisar e sumarizar os trabalhos encontrados na literatura recente, relacionados ao tema exposto. Para isso, utilizou-se o protocolo descrito em [Keele et al. 2007], com a finalidade de: encontrar problemas em aberto ou contribuições científicas para o tema explorado; bem como, sintetizar (por meio de uma base de conhecimento, resultante da Revisão Sistemática da Literatura, RSL) os principais dados empíricos presentes na literatura que abordam o tema, conforme detalhado nas próximas seções.

Ademais, apresentamos aqui uma breve análise sobre a relação do presente trabalho com os "Grandes Desafios de Pesquisa em Sistemas de Informação no Brasil 2016 a 2026". Neles, observou-se uma frequente preocupação em: interoperabilidade de dados; e, dados abertos. Nesse sentido, os trabalhos: *"Open Perspectives on Cloud Computing Adoption: Realities and Challenges for Information System Practitioners in Brazil*" e "Sistemas de Informação baseados em Dados Abertos (Conectados): De Abertura à Inovação", reforçam a importância das empresas utilizarem os benefícios da CN; bem como, se preocuparem com a segurança da informação; assuntos esses diretamente relacionados com o presente trabalho.

2. Objetivos

Mais especificamente, no presente trabalho, investigou-se a literatura recente visando encontrar brechas de pesquisa na interseção entre os temas abordados em [Mishra et al. 2018], [Subramanian and Jeyaraj 2018] e [Moustafa et al. 2019b]. Com isso, foi possível: elencar os principais critérios de taxonomia de problemas e soluções apresentados por tais autores; criar uma base de conhecimento atualizada sobre segurança de dados em CN com uso de ML; levantar questões de pesquisas claras e objetivas; para então, sumarizar tais trabalhos e apresentá-los da forma mais esclarecedora possível, por

meio de uma RSL.

3. Trabalhos Relacionados

Segundo [Subramanian and Jeyaraj 2018], um dos maiores problemas na implementação de CN é a virtualização, devido a falta de controle sobre o ambiente virtualizado. Ou seja, os problemas com o compartilhamento entre *Virtual Machines* (VMs) de recursos comuns, ou com o compartilhamento de infraestrutura, permitem que os ataques se espalham rapidamente, e os usuários não saibam exatamente onde seus dados confidenciais estão localizados, pois provedores de serviços em nuvem gerenciam seus *data centers* em locais distribuídos geograficamente. [Subramanian and Jeyaraj 2018] ainda destacam que, estratégias modernas de proteção, como *firewalls*, antivírus e sistemas de detecção de intrusão, não fornecem a segurança adequada aos aspectos pertinentes presentes na CN.

Ainda segundo [Subramanian and Jeyaraj 2018], a camada de virtualização permite que diferentes VMs sejam implantadas e executadas simultaneamente no mesmo *host* físico. Isso é feito com um *software* específico conhecido como VMM (*Virtual Machine Monitor*), que atribui e preserva a separação de recursos entre instâncias da VM. Para permitir a comunicação entra as VMs é adicionado uma rede virtual, permitindo a comunicação através de um comutador virtual[1]. Além disso, ataques de *hardware* também pode ser um desafio; [Subramanian and Jeyaraj 2018] cita que, caso o atacante consiga se autenticar através de falhas no *Hypervisor*, ele pode chegar aos níveis de acesso da máquina, e com isso, realizar ataques de negação de serviço em múltiplas instâncias. Os principais desafios na camada de *hardware* são *backup*, localização do servidor, manutenção do *firewall* e o monitoramento da integridade do *hardware*.

[Subramanian and Jeyaraj 2018] ainda cita, que o vazamento de dados é um dos principais desafios na segurança em CN e são considerados os atributos fundamentais para sua funcionalidade estável e segura. A união de ambientes caracterizados por segregação geográfica, difícil localização e tenacidade múltipla incrementam os riscos de vazamento de dados. Assim, os dados presentes no serviço podem ser categorizados em: *data-in-rest*[2] e *data-in-transit*[3].

Por outro lado, [Mishra et al. 2018] realizaram uma investigação detalhada sobre soluções para a detecção de intrusão em CN baseadas em ML. Eles classificaram os principais tipos de ataques e suas características; onde, por meio de *features*[4] foi possível extrair cada categoria de ataque; e, posteriormente, efetuar uma sumarização das vantagens na utilização de técnicas de ML. Os autores afirmaram que um IDS (*Intrusion Detection System*) baseado em ML pode classificar falhas por meio de uma base de conhecimento de detecção de má utilização; ou, detecção de anomalias.

Eles também apresentaram uma categorização dos ataques de acordo com dois

[1]Um comutador virtual permite a comunicação entre máquinas virtuais e máquinas físicas através de recursos gerenciáveis disponibilizado pelo *Hypervisor*.

[2]Remete à segurança e integridade dos dados armazenados, como por exemplo, recuperabilidade, segregação geográfica, localização dos dados e a remoção eficiente de dados.

[3]Atualmente, para efetivar o transporte seguro de dados é utilizado o protocolo *Transport Layer Security* (TLS), além disso, outro meio de autenticar os dados em transporte é: utilizar métodos para verificar a origem dos dados e o caminho percorrido pelo mesmo de forma a garantir e auditar a integridade e segurança dos dados.

[4]Características observáveis em um conjunto de informações, onde é possível mensurar e classificá-lo.

datasets: KDD99 e UNSW-NB; e utilizaram a seguinte taxonomia: *Denial of Service Attacks*, *Scanning Attacks*, *User to Root Attacks* e *Remote to User Attacks* para os ataques baseados no *dataset* KDD99; e, *Fuzzers*, *Analysis*, *Backdoor*, *Exploits*, *Generic*, *Reconnaissance*, *Shellcode* e *Worms* aos relacionados ao *dataset* UNSW-NB. Além disso, [Mishra et al. 2018] ainda elencaram as principais ferramentas[5] utilizadas para realizar buscas de falhas e explorá-las.

[Moustafa et al. 2019b] realizaram testes de *benchmarking* com o objetivo de comparar técnicas de detecção de anomalias em diferentes *datasets*. Para os autores, a filosofia de um invasor é quase invariante, e pode ser simplificada em duas fases: a primeira, denominada fase de exploração, é um método para controlar a execução do fluxo do programa alvo. Em seu nível abstrato, isso pode ser alcançado utilizando métodos de *stack/heapbased buffer overflow*. A segunda, consiste em enviar *payloads* para o alvo, com o propósito de obter acesso aos recursos do sistema através das falhas obtidas na primeira fase.

Os trabalhos até então apresentados ilustram a grande diversidade de temas e propostas de pesquisa ligadas à segurança da informação em CN. Assim, para apoiar tal análise, as seções seguintes apresentam: o protocolo metodológico; uma descrição e discussão dos resultados obtidos; bem como algumas conclusões e considerações finais.

4. Metodologia

Segundo [Brereton et al. 2007, Keele et al. 2007], a RSL (Revisão Sistemática da Literatura) pode ser utilizada como ferramenta para encontrar, sumarizar e analisar pesquisas relevantes dentro de uma determinada área. Tal processo envolve várias etapas e aspectos técnicos [Abdalla et al. 2015]; assim, as seções a seguir, descrevem os aspectos utilizados na presente RSL: as ferramentas utilizadas; termos de busca; bases de pesquisa; critérios de inclusão e exclusão; avaliação da qualidade; e, questões de pesquisa.

4.1. Ferramentas

Para apoiar a condução da RSL é importante utilizar ferramentas que auxiliem o processo de: organização dos arquivos; armazenamento de resultados; e, favoreçam a análise gráfica. Para isso, utilizou-se: *Google Sheets*, para manter o controle de quais artigos estavam sob análise, rejeitados ou aceitos, e armazenar informações sobre, critérios, dados a extrair e etapas a serem executadas; Zotero[6], para manter os metadados dos artigos (e.g., autor, título, data e veículo); e, *Google Drive*, para armazenar artigos, documentos, revisões e análises desenvolvidas, bem como, as planilhas para planejamento e análise dos dados.

4.2. Termos de Busca

A *string* de pesquisa foi criada utilizando os termos frequentes nos trabalhos relacionados, de forma que, cada base pudesse ser utilizada independente de suas particularidades de consulta:

[5]São elas: *Nmap, scapy, Metasploit, Armitage, Dsniff, Tcpdump, Net2pcap, Snoop, Ettercap, Nstreams, Argus, Karpski, Ethereal, Amap, Vmap, TTLscan e Paketto*.

[6]Disponível em: `https://www.zotero.org`.

(machine-learning OR machine learning) AND (cloud-computing OR cloud computing) AND (data security OR data-security)

4.3. Bases de Pesquisa

Foram utilizadas quatro bases de pesquisas, selecionadas devido o alto impacto dos trabalhos nelas indexados[7]. Um segundo fator para a escolha das bases é a experiência do autor com os respectivos repositórios: ACM Digital Library[8]; IEEE Xplorer Digital Library[9]; Google Scholar[10]; e, Science Direct[11].

4.4. Critérios de Inclusão e Exclusão

A realização da busca por artigos relacionados pode resultar em centenas de centenas de trabalhos que não pertencem ao escopo principal do tema analisado, desta forma, para remover tais trabalhos foram adotados os seguintes critérios de exclusão: (CE1) trabalhos que não sejam artigos científicos, como: manuais, listas, livros; (CE2) trabalhos que não estão incluídos ou relacionados ao escopo, tema ou domínio pesquisado; e, (CE3) trabalhos não reproduzíveis, depreciados, não testados ou não verificáveis. Os critérios de inclusão foram: (CI1) trabalhos que contêm dados quantitativos e atualizados sobre intrusão a sistemas hospedados na nuvem; (CI2) trabalhos que abordem problemas ou propostas de solução envolvendo segurança de dados utilizando *Machine Learning*; (CI3) trabalhos publicados, no máximo, há cinco anos[12].

4.4.1. Fator de Impacto

Foi necessário construir de uma equação para determinar a relação entre a proporção de citações por ano e o limite de tempo na qual o artigo foi publicado[13]. A Equação 1, representa tal restrição, onde α é a quantidade de citações; e, β é o ano de publicação do trabalho. Logo, F busca apontar os artigos com maior fator de impacto para o tema em questão. Tal fator foi empregado para realizar uma ordenação dos trabalhos, de forma que os artigos com um fator de impacto abaixo de 1 foram descartados da revisão. $F = \frac{\alpha}{2020-\beta}$ (1).

[7]Por exemplo, cada base apresenta a seguinte descrição sobre seu conteúdo: "A ACM Digital Library (DL) é o banco de dados mais abrangente do mundo, com artigos em texto completo e literatura bibliográfica que aborda computação e tecnologia da informação". Ademais, "O IEEE Xplore fornece acesso via web a mais de quatro milhões de documentos em texto completo de algumas das publicações mais citadas do mundo em engenharia elétrica, ciência da computação e eletrônica". E, "ScienceDirect é um site que fornece acesso baseado em assinatura a um grande banco de dados de pesquisas científicas e médicas. Hospeda mais de 12 milhões de conteúdos de 3.500 periódicos acadêmicos e 34.000 e-books". Por fim, "O Google Scholar é um mecanismo de pesquisa na Web de acesso livre que indexa o texto completo ou os metadados da literatura acadêmica em uma variedade de formatos e disciplinas de publicação".

[8]`http://dl.acm.org/`

[9]`http://ieeexplore.org/`

[10]`https://scholar.google.com.br`

[11]`https://www.sciencedirect.com`

[12]A redução do escopo de tempo visa manter o trabalho dentro de uma estimativa de tempo viável.

[13]Tal equação busca garantir a presença de artigos com um maior fator do impacto, ou seja, que possuem um menor tempo de publicação e um maior número de citações.

4.5. Avaliação da Qualidade

[Keele et al. 2007] afirma que a avaliação dos artigos selecionados pode ser utilizada para maximizar a validade dos estudos encontrados e guiar novas pesquisas. Assim, os critérios aqui utilizados foram: (CQ1) "Os autores definem de forma clara o escopo e objetivo do trabalho?"; (CQ2) "O trabalho expõe dados de experimentos realizados?"; (CQ3) "O trabalho é reproduzível?"; (CQ4) "Os resultados do trabalho são expostos de forma clara?"; e; (CQ5) "O trabalho indica futuras continuações e/ou pesquisas posteriores?". Para todas as questões foram atribuídos valores: "0 para Não; 1 para Parcialmente; e, 2 para Sim".

Os artigos não qualificados dentro dos critérios de inclusão e exclusão, ou por algum motivo que invalide a inclusão dentro dos trabalhos válidos da pesquisa, são rotulados como: "Depreciado", se não há mais suporte para técnicas utilizadas ou as ferramentas propostas não são mais acessíveis; "Não Artigo" se foi publicado como livro, manual ou qualquer outro tipo que não seja artigo; "Inacessível" se não apresenta acesso, ou link de acesso válido; "Fora de Escopo" se foi rejeitado, pois não satisfaz os critérios de inclusão e exclusão relacionados ao escopo; "Antigo" se ultrapassa o tempo inicial aqui delimitado; "Baixo número de citações" se não atinge o valor mínimo definido como fator de impacto.

4.6. Questões de Pesquisa

Segundo [Keele et al. 2007], Questões de Pesquisa (QP's) constituem a parte essencial da revisão sistemática. De forma geral, a sintetização dos trabalhos encontrados e a extração de dados são processos que convergem em respostas para as questões de pesquisa. No presente trabalho, foram desenvolvidas cinco QP's em sua totalidade, de maneira que cada QP contenha um objetivo específico. São elas: (QP1) Quais algoritmos de *Machine Learning* são utilizados para solucionar problemas envolvendo segurança de dados?; (QP2) Qual a susceptibilidade de ataque a sistemas hospedados em nuvem?; (QP3) Quais tipos de ataques são mais abordados nos estudos?; (QP4) Quais os impactos negativos da utilização de *Machine Learning* na defesa de sistemas em nuvem? (QP5) Quais técnicas além de *Machine Learning* são utilizadas para identificar anomalias no tráfego de dados pela rede?

5. Descrição e Discussão dos Resultados Obtidos

Com os resultados obtidos foi possível sumarizar e organizar os dados, de forma a explorar quais lacunas podem ser investigadas, conforme apresentado nas seções a seguir.

5.1. Distribuição dos trabalhos pelas bases de busca

Através da sumarização foi possível verificar que o IEEE detêm a maior quantidade de artigos (62 do total) por ano, onde os maiores picos de artigos aceitos, rejeitados e em possível reanálise aconteceram em 2017 e 2018. Além disso, 2017 foi o ano com maior quantidade de artigos encontrados nas plataformas pesquisadas sobre o tema em questão.

Ao total, 69.81% dos trabalhos pesquisados não foram aceitos (ou seja, não passaram pelos critérios de inclusão e exclusão) e 16.98% dos artigos ficaram classificados para uma possível reanálise; o que demonstra que existe uma grande possibilidade de explorar outros aspectos da CN, ML e segurança de dados, que não foram abordados no presente trabalho.

5.2. Respostas à Questão de Pesquisa QP1

As técnicas utilizadas nos trabalhos encontrados, na maioria dos casos, seguem as etapas de: aplicar um classificador para encontrar o modelo mais otimizado para um determinado problema; e, posteriormente, utilizar o modelo para realizar a classificação de anomalias ou tipos de ataques. [Ghosh and Mitra 2015] descrevem a utilização de *Genetic Algorithms* (GA) para a seleção de *features* em um *dataset* aplicado a um *Best Feature Set Selection* (BFSS) para a seleção das melhores *features* encontradas. Por fim, é utilizado um classificador baseado em *Logistic Regression* (LR) e *Gradient Descent* (GD) para realizar a classificação de anomalias baseadas nas *features* selecionadas. [Idhammad et al. 2018] utiliza *Naive Bayes classifier* construído sob o *dataset* CIDDS-01 com o intuito de detectar anomalias no tráfego de rede. Após identificar o tráfego anômalo é utilizado *Random Forest* para realizar a classicação de qual tipo de ataque está ocorrendo. [Zekri et al. 2017] e [Win et al. 2017] utilizam *DecisionTree* a qual é encontrada em diversos trabalhos como classificador de anomalias, tal que em sua maioria é acompanhado de outro algoritmo, com o intuito de encontrar a menor árvore de decisão possível e então, executar a classificação de forma otimizada. *Logistic Regression* e *Belief Propagation* são utilizados para calcular a probabilidade condicional de um ataque acontecer baseado nas características extraídas.

5.3. Respostas à Questão de Pesquisa QP2

[Idhammad et al. 2018] relata que as vulnerabilidades presentes (que em sua maioria são alvos de escaneamentos e ataques a sistemas em nuvem) são resultados da transição de paradigmas computacionais, ou seja, a herança de múltiplas tecnologias que dão origem aos modelos em nuvem. [Win et al. 2017] descrevem que ambientes virtualizados são atrativos para ataques que visam explorar falhas de segurança, pois tais ambientes têm a capacidade de agrupar diferentes recursos de computação, bem como habilitar o escalonamento de recursos sob demanda. Fato este, que é evidenciado também em [Al Haddad et al. 2016] ao apresentar que os serviços em nuvem em sua maioria são alvos de ataques devido a quantidade de tecnologias envolvidas.

Segundo [Kumara and Jaidhar 2018], soluções convencionais de segurança em sistemas virtualizados são insuficientes para proteger os usuários dos sistemas operacionais contra *malwares* avançados. [Masetic et al. 2017] afirma que os provedores de computação em nuvem não fornecem segurança necessária durante o transporte de dados para nuvem, deixando este trabalho para os clientes, que muitas vezes não possuem o conhecimento necessário para construir um sistema de segurança consistente.

5.4. Respostas à Questão de Pesquisa QP3

Os trabalhos analisados demonstram que o maior objetivo em utilizar ML na segurança de serviços em nuvem é monitorar possíveis anomalias na rede. Em segundo lugar, os problemas mais abordados são os de negação de serviço, seguidos de Probe, R2L e U2R. Tais ataques são os que mais causam danos aos serviços em nuvem, seja por interrupção ao acesso do sistema, perda de dados ou vazamento de dados.

5.5. Respostas à Questão de Pesquisa QP4

Alguns autores, como [Ghosh and Mitra 2015], apontam que o principal problema na utilização de ML são a quantidade de memória e processamento de dados necessários

para a realização de treinamento e classificação. Além disso, [Huang et al. 2017], [Al Haddad et al. 2016] e [Cui and He 2016] apontam que a eficiência das técnicas baseadas em ML também é um limitador devido ao aumento da complexidade em decorrência da expansão de dados gerados no tráfego, e que em alguns casos, há uma quantidade considerável de erros ocasionados por falsos-positivos. [Moustafa et al. 2019a] afirma que a solução para garantir segurança deve ser capaz de lidar com grande volume, alta velocidade e alta dimensionalidade de dados, ou seja, *Big Data*.

5.6. Respostas à Questão de Pesquisa QP5

Em geral, [Ghosh and Mitra 2015] e [Huang et al. 2017] relatam que técnicas baseadas em IDS, IPS, *Firewall* são utilizadas em sistemas baseados em regras. Alguns autores como [Ghosh and Mitra 2015], apontam a utilização das variantes: *Host Based Intrusion Detection System (HIDS)* e *Network Based Intrusion Detection System (NIDS)* para a proteção de sistemas.

[Zekri et al. 2017] apontam a existência de outras técnicas para a realização de detecção de intrusão, como: *Signature-based Detection* (SD), *Anomaly-based Detection* (AD) e *Stateful Protocol Analysis* (SPA). Os trabalhos investigados sugerem que, para prover a segurança de serviços em nuvem, a utilização de *Machine Learning* e técnicas derivadas, em sua maioria, são mais eficientes para tratar vulnerabilidades já conhecidas, no entanto, elas falham no tratamento de ataques como *zero day attacks*, isto é, ataques que ainda não estão vinculados e registrados em bases de dados de soluções tradicionais de segurança como: *firewall*, antivírus, IDS e IPS baseados em regras.

5.7. Possíveis Lacunas de Pesquisa a serem Exploradas

[Kumar et al. 2017] ilustram alguns problemas relacionados à geração de dados. Conforme relatado no presente trabalho, existem poucos *datasets* na literatura que podem ser utilizados durante a etapa de extração e validação de características, para então, iniciar a etapa de classificação. Portanto, alguns problemas tendem a aparecer, devido ao fato de que a geração de ataques de forma simulada em ambientes controlados não representa o tráfego real de dados oriundos de uma rede em nuvem.

Segundo [Lipton et al. 2015] *Recurrent Neural Networks* (RNN's) são redes neurais aprimoradas pela inclusão de um conceito temporal ao modelo, fornecendo a possibilidade de memorizar etapas passadas. Tais características fornecem às RNN's a possibilidade de desenvolvimento de modelos para o reconhecimento de fala, entre outras soluções que dependem da variação de tempo. Desta forma, um possível trabalho futuro, seria investigar a utilização de modelos recorrentes para aprimorar a segurança de dados na CN.

Uma das soluções que surgem de forma emergente para os problemas supracitados é o SecaaS (*Securituy as a Service*), conforme apresentado em [Kim et al. 2019]. SecaaS tem o papel de proporcionar uma segurança inteligente, sob demanda, facilmente configurável e provida pela Internet, assim como os outros níveis disponibilizados (PaaS, SaaS e IaaS).

6. Conclusões e Trabalhos Futuros

A busca por sistemas de detecção de intrusão baseados em *Machine Learning* (ML) têm sido tema de pesquisas encontradas em vários trabalhos recentes da literatura. Os resulta-

dos obtidos revelaram várias formas de mitigar falhas de segurança da CN, no entanto, é importante levar em consideração que, os fatores intrínsecos ao contexto da CN expõem tais sistemas a ataques, como a exploração de vulnerabilidades na virtualização. Como trabalhos futuros é possível explorar as lacunas detalhadas na seção anterior, e explorar novas contribuições, como o SecaaS, buscando o aperfeiçoamento de seus aspectos técnicos.

Referências

Abdalla, G., Damasceno, C. D. N., and Nakagawa, E. Y. (2015). A systematic literature review on systems-of-systems knowledge representation. *Inst. Math. Comput. Sci., Univ. Sao Paulo, Sao Carlos, Brazil, Tech. Rep*.

Al Haddad, Z., Hanoune, M., and Mamouni, A. (2016). A collaborative framework for intrusion detection (c-nids) in cloud computing. In *2016 2nd International Conference on Cloud Computing Technologies and Applications (CloudTech)*, pages 261–265. IEEE.

Ashktorab, V., Taghizadeh, S. R., et al. (2012). Security threats and countermeasures in cloud computing. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, 1(2):234–245.

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., and Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of systems and software*, 80(4):571–583.

Cui, B. and He, S. (2016). Anomaly detection model based on hadoop platform and weka interface. In *2016 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*, pages 84–89. IEEE.

Ghosh, P. and Mitra, R. (2015). Proposed ga-bfss and logistic regression based intrusion detection system. In *Proceedings of the 2015 Third International Conference on Computer, Communication, Control and Information Technology (C3IT)*, pages 1–6. IEEE.

Huang, C., Min, G., Wu, Y., Ying, Y., Pei, K., and Xiang, Z. (2017). Time series anomaly detection for trustworthy services in cloud computing systems. *IEEE Transactions on Big Data*.

Idhammad, M., Afdel, K., and Belouch, M. (2018). Distributed intrusion detection system for cloud environments based on data mining techniques. *Procedia Computer Science*, 127:35–41.

Keele, S. et al. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, Ver. 2.3 EBSE Technical Report. EBSE.

Kim, H., Kim, J., Kim, Y., Kim, I., and Kim, K. J. (2019). Design of network threat detection and classification based on machine learning on cloud computing. *Cluster Computing*, 22(1):2341–2350.

Kumar, R. S. S., Wicker, A., and Swann, M. (2017). Practical machine learning for cloud intrusion detection: challenges and the way forward. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 81–90.

Kumara, A. and Jaidhar, C. (2018). Automated multi-level malware detection system based on reconstructed semantic view of executables using machine learning techniques at vmm. *Future Generation Computer Systems*, 79:431–446.

Lin, A. and Chen, N.-C. (2012). Cloud computing as an innovation: Percepetion, attitude, and adoption. *International Journal of Information Management*, 32(6):533–540.

Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.

Masetic, Z., Hajdarevic, K., and Dogru, N. (2017). Cloud computing threats classification model based on the detection feasibility of machine learning algorithms. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1314–1318. IEEE.

Mell, P., Grance, T., et al. (2011). The nist definition of cloud computing. *National Institute of Science and Technology Cloud, Hybrid Special Publication*.

Mishra, P., Varadharajan, V., Tupakula, U., and Pilli, E. S. (2018). A detailed investigation and analysis of using machine learning techniques for intrusion detection. *IEEE Communications Surveys & Tutorials*, 21(1):686–728.

Modi, C., Patel, D., Borisaniya, B., Patel, A., and Rajarajan, M. (2013). A survey on security issues and solutions at different layers of cloud computing. *The journal of supercomputing*, 63(2):561–592.

Moustafa, N., Choo, K.-K. R., Radwan, I., and Camtepe, S. (2019a). Outlier dirichlet mixture mechanism: Adversarial statistical learning for anomaly detection in the fog. *IEEE Transactions on Information Forensics and Security*, 14(8):1975–1987.

Moustafa, N., Hu, J., and Slay, J. (2019b). A holistic review of network anomaly detection systems: A comprehensive survey. *Journal of Network and Computer Applications*, 128:33–55.

Patil, G. R. M. D. A. (2018). Data breaches as top security concern in cloud computing. *International Journal of Pure and Applied Mathematics*, 119(14):19–28.

Pereira, A. L., Penha, E. W. M., Gomes, N. A., and Freitas, R. R. (2016). Computação em nuvem: a segurança da informação em ambientes na nuvem e em redes físicas. *Brazilian Journal of Production Engineering-BJPE*, 2(1):12–27.

Subramanian, N. and Jeyaraj, A. (2018). Recent security challenges in cloud computing. *Computers & Electrical Engineering*, 71:28–42.

Tan, Z., Nagar, U. T., He, X., Nanda, P., Liu, R. P., Wang, S., and Hu, J. (2014). Enhancing big data security with collaborative intrusion detection. *IEEE cloud computing*, 1(3):27–33.

Win, T. Y., Tianfield, H., and Mair, Q. (2017). Big data based security analytics for protecting virtualized infrastructures in cloud computing. *IEEE Transactions on Big Data*, 4(1):11–25.

Zekri, M., El Kafhali, S., Aboutabit, N., and Saadi, Y. (2017). Ddos attack detection using machine learning techniques in cloud computing environments. In *2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech)*, pages 1–7. IEEE.